A proteomics approach to discovering natural products and their biosynthetic pathways

Stefanie B Bumpus1,3,4, Bradley S Evans2–4, Paul M Thomas1,3, Ioanna Ntai1,3 & Neil L Kelleher1–3

Many natural products with antibiotic, anticancer and antifungal properties are synthesized by nonribosomal peptide synthetases (NRPSs) and polyketide synthases (PKSs). Although genome sequencing has revealed the diversity of these enzymes, identifying new products and their biosynthetic pathways remains challenging. By taking advantage of the size of these enzymes (often >2,000 amino acids) and unique marker ions derived from their common phosphopantetheinyln cofactor, we adapted mass spectrometry–based proteomics to selectively detect NRPS and PKS gene clusters in microbial proteomes without requiring genome sequence information. We detected known NRPS systems in members of the genera Bacillus and Streptomyces, and screened 22 environmental isolates to uncover production of unknown natural products from the hybrid NRPS-PKS zwitermicin. A biosynthetic gene cluster. We also discovered an NRPS cluster that generates a seven-residue lipopeptide. This ‘protein-first’ strategy complements bioassay- and sequence-based approaches by finding expressed gene clusters that produce new natural products.

Over half of the nearly 1,000 new chemical entities introduced as antibacterial or anticancer drugs over the 25 years from 1981 to 2006 are natural products or derivatives thereof. However, traditional bioassay-guided discovery strategies, where an iterative cycle of metabolite fractionation and bioassay panels attempts to isolate the chemical compound responsible for the observed bioactivity, tend to rediscover known compounds most of the time. Systems–biology approaches—such as genomics, transcriptomics and metabolomics—are now being adapted to update natural product discovery platforms and bypass this so-called dereplication bottleneck. These strategies continue to evolve as more microbial genomes become available; the availability of more sequenced genomes facilitates use of bioinformatics to predict the organisms’ biosynthetic potentials.

There have been a few successful attempts where sequence-based approaches have successfully guided the search for new natural products. Nonetheless, there remains a great disparity between the genetic potential for natural product production and the actual expression of biosynthetic gene clusters under laboratory culture conditions to produce a useful metabolite. Although several reports demonstrate diverse methods to force expression of ‘cryptic’ gene clusters, accessing novel compounds and the enzymes that make them is still a low-throughput affair. Discovery of secondary metabolites with new scaffolds and/or clinically relevant antibiotic, immunosuppressive and antiproliferative properties remains the primary constraint to progress in this field.

In developing a complementary approach to address these problems, we set out to circumvent the bias that arises when screening against only one drug target or indicator cell line by bioassay-based screening, and to develop a strategy that identifies robustly expressed genes without the requirement for DNA sequence information a priori. Our method, called PrISM (Proteomic Investigation of Secondary Metabolism; Fig. 1), permits targeted detection of peptides produced nonribosomally (NRPs) and polyketides (PKs) produced by NRPS and/or PKS systems with simultaneous identification of the gene cluster responsible for synthesis of the natural product. Tandem discovery of a gene cluster and its associated secondary metabolite can expedite the downstream goal of pathway engineering to improve yield, bioactivity or bioavailability.

Multimodular NRPS and PKS enzymes are enormous (often >200 kDa) and have many domains that act as a molecular assembly-line to create complex natural product scaffolds. The various domains are responsible for substrate activation, condensation and tailoring, while the growing natural product is covalently tethered to carrier regions, also called thiolation (T) domains, that uniformly harbor a phosphopantetheinyl (Ppant) cofactor bound to a serine residue in their active sites.

In even the earliest implementations of mass spectrometry (MS) to detect covalent intermediates on the thioetemplate family of NRPS and PKS enzymes, the facile release of the Ppant cofactor was noted; its phosphodiester linkage is labile during tandem MS (MS/MS) (Supplementary Figs. 1 and 2), which is conceptually similar to the ion chemistry used in modern phosphoproteomics. More recently, mass spectrometry has revealed a great diversity of covalent chemistry occurring on NRPS and PKS enzymes in vitro, with the Ppant ejection assay now established for single enzymes using liquid chromatography–mass spectrometry (LC-MS) (Supplementary Fig. 1). The Ppant ejection assay specifically transforms covalent intermediates attached to large enzymes into small ions where detection by mass spectrometry is both sensitive and accurate. This facilitates the unambiguous assignment of empirical formulae. Benchtop and high-performance Fourier transform mass spectrometers (FTMS)
have both been used to investigate NRPS and PKS intermediates in reconstituted systems. A mass accuracy of <2 p.p.m. of FTMS for ions diagnostic of Ppant (that is, m/z 261.1267 and m/z 359.1036) translates into high selectivity for detection of Ppant-containing peptides in complex proteomes (Fig. 1c, middle) and forms one aspect of the integrated approach described here. Combining the Ppant ejection assay with a gel-based method for selecting high-molecular-weight proteins enables us to directly target NRPS and/or PKS biosynthetic systems in a complex proteome.

In developing the overall PriSM workflow (Fig. 1), we interrogated three systems with increasing complexity (Supplementary Discussion). We first analyzed a di-domain enzyme from the gramicidin S system (PheAT from gramicidin S synthetase A, 70 kDa)\(^{17}\), using shotgun proteomics in conjunction with the Ppant ejection assay. The single tryptic peptide (Asp\(^{564}\)-Lys\(^{575}\), 1,638.70 Da) from PheAT harboring the Ppant arm was detected against the full complement of Escherichia coli proteins (Supplementary Fig. 3). We next analyzed the native producer of gramicidin S, Bacillus brevis, as it entered early stationary phase\(^{17}\). Prior LC-FTMS of a crude extract of B. brevis had verified production of this 10-mer NRPS at this stage of the life cycle. Using this system, we detected four of the five carrier peptides from the GsRA (127 kDa) and GsRB (510 kDa) proteins by shotgun proteomics (Supplementary Fig. 4). High molecular weight bands from SDS-PAGE gels of B. brevis were also analyzed by in-gel digestion and nanocapillary LC-MS/MS (nanoLC-MS) to identify NRPSs encoded by gsaR and gsrb (Supplementary Table 1). A similar overall result was obtained for the phosphinothricin tripeptide system in the substantially more complicated proteome background of the native producer, Streptomyces viridochromogenes DSM 40736 (ref. 18) (Supplementary Fig. 5).

With proof-of-concept experiments in hand for the first half of the PriSM workflow (Fig. 1a–c) in both Gram-positive (e.g., B. brevis and S. viridochromogenes) and Gram-negative bacteria (e.g., E. coli), we set out to apply this method to strains isolated from the environment without the benefit of DNA sequence or other information a priori. We isolated Bacillus spores from heat-treated soil samples\(^{19}\), and stored 22 isolates for proteomic analysis after 16S rDNA sequencing (Supplementary Table 2). The ten NK2018 proteome by strong cation exchange chromatography (LTQ-FT) operating at 12 T. Such analysis of NK2018 detected Ppant-containing peptides, the most prominent being a 2+ peptide at m/z 1,038.98 (2,075.94 Da) that showed all expected Ppant elimination marker ions during MS/MS analysis\(^{14}\). These MS\(^{5}\) experiments on the Ppant-producing peptide provided sufficient de novo sequence information for its identification as the acyl carrier protein (ACP) active-site peptide from fatty acid biosynthesis (Supplementary Fig. 6). The ten amino acid–sequence generated, \([\text{GAD}_{n}\text{pant}(\text{I/L})\text{DVVE(\text{I/J})}}\], was sufficient for differentiation of this peptide as a fatty acyl ACP (AcP), because the sequence motif flanking the active-site serine is distinct from that found in either NRPSs or PKSs\(^{20}\). The identification of this peptide...
provides a positive control for identification of phosphopantetheinylated peptides and de novo generation of long stretches of peptide sequence information, a critical step for design of good primers for PCR.

For the targeted analysis of proteins >200 kDa, we subjected SDS-PAGE gel bands harboring high molecular weight proteins of interest from NK2018 to in-gel trypsin digestion and collected nanoLC-MS data using a 12 T LTQ-FT for high-resolution detection of intact peptides and the phosphopantetheinyl ejection ions. We collected unit-resolution MS/MS data in a data-dependent fashion for the six peptides and the phosphopantetheinyl ejection ions. We collected data using a 12 T LTQ-MS/MS for high-resolution detection of intact proteins from NK2018 to in-gel trypsin digestion and collected nanoLC-MS data.

When we searched MS/MS data collected from all peptides (not just those harboring the Ppant modification), the top three predicted protein identifications were NRPS and hybrid NRPS-PKS proteins from Bacillus cereus AH1134. The genome of this strain was recently sequenced and the annotations are available (accession number NZ_ABDA0000000). The peptides identified arose from expression of two separate gene clusters, called here cluster no. 1 (C1) and cluster no. 2 (C2), producing at least three different NRPSs and or PKSs (C1S2 (ZmaA), C1S6 (ZmaK) and C2S2) (Supplementary Figs. 9 and 10) on two separate contigs (C1 from contig GenBank accession number ABDA02000035 and C2 from contig GenBank accession number ABDA02000007). All peptides with homology to B. cereus AH1134 and predicted to derive from NRPS, PKS or hybrid NRPS-PKS gene products are listed in Supplementary Table 2. The best database search results were manually validated (e.g., Fig. 2d) and provided enough sequence information to design degenerate PCR primers (Supplementary Tables 3 and 4) that ultimately obtained stretches of DNA sequence that were >94% identical to B. cereus AH1134 (Supplementary Table 5).

Figure 2c shows the results of 12 representative PCRs, showing ample microsequence to convert peptide MS/MS data into DNA sequence of expressed NRPS and hybrid NRPS-PKS gene clusters (see Supplementary Table 4 for complete list of all PCR-amplified products and the PCR numbers corresponding to the lanes in c).

The integrated data from PrISM along with targeted PCR (Fig. 2e) show direct evidence for expression of two gene clusters from Bacillus strain NK2018 with the basic architecture of those observed in the newly sequenced B. cereus AH1134; the high sequence identity between the two strains was sufficient for the assumption that the two clusters in this strain are orthologous. The individual domains within the synthetases identified by LC-MS/MS are highlighted in red in Figure 3a and Supplementary Figure 9. A large number of the genes in cluster no. 1 are orthologous to those that produce the aminopolyol antibiotic zwiterminic A. Zwitermicin A has a broad spectrum of activity against both Gram-positive and Gram-negative bacteria as well as certain eukaryotes. A total synthesis was recently reported and the sequence of individual biosynthetic proteins has been reported over the past 4 years. A 2009 report based on the analogous B. cereus AH1134 sequence we found expressed in NK2018 revealed that the zwitermicin A biosynthetic gene cluster is...
Figure 3 Identification of new lipopeptidapeptides in NK2018. (a) Domain organization of cluster no. 2 based upon the gene sequence in *B. cereus* AH1134. Amino acid substrates were selected based upon bioinformatic analysis and the structure of the detected peptides. Peptides from the domains in red were identified by nanoLC-MS/MS. C, condensation domain; A, adenylation domain; T, thiolation domain; TE, thioesterase. (b) Base peak chromatogram (top) of an NK2018 culture supernatant sample and a selected ion chromatogram (SIC; red, bottom) for the species at *m/z* 908.4845. The mass spectrum of the 1+ charge state of the ion is shown (inset). (c) Putative structure for the lipopeptidapeptides detected, including the species detected in b. The structures are predicted to differ in the length of the fatty acid tail and in the formation of a lactone ring (table inset).

**Letters**

much larger than expected and predicted to produce three zwittricin A–related small molecules. Targeted searching for such molecules allowed detection of zwittricin A and a previously undetected methionine-containing NRPS product (Supplementary Fig. 9), thus completing the PrISM approach (cf. Fig. 1f). Additionally, two of the phosphopantetheinylated peptides observed during nanoLC-MS and identified by manual analysis of the high molecular weight peptide in-gel digestion match within 3 p.p.m. to thiolation domain active-site peptides from ZmaB and ZmaK (Supplementary Fig. 8).

We uncovered evidence for expression of a second cluster from *Bacillus* strain NK2018 by detecting peptides from the two-module NRPS protein depicted as C2S2 in Figure 3a. Annotation of the flanking ~50 kb of sequence from *B. cereus* AH1134 around the gene for the protein identified from this cluster is shown in Figure 3a and Supplementary Figure 10. Three genes predicted to encode NRPSs are present in this gene cluster, along with the nearby efflux protein, a phosphopantetheinyl transferase and a type II thioesterase. There are homologs of this gene cluster known in other *B. cereus* strains, such as B4264 and G9842, and multiple *Bacillus weihenstephanensis* strains, but it is clearly an orphan with no corresponding natural product known. As discussed below, the PrISM platform has enabled identification of this new gene cluster by detection of the expressed gene products and linked this expression with secondary metabolite production.

We previously described NRPS rules that make predictions for substrate specificity of adenylation domains based upon characterized domains to predict a seven-residue, NRPS-type natural product with amino acids including serine, alanine and threonine, in addition to two glutamine or glutamic acid residues at the C terminus. Targeted analysis of NK2018 extracts for peptides of this type uncovered a set of six related species at *m/z* 908.4845, 922.5007, 926.4951, 936.5165, 939.5105 and 953.5192 (Supplementary Fig. 12) that were analyzed and sequenced using MS/MS (Supplementary Figs. 13–16). *De novo* sequencing generated a sequence of six amino acids common for substrate specificity of adenylation domains based upon characterization. The intact and fragment masses support the assignments of ring open form (gain of 18 Da) and the lactone ring as drawn in Figure 3c and reported for the homologous kurtоказins (Supplementary Fig. 17). Further, amino acid analogs were present in the threonine residue localizes the 14-Da variations and the hydroxyl group to the fatty acid tail, with the hydroxyl group assigned to position 3 based on the precedents from other *Bacillus* lipopeptides such as surfactin. However, the hydroxyl group is not involved with lactone ring formation, a conclusion supported by detailed interpretation of MS/MS data (Supplementary Figs. 14–16). The structure of these reported natural products and associated bioinformatic analysis strongly support the assignment of cluster no. 2 as the previously unreported biosynthetic gene cluster for these compounds. The unique domain organization of this cluster involves two extra condensation domains, one which we hypothesize to facilitate NRPS initiation by loading the fatty acyl chain in conjunction with the type II thioesterase, analogous to the SrfD protein in surfactin biosynthesis.

MS/MS clearly shows these six species (Supplementary Fig. 12), which differ by exactly 18.0103 Da and 14.0162 Da and have highly related fragmentation patterns (Supplementary Figs. 14–16). These mass differences are unambiguously due to differences of CH_3 and H_2O, which are best explained by incorporation of longer fatty acid chains and lactone ring formation, respectively. The intact and fragment masses support the assignments of ring open form (gain of 18 Da) and the lactone ring as drawn in Figure 3c and reported for the homologous kurtоказins (Supplementary Fig. 17). Further, amino acid analogs were present in the threonine residue localizes the 14-Da variations and the hydroxyl group to the fatty acid tail, with the hydroxyl group assigned to position 3 based on the precedents from other *Bacillus* lipopeptides such as surfactin. However, the hydroxyl group is not involved with lactone ring formation, a conclusion supported by detailed interpretation of MS/MS data (Supplementary Figs. 14–16). The structure of these reported natural products and associated bioinformatic analysis strongly support the assignment of cluster no. 2 as the previously unreported biosynthetic gene cluster for these compounds. The unique domain organization of this cluster involves two extra condensation domains, one which we hypothesize to facilitate NRPS initiation by loading the fatty acyl chain in conjunction with the type II thioesterase, analogous to the SrfD protein in surfactin biosynthesis.

This report extends prior work on isolated NRPS and PKS enzymes and initial reports of microbial proteomics or selective labeling of phosphopantetheinylated proteins into a general method for targeted proteome analysis; NRPS, PKS and fatty acid biosynthetic gene products from diverse organisms are detected with antibody-like specificity for Ppant–containing proteins common to all thiotemplate systems. Detection of a biosynthetic enzyme in a microbial proteome provides a valuable entrée into an unsequenced genome at the protein level, enabling identification of NRPS- and PKS-type natural products and the genes responsible for their synthesis. The integrated PrISM approach augments genomic methods and directs efforts toward expressed genes, a strong indication that the corresponding natural product is also being produced.
Given the potential of different systems biology–based approaches to direct natural products discovery in ways different from classical bioassay-based discovery, it is important to place PrISM in the context of comparable DNA-, RNA- and small molecule–based methods. The rapid increase in genome sequence information unquestionably opens the door for developing sequence-based discovery and characterization methods, such as those involving genome mining, RNA-based analysis and heterologous expression of full gene clusters. But although genome mining provides information on the biosynthetic potential of an organism, it does not reveal which secondary metabolic pathways are actually expressed. In contrast, proteomics permits direct observation of gene expression and detects whether enzymes are correctly post-translationally modified. Although induction of ‘cryptic’ gene clusters and heterologous expression of full pathways has been achieved, these approaches are fairly challenging. Well-developed RNA-based methods for monitoring gene expression, such as RT-PCR or transcriptomics, are not generally used in the context of an unsequenced genome. A structure-based approach using direct chemical screening can reveal small molecules produced by bacterial strains under a variety of growth conditions. It does not, however, provide information on the biosynthetic machinery for those metabolites. We believe that, as technology improves for all-omic analyses, PrISM will fill an important gap and accelerate development of complementary approaches in the systems biology of natural products research.

Challenges in our implementation of PrISM included the translation of sequence information from MS into PCR products and a self-imposed bias toward >100-kDa NRPs, PKSs, and hybrid systems. Future improvements in sample processing and MS should provide increased data quality to facilitate design of degenerate primers. We also note that a global proteomics approach can, in principle, provide a more complete picture of an organism’s biosynthetic capacity. All challenges familiar to natural product structure elucidation still apply to PrISM, including elucidation of complex polyketide structures.

It will be interesting to see how molecular screening methods contribute to natural product discovery as they ramp up for application to hundreds or even thousands of strains. We project that PrISM will be most valuable for screening conditions and strains where novel NRPs/PKS scaffolds are produced. PrISM achieves dereplication at the biopolymer level where relatives of well-characterized systems are quickly flagged (because of high sequence identity to known domains and modules). For such cases, one will elect to simply not continue down the PrISM workflow (Fig. 1) or choose to spot-check the putatively homologous cluster with a small set of easy PCRs.

With PrISM, proteomics performs the initial survey in the natural product discovery pipeline. Subsequent small–molecule–detection modalities, such as bioassays, MS or NMR, can then be used to assess structure and activity. Therefore, we have achieved a net reversal of the traditional “small molecule first” discovery process. Streamlined versions of PrISM will realize efficiencies of scale and detect enzyme fingerprints from cyanobacteria, fungi and even environmental samples subjected to metaproteomics, with extension to all types of secondary metabolism also possible.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturebiotechnology/.

Accession code. The data collected for this manuscript are open access according to the Science Commons CC0 license and can be downloaded from the Tranche network (https://proteomecommons.org/tranche/) using the hashes provided in Supplementary Table 6. These hashes may be used to prove exactly what files were published as part of this manuscript’s data set, and the hashes may also be used to check that the data have not changed since publication.

Note: Supplementary information is available on the Nature Biotechnology website.

ACKNOWLEDGMENTS

We thank W. Metcalf for providing the S. viridochromogenes DSM 40736, B. subtilis ATCC 6633 strains and Taq polymerase, and for, along with W. van der Donk and P. Yau, technical assistance. We would also like to thank D. Dlott and the following members of the Kelleher Research Group for their assistance in this work: L. Zannidou, J. Osugi, J. Norris, J. Anderson and H. Hannon. This work was supported in part by National Institutes of Health (NIH) grants (N.L.K., R01 GM 66725–07, P01 GM 077596–03), NIH Chemistry Biology Interface Training grant (P.M.T.), NIH Molecular Biophysics Training Grant (B.S.E.) and NIH Cell & Molecular Biology Training Grant (S.B.B.). P.M.T. was also supported by an Analytical Chemical Society (ACS)-Division of Analytical Chemistry Fellowship sponsored by Eli Lilly and Co. S.B.B. is currently supported by an ACS-Division of Analytical Chemistry Fellowship sponsored by the Society for Analytical Chemists of Pittsburgh.

AUTHOR CONTRIBUTIONS

S.B.B. designed and performed proteomic analyses, performed gel-based analyses, identified the natural products discussed herein, conducted LC-MS analyses and wrote the paper. B.S.E. isolated and characterized strains, performed gel-based analyses performed LC-MS analyses, designed and executed genomic analyses of NK2018 and wrote the paper. P.M.T. assisted in experimental design and performed gel-based analyses and wrote the paper. L.N. assisted in experimental design and conducted LC-MS analyses. N.L.K. designed experiments and wrote the paper. S.B.B. and B.S.E. contributed equally to this study. All authors discussed the results and commented on the manuscript.

Published online at http://www.nature.com/naturebiotechnology/. Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/.

1. Fischbach, M.A. & Walsh, C.T. Assembly-line enzymology for polyketide and nonribosomal Peptide antibiotics: logic, machinery, and mechanisms. Chem. Rev. 106, 3468–3496 (2006).
2. Zerky, M. & Chalis, G.L. Strategies for the discovery of new natural products by genome mining. Chem. Bio. Chem. 10, 625–633 (2009).
3. Kevany, B.M., Rasko, D.A. & Thomas, M.G. Characterization of the complete zwittiercin A biosynthetic gene cluster from Bacillus cereus. Appl. Environ. Microbiol. 75, 1144–1155 (2009).
4. Newman, D.J. & Cragg, G.M. Natural products as sources of new drugs over the last 25 years. J. Nat. Prod. 70, 461–477 (2007).
5. Weinstein, M.J. & Wagman, G.H. Antibiotics: Isolation, Separation, and Purification (Elsevier Scientific Publishing Company, New York, 1978).
6. Lauru, S., Deeth, R.J., Bailey, L.M. & Chalis, G.L. Discovery of a new peptide natural product by Streptomyces coelicolor genome mining. Nat. Chem. Biol. 1, 265–269 (2005).
7. Knappe, T.A. et al. Isolation and structural characterization of capistruin, a lasso peptide predicted from the genome sequence of Burkholderia thailandensis E264. J. Am. Chem. Soc. 130, 11446–11454 (2008).
8. Zazopoulos, E. et al. A genomics-guided approach for discovering and expressing cryptic metabolic pathways. Nat. Biotechnol. 21, 187–190 (2003).
9. Bode, H.B., Behne, B., Hofs, R. & Zeeck, A. Big effects from small changes: possible ways to explore nature’s chemical diversity. ChemBioChem 3, 619–627 (2002).
10. Brakhage, A.A. et al. Activation of fungal silent gene clusters: a new avenue to drug discovery. Prog. Drug Res. 66, 1–32 (2008).
11. Lambalot, R.H. et al. A new enzyme superfamily-the phosphopantetheinyl transferases. Chem. Biol. 3, 923–936 (1996).
12. Stein, T. et al. Detection of 4’-phosphopantetheinylase at the thyoester binding site for L-valine of gramicidin S synthetase 2. FEBS Lett. 340, 39–44 (1994).
13. Dorrestein, P.C. & Keleher, N.L. Dissecting non-ribosomal and polyketide biosynthetic machineries using electrospray ionization Fourier-Transform mass spectrometry. Nat. Prod. Rep. 23, 893–918 (2006).
14. Dorrestein, P.C. et al. Facile detection of acyl and peptidyl intermediates on thioesterolate carrier domains via phosphopantetheinyl elimination reactions during tandem mass spectrometry. Biochemistry 45, 12756–12766 (2006).
15. Meluzi, D., Zheng, W.H., Hensler, M., Nizet, V. & Dorrestein, P.C. Top-down mass spectrometry on low-resolution instruments: characterization of phosphopantetheinylated carrier domains in polyketide and non-ribosomal biosynthetic pathways. Biomed. Chem. Lett. 18, 3107–3111 (2008).
16. Crawford, J.M. et al. Deconstruction of iterative multidomain polyketide synthase function. *Science* **320**, 243–246 (2008).
17. Matteo, C.C., Glade, M., Tanaka, A., Piret, J. & Demain, A.L. Microbiological studies on the formation of Gramicidin S synthetases. *Biotechnol. Bioeng.* **17**, 129–142 (1975).
18. Blodgett, J.A., Zhang, J.K. & Metcalf, W.W. Molecular cloning, sequence analysis, and heterologous expression of the phosphinothricin tripeptide biosynthetic gene cluster from *Streptomyces viridochromogenes* DSM 40736. *Antimicrob. Agents Chemother.* **49**, 230–240 (2005).
19. Treweeke, R.S., Martin, P.A. & Rechelderfer, C.F. Selective process for efficient isolation of soil *Bacillus* spp. *Appl. Environ. Microbiol.* **53**, 1263–1266 (1987).
20. Mercer, A.C. & Burkart, M.D. The ubiquitous carrier protein—a window to metabolite biosynthesis. *Nat. Prod. Rep.* **24**, 750–773 (2007).
21. Geer, L.Y. et al. Open mass spectrometry search algorithm. *J. Proteome Res.* **3**, 958–964 (2004).
22. Silo-Suh, L.A., Stabb, E.V., Raffel, S.J. & Handelsman, J. Target range of zwittermicin A, an aminopolyol antibiotic from *Bacillus cereus*. *Curr. Microbiol.* **37**, 6–11 (1998).
23. Rogers, E.W., Dalisay, D.S. & Molinski, T.F. (+)-Zwittermicin A: assignment of its complete configuration by total synthesis of the enantiomer and implication of D-serine in its biosynthesis. *Angew. Chem. Int. Edn Engl.* **47**, 8086–8089 (2008).
24. 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).
25. Rausch, C., Weber, T., Kohlbacher, O., Wohlleben, W. & Huson, D.H. Specificity prediction of adenylation domains in nonribosomal peptide synthetases (NRPS) using transductive support vector machines (TSVMs). *Nucleic Acids Res.* **33**, 5799–5808 (2005).
26. 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).
27. Steller, S. et al. Initiation of surfactin biosynthesis and the role of the SrfD-thioesterase protein. *Biochemistry* **43**, 11331–11343 (2004).
28. Schley, C., Altmeier, M.O., Swart, R., Muller, R. & Huber, C.G. Proteome analysis of *Myxococcus xanthus* by off-line two-dimensional chromatographic separation using monolithic poly-(styrene-divinylbenzene) columns combined with ion-trap tandem mass spectrometry. *J. Proteome Res.* **5**, 2760–2768 (2006).
29. Rodriguez-Garcia, A., Barreiro, C., Santos-Benet, F., Sola-Landa, A. & Martin, J.F. Genome-wide transcriptomic and proteomic analysis of the primary response to phosphate limitation in *Streptomyces coelicolor* M145 and in a *DeltaphoP* mutant. *PloS One* **2**, e2410–e2429 (2007).
30. Meier, J.L., Mercer, A.C. & Burkart, M.D. Fluorescent profiling of modular biosynthetic enzymes by complementary metabolic and activity based probes. *J. Am. Chem. Soc.* **130**, 5443–5445 (2008).
31. Lo, I. et al. Strain-resolved community proteomics reveals recombining genomes of acidophilic bacteria. *Nature* **446**, 537–541 (2007).
**Online Methods**

**Materials.** Trypsin (TRL3) for digests of bacterial proteomes in shotgun proteomics was purchased from Worthington Biochemicals. Sequencing-grade trypsin (Promega) was used for all in-gel digestions. *E. coli* BL21(DE3) cells were purchased from EMD Biosciences. All other chemicals used were purchased from either ThermoFisher Scientific or Sigma-Aldrich unless otherwise noted.

**Cloning of GrsA PheAT.** All cloning was performed in *E. coli* strain DH5α. All PCR used Phusion Hot Start Polymerase (Finnzymes) and PCR-grade dNTPs (Invitrogen). Restriction enzymes were obtained from Invitrogen and T4 DNA ligase was from New England Biolabs. PCR products and restriction-digested DNA were purified with Qiaquick gel extraction and PCR cleanup kits (Qiagen). Sfp was amplified with primers F 5′-CCATATGATGAAGATTTA CGGAAATTATATGACGAC-3′ and R 5′-CCTGCGATTCATATTTAAAAAGACTTGTCG TACGAGACC-3′ containing the NdeI and KpnI restriction sites, respectively, (underlined) using the plastid pUC-8 Sfp as the template. The PCR product was cleaned up before digestion with NdeI and KpnI. The linear fragment was purified from a 1% agarose gel before it was ligated to similarly cut and purified PET-Duet-1 previously modified to contain the BamHI to HindIII fragment of pQE-60 (Qiagen) to yield PET-Duet-1-Sfp. PheAT was amplified from plasmid pQE-60 PheATE using primers F 5′-ATATCCATGTTAAACAGCTTCAA-3′ and R 5′-ATGACGTTCATTTGGCTTACCA-3′ containing the NcoI and BamHI restriction sites, respectively (underlined). The PCR product was cleaned up before digestion with NcoI and BamHI and gel purified before it was ligated to similarly cut PET-Duet-1 Sfp to yield PET-Duet-1 PheAT-His6_Sfp. Sequence was confirmed by sequencing at the UIUC Core DNA Sequencing Facility.

**Preparation of samples for proteomic investigations of PheAT.** 100 ml of Luria-Bertani (LB) broth supplemented with ampicillin (final concentration 100 μg/ml) was inoculated with one colony of *E. coli* BL21(DE3) transformed with PET-Duet-1_PheAT-His6_Sfp and grown overnight at 37 °C with shaking at 225 rpm. Ten ml of the starter culture was added to 1 liter of LB supplemented with ampicillin (final concentration 100 μg/ml) and plated at 37 °C with shaking at 225 rpm until an OD600 of ~0.6. At this time, the incubation temperature was dropped to 20 °C and IPTG was added to a final concentration of 1 mM. The culture was incubated with shaking at 18 °C for an additional 20 h, at which time 500 ml cell culture was collected by centrifugation (10 min; 4 °C; 4,400g). Cell pellets were resuspended in lysis buffer (25 mM Tris-HCl, pH 7.5–7.8) and lysed by sonication (4 cycles of 30 s sonication on ice followed by 30 s incubation on ice). The soluble lysate was collected by centrifugation (15 min; 4 °C; 47,810g). Protein concentration of the soluble lysate was determined by the Bradford assay. Five mg of protein was digested in a reaction mixture of 0.05 M NH₄HCO₃ (pH 7.8), 3 M urea and trypsin (ratio of 1:10 trypsin:substrate) by incubating at 30 °C for 20 min, and the reaction was quenched by freezing at –80 °C or addition of SCX solvent A.

**Preparation of S. viridochromogenes samples.** *S. viridochromogenes* DSM 40736 was grown on solid ISP2 medium (Difco) for 4–5 d at 30 °C. One colony was selected and added to 15 ml of MYG media (1 liter contains 10 g malt extract, 4 g yeast extract, 4 g glucose, pH 7.3) in a baffled flask for 4–5 d at 30 °C with shaking at 225 rpm. Seven ml of the starter culture was fully homogenized using a sterile glass homogenizer and added to 1 liter MYG. The culture was incubated at 30 °C with shaking at 225 rpm until significant phosphorothionitrin tripeptide (PTT) production was observed by bioassay. Protocols for performing the bioassay for PTT production have been described previously. In brief, *B. subtilis* ATCC 6633 was grown in minimal media (1 liter contains 3 g KH₂PO₄, 7 g KH₂PO₄, 0.5 g sodium citrate-dihydrate, 0.1 g MgSO₄-7H₂O, 1 g (NH₄)₂SO₄ and 2 g glucose) at 37 °C until an OD₆00 of ~0.4. 200 μl of the culture was plated on minimal media (same recipe as above, with addition of 12 g agar/liter). Six-mm paper disks were placed on top of the plated lawn, and 9 μl of supernatant from *S. viridochromogenes* growth was placed on the disk (9 μl of MYG media was used as a control). The plates were placed at 37 °C for overnight growth and monitoring of PTT production. The bioassays were performed daily after the first overnight growth of the 1 liter cultures. After PTT production was observed, cells were harvested by centrifugation (20 min; 4 °C; 17,600g) and resuspended in lysis buffer (25 mM Tris-HCl, pH 7.5–7.8). Cells were lysed by two passages through a French press operating at high pressure, and the soluble lysate was collected by centrifugation (45 min; 4 °C; 47,810g). Trypsin digests were performed as previously described.

**Isolation of strains of Bacillus and preparation of proteome samples.** Soil was collected from Haughton, Louisiana, USA, and *Bacillus* strains were isolated by heat treatment and dilution plating on nutrient agar. Strain NK2018 was chosen for further analysis based on presence of high molecular weight bands on an SDS-PAGE gel (Supplementary Fig. 7). A starter culture of nutrient broth (50 ml) was inoculated with a single colony of NK2018 and grown overnight. An aliquot (5 ml) was used to inoculate 1 liter nutrient broth for 24 h or 48 h growth at 30 °C. Cells were isolated by centrifugation (10,000g, 10 min, 4 °C) and resuspended in a minimal amount of 100 mM NH₄HCO₃, pH 7.8. Cells were lysed by sonication (5 cycles of 30 s sonication on ice followed by 30 s incubation on ice) and debris was cleared by centrifugation (20 min; 4 °C; 47,810g). Protein concentration was estimated by the BCA assay (ThermoFisher).

**Proteomic analysis of PheAT, B. brevis and S. viridochromogenes.** SCX chromatography was carried out using a Shimadzu Prominence high performance (HP)LC. The column used for SCX analysis was a Polysulfoethyl A column (PolyLC) with 4.6 mm inner diameter and length of 200 mm. An entire tryptic digest was loaded onto the SCX column and eluted with a step gradient (Supplementary Table 7) using solvents A (20 mM citric acid (pH 2.65), 25% acetonitrile) and B (20 mM citric acid (pH 2.65), 1 M NH₄Cl, 25% acetonitrile) flowing at 0.5 ml/min. Fractions were collected every 2 min in 96-well plates and analyzed by FTMS immediately or stored at −20 °C until further analysis. SCX fractions were subjected to reversed-phase (RP)LC-MS/MS according to the following method. Online RPLC-MS/MS data were collected using a ThermoFisher 12 T LTQ-FT Ultra coupled to an Agilent autosampler and Agilent HP1100 binary pump HPLC system. The column used for all RPLC analysis was a Jupiter C18 or C4 1 mm × 150 mm column (Phenomenex). The gradient used for all RPLC analysis is provided in Supplementary Table 8, with solvent A being water + 0.1% formic acid and solvent B being acetonitrile + 0.1% formic acid flowing at 100 μl/min. 150–300 μl of each SCX fraction was injected onto the RPLC column for MS analysis.

**Parameters for MS analysis.** All MS methods included the following events: (i) FT scan, m/z 500–2,000, (ii) FT scan, source-induced dissociation (SID) = 75, detect m/z 200–600, (iii) data-dependent MS/MS on the top X (X = 3 for high-resolution MS/MS data collection and X = 6 or 10 for unit resolution MS/MS data collection) peaks in a given spectrum using collision-induced dissociation (CID) or infrared multi-photon dissociation (IRMPD).
Data analysis and peptide identification for full proteome analysis. All data were analyzed using QualBrowser, part of the Xcalibur software packaged with the ThermoFisher LTQ-FT and custom in-house software. Selected ion chromatograms were generated for the PfPnt ejection ions of interest. Based upon the elution of the PfPnt ejection ion, the time of elution was analyzed for the presence of predicted active-site peptides of the proteins in question (masses calculated based upon published sequences). MS/MS data generated by CID or IRMPD were analyzed manually.

Gel-based proteomic analysis. An aliquot of the soluble proteome (350 µl) was added to 100 µl 2× SDS-PAGE loading buffer and incubated at 95 °C for 5 min before loading onto a BioRad SDS-PAGE gel (Tris-HCl gradient gel, 4–20%, 10 × 30 µl wells). The gel was stained with colloidal Coomassie G-250 and the band at ~225 kDa was excised with a razor blade and chopped into pieces smaller than 2 mm² (Supplementary Figs. 7 and 18). These gel samples were then destained, reduced with DTT, alkylated with iodoacetamide and digested with trypsin. Peptides were extracted, lyophilized, rehydrated with 0.1% acetic acid and bond-loaded onto a self-packed C₄ nano- LC guard column (75 µm × 10 cm, 10- to 20-µm particle size). This guard column was then placed upsteam of a ProteoPepII C₁₈ column (75 µm × 10 cm, New Objective) and peptides were eluted over an 90 min linear gradient of water and acetonitrile with 0.1% formic acid at a flow rate of 300 nl/min (produced by an Eksigent 1D nano-LC) into a 12 T ThermoFisher LTQ-FT Ultra. Samples were analyzed using the online PfPnt ejection assay as well as data-dependent low resolution CID on the top six precursors. LC/MS/MS data from each in-gel digestion run were processed into DTA files with BioWork3 3.2 (ThermoFisher) and concatenated into encapsulated XML. These data were automatically searched (using a custom Perl script) against the nonredundant protein database with the open mass spectrometry search algorithm described in the main text, using standard settings for the detection of intact peptides at high resolution (FTMS) and MS/MS fragment ions with unit resolution (FTMS) (0.01 Da intact peptide tolerance, 0.4 Da fragment ion tolerance). Individual files were then combined into one master file with a custom Perl script for viewing with the open mass spectrometry search algorithm browser. Phosphopantetheinylated peptides were observed as the carboxyamidomethylated-pantetheinyl (Ppant-Cam) ejection products (~318.1482 Da), generated from alkylation of the free Ppant thiol with iodoacetamide during the standard in-gel digestion procedure.

Genomic analysis of NK2018. Genomic DNA was isolated from 5 ml cells (overnight culture at 30 °C in nutrient broth) using a Qiagen DNeasy Blood and Tissue DNA Kit. PCR was performed on BioRad DNA Engine thermal set and primers with the most likely codon usage (predicted from sequenced B. cereus AA1134 (GenBank accession numbers ABDA02000035 and ABDA02000007) and the second being the peptide sequences reported in Supplementary Table 2. Supplementary Table 4 summarizes the 26 PCRs completed in this study, whereas Supplementary Figure 11 is an agarose gel separation of the products of the PCRs. In Supplementary Tables 3 and 4, those rows in white correspond to primers and PCRs completed using B. cereus AA1134 sequence and those rows in gray correspond to primers and PCRs completed using sequence information generated from nano LC-MS analysis. See Supplementary Table 5 for the results of sequencing of selected PCR products.

Identification of NK2018 natural products. A 100 ml starter culture of NK2018 was grown in supplemented M9 minimal medium (per 1 liter: 800 ml H₂O, 200 ml M9 salts (a 1 liter 10x solution contains 64 g Na₂HPO₄·7H₂O, 15 g KH₂PO₄, 2.5 g NaCl and 2.5 g NH₄Cl), 20 ml 20% glucose, 1 ml 1 M MgSO₄·7H₂O, 100 ml 1 M CaCl₂ and 10 ml Difco nutrient broth)) at 30 °C for 2 d. An aliquot of this starter culture was used to inoculate (at a 1:100 dilution) either a 100 ml or 500 ml volume of M9 minimal media (minus nutrient broth) and was grown for 3–10 d at 30 °C. Culture purity was assessed by examination by light microscope and by conducting PCRs no. 3, no. 6, no. 9, no. 14 and/or no. 26 (from Supplementary Table 4, data not shown) on genomic DNA purified from each culture as described. Cultures were harvested by centrifugation (10,000 × g, 10 min) and the culture supernatant filter sterilized. The cell pellets were used to prepare protein samples as described previously for PrISM analysis. These samples were used to confirm the presence of high molecular weight proteins by SDS-PAGE (Supplementary Fig. 18) and for PrISM proteome analysis (using the gel-based proteomic analysis as described above) to confirm the species as NK2018 and to assay production of NRPS- and hybrid NRPS-PKS–related proteins (data not shown). Culture supernatant was concentrated 15– to 20-fold through rotary evaporation and stored at –80 °C until further analysis. LC-MS analysis was conducted on all culture supernatants using a Phenomenex Gemini-NX C18 column (5 µm particle size, 110 Å pore size) with either a 2 mm or 4.6 mm inner diameter. All LC-MS analysis was conducted on a ThermoFisher LTQ-FT operating at 7 T and connected in-line with Surveyor MS pump and autosampler. An external fraction collector was added if HPLC fractions were to be collected simultaneously with MS analysis. For LC-MS analysis, a shallow gradient was used as shown in Supplementary Table 9, where solvent A was H₂O + 0.1% HCOOH and solvent B was acetonitrile + 0.1% HCOOH. Each FTMS analysis included a full scan (m/z 300–2,000) with data-dependent MS/MS on the top three ions in each full scan. The MS/MS data were analyzed using an in-house software package and ThermoFisher Xcalibur QualBrowser. Additional LC-MS separation of these samples was conducted, collecting HPLC fractions at 1-min intervals and drying the fractions under vacuum. Fractions were resuspended in electrospray solution (94% H₂O, 4% methanol, 2% formic acid) and analyzed by direct infusion into a ThermoFisher LTQ-FT Ultra operating at 12 T using a TriVersa robot system for sample delivery. Extensive MS² analysis was conducted on the species of interest to gain as much structural information as possible; this information is summarized in Supplementary Figures 13–16 and Supplementary Table 10.