nature chemical biology

ARTICLE

PUBLISHED ONLINE: 31 OCTOBER 2016 | DOI: 10.1038/NCHEMBIO.2219

Dereplication of peptidic natural products through database search of mass spectra

Hosein Mohimani[1], Alexey Gurevich[2], Alla Mikheenko[2], Neha Garg[3], Louis-Felix Nothias[3], Akihiro Ninomiya[4], Kentaro Takada[4], Pieter C Dorrestein[3,5] & Pavel A Pevzner[1,2]*

Peptidic natural products (PNPs) are widely used compounds that include many antibiotics and a variety of other bioactive peptides. Although recent breakthroughs in PNP discovery raised the challenge of developing new algorithms for their analysis, identification of PNPs via database search of tandem mass spectra remains an open problem. To address this problem, natural product researchers use dereplication strategies that identify known PNPs and lead to the discovery of new ones, even in cases when the reference spectra are not present in existing spectral libraries. DEREPLICATOR is a new dereplication algorithm that enables high-throughput PNP identification and that is compatible with large-scale mass-spectrometry-based screening platforms for natural product discovery. After searching nearly one hundred million tandem mass spectra in the Global Natural Products Social (GNPS) molecular networking infrastructure, DEREPLICATOR identified an order of magnitude more PNPs (and their new variants) than any previous dereplication efforts.

After a long decline in the pace of antibiotics discovery in the 1990s, natural products are again at the center of attention, as exemplified by the recent discoveries of novel classes of natural product drugs[1–4]. The key condition for enabling the renaissance of the natural product research is a turn from the current one-off process of analyzing natural products to high-throughput discovery pipelines. Thus, in addition to development of new experimental technologies, future studies of natural products will also require new computational approaches[5–7]. The recent launch of the GNPS molecular networking infrastructure[8] brought together over a hundred laboratories that have already generated an unprecedented amount of publicly available mass spectra of natural products. However, to transform natural product discovery into a high-throughput technology and to fully realize the promise of the GNPS project, new algorithms for natural product discovery are needed[6,9,10]. Indeed, although spectra in the GNPS molecular network represent a gold mine for future discoveries, their interpretation remains a bottleneck.

In this paper we focus on PNPs, which are produced by two types of biosynthetic machineries: nonribosomal peptide (NRP) synthetases[11] and ribosomally synthesized and post-translationally modified peptide (RiPP) synthetases[12], which synthesize NRPs and RiPPs, respectively. NRPs are not directly inscribed in genomes but are made by large multimodular NRP synthetases using nonribosomal code. Although RiPPs are encoded in the genome, the RiPP-encoding genes are often short, making it difficult to annotate them[13].

Development of reference spectral libraries of tandem mass spectra (MS/MS) has enabled identification of metabolites by searching spectra against these libraries as an alternative to the searches of candidate molecules in chemical databases[14]. However, in the case of PNPs, such libraries are small because, until recently, there was no centralized effort to annotate spectra of various PNPs. Although this situation had changed with the release of GNPS[8], the utility of data in this network needs to be enhanced with additional tools that can be applied to large extract collections in therapeutic discovery programs for identification of the previously described natural products and their variants. Such dereplication tools should be fast so that they can be applied to all GNPS spectra.

Natural product researchers face the challenge of maximizing the discovery of new compounds while minimizing the reevaluation of known compounds. The process of using the information about the chemical structure of a previously characterized compound to identify this compound in an experimental sample (without having to repeat the entire isolation and structure-determination process) is called dereplication[15]. Another challenge is finding variants of known compounds because those variants are sometimes more effective in clinical applications. For example, caspofungin is one of many examples of a variant PNP that proved to be effective in clinical applications[16]. Although many low-abundance variants of PNPs have been reported in the last two decades, it is difficult to identify all variants without dedicated computational tools. In this paper, we present a dereplication algorithm that identified hundreds of previously unknown variant PNPs.

In the case of PNPs, MS-based dereplication refers to matching MS/MS data against PNPs in a chemical library such as AntiMarin[17]. Similarly to database search tools in proteomics (for example, Sequest[18]), dereplication algorithms search for peptide-spectrum matches (PSMs) and score them based on similarities between theoretical spectra derived from peptides in the chemical library and experimental tandem spectra. The matched peptide that forms a statistically significant PSM with the highest score (against a given spectrum) is reported as a putative annotation. In many cases, a PNP in the new sample is absent in the database of known PNPs, but its variant is present in this database (for example, with a substitution, a modification or an adduct). Identification of an unknown PNP from its known variants is called the variable dereplication (as opposed to the standard dereplication when a PNP is present in the chemical database).

[1]Department of Computer Science and Engineering, University of California, San Diego, La Jolla, California, USA. [2]Center for Algorithmic Biotechnology, Institute of Translational Biomedicine, St. Petersburg State University, St. Petersburg, Russia. [3]Collaborative Mass Spectrometry Innovation Center, Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California, San Diego, La Jolla, California, USA. [4]Laboratory of Aquatic Natural Products Chemistry, School of Agricultural and Life Sciences, The University of Tokyo, Tokyo, Japan. [5]Department of Pharmacology, University of California, San Diego, La Jolla, California, USA. *e-mail: ppevzner@ucsd.edu

This description hides many details that make PNP identification difficult. First, in difference from peptides analyzed in traditional proteomics (that are assembled from 20 proteinogenic amino acids), PNPs are assembled from hundreds of amino acids. Second, PNP architectures are not limited to linear peptides but include cyclic, branched-cyclic and even more complex configurations. Third, although there was large effort invested in analyzing statistical significance of PSMs in traditional proteomics, methods for evaluating statistical significance of PNPs are still in infancy. Fourth, search for substituted and modified variants of known PNPs requires complex, blind database searches[19] because the set of possible substitutions and modifications is not known in advance.

Previously developed dereplication approaches include NRP-Dereplication[20] algorithm for cyclic peptides and informatic search algorithm for natural products (iSNAP)[21] algorithm for both cyclic and branch-cyclic peptides. However, in difference from NRP-Dereplication, iSNAP does not perform variable dereplication. DEREPLICATOR overcame the limitations of both NRP-Dereplication (cyclic peptides only) and iSNAP (standard dereplication only) and further addressed the problem of evaluating the statistical significance (*P* values) of PSMs formed by PNPs. By applying spectral networks[22,23] to perform variable dereplication, it enabled to our knowledge the first high-throughput PNP identification effort in the field of natural products that resulted in the discovery of many new variant PNPs.

RESULTS

Outline of the DEREPLICATOR algorithm

In **Figure 1** and **Supplementary Figure 1**, we illustrate the DEREPLICATOR pipeline, which includes the following steps described in the Online Methods: (i) generating decoy database of PNPs, (ii) constructing theoretical spectra for all PNPs in the database, (iii) generating and scoring PSMs, (vi) computing *P* values of PSMs and generating the set of statistically significant PSMs, (v) computing false discovery rate (FDR), and (vi) enlarging the set of found PSMs through variable dereplication via spectral networks.

The concept of spectral networks[22] (also known as molecular networks[24] when applied to metabolites and natural products) was introduced to reveal spectra of related peptides in a proteomic data set without knowing what these peptides are. Nodes in a spectral network correspond to spectra, and edges connect spectra that are generated from related peptides, for example, peptides differing by a single substitution, modification (such as oxidation, acetylation, methylation, etc.), or adduct (such as proton, sodium, potassium, etc.). Spectral networks enable variable dereplication of new variants of known PNPs via propagation of PSMs through a spectral network[25] and allow one to generate a hypothesis regarding the nature of the structural relatedness of peptides represented by the spectra within the network. Spectral networks are well suited for analyzing PNPs because most PNPs form families of related peptides through biosynthetic promiscuity, incomplete biosynthetic processing, non-enzymatic reactions or mutations between different species (**Supplementary Results**, **Supplementary Fig. 2**).

Benchmarking DEREPLICATOR

To benchmark DEREPLICATOR, we used the AntiMarin database[17] to dereplicate all spectra from the following GNPS data sets: $\text{Spectra}_{\text{GNPS}}$ (all spectra in GNPS), Spectra_{4} (four low-resolution GNPS data sets from *S. roseosporus*, *Bacillus* and *Pseudomonas* cultures, and two wild-type isolates), $\text{Spectra}_{\text{High}}$ (high-resolution GNPS data sets $\text{Spectra}_{\text{Fungi}}$, $\text{Spectra}_{\text{Acti}}$, $\text{Spectra}_{\text{Pseu}}$ and $\text{Spectra}_{\text{Cyan}}$ containing spectra from Fungi, Actinomycetales, Pseudomonas and Cyanobacteria, respectively), and $\text{Spectra}_{\text{Acti36}}$ (36 subsets of the $\text{Spectra}_{\text{Acti}}$ data set that contain bacterial extracts from 36 strains with known genome). Details of these data sets and the number of

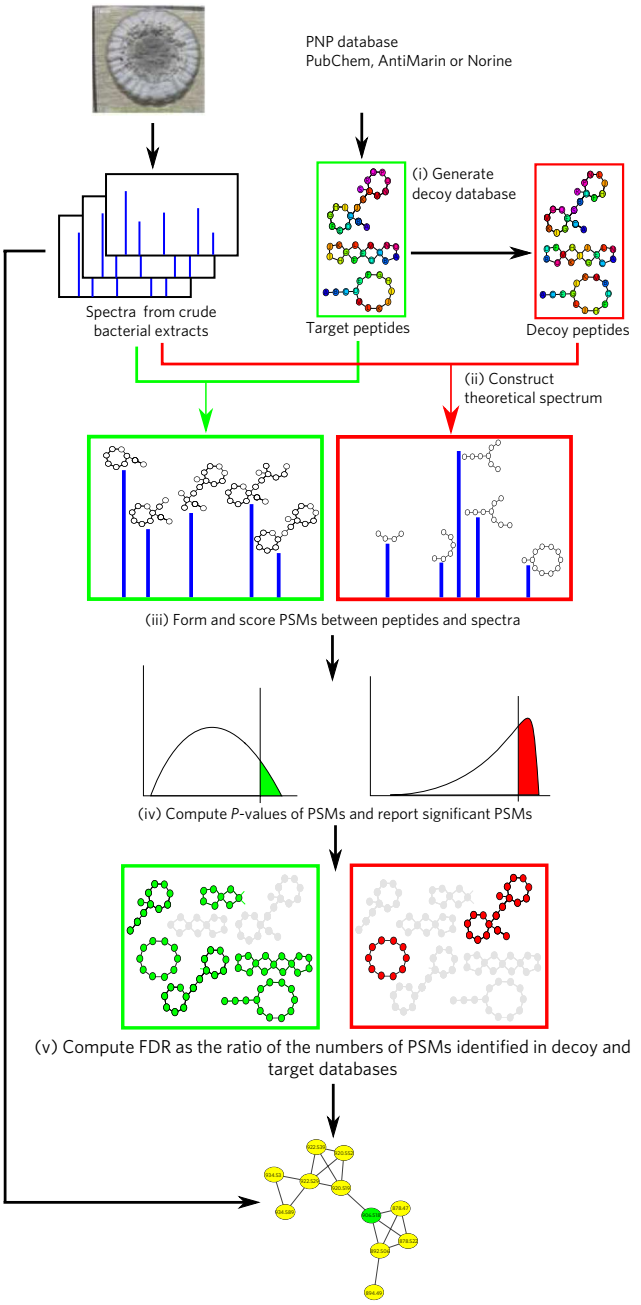

Figure 1 | DEREPLICATOR pipeline. DEREPLICATOR pipeline includes the following steps: (i) generating decoy database of PNPs (ii) constructing theoretical spectra for all PNPs in the database, (iii) generating and scoring PSMs, (vi) computing *P* values of PSMs and generating the set of statistically significant PSMs, (v) computing false discovery rate, and (vi) enlarging the set of found PSMs through variable dereplication via spectral networks. Various steps related to target and decoy databases are shown in green and red boxes, respectively. Six peptides identified in target database and two peptides identified in decoy database are shown in green and red, respectively.

PNPs in various chemical databases are in **Supplementary Tables 1** and **2**, and **Supplementary Figure 3**.

Analyzing statistical significance of identified PNPs

The crucial element of any MS/MS database search is analysis of statistical significance by computing *P* values (for individual PSMs) and

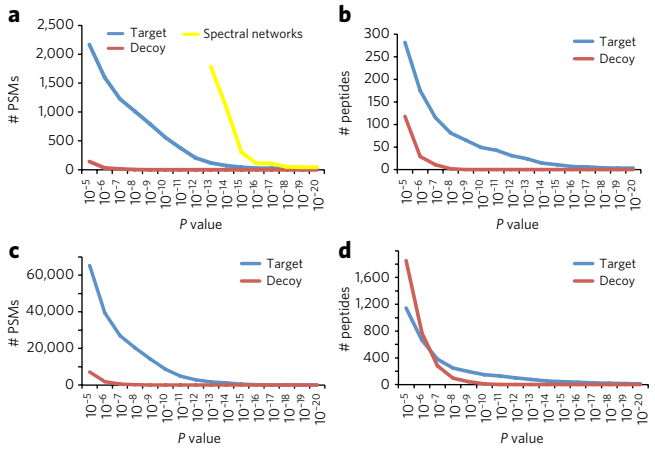

Figure 2 | Number of PSMs and peptides identified by DEREPLICATOR. (**a**,**b**) For each *x* (shown as *P* value along the *x* axis), the plots show the number of identified PSMs or peptides with *P* values below *x*. Number of PSMs (**a**) and peptides (**b**) for the target AntiMarin and decoy databases in the search of Spectra$_4$. 1,787 PSMs and 180 unique PNPs with *P* value below 10^{-13} were dereplicated via spectral networks. (**c**,**d**) Number of PSMs (**c**) and peptides (**d**) for the target AntiMarin and decoy databases in the search of Spectra$_{GNPS}$. All searches were performed with the precursor mass tolerance 0.05 Da.

FDRs (for the entire set of identified PSMs). To compute *P* values, DEREPLICATOR uses the MS-DPR (mass spectrometry direct probability redistribution) algorithm[26], motivated by a similar approach in particle physics. To compute the FDR, DEREPLICATOR uses the concept of decoy database and extends it to nonlinear peptides (Online Methods). We note that FDR in proteomics is estimated as the ratio of the numbers of identified PSMs (rather than peptides) in the decoy database and the target database. In this paper, we took a more conservative approach by reporting the ratio of the number of identified unique peptides in the decoy and the target database.

We analyzed the distribution of *P* values of PSMs and peptides identified by DEREPLICATOR in the search of Spectra$_4$ and Spectra$_{GNPS}$ against the target AntiMarin database and decoy database of the same size (**Fig. 2**). For Spectra$_{GNPS}$ data set and *P* value threshold of 10^{-10}, DEREPLICATOR identified 8,622 PSMs (150 unique peptides) in the target database and 22 PSMs (11 unique peptides) in the decoy database. This translates into 0.2% FDR at the PSM level and 7.3% FDR at the peptide level. The *P* value cutoff of 10^{-10} was two orders of magnitude more stringent than the median *P* value of the manually curated PNP spectra in the GNPS spectral library[8], the reference library of all annotated MS/MS data in GNPS. We thus project that hundreds of PNPs reported below represent a fraction of PNPs whose reference spectra have been already deposited to GNPS.

Although none of the decoy PSMs in the Spectra$_4$ data set had a *P* value below 10^{-11} (estimated FDR is zero), there were 374 PSMs in the target database with *P* values varying from 10^{-11} to 10^{-27}. These PSMs correspond to 37 unique PNPs (**Table 1**). We found only two PSMs in the decoy database with *P* values below 10^{-8} as compared to 904 PSM (and 78 peptides) in the target database with *P* values below 10^{-8}. Although 78 identified PNPs were also represented by reliable PSMs with low FDR of 0.2%, we limited analysis to even more statistically significant 37 dereplicated peptides in **Table 1** and conducted a literature search for all these peptides.

Validation of DEREPLICATOR results

Using the conservative FDR cutoff, we validated the results of DEREPLICATOR by (i) comparing them with results reported in literature, (ii) comparing them with available spectra from known PNPs, and (iii) analyzing the biosynthetic capacity of the producing organisms.

In **Table 1**, we list 37 PNPs (13 cyclic and 24 branch cyclic) from PSMs identified by DEREPLICATOR with *P* values below 10^{-11}. To validate them, we surveyed the papers reporting spectra of these PNPs. As spectra for most of these PNPs are only available as images in journal papers (rather than computer files), we were limited to comparing these images with spectra in Spectra$_4$ data set by eye. For 35 of 37 PNPs, an MS/MS of the peptide had been published in the literature, and a visual comparison confirmed that the dereplicated PNPs were correct.

We further analyzed the species that gave rise to the PNPs in **Table 1**. If these species were evolutionary close to the PNP-producing species reported in the published papers, we considered that as additional evidence supporting the dereplicated PNPs. For 31 of 37 PNPs, information about the PNP producer was available, and for all of them, the referenced paper reported that these PNPs are produced by an evolutionary close bacterial species. Overall, 36 of 37 PNPs in **Table 1** are supported by at least one of these two tests. Presence of multiple PNPs from the same PNP family in **Table 1** (for example, eight variants of surfactin) was additional evidence that these PNPs were correctly identified.

To further evaluate PSMs identified by DEREPLICATOR, we compared spectra for these PSMs to the annotated spectra in the GNPS spectral library[8] that currently contains only 81 PNPs and includes only 21 of 37 PNPs listed in **Table 1**. Moreover, only 18 of these 21 PNP in GNPS spectral library had spectra generated with the same type of instrument (linear trap quadrupole, Fourier transform ion cyclotron resonance; LTQ-FTICR) as the spectra in Spectra$_4$ data set, and all these 18 spectra in GNPS turned out to be similar to spectra in the Spectra$_4$ data set, with cosine values varying from 0.4 to 0.8 (cosine value for spectra from different peptides is expected to be close to 0).

Dereplication of the entire GNPS molecular network

As GNPS is often missing information on whether a specific spectral data set was acquired using a low- or high-resolution instrument, we analyzed all spectra in GNPS in the low-resolution mode. For Spectra$_{GNPS}$ data set and the *P* value threshold 10^{-11}, DEREPLICATOR identified 4,892 PSMs (129 peptides) in the target database and 8 PSMs (3 peptides) in the decoy database. In **Supplementary Table 3**, we list the 129 identified PNPs (71 cyclic, 41 branch cyclic and 17 linear) that include 47 peptides, 81 lipopeptides and a hybrid polyketide-peptide. When we used the spectral network for variable dereplication, the number of identified PSMs for the entire GNPS data set (at *P* value 10^{-11}) increased to 69,995 (see **Supplementary Fig. 2** for examples). About 75% of PNPs from AntiMarin listed in **Supplementary Table 3** have variant PNPs (as revealed by neighbors in the GNPS molecular network), bringing to light a remarkable diversity of previously unreported PNP variants (**Supplementary Table 4**).

To further evaluate the PNPs found via variable dereplication, we analyzed the mass shifts of PNP variants from **Supplementary Table 4** as compared to known PNPs from **Supplementary Table 3**. If the new PNP variants are correct, then we expect them to have many characteristic mass shifts such as 14 Da, a change of CH_2 (ref. 27). The histogram of mass shifts of PNP variants (**Supplementary Fig. 4**) illustrates that a large fraction of them (~40%) have characteristic mass shifts of 14 Da, 17 Da, 18 Da, 28 Da, 30 Da, 42 Da and 113 Da. The spectral network of stenothricins illustrates how analysis of spectral networks and characteristic mass shifts revealed new members of the PNP families (**Supplementary Fig. 2**). Indeed, the mass shift 7 Da connecting the three known stenothricins, also connects stenothricin IV with a node 573.808 Da. As the spectra in this connected component originate from doubly charged ions, 7 Da corresponds to the characteristic mass shift of 14 Da.

Table 1 | 37 PNPs (in the increasing order of *P* values) identified by DEREPLICATOR in the search of Spectra$_4$ against AntiMarin database for *P* value threshold 10^{-11}

#	Organism	GNPS identifier	PNP	Structure	Category	*P* value	SPC score	# peaks	# bonds	# variants	Library search (instrument)	Producer/ref.
1	*B. clausii*	78552	Bacitracin A	Bcyc	Peptide	2.0×10^{-26}	25	100	11	1	n/a	*Bacillus*[39]
2	*P. tolaasii* CH36	78552	Tolaasin I	Bcyc	Lipo	3.4×10^{-22}	21	76	18	1	qTof	*P. tolaasii*[40]/[41]
3	*P. tolaasii* CH36	78552	Tolaasin B	Bcyc	Lipo	2.5×10^{-21}	22	149	18	1	qTof	*P. tolaasii*[41]/[41]
4	*S. roseosporus*	78577	Daptomycin	Bcyc	Lipo	6.3×10^{-19}	25	125	13	1	0.55 (LTQ)	*S. roseosporus*[27]
5	*B. subtilis* NCIB 3610	78552	Surfactin B	Cyc	Lipo	1.8×10^{-18}	18	70	7	3	0.77 (LTQ)	*B. subtilis*[42]
6	*Streptomyces*	78557	Surfactin variant	Cyc	Lipo	5.6×10^{-18}	18	149	7	1	0.75 (LTQ)	-/[42]
7	*P. tolaasii* CH36	78552	Tolaasin C	Bcyc	Lipo	1.9×10^{-17}	15	155	19	1	n/a	*P. tolaasii*[41]/[41]
8	*B. subtilis* subsp. spizizenii	78552	Mycosubtilin III	Cyc	Lipo	1.4×10^{-16}	14	75	8	1	n/a	*B. subtilis*[43]/-
9	*S. roseosporus*	78577	Stenothricin IV	Bcyc	Lipo	1.7×10^{-16}	24	90	9	4	0.53 (LTQ)	*Streptomyces*[44]/[27]
10	*B. subtilis* NCIB 3610	78552	Surfactin variant	Cyc	Lipo	3.4×10^{-16}	19	70	9	3	0.77 (LTQ)	*B. subtilis*[42]
11	*B. subtilis* NCIB 3610	78552	Plipastatin variant	Bcyc	Lipo	3.9×10^{-16}	24	115	10	1	n/a	*B. subtilis*/[45]
12	*Streptomyces*	78557	Glumamycin	Bcyc	Lipo	1.2×10^{-15}	25	90	12	2	n/a	-/-
13	*B. subtilis* NCIB 3610	78552	Surfactin A1	Cyc	Lipo	4.5×10^{-15}	15	70	7	1	0.77 (LTQ)	*B. subtilis*[42]
14	*Streptomyces*	78557	Valinomycin	Cyc	Peptide	6.3×10^{-15}	6	75	6	12	0.71 (hFT)	-/[46]
15	*B. subtilis* NCIB 3610	78552	Plipastatin variant	Bcyc	Lipo	1.2×10^{-14}	26	115	10	1	n/a	*B. subtilis*[45]/[45]
16	*B. subtilis* NCIB 3610	78552	Surfactin D	Cyc	Lipo	2.3×10^{-14}	17	75	7	3	n/a	*B. subtilis*[42]
17	*B. subtilis* NCIB 3610	78552	Surfactin variant	Cyc	Lipo	2.7×10^{-14}	16	70	7	3	0.60 (LTQ)	*B. subtilis*[42]
18	*S. roseosporus*	78577	A21978 C2	Bcyc	Lipo	2.8×10^{-14}	24	140	13	2	0.51 (LTQ)	*S. roseosporus*[27]
19	*S. roseosporus*	78577	Stenothricin I	Bcyc	Lipo	3.0×10^{-14}	21	90	9	4	0.43 (LTQ)	*Streptomyces*[44]/[27]
20	Unknown	78607	Kurstakin 2	Bcyc	Lipo	4.2×10^{-14}	7	60	7	7	n/a	-/[47]
21	*S. roseosporus*	78577	A21978 C3	Bcyc	Lipo	4.3×10^{-14}	18	120	13	2	0.51 (LTQ)	*S. roseosporus*[27]
22	*B. subtilis* NCIB 3610	78552	Surfactin variant	Cyc	Lipo	5.2×10^{-14}	16	70	7	1	0.77 (LTQ)	*B. subtilis*[42]
23	*S. roseosporus*	78577	Stenothricin III	Bcyc	Lipo	5.2×10^{-14}	23	90	9	1	0.64 (LTQ)	*Streptomyces*[44]/[27]
24	*S. roseosporus*	78577	A21978 C1	Cyc	Lipo	5.7×10^{-14}	30	135	13	2	0.54 (LTQ)	*S. roseosporus*[27]
25	*B. subtilis* NCIB 3610	78552	Surfactin variant	Bcyc	Lipo	1.3×10^{-13}	14	65	7	1	0.77 (LTQ)	*B. subtilis*[42]
26	*P. fluorescens* BW10S2	78552	Massetolide F	Bcyc	Lipo	1.8×10^{-13}	14	90	9	1	qTof	*P. fluorescens*[48]/[49]
27	*B. licheniformis*	78552	Bacitracin B3	Bcyc	Peptide	3.5E-13	21	115	11	1	n/a	*Bacillus*[39]
28	*B. subtilis* NCIB 3610	78552	Surfactin variant	Cyc	Lipo	3.9×10^{-13}	14	70	7	3	n/a	*B. subtilis*[42]
29	Unknown	78607	Kurstakin 1	Bcyc	Lipo	8.7×10^{-13}	7	60	7	7	n/a	-/[47]
30	*B. cereus*	78552	Kurstakin 4	Bcyc	lipo	1.6×10^{-12}	7	108	7	5	n/a	*Bacillus*[50]/[47]
31	*Streptomyces*	78557	Lichenysin G5a	Cyc	Lipo	1.9×10^{-12}	16	120	7	3	n/a	-/[42]
32	*B. pamilus*	78552	Surfactin variant	Cyc	Lipo	2.7×10^{-12}	15	75	7	1	n/a	*B. subtilis*[42]
33	*B. subtilis* NCIB 3610	78552	Plipastatin B2	Bcyc	Lipo	3.1×10^{-12}	25	122	10	1	0.80 (LTQ)	*B. subtilis*[45]/[45]
34	*S. roseosporus*	78577	Stenothricin II	Bcyc	Lipo	3.4×10^{-12}	22	90	9	4	0.40 (LTQ)	*Streptomyces*[44]/[27]
35	*B. subtilis* NCIB 3610	78552	Plipastatin variant	Bcyc	Lipo	3.8×10^{-12}	26	115	10	1	n/a	*B. subtilis*[45]/[45]
36	*B. amyloliquefaciens* FZB42	78552	Plipastatin A2	Bcyc	Lipo	5.8×10^{-12}	24	120	10	1	0.75 (LTQ)	*B. subtilis*[45]/[45]
37	*B. subtilis* NCIB 3610	78552	Plipastatin A1	Bcyc	Lipo	6.8×10^{-12}	23	115	10	1	n/a	*B. subtilis*[45]/[45]

The precursor mass tolerance was set to 0.05 Da. The 'organism' column indicates the species present in one of four GNPS data sets contributing to Spectra$_4$ (if known). GNPS data sets MSV000078552 (*Bacillus* and *Pseudomonas* cultures), MSV000078557 (Chinese marine strains), MSV000078577 (*S. roseosporus*) and MSV000078607 (Cubist strains) are referred to as data sets 78552, 78557, 78577 and 78607, respectively. Genomes of the producer organisms are known for the first two data sets but are not available for the last two data sets. *B.*, *P.* and *S.* stand for *Bacillus*, *Pseudomonas* and *Streptomyces*, respectively. The remaining columns specify the PNP from AntiMarin, structure (cyclic (cyc) or branch cyclic (bcyc)), category (peptide or lipopeptide), *P* value, shared peak count (SPC) score, the number of peaks in the spectrum, the number of generalized peptide bonds, the number of PNP variants identified through analysis of the spectral network, and information about the GNPS spectral library search that includes the cosine value and the instrument type (if PNP is present in the spectral library). The final column provides a reference to a paper that contains an image of a spectrum from the PNP (if available) and information from that reference about the species producing this PNP (if available). Since for tolaasins and massetolide (rows 2, 3 and 26), spectra in Spectra$_4$ data set and GNPS spectral library were collected with different instruments (LTQ-FTICR and qTof, respectively), we did not report their cosines. LTQ-FTICR and hybrid FT are abbreviated as LTQ and hFT, respectively. All spectra in Spectra$_4$ were collected on ThermoFinnigan LTQ instrument with electrospray ionization, linear ion trap analyzer, CID activation, and electron multiplier detector. Data not available are indicated by n/a.

Dereplication of the GNPS spectral library

To further validate DEREPLICATOR, we analyzed all 81 annotated and manually curated spectra of PNPs in the GNPS spectral library. 40 of 81 PSMs formed by PNPs in this library had low *P* values (below 10^{-8}) that DEREPLICATOR usually considers as reliable PSMs (21 of 81 are represented by very low-quality spectra with *P* values above 10^{-4}). Thus, all PSMs reported in this paper represent much higher quality spectra than most (41 of 81) spectra in the manually curated GNPS spectral library (their *P* values were at most 10^{-11}, three orders of magnitude lower than the median *P* value in the GNPS spectral library).

DEREPLICATOR correctly identified all 40 high-quality spectra in the GNPS spectral library. Even with extremely high *P* value threshold of 10^{-4}, DEREPLICATOR correctly identified 58 of 81 spectra in the GNPS spectral library. This analysis illustrates that the low *P* value threshold 10^{-11} that we used in this analysis is conservative and that GNPS is likely to contain spectra representing thousands more variant PNPs.

Dereplication of short PNPs

As spectra of short peptides have less information content (smaller number of fragment ions matching theoretical spectra) than long peptides, their *P* values are typically larger. As the result, a typical cutoff for the size of the peptide in proteomics is 7 amino acids (6 amide bonds) as otherwise the FDR exceeds the acceptable threshold. As illustrated in **Supplementray Table 2**, for $Spectra_{High}$ data set with the FDR threshold set to 0%, the default mode of DEREPLICATOR identified 6, 11, 19, 51 and 213 PNPs with 2, 3, 4, 5 and 6 or more bonds, respectively (**Supplementary Fig. 5**).

To improve identification of short PNPs, we implemented a special mode of DEREPLICATOR optimized for identification of short PNPs (Online Methods). As the result, the number of identified short PNPs with less than 6 bonds increased from 125 to 193 PNPs at FDR 15%, and the percentage of AntiMarin compounds discovered in the $Spectra_{High}$ data set (analyzed in **Supplementary Table 5**) increased to 9% for 6 bonds or more, 14% for 5 bonds, 3% for 4 bonds and 2% for 3-bond compounds, out of all AntiMarin compounds. Note that DEREPLICATOR generates a theoretical spectrum for each PNP (including short PNPs) by considering generalized peptide bonds that include N-C-O linkage amide bonds as well as C-C-O linkage bonds between thiazoles/oxazoles and dehydroalanines/dehydrobutyrines and other amino acids (**Supplementary Fig. 6**).

High-resolution versus low-resolution MS/MS for PNP discovery

We searched the $Spectra_{High}$ data set against the target AntiMarin database and decoy database of the same size and identified 5,109 PSMs (325 PNPs) in the target database and 59 PSMs (42 PNPs) in the decoy database at the *P* value threshold 10^{-10}. Note that, for the same data set, the number of identified PNPs in the low-resolution mode reduced from 325 to 79 as compared to the high-resolution mode (with 2 PSMs and 2 peptides identified in the decoy database in the low-resolution mode). *P* values in the high-resolution mode were typically at least five orders of magnitude lower than *P* values in the low-resolution mode (**Supplementary Table 6**). The fact that the high-resolution spectra are vastly superior to the low-resolution spectra with respect to nonlinear PNP identification (fourfold increase in the number of identified PNPs) is surprising because the difference between the high-resolution and the low-resolution spectra with respect to identification of linear peptides in proteomics is not so large (20–30% increase[28]).

To validate PNPs identified in the $Spectra_{High}$ data set, we analyzed their distributions between Fungi, Actinomycetales, Pseudomonas and Cyanobacteria. According to AntiMarin, most (167 of 180) of the PNPs identified from the $Spectra_{Fungi}$ data set had been first reported from fungal sources. Similarly, most (53 of 64) of the peptides identified from the $Spectra_{Cyan}$ data set had been first reported from cyanobacterial sources. Some of 13 peptides identified from $Spectra_{Fungi}$ data set and forming PSMs with nonfungal sources, are clearly not false identifications; for example, all four Pseudomonas peptides are variants of massetolide (it is unlikely that four spurious PSM originate from the same PNP family). There are a few reasons why spectra from $Spectra_{Fungi}$ data sets form PSMs with peptides from bacterial sources apart from being false PSMs: for example, laboratory contamination and morphology misidentification as many collections contain misidentified organisms.

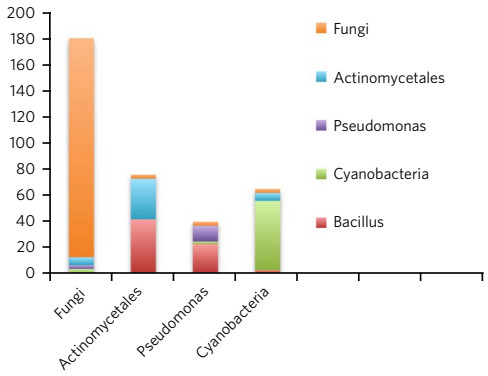

Figure 3 | Number of peptides identified by DEREPLICATOR in $Spectra_{High}$ data set. The number of unique peptides identified from Fungal, Actinomycetales, Pseudomonas or Cyanobacteria spectral data sets (horizontal axis), coming from Fungal, Actinomycetales, Cyanobacteria or Bacillus sources (color coding). As *B. subtilis* was added to the extracts from the samples $Spectra_{Acti}$ and $Spectra_{Pseu}$, 42 and 22 peptides from *Bacillus* sources identified in $Spectra_{Acti}$ and $Spectra_{Pseu}$ represent contaminants. Since *Bacillus* growth medium is similar to that of *Actinomycetes* and *Pseudomonas*, samples from *Actinomycetes* and *Pseudomonas* often have small *Bacillus* contaminations that originate from pre-autoclaving growth in the media.

A similar analysis of $Spectra_{Acti}$ and $Spectra_{Pseu}$ data sets should be done with caution as *Bacillus subtilis* was added as true positive to these samples. As a result, 42 and 22 peptides from *Bacillus* sources were identified in $Spectra_{Acti}$ and $Spectra_{Pseu}$, respectively. After removing surfactins (typically associated with *Bacillus* species), most of the peptides identified in $Spectra_{Acti}$ (31 of 35) and $Spectra_{Pseu}$ (12 of 18) had Actinomycetales and Pseudonomas sources, respectively (**Fig. 3**). It further suggests that metabolite origin tracking using DEREPLICATOR can become a useful tool for capturing contamination or incorrect sample labeling.

Using DEREPLICATOR to optimize sample preparation

The data set $Spectra_{Acti36}$ was collected under three different growth conditions and extracted in three different ways. DEREPLICATOR can screen the output of the experiment and reveal promising versus not-so-promising experimental conditions (microorganisms can produce different PNPs under different conditions). We used DEREPLICATOR to investigate which of nine combinations of growth conditions and extraction methods performs the best for the PNP discovery. In addition to nine pairs (strain, peptide) shown as blue squares in **Supplementary Figure 7**, DEREPLICATOR also found surugamide in 2 of 36 strains bringing the maximum possible number of pairs (strain, peptide) to 11 for each of nine possible conditions. In **Supplementary Figure 8** we illustrate that butanol extract from A1 agar led to the recovery of 10 of 11 (90%) such pairs, making it the most efficient combination.

Cross-validating genome mining and peptidomics results

We further cross-validated PNPs identified by DEREPLICATOR from $Spectra_{Acti36}$ data set partitioned into 36 subsets[13,29–31]. Since we had two independent approaches (mass spectrometry and

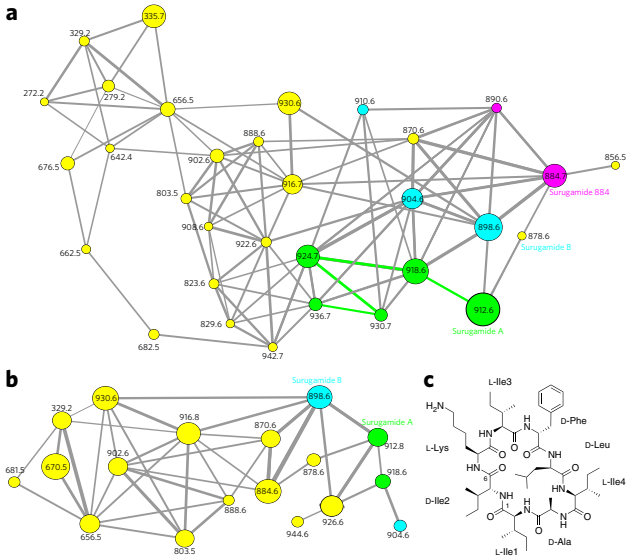

Figure 4 | Spectral networks illustrating the results of the SILAC experiment. (**a**) Spectral network of surugamides from *S. albus* J1074 when the strain was labeled by $^{13}C_6$ isoleucines. A path connecting five green nodes reveals surugamide A (911.621 Da, observed at *m/z* of 912.610) and four SILAC incorporations into isoleucine with characteristic 6 Da mass shifts (surugamide A has four isoleucines, which are observed as addition of 6 Da, 12 Da, 18 Da and 24 Da to the precursor ion). Blue nodes reveal incorporations in surugamide B with three isoleucines (897.605 Da, observed at *m/z* of 898.611), and purple nodes reveal incorporations in a previously unknown surugamide variant with two isoleucines (*m/z*, 884.589). (**b**) Spectral network of surugamides from *S. albus* J1074 when the strain was labeled by $^{13}C_6$ lysine. Green and blue nodes reveal SILAC incorporations into a single lysine in surugamides A and B. Sizes of the nodes reflect relative abundance based on total intensity of the ion that was fragmented. Width of the edges connecting the nodes reflects the similarity (cosine score) between corresponding spectra. As we used a stringent cosine threshold 0.7, some related spectra are not connected by edges. (**c**) Structure of surugamide A.

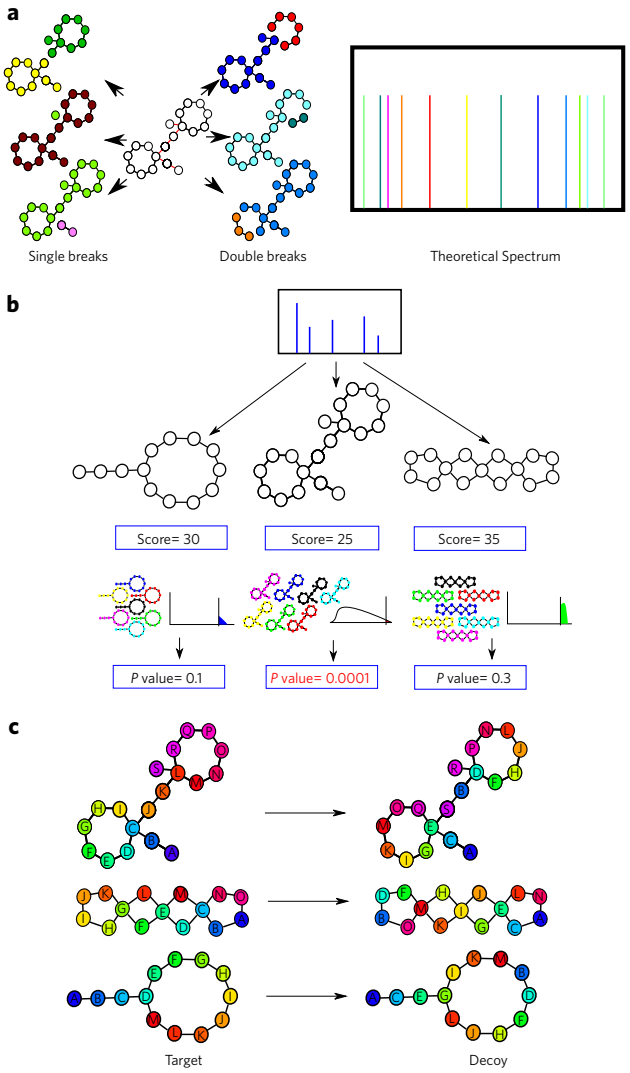

Figure 5 | Generating theoretical spectra and computing *P* values of PSMs formed by PNPs with various architectures. (**a**) Generating the theoretical spectrum of a branch-cyclic peptide (only 12 of 90 peaks in the theoretical spectrum are shown). Nodes and edges in the PNP graph are shown as circles and lines. Bridges are shown as red edges. Intensities of all peaks in the theoretical spectrum are the same since prediction of intensities remains an open problem. (**b**) MS-DPR[26] explores a large set of peptides (enriched for high-scoring peptides) to accurately estimate *P* values. Each such set is illustrated as a collection of seven peptides, each with a different shuffled sequence of amino acids. (**c**) Constructing decoy database of PNPs by randomly rearranging amino acids while preserving the architecture of a PNP.

genome mining) to check whether a given strain produces a given PNP, we could cross-validate the results. At a *P* value threshold of 10^{-10}, DEREPLICATOR identified 9 PNPs in 8 of 36 strains in these data sets (grisemycin, calcium-dependent antibiotic (CDA), daptomycin, actinomycin, stendomycin, cyclomarin, salinamide, arylomycin and surugamide).

We extracted the biosynthetic gene cluster for 8 of these 9 PNPs from the database minimum information about a biosynthetic gene cluster (MIBiG)[32] (the biosynthetic gene cluster for surugamide remains unknown) and performed a BLAST search of the 36 actinomycetales against these gene clusters. This search revealed that, in the majority of cases, when DEREPLICATOR reported evidence for production of a chemotype in a specific strain, genome mining also predicted the corresponding genotype in the same strain, thus providing additional support for both peptides identified by DEREPLICATOR and for MIBiG predictions (**Supplementary Fig. 7**).

DEREPLICATOR found surugamide in four GNPS data sets from *Streptomyces albus* J1074 generated by independent studies[13,29,30], and in a data set from *Streptomyces* sp. CNY228. The utility of DEREPLICATOR is illustrated by the surprising fact that all previous studies did not identify surugamides in *S. albus* J1074, a workhorse strain for *Streptomyces* synthetic biology and heterologous expression[33].

Validating surugamide compounds

Surugamide[34] and the related molecules champacyclin[35] and reginamide[25] are recently discovered NRPs from marine streptomyces that share the same amino acid sequences. Because multiple pieces of bioinformatics evidence pointed to production of surugamides in *S. albus* J1074, we set out to validate them experimentally. Although the NRP synthetase responsible for synthesizing surugamides remains unknown, our analysis identified a putative surugamide-encoding NRP synthetase using a peptidogenomics approach[36] (this computational hypothesis needs to be experimentally validated; see **Supplementary Note** for details).

To demonstrate that the molecules corresponding to the identified spectra are indeed surugamides, we conducted a stable isotope labeling with amino acids in cell culture (SILAC) experiment with *S. albus* J1074 sample and analyzed the resulting spectral network. The SILAC experiments revealed each incorporated amino

acid as a characteristic 6 Da mass shift in the corresponding mass spectrum.

When we cultured *S. albus* J1074 in the presence of $^{13}C_6$ -labeled lysine, we observed the addition of $^{13}C_6$ -labeled lysine for surugamide A and surugamide B, supporting that these nodes in the spectral network represent surugamides (**Fig. 4a**). When we cultured *S. albus* J1074 in the presence of $^{13}C_6$-labeled isoleucine, we observed four additional nodes in the spectral network corresponding to the addition of $^{13}C_6$-labeled isoleucines (**Fig. 4b**). A path consisting of four edges (**Fig. 4b**) revealed incorporation of up to four $^{13}C_6$-labeled isoleucines represented by nodes with *m/z* 918.647, 924.668, 930.688 and 936.686. Further inspection of these spectra revealed incorporation into each of the four isoleucine positions in the surugamide A structure (**Fig. 4c**). For surugamide B with three isoleucines, the spectral network revealed addition of up to two $^{13}C_6$-labeled isoleucines.

In summary, our SILAC experiments supported the incorporation of the four isoleucines and the lysine, together with the adenylation domain specificity and location of the epimerase domains in the biosynthetic gene cluster that we predicted. These experiments, together with the fact that no other gene cluster in *S. albus* J1074 has propensity to produce surugamide, support DEREPLICATOR identifications.

As further confirmation that the identified spectrum in the extract of *S. albus* J1074 was surugamide A, we compared the retention time and the spectrum of *m/z* 912.627 observed in the extract of *S. albus* J1074 with the previously purified and NMR-spectroscopy-characterized authentic standard of surugamide A[34]. Both the retention time and the spectra of authentic surugamide A and putative surugamide A detected in the extract of *S. albus* J1074 were nearly identical (**Supplementary Fig. 9**). Furthermore, when we added the authentic surugamide A to the extract of *S. albus* J1074, we observed a single peak at *m/z* 912.627, further supporting that the detected molecule in the extract of *S. albus* J1074 was surugamide A.

DISCUSSION

Although molecular networks for PNP discovery recently gained a lot of momentum[24,37], they require time-consuming manual follow-up analysis to transform cryptic information into identified spectra of known compounds or their variants. Thus, the shortage of computational tools for PNP analysis is the key bottleneck for taking advantage of the wealth of PNPs in various species.

Currently, over 98% of spectra in the GNPS molecular networking infrastructure represent 'dark matter of metabolomics'[38] since they evaded all attempts to interpret them[8]. However, much of this dark matter is likely formed by spectra from known molecules present in chemical databases. As the result, there is a contrast between the large number of known structures of natural products and rather small number of their annotated spectra in the GNPS spectral library. Therefore, to fully use the potential of the GNPS project, the development of algorithms for matching millions (and soon billions) of spectra of natural products against chemical databases is needed. In the 'living data' concept, public data is periodically reanalyzed and new findings are relayed back to biologists who contributed specific data sets. Although DEREPLICATOR can be run as a standalone search through GNPS by generating theoretical spectra and computing *P*-values of PSMs formed by PNPs (**Fig. 5**), it is now also run on each newly deposited public data set in the GNPS to perform both standard and variable dereplication, making it a part of the 'GNPS living data'.

Because it is impractical to validate annotations of millions of spectra with isolation and NMR spectroscopy analysis, the only feasible way forward is to develop a measure of statistical confidence of PSMs with respect to the core structure of PNPs (as MS is blind to stereochemistry). Although such measures are widely used in proteomics and genomics, they are currently missing in the field of natural products. To address the challenge of evaluating the statistical significance of PSMs identified by DEREPLICATOR, we complemented it with *P* values and demonstrated that PSMs with low *P* values represent confident spectral identifications with low FDR.

DEREPLICATOR is to our knowledge the first software tool in the field of natural products that is compatible with high-throughput analysis of millions of spectra and aimed at reducing the peptidic fraction of the 'dark matter of metabolomics'. Although it has limitations with respect to analyzing short PNPs, it has already increased the size of the publicly available GNPS spectral library of PNPs by an order of magnitude. We envision that DEREPLICATOR will be used to prioritize strains and molecules in natural-product discovery programs, to discover analogs of known natural products, and to reveal biosynthetic promiscuity, intermediates and shunt products.

URLs. DEREPLICATOR is available as both a stand-alone tool (http://cab.spbu.ru/software/dereplicator) and a web application (http://gnps.ucsd.edu).

Received 3 December 2015; accepted 17 August 2016; published online 31 October 2016

METHODS

Methods and any associated references are available in the online version of the paper.

References

1. Li, J.W. & Vederas, J.C. Drug discovery and natural products: end of an era or an endless frontier? *Science* **325**, 161–165 (2009).
2. Fischbach, M.A. & Walsh, C.T. Antibiotics for emerging pathogens. *Science* **325**, 1089–1093 (2009).
3. Ling, L.L. *et al.* A new antibiotic kills pathogens without detectable resistance. *Nature* **517**, 455–459 (2015).
4. Harvey, A.L., Edrada-Ebel, R. & Quinn, R.J. The re-emergence of natural products for drug discovery in the genomics era. *Nat. Rev. Drug Discov.* **14**, 111–129 (2015).
5. Donia, M.S. & Fischbach, M.A. Small molecules from the human microbiota. *Science* **349**, 1254766 (2015).
6. Medema, M.H. & Fischbach, M.A. Computational approaches to natural product discovery. *Nat. Chem. Biol.* **11**, 639–648 (2015).
7. Walsh, C.T. A chemocentric view of the natural product inventory. *Nat. Chem. Biol.* **11**, 620–624 (2015).
8. Wang, M. *et al.* Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nat. Biotechnol.* **34**, 828–837 (2016).
9. Vaniya, A. & Fiehn, O. Using fragmentation trees and mass spectral trees for identifying unknown compounds in metabolomics. *Trends Analyt. Chem.* **69**, 52–61 (2015).
10. Mohimani, H. & Pevzner, P.A. Dereplication, sequencing and identification of peptidic natural products: from genome mining to peptidogenomics to spectral networks. *Nat. Prod. Rep.* **33**, 73–86 (2016).
11. Marahiel, M.A., Stachelhaus, T. & Mootz, H.D. Modular peptide synthetases involved in nonribosomal peptide synthesis. *Chem. Rev.* **97**, 2651–2674 (1997).
12. Arnison, P.G. *et al.* Ribosomally synthesized and post-translationally modified peptide natural products: overview and recommendations for a universal nomenclature. *Nat. Prod. Rep.* **30**, 108–160 (2013).
13. Mohimani, H. *et al.* Automated genome mining of ribosomal peptide natural products. *ACS Chem. Biol.* **9**, 1545–1551 (2014).
14. Smith, C.A. *et al.* METLIN: a metabolite mass spectral database. *Ther. Drug Monit.* **27**, 747–751 (2005).
15. Yang, J.Y. *et al.* Molecular networking as a dereplication strategy. *J. Nat. Prod.* **76**, 1686–1699 (2013).
16. Balkovec, J.M. *et al.* Discovery and development of first in class antifungal caspofungin (CANCIDAS®)—a case study. *Nat. Prod. Rep.* **31**, 15–34 (2014).
17. Blunt, J., Munro, M. & Laatsch, H. Antimarin database. University of Canterbury; Christchurch, New Zealand: University of Gottingen; Gottingen, Germany, (2007).
18. Eng, J.K., McCormack, A.L. & Yates, J.R. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **5**, 976–989 (1994).
19. 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).

20. Ng, J. *et al.* Dereplication and *de novo* sequencing of nonribosomal peptides. *Nat. Methods* **6**, 596–599 (2009).
21. Ibrahim, A. *et al.* Dereplicating nonribosomal peptides using an informatic search algorithm for natural products (iSNAP) discovery. *Proc. Natl. Acad. Sci. USA* **109**, 19196–19201 (2012).
22. Bandeira, N., Tsur, D., Frank, A. & Pevzner, P.A. Protein identification by spectral networks analysis. *Proc. Natl. Acad. Sci. USA* **104**, 6140–6145 (2007).
23. Bandeira, N. Spectral networks: a new approach to de novo discovery of protein sequences and posttranslational modifications. *Biotechniques* **42**, 687–691 (2007).
24. Watrous, J. *et al.* Mass spectral molecular networking of living microbial colonies. *Proc. Natl. Acad. Sci. USA* **109**, E1743–E1752 (2012).
25. Mohimani, H. *et al.* Multiplex *de novo* sequencing of peptide antibiotics. *J. Comput. Biol.* **18**, 1371–1381 (2011).
26. Mohimani, H., Kim, S. & Pevzner, P.A. A new approach to evaluating statistical significance of spectral identifications. *J. Proteome Res.* **12**, 1560–1568 (2013).
27. Liu, W.T. *et al.* MS/MS-based networking and peptidogenomics guided genome mining revealed the stenothricin gene cluster in Streptomyces roseosporus. *J. Antibiot. (Tokyo)* **67**, 99–104 (2014).
28. Kim, S. & Pevzner, P.A. MS-GF+ makes progress towards a universal database search tool for proteomics. *Nat. Commun.* **5**, 5277–5286 (2014).
29. Duncan, K.R. *et al.* Molecular networking and pattern-based genome mining improves discovery of biosynthetic gene clusters and their products from Salinispora species. *Chem. Biol.* **22**, 460–471 (2015).
30. Traxler, M.F., Watrous, J.D., Alexandrov, T., Dorrestein, P.C. & Kolter, R. Interspecies interactions stimulate diversification of the Streptomyces coelicolor secreted metabolome. *MBio* **4**, e00459–13 (2013).
31. Penn, K. & Jensen, P.R. Comparative genomics reveals evidence of marine adaptation in Salinispora species. *BMC Genomics* **13**, 86 (2012).
32. Medema, M.H. *et al.* Minimum information about a biosynthetic gene cluster. *Nat. Chem. Biol.* **11**, 625–631 (2015).
33. Zaburannyi, N., Rabyk, M., Ostash, B., Fedorenko, V. & Luzhetskyy, A. Insights into naturally minimised Streptomyces albus J1074 genome. *BMC Genomics* **15**, 97 (2014).
34. Takada, K. *et al.* Surugamides A-E, cyclic octapeptides with four D-amino acid residues, from a marine streptomyces sp.: LC-MS-aided inspection of partial hydrolysates for the distinction of D- and L-amino acid residues in the sequence. *J. Org. Chem.* **78**, 6746–6750 (2013).
35. Pesic, A. *et al.* Champacyclin, a new cyclic octapeptide from Streptomyces strain C42 isolated from the Baltic Sea. *Mar. Drugs* **11**, 4834–4857 (2013).
36. Kersten, R.D. *et al.* A mass spectrometry-guided genome mining approach for natural product peptidogenomics. *Nat. Chem. Biol.* **7**, 794–802 (2011).
37. Bouslimani, A. *et al.* Molecular cartography of the human skin surface in 3D. *Proc. Natl. Acad. Sci. USA* **112**, E2120–E2129 (2015).
38. da Silva, R.R., Dorrestein, P.C. & Quinn, R.A. Illuminating the dark matter in metabolomics. *Proc. Natl. Acad. Sci. USA* **112**, 12549–12550 (2015).
39. Govaerts, C. *et al.* Sequencing of bacitracin A and related minor components by liquid chromatography/electrospray ionization ion trap tandem mass spectrometry. *Rapid Commun. Mass Spectrom.* **17**, 1366–1379 (2003).
40. Nutkins, J.C. *et al.* Structure determination of tolaasin, an extracellular lipodepsipeptide produced by the mushroom pathogen, *Pseudomonas tolaasii* Paine. *J. Am. Chem. Soc.* **113**, 2621–2627 (1991).
41. Bassarello, C. *et al.* Tolaasins A–E, five new lipodepsipeptides produced by *Pseudomonas tolaasii*. *J. Nat. Prod.* **67**, 811–816 (2004).
42. Gonzalez, D.J. *et al.* Microbial competition between *Bacillus subtilis* and *Staphylococcus aureus* monitored by imaging mass spectrometry. *Microbiology* **157**, 2485–2492 (2011).
43. Peypoux, F. *et al.* Revised structure of mycosubtilin, a peptidolipid antibiotic from *Bacillus subtilis*. *J. Antibiot. (Tokyo)* **39**, 636–641 (1986).
44. Hasenböhler, A., Kneifel, H., König, W.A., Zähner, H. & Zeiler, H.J. 134. Mitteilung. Stenothricin, ein neuer Hemmstoff der bakteriellen Zellwandsynthese (Metabolic products of microorganisms. 134. Stenothricin, a new inhibitor of the bacterial cell wall synthesis.). *Arch. Microbiol.* **99**, 307–321 (1974).
45. Tsuge, K., Ano, T., Hirai, M., Nakamura, Y. & Shoda, M. The genes degQ, pps, and lpa-8 (sfp) are responsible for conversion of *Bacillus subtilis* 168 to plipastatin production. *Antimicrob. Agents Chemother.* **43**, 2183–2192 (1999).
46. Sheil, M., Kilby, G., Curtis, J., Bradley, C. & Derrick, P. Low-energy tandem mass spectra of the cyclic depipeptide valinomycin—a comparison with four-sector tandem mass spectra. *J. Mass Spectrom.* **28**, 574–576 (2005).
47. Bumpus, S.B., Evans, B.S., Thomas, P.M., Ntai, I. & Kelleher, N.L. A proteomics approach to discovering natural products and their biosynthetic pathways. *Nat. Biotechnol.* **27**, 951–956 (2009).
48. Gerard, J. *et al.* Massetolides A-H, antimycobacterial cyclic depsipeptides produced by two pseudomonads isolated from marine habitats. *J. Nat. Prod.* **60**, 223–229 (1997).
49. Reybroeck, W. *et al.* Cyclic lipodepsipeptides produced by *Pseudomonas* spp. naturally present in raw milk induce inhibitory effects on microbiological inhibitor assays for antibiotic residue screening. *PLoS One* **9**, e98266 (2014).
50. Hathout, Y., Ho, Y.P., Ryzhov, V., Demirev, P. & Fenselau, C. Kurstakins: a new class of lipopeptides isolated from *Bacillus thuringiensis*. *J. Nat. Prod.* **63**, 1492–1496 (2000).

Acknowledgments

We thank M. Wang and N. Bandeira for insightful suggestions on using molecular networking and spectral library search, and M. Medema for guidelines on running ant-iSMASH. The work of H.M., P.D. and P.A.P. was supported by the US National Institutes of Health (grant 2-P41-GM103484). P.D. is supported by GM097509. A.G., A.M. and P.A.P. were supported by Russian Science Foundation (grant 14-50-00069).

Author contributions

H.M. and A.G. implemented DEREPLICATOR algorithm. H.M., A.G. and A.M. designed the webserver. N.G. and L.-F.N. collected and analyzed mass spectrometry data and conducted SILAC experiments. A.N. and K.T. purified standard surugamide. P.C.D. and P.A.P. designed and directed the work. H.M. and P.A.P. wrote the manuscript.

Competing financial interests

The authors declare competing financial interests: details are available in the online version of the paper.

Additional information

Any supplementary information, chemical compound information and source data are available in the online version of the paper. 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.A.P.

ONLINE METHODS

Constructing theoretical spectra of PNPs. DER\EPLICATOR generates a theoretical spectrum for each PNP by first constructing a PNP graph with amino acids as nodes and generalized peptide bonds as edges (**Fig. 5a**). Here generalized peptide bonds include N-C-O linkage amide bonds as well as C-C-O linkage bonds common among peptides containing thiazole, oxazole, dehydroalanine and dehydrobutyrines (**Supplementary Fig. 6**). The notion of generalized peptide bonds is useful as illustrated by identification of the thiazole-oxazole-containing PNP plantazolicin from *B. amyloliquefaciens*, lanthipeptide SapB from *S. coelicolor* and complex PNPs such as two-ring-containing actinomycin from *Streptomyces* sp. CNS654 (**Supplementary Fig. 6**).

A generalized peptide bond is called a bridge if removing the bond disconnects the PNP graph. Although theoretical spectra of linear peptides are generated by removing all bridges (single bonds), spectra of nonlinear peptide are generated by removing some bond pairs. A pair of bonds is called a 2-cut if none of them are bridges but removing both of them disconnects the graph. The theoretical spectrum of a peptide consists of the masses of all subgraphs resulting from removal of a bridge or a 2-cut from the PNP graph. We refer to the theoretical spectrum of '*Peptide*' as '*Spectrum*(*Peptide*)'.

Generating and scoring PSMs. DEREPLICATOR compares each spectrum in the spectral data set against each peptide in the chemical database. It further forms a PSM if the precursor mass of the spectrum matches the molecular mass of the peptide (up to a predefined maximum error). As DEREPLICATOR only compares a spectrum against all peptides with similar mass, the number of peptides to compare against is much smaller than the PNP database size. We scored a PSM formed by *Peptide* and *Spectrum* using PepNovo[51] and Mass Spectrometry Generating Function (MS-GF+)[28]. Note that *Spectrum* represents the experimental mass spectrum, in contrast to *Spectrum(Peptide)*, which represents the theoretical spectrum of *Peptide*.

SPC score(*Peptide*, *Spectrum*) was defined as the shared peak count, the number of peaks shared between *Spectrum*(*Peptide*) and the filtered version of *Spectrum* as defined by PepNovo[52,53]. Two peaks are shared if their masses are within a predefined threshold. Although we used SPC score to summarize the statistics of found PSMs, DEREPLICATOR uses a more advanced MSGF score[28]. Admittedly, as 'score' was initially developed for linear peptides, it does not take into account the specifics of fragmentation of nonlinear peptides. However, it performed better than SPC score in our database searches.

Computing *P* values of PSMs. The PSM scores poorly correlated with *P* values of PSMs[23] and thus should not be used for evaluating the statistical significance of found PSMs. Indeed, the PSM scores do not remove the bias toward large PNPs, PNPs with different architectures (for example, linear versus cyclic), or spectra with many peaks. Although methods for evaluating statistical significance of linear PSMs are well developed[54], they do not extend to the evaluation of the statistical significance of nonlinear PSMs.

Estimating *P* values of PSMs is a difficult instance of a general problem of estimating the probabilities of extremely rare events. For linear peptides, the generating function approach[55] efficiently explores the huge set of all possible peptides (rather than relatively small set of all peptides in the database[23]) to derive *P* values for PSMs. MS-DPR algorithm[26] for computing *P* values for PSMs formed by nonlinear peptides is motivated by a similar approach in particle physics[56]. MS-DPR evaluates *P* values based on exploring various peptides that are not present in the peptide database and addresses an important problem of deciding whether a spectrum was generated by a linear, cyclic or branch-cyclic peptide[26] (**Fig. 5b**). This feature (that was missing in previous approaches) is important for analyzing large data sets containing spectra of PNPs with various structures (linear, cyclic and branch cyclic). DEREPLICATOR reports PSMs with *P* values below a predefined threshold and informally defines the *P* value of a peptide as the minimum *P* value of all PSMs formed by this peptide.

Generating decoy database of PNPs. To compute the FDR, DEREPLICATOR uses the concept of decoy database[57] and extends it to nonlinear peptides. For each PNP in the chemical database (denoted *Peptides*), DEREPLICATOR constructs a decoy PNP with the same topology but randomly rearranged amino acids (**Fig. 5c**). The resulting set of PNPs forms a decoy database '*DecoyPeptides*'.

Computing false discovery rate. Given a chemical database *Peptides*, a spectral data set *Spectra*, and a score threshold *T*, DEREPLICATOR finds all high-scoring PSMs, i.e., all PSMs formed by a peptide *P* from *Peptides* and a spectrum *S* from *Spectra* with MSGF score $(P, S) \geq T$. This approach is analogous to the peptide identification approach in proteomics. DEREPLICATOR further computes *P* values of all high-scoring PSMs using MS-DPR and forms the list of the high-scoring PSMs in the increasing order of their *P* values. Given a *P* value threshold θ, we defined PSM_θ (*Peptides*, *Spectra*) as the set of all PSMs in this list with *P* values below θ.

To evaluate the statistical significance of PSMs found in proteomics searches, researchers report FDR that estimates the fraction of false PSM among all reported PSMs. The target-decoy approach[57] for estimating FDR is based on generating a decoy proteome and searching all spectra against both the target and decoy proteomes. The target-decoy approach further uses the number of PSMs found in the decoy proteome to evaluate FDR. As the decoy proteome is generated randomly, we expect to find very few PSMs in PSM_θ (*DecoyPeptides*, *Spectra*) for an appropriately chosen *P* value threshold θ. We thus compute the FDR as the ratio of the number of identified PSMs in the decoy and target proteomes:

$$\mathrm{FDR}_\theta = |\mathrm{PSM}_\theta(DecoyPeptides, Spectra)| / |\mathrm{PSM}_\theta(Peptides, Spectra)|$$

Variable dereplication of PNPs via spectral networks. Ideally, each PNP family corresponds to a connected component in the spectral network. However, spurious edges in spectral networks often connect unrelated spectra from different PNP families, making it difficult to perform variable dereplication. To minimize the number of spurious edges, DEREPLICATOR uses a stringent threshold for defining spectral pairs (edges in the spectral network).

DEREPLICATOR constructs the spectral network of a spectral data set and finds connected components in this network (**Supplementary Fig. 2**). We refer to a connected component in a spectral network as a PNP component if one of the spectra (nodes) in this component was identified as a statistically significant PSM. We further use such PSMs to perform the variable dereplication of all spectra in the PNP component[25]. For each PNP derived via variable dereplication, we used MS-DPR to compute its *P* value. The variable dereplication is accepted if the resulting *P* value does not exceed the threshold θ.

Characteristic shifts in spectral networks. As shown in **Supplementary Figure 2**, many edges in the PNP components correspond to the mass shift 14 Da (7 Da for doubly charged ions). Nodes separated by the mass shift of 14 Da is a common feature of molecular networks that often reveals new variants of known compounds[27] (for example, substitutions of isoleucine for valine). This and other common shifts reveal analogs with amino acid substitutions, truncations (in the case of branch cyclic peptides), hydrolysis products, differently sized lipid side chains, glycosylation, methylation and other variant PNPs[27].

For example, the mass shift 14 Da is a characteristic feature of the kurstakin family (and many other PNP families) because it connects some known variants of kurstakins (**Supplementary Fig. 2**). Thus, as kurstakin 4 is connected by the 14 Da shift to a node with mass 920.519 in the spectral network, this node likely represents a still unknown variant of kurstakin. Indeed, as spurious edges in the connected component have spurious mass shifts, it is extremely unlikely that such spurious edges will have mass shifts characteristic for a specific PNP family, The node with mass 934.589 (with the mass shift 14 Da from the node with mass 920.519) may represent a yet another unknown variant of kurstakin.

Identification of short PNPs. In the first approximation, the FDR equals to the *P* value threshold multiplied by the database size to account for multiple-hypothesis testing[54]. For example, in practice, to avoid false identifications, existing MS/MS database search pipelines often discard all PSMs formed by peptides shorter than 7 amino acids while searching bacterial proteomes. It does not mean that identification of such peptides is impossible but rather

means that researchers have no choice but to consider a few such identifications (to be within the given FDR) or relax the FDR beyond the traditional 1–3%. Our computational analysis illustrates that short PNPs are indeed difficult to identify via database search owing to low information content resulting in high *P* values.

To improve identification of short PNPs, we compared the characteristics of PSMs formed by short PNPs identified in AntiMarin with characteristics of PSM identified in the decoy database. This comparison revealed the striking difference: most PSM from short AntiMarin peptides originated from spectra with charge +1 and isotopic shift 0 Da, whereas most PSM from short decoy peptides originated from spectra with charge +2 and +3 and isotopic shifts +1 Da, and +2 Da. Thus, although the search for multicharged spectra and spectra with nonzero isotopic shifts makes sense for long peptides (it increases the number of identified PSMs at the expense of a modest increase in FDR), it is counterproductive for short PNPs (for example, we do not expect short PNPs to result in spectra with isotopic shifts). We thus modified DEREPLICATOR to limit analysis of PSMs formed by short PNPs to only spectra of charge +1 and isotopic shift 0 Da. After this change, most decoy PSMs formed by short PNPs disappeared (without significantly reducing the number of target PSM formed by short PNPs). As the result, at the FDR threshold of 15%, DEREPLICATOR identified 47, 36 and 110 compounds with 3, 4 and 5 bonds, respectively.

Experimental validation of PNPs. We performed SILAC experiments to validate some PNPs identified by DEREPLICATOR. *S. albus* J1074 and *S. albus* ATCC 21838 strains were cultured on ISP2, A1 and R5 agar medium (10 mL) with and without 1 mM of $^{13}C_6$-labeled isoleucine for 6 d at 30 °C. A similar experiment was conducted for $^{13}C_6$ -labeled lysine. Mass spectra from resulting samples were acquired in positive ion mode over a mass range of 100–1,500 *m*/*z* using a QExactive (Thermo Scientific) mass spectrometer with HESI-II probe source (**Supplementary Note**).

Validation of surugamide A. The putative identification of surugamide A, annotated by the DEREPLICATOR in *S. albus* J1074 extract, was validated by comparison of the MS/MS spectrum and retention time with an authentic standard of surugamide A[34], analyzed by liquid chromatography (L)-MS/MS under the same analytical conditions. Furthermore, a comigration assay was performed to control any matrix effect, by spiking the extract of *S. albus* J1074 with the authentic standard of surugamide A. The **Supplementary Note** includes a description of the experimental details.

Revealing the biosynthetic gene cluster for surugamides. The NRP synthetase responsible for synthesizing surugamides remains unknown. Below we describe a method that combines peptidogenomics[36] with DEREPLICATOR to point to the elusive NRP synthetase responsible for surugamide.

Although DEREPLICATOR identified surugamide in multiple *Streptomyces* strains, only one of them (*S. albus* J1074) was assembled into a single scaffold (most other strains were split into over 100 contigs). However, the assembly was performed using a non-reproducible computational protocol, making it difficult to estimate the number of missassemblies. We thus faced the challenge of finding a surugamide-producing NRP synthetase in a genome with potential assembly errors.

Although the NRP synthetase predictor 2 (NRPSpredictor2)[58] identified 36 adenylation domains in *S. albus* J1074, it is unclear which of them code for surugamide. To account for possible assembly artifacts, we focused on triples of consecutive adenylation domains in the genome and further added a constraint that the genomic distance between consecutive domains in a triple should not exceed 20 kb. For each of the 22 triples $A_1A_2A_3$ of adenylation domains satisfying this constraint and for each of 8,000 3-mers $X_1X_2X_3$ of proteinogenic amino acids, we computed $\text{Score}(A_1A_2A_3, X_1X_2X_3) = \text{Score}(A_1, X_1) + \text{Score}(A_2, X_2) + \text{Score}(A_3, X_3)$, where $\text{Score}(A, X)$ is the NRPSpredictor2 score of an adenylation domain A against an amino acid X (the percentage of matches between the 10-residue specificity code of the adenylation domain A and the 'ideal' specificity code of an amino acid X as defined by NRPSpredictor2).

For each of 8,000 3-mers $X_1X_2X_3$, we found the triple of consecutive adenylation domains $A_1A_2A_3$ (among 22 such triples) with maximum score resulting in the histogram shown in **Supplementary Figure 10**. We further defined the *P* value of a 3-mer as the fraction of 3-mers (among 8,000) with this or higher score. For example, the *P* value of Ile,Phe,Leu was 164/8,000 = 0.0205 as its score (250) had rank 164 among the 8,000 3-mers.

The amino acid sequences of surugamide A and surugamide B are IAIIKIFL and IAVIKIFL, respectively. As shown in **Supplementary Figure 10**, it was somewhat surprising, the *P* values of all eight 3-mers forming IAIIKIFL were below the mean *P* value 1/2 (similar result held for IAVIKIFL). To quantify this statistical bias, we defined the bias of a 3-mer as its *P* value divided by 2 and the bias of a peptide as the product of biases of its 3-mers. The bias of IAIIKIFL was 7.4×10^{-7}, and the bias of a random peptide was close to 1, implying that IAIIKIFL is likely to be coded by the adenylation domains in *S. albus* J1074 that generate the high-scoring 3-mers shown by red bars in **Supplementary Figure 10**. Further analysis revealed that these adenylation domains are clustered at the genomic location 2863086-2868922 of *S. albus* J1074.

51. Frank, A.M. Predicting intensity ranks of peptide fragment ions. *J. Proteome Res.* **8**, 2226–2240 (2009).
52. Frank, A. & Pevzner, P. PepNovo: de novo peptide sequencing via probabilistic network modeling. *Anal. Chem.* **77**, 964–973 (2005).
53. Frank, A.M. A ranking-based scoring function for peptide-spectrum matches. *J. Proteome Res.* **8**, 2241–2252 (2009).
54. Gupta, N., Bandeira, N., Keich, U. & Pevzner, P.A. Target-decoy approach and false discovery rate: when things may go wrong. *J. Am. Soc. Mass Spectrom.* **22**, 1111–1120 (2011).
55. Kim, S., Gupta, N. & Pevzner, P.A. Spectral probabilities and generating functions of tandem mass spectra: a strike against decoy databases. *J. Proteome Res.* **7**, 3354–3363 (2008).
56. Kahn, H. & Harris, T. Estimation of particle transmission by random sampling. in *Handbook of Mathematical Functions* Vol. 12 (ed. Abramowitz, M.) 27–30 (National Bureau of Standards, 1951).
57. Elias, J.E. & Gygi, S.P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods* **4**, 207–214 (2007).
58. Röttig, M. *et al.* NRPSpredictor2–a web server for predicting NRPS adenylation domain specificity. *Nucleic Acids Res.* **39**, W362–W367 (2011).

doi:10.1038/nchembio.2219