Supplementary Information

A Proteomics Approach to Discovery of Natural Products and Their Biosynthetic Pathways

Stefanie B. Bumpus[1, 3, 4], Bradley S. Evans[2, 3, 4], Paul M. Thomas[1, 3], Ioanna Ntai[1, 3] and Neil L. Kelleher[1, 2, 3]

[1]Department of Chemistry, [2]Department of Biochemistry & [3]The Institute for Genomic Biology, University of Illinois at Urbana-Champaign, 600 South Mathews Avenue, Urbana, IL 61801. [4]These authors contributed equally to this work.

Supplementary Discussion

Pseudo MS^3 assay

In the course of sample analysis, we have seen occasional peptides which will eject an ion with *m/z* 261.1267 within the SID scan, yet are not related to NRPS/PKS proteins. Although high mass accuracy is sufficient to discriminate most of the di-peptidyl fragments at nominal *m/z* 261, the methionine (Met)-sulfoxide-(leucine (Leu)- isoleucine (|le)) b_2 (or $y_2 – H_2O$) ion has the exact same molecular formula as the Ppant product. To discriminate against detecting this ion, a pseudo MS^3 approach was developed. Supplementary Figure 2 shows the MS^3 fragment ion spectrum of the Ppant ion ejected from an injection of coenzyme A. To decrease the false discovery rate for phosphopantetheinylated peptides even lower, each SID scan is followed by a low resolution SID scan (for speed) where the species at *m/z* 261 is fragmented and selected ion monitoring is employed for the fragment ions at *m/z*'s 243, 184 and 159. Supplementary Figure 2 illustrates this process as applied to the PheAT construct[1]. Supplementary Figure 2 shows co-elution of the low resolution MS^3 ions with the high resolution MS^2 ion seen. A similar assay has been reported for the analysis of recombinant T domain containing protein constructs for analysis in low-resolution MS instruments[2].

Detection of a phosphopantetheinylated peptide from an overexpressed target in the E. coli proteome

As proof-of-concept, a NRPS adenylation-thiolation di-domain (PheAT) involved in activation and incorporation of a phenylalanine residue during biosynthesis of the natural product gramicidin S was co-expressed with the *Bacillus subtilis* phosphopantetheinyl transferase, Sfp[3], in *Escherichia coli*. The resulting *holo* PheAT protein was left unpurified in an *E. coli* whole cell extract, and an aliquot of this proteome was subjected to Bottom Up proteomic analysis using MudPIT-like protocols[4] initiated with a 20 min trypsin digestion with a 1:10 protease:substrate ratio. The entire peptide mixture was separated in the first dimension by strong cation exchange chromatography (SCX), and each SCX fraction analyzed by reverse-phase liquid chromatography (RPLC)-MS using a 12 T hybrid linear ion trap-FTMS (ThermoFisher LTQ-FT). Each LC-MS run has automated fragmentation[5, 6] of ~5 peptides each 10 s, and those ejecting the Ppant cofactor produce a distinctive ion at 261.1267 Da or 359.1031 Da. In one of 20 SCX fractions analyzed a Ppant ejection product was observed that matched within 1 ppm to the predicted mass for a *holo* Ppant ejection ion (Supplementary Fig. 3). When summing across all intact mass scans collected during this time, a peptide was observed with a mass of 1,638.70 Da, matching within 6 ppm of the theoretical mass for the PheAT active site tryptic peptide. Note the baseline of Supplementary Figure 3b. This is a depiction of the specificity of this methodology.

Application of PrISM to gramicidin S biosynthesis in Bacillus brevis

Moving from an overexpressed system in the *E. coli* proteome to a native system in a *Bacillus,* the organism *Bacillus brevis* ATCC 9999 was selected, and has served for years as the prototypical NRPS system[7, 8]. This organism is a native producer of gramicidin S, a cyclic decapeptide synthesized by two NRPSs, GrsA (126 kDa) and GrsB (510 kDa) (Supplementary Fig. 4)[9, 10]. GrsA contains one thiolation domain and incorporates one amino acid monomer, while GrsB contains four thiolation domains and is responsible for incorporation of the remaining amino acids and release of the final natural product (Supplementary Fig. 4a). *B. brevis* was grown in supplemented YP medium at 30°C for 7-8 h[11]. The soluble *B. brevis* proteome was subjected to Ppant-proteomics as above, and each SCX fraction mined for *holo* active site peptides from GrsA and GrsB[4]. Ppant ejection ions were identified for four of the five thiolation domains in these two proteins, and the corresponding active site peptides were identified with 5-15 ppm mass accuracy (Supplementary Fig. 4b-d). These results confirm the ability to identify thiolation domain peptides expressed at endogenous levels in the context of a native proteome. Gramicidin S production at the time of proteomic sampling by PrISM was confirmed by small molecule MS of the culture supernatant (Supplementary Fig. 4e).

Application of PrISM to phosphinothricin tripeptide biosynthesis in Streptomyces viridochromogenes

Moving from the Bacillus genome to an actinomycete, the PrISM platform was applied to the soluble proteome of *Streptomyces viridochromogenes* DSM 40736, a native producer of the herbicide phosphinothricin tripeptide (PTT)[12]. *S. viridochromogenes* was grown in liquid media for 4-5 days under previously determined conditions[13], and PTT production was detected by bioassay against *Bacillus subtilis* ATCC 6633 grown on minimal medium (Supplementary Fig. 5d). The proteome of the organism was collected and subjected to rapid trypsin digestion, followed by SCX and RPLC-MS/MS. Three enzymes with four thiolation domains (PhsA, PhsB and PhsC) are known to be involved in tripeptide formation during PTT biosynthesis[13, 14]. As shown in Supplementary Figure 5b-c, the thiolation domain from PhsA and the N-terminal thiolation domain from PhsB were readily detected. These results illustrate the scalability of PrISM to complex proteomes and its ability to connect the production of a natural product with the expression of its biosynthetic gene cluster, as subsequent small molecule MS also showed detection of PTT (Supplementary Fig. 5e).

A positive control: identification of the fatty acyl ACP (AcpP) in NK2018

In two SCX fractions from NK2018, a 2+ peptide at *m/z* 1,038.98 (2,075.94 Da) was identified as harboring a Ppant arm (Supplementary Fig. 6). To identify this peptide, an aliquot of the SCX fraction was separated by off-line RPLC using the gradient in Supplementary Table 6. Fractions corresponding to the elution time of the Ppant ejection ion observed in the online LC-MS run were mined for a peptide that generated the Ppant ejection marker ions when subjected to MS/MS. The fraction was scanned across the region of *m/z* 500-1,500 for peptides generating the Ppant ejection ion, and isolation of the peptide at *m/z* 1,038.98 generated not only the small molecule Ppant ejection ions of interest but also the two expected peptide marker ions at *m/z* 1,816.9 (*apo* + 80 Da) and *m/z* 1,718.9 (*apo* – 18 Da) (Supplementary Fig. 6a). The species at *m/z* 1,817 was isolated in the ion trap mass spectrometer (ITMS) (Supplementary Fig. 6b), and subjected to MS^3 to generate loss of the phosphate group and leave the *apo* – 18 Da species at *m/z* 1,718 (Supplementary Fig. 6c). The *apo* – 18 Da species was isolated in the ITMS (Supplementary Fig. 6d) and subjected to MS^4 with detection of all fragment ions in the ITMS. Analysis of the resulting spectra, in combination with fragment ions generated from the initial MS^2 of the parent peptide, resulted in *de novo* sequencing of the ten amino acid sequence shown (Supplementary Fig. 6e).

Identification of phosphopantetheinylated peptides from NK2018 by nanoLC-MS

As shown in Figure 2 and Supplementary Figure 8, a phosphopantetheinylated peptide was identified during the nanoLC-MS of the 200 kDa band from the SDS-PAGE separation of the NK2018 soluble proteome. This peptide was identified by analysis of the MS/MS spectra of all peptides co-eluting with the Ppant ejection ion and selected for data dependent MS/MS. One peptide, a 3+ peptide at *m/z* 1,083.53 (3,427.57 Da), showed all four of the predicted marker ions for Ppant ejection, including the peptide marker ions at *m/z* 1,466.4 (the *apo* + 80 Da species) and *m/z* 1,417.8 (the *apo* – 18 Da species). The two small molecule marker ions were observed at *m/z* 318.2 and 416.2 (see Supplementary Fig. 1, masses shifted due to the modification of the available sulfhydryl group with iodoacetemide during preparation for in-gel trypsin digestion). Based upon intact peptide mass values and the presence of the Ppant ejection marker ions, this peptide was identified as a T domain active site peptide from the NK2018 ZmaB-homolog; a second Ppant ejection ion in the same LC-MS analysis was identified as a T domain active site peptide from the NK2018 ZmaK-homolog. This is a standard data analysis procedure that can be used to identify the parent ion of an observed Ppant ejection product, which can then be isolated for more targeted MS^n analysis.

Identification of zwittermicin A and a putative zwittermicin-A related product produced by NK2018

For the identification of zwittermicin A (ZmA) and related products, the LC-MS data files were searched manually for the theoretical masses of ZmA (1+, *m/z* 397.2041) and the ZmA-related product as predicted in recent reports (1+, *m/z* 333.1479)[15]; representative results of these analyses are shown in Supplementary Figure 9c. The masses corresponding to ZmA and the related compound shown were detected in concentrated supernatant from NK2018 grown in M9 minimal medium for 7-10 days, in addition to half-strength trypticase soy broth (0.5X TSB) as previously reported[15].

Analysis of MS/MS data for structure elucidation of NK2018 lipoheptapeptides

Manual interrogation of LC-MS files of culture supernatant from NK2018 grown in M9 minimal medium revealed a set of species (*m/z* 908.4845, 922.5007, 936.5165, 926.4951, 940,5112, and 954.5272) differing in mass by either 14 or 18 Da. As discussed below, detailed MS/MS analysis was conducted on this series of 6 related peptides and implicates the structure shown in Figure 3c. The predicted chemical structures, based on MS/MS data and previous reports of other *Bacillus* lipopeptides, are presented in Supplementary Figure 12. The MS/MS data discussed below are presented in Supplementary Figures 13-16. The parent ion is denoted as M. Please refer to Supplementary Figure 13f for a fragmentation map of the parent ion at *m/z* 926.4951. In addition to the new compounds reported here, the previously reported kurstakins were detected after NK2018 growth in supplemented minimal medium (at all time points analyzed); the intact masses matched to the previously reported lipopeptides and on-line LC-MS/MS analysis of these compounds showed the characteristic marker fragment ion at *m/z* 609[16].

Structure elucidation of the amino acid components

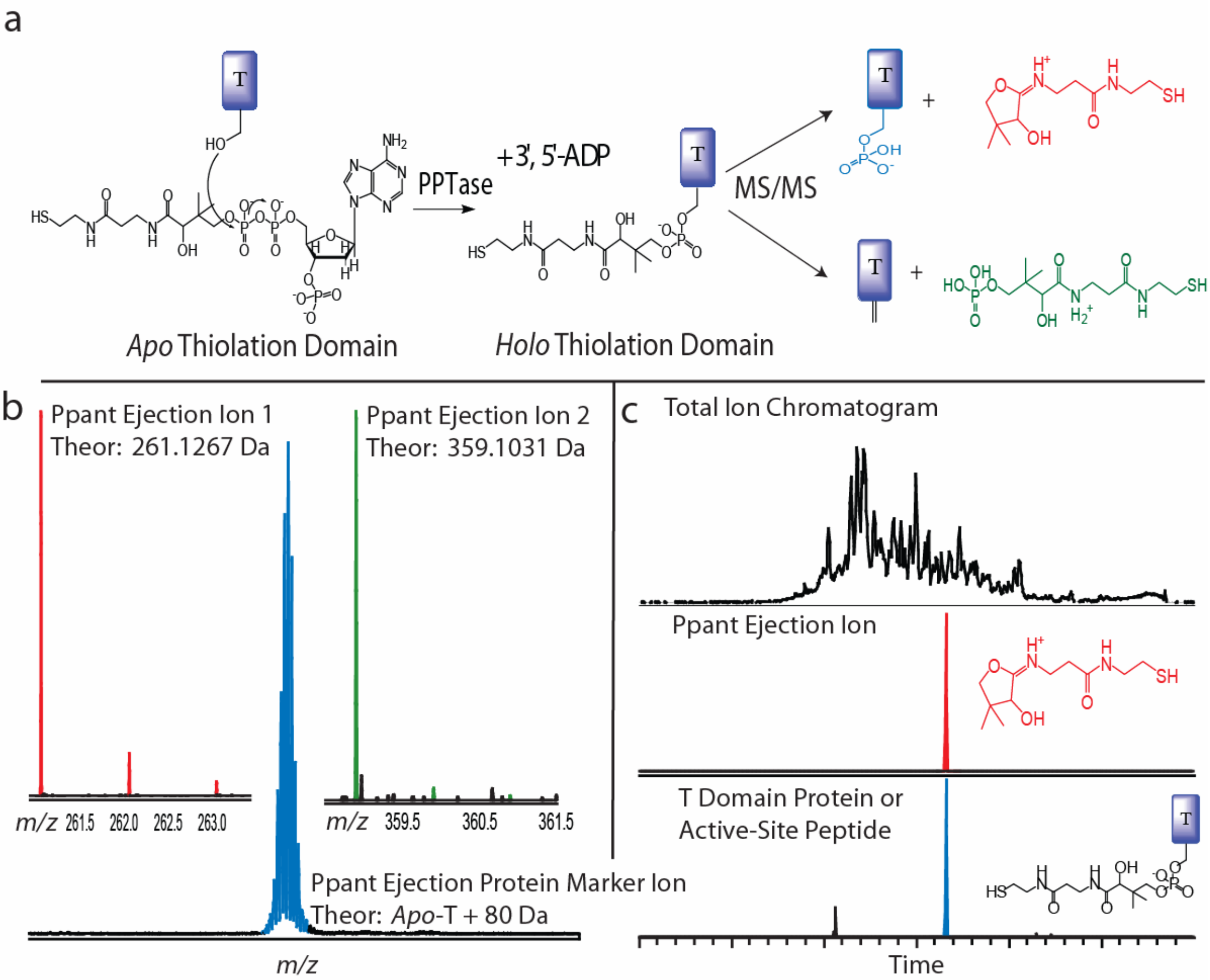
MS/MS on the 6 parent ions (Supplementary Fig. 12a) resulted in several fragment ions, shown in Supplementary Figure 13e. It is important to note that MS/MS data revealed fragment ions with several losses of 17 Da, 18 Da and 28 Da from the loss of H_2O, NH_3 and CO, respectively. This is indicative of a peptide species containing amino acid side chains capable of water (*e.g.* serine (Ser), threonine (Thr)) and/or ammonia (*e.g.* glutamine (Gln)) loss. An abundant ion consistently resulted from the loss of water from the parent ion (M-H_2O), followed by additional losses as described previously. Fragment ions were identified that were equal in mass to M-Gln (b_6-H_2O) and M-Gln-Gln (b_5-H_2O) (in addition to water and ammonia losses), suggesting that these species contain two consecutive Gln residues. These large ions also showed a variance of 14 Da between species; the significance of this will be discussed below.

An additional fragment ion (y_6) was observed at either *m/z* 609.2704 (M = *m/z* 908.4845, 922.5007, or 936.5165) or *m/z* 627.2813 (M = *m/z* 926.4951, 940.5112, 954.5272). When the fragment ions smaller than this ion were examined, a series of ions were observed that were separated by the mass shifts of histidine (His)-Ser-alanine (Ala)-glycine (Gly). The mass of the smallest ion in the MS/MS spectrum from *m/z* 926.4951, resulting from loss of His, had a mass equal to the mass of two glutamine residues. When the chemical structure of an amino acid chain built of these six amino acids (Gln-Gln-His-Ser-Ala-Gly) was examined, it was observed that the masses of the theoretical *y* ions of this peptide matched exactly to those observed in the mass spectra of the parent ions at *m/z* 926.4951, 940.5112, 954.5272. This is clear evidence that these species contain this amino acid sequence. The parent ions at *m/z* 908.4845, 922.5007, 936.5165 are all 18 Da less the other 3 parent ions, indicating that a water loss has occurred in this amino acid sequence (Supplementary Fig. 15). We assign that water loss to the formation of a lactone ring between the Ser residue and the C-terminal Gln residue, similar to that in the kurstakins[16]. The observance of the smallest ion containing 2 Gln residues indicates that these residues correspond to the large fragment ions observed (b_5-H_2O and b_6-H_2O). This leads to the assignment of the low molecular weight ions as a *y*-ion series that can be best explained by a molecule containing the following amino acid sequence: N_t-Gly-Ala-Ser-His-Gln-Gln-C_t.

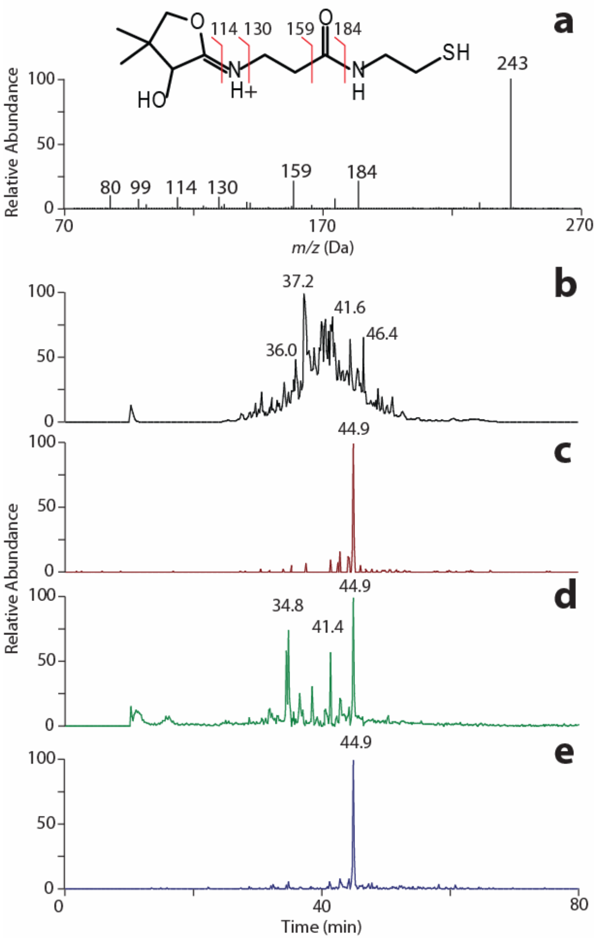
Structure elucidation of the fatty acid-tail containing component

As mentioned previously, the large *b* ions observed in the MS/MS spectra for parent ions at *m/z* 926.4951, 940.5112 and 954.5272 showed a shift of 14 Da (Supplementary Fig. 14); similar shifts were observed for the species at *m/z* 908.4845, 922.5007, and 936.5165. Upon initial observation, this mass shift was attributed to either methylation of the peptide species, incorporation of alternate amino acids that differ by a methylene group, or the addition of a fatty acid tail. Bioinformatics analysis of C2 revealed that many of the amino acids in the eludicated sequence are predicted to be incorporated into the C2 product, yet there was not the presence of the predicted N-terminal threonine. It was observed in the MS/MS spectra of the parent ion *m/z* 926.4951 that the mass of the y_6 ion was equal to the mass of the parent ion minus the total mass of a threonine plus a C_{12} hydroxylated fatty acid (FA). This is consistent with the MS/MS data, as the loss of this terminal portion would result in the *y*-ion series observed. The retention of the 14 Da mass shift by the large *b* ions observed is also consistent with the incorporation of fatty acid chains differing by a methylene group, as the fatty acid tail would be retained in these fragment ions. The chemical structure of the fatty acids tails incorporated into the lipopeptide series is reported in Figure 3c and Supplementary Figure 12.

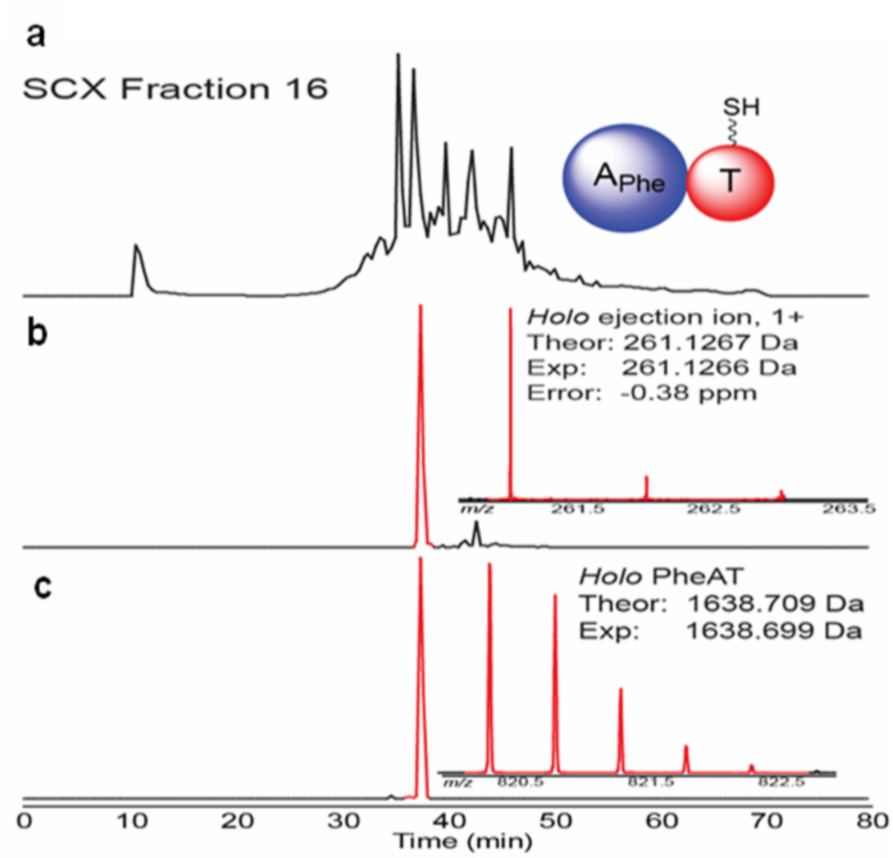
Supplementary Figures



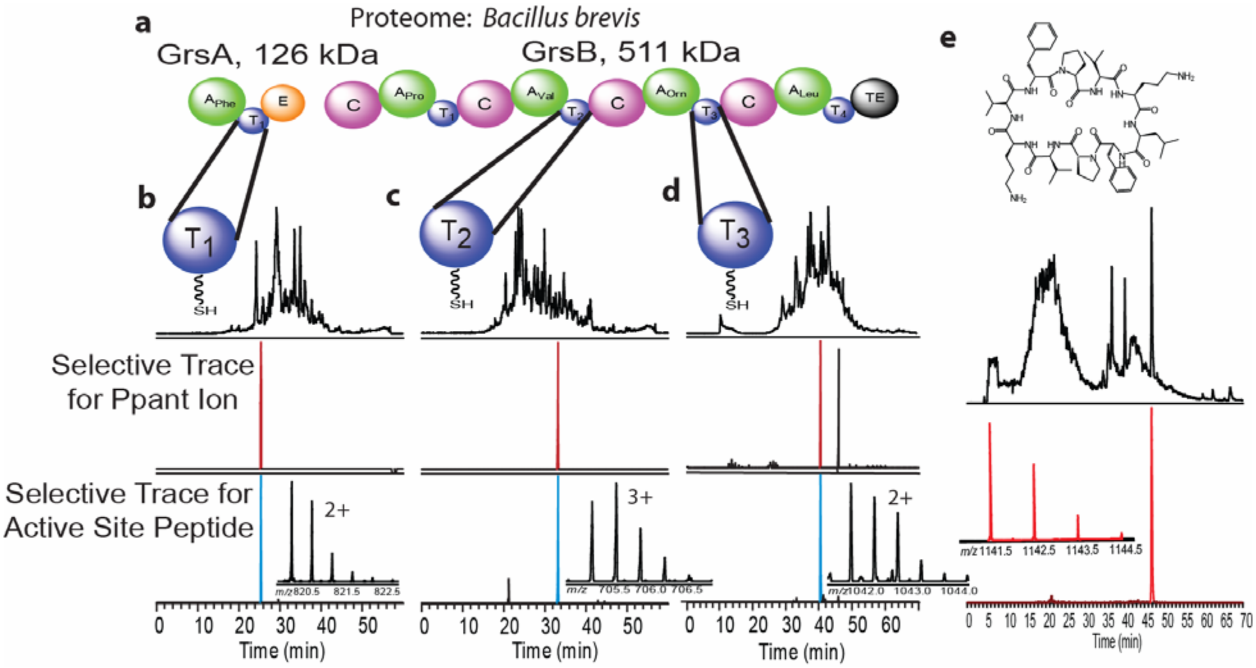
Supplementary Figure 1. The phosphopantetheinyl (Ppant) ejection assay. **a**, An *apo* thiolation domain is converted to the *holo* form by the action of a phosphopantetheinyl transferase (PPTase). When the *holo* protein (or active site peptide) is subjected to tandem MS, the Ppant arm is ejected off in one of two hypothesized mechanisms, resulting in two small molecule and two peptide (a charge-reduced *apo* + 80 Da and an *apo* - 18 Da) marker ions. **b**, Two small molecule marker ions are generated during Ppant ejection (shown here with masses for the *holo* form). A charge-reduced peptide marker ion is shown for the measured mass of the *apo* protein + 80 Da, resulting from the loss of the smaller Ppant ejection small molecular ion. **c**, Representative data from the on-line Ppant ejection assay, where the elution of the Ppant ejection ion (shown in red in the selected ion chromatogram for *m/z* 261.1263-261.1273) leads to identification of the active site containing peptide.



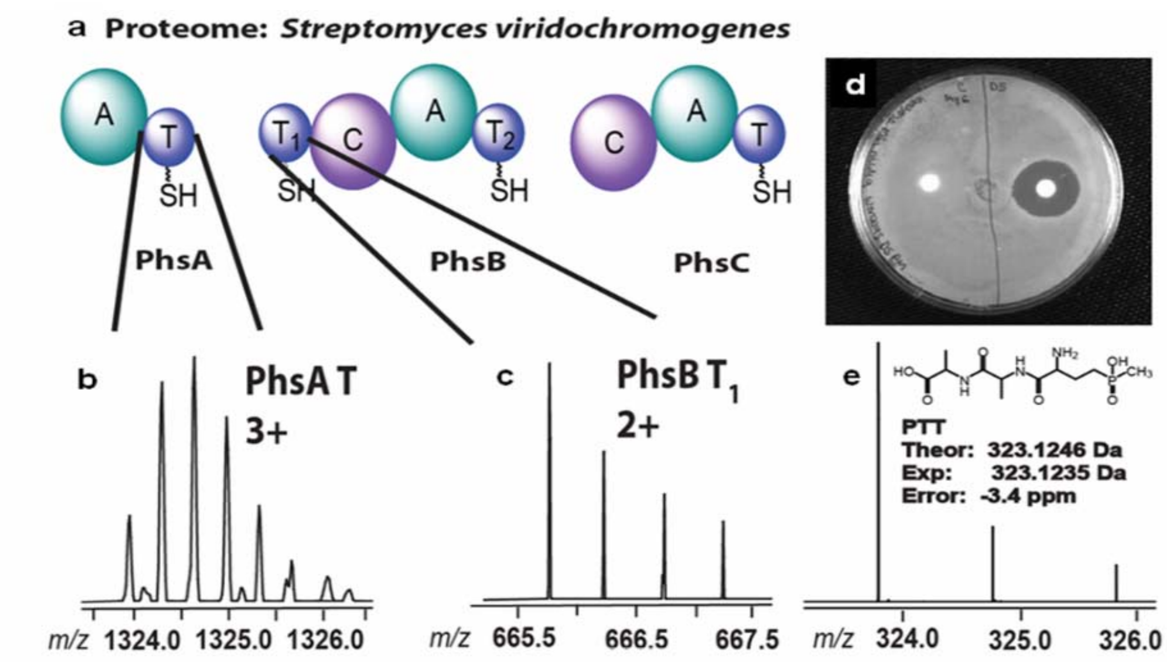
Supplementary Figure 2. Confirmation of Ppant ejection by a pseudo-MS^3 assay. **a**, MS^3 spectrum of the phosphopantetheinyl component of coenzyme A. **b**, Total ion chromatogram (TIC) for LC-MS analysis of a digest of the phosphopantetheinylated protein PheAT. **c**, Selected ion chromatogram (SIC) for the Ppant ejection ion generated by SID. **d**, Low-resolution SIC for the MS^3 fragment at *m/z* 243. **e**, Low resolution SIC for the MS^3 fragment at *m/z* 184.



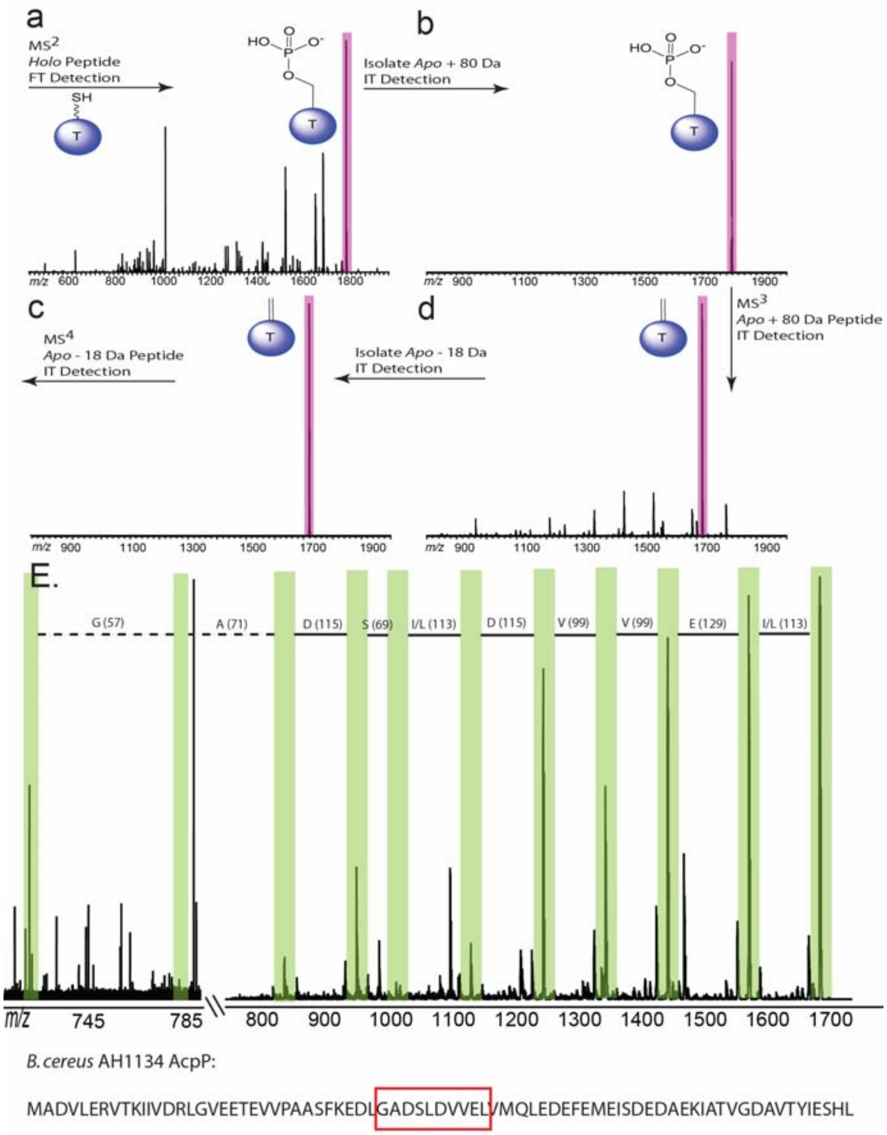
Supplementary Figure 3. Identification of the PheAT active site peptide in a complex peptide mixture. **a**, TIC for SCX fraction 16. Inset: representation of PheAT showing the phosphopantetheinylated T domain. **b**, SIC for *holo* Ppant ejection ion generated by SID. Inset: ejection ion observed. **c**, SIC for *holo* PheAT phosphopantetheinylated T domain active site peptide. Inset: PheAT *holo* T domain active site peptide.



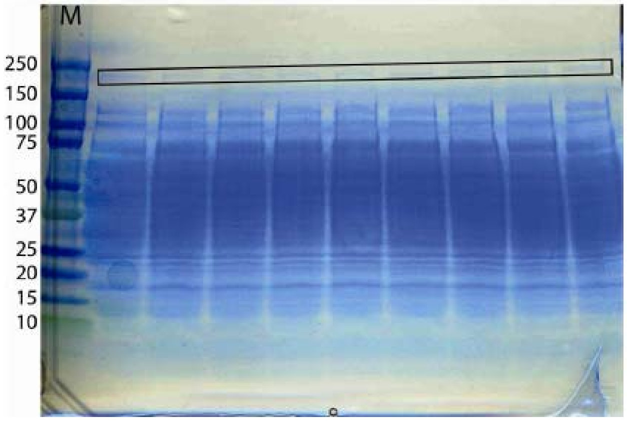
Supplementary Figure 4. Application of PrISM to detection of expressed NRPSs involved in gramicidin S biosynthesis in *Bacillus brevis* ATCC 9999. **a**, Schematic of NRPSs involved in gramicidin S biosynthesis in *B. brevis.* In Supplementary Figure 4b-d, the top panel corresponds to the TIC for the SCX fraction, the middle panel corresponds to the SIC for the Ppant ejection ion (*m/z* 261.1267), the bottom panel is the SIC for the phosphopantetheinylated active site peptide, and the inset in the bottom panel is the mass spectrum for a single charge state of the peptide. **b**, Active site peptide from GrsA T_1. **c**, Active site peptide from GrsB T_2. **d**, Active site peptide from GrsB T_3. **e**, Identification of gramicidin S in culture supernatant by LC-MS.



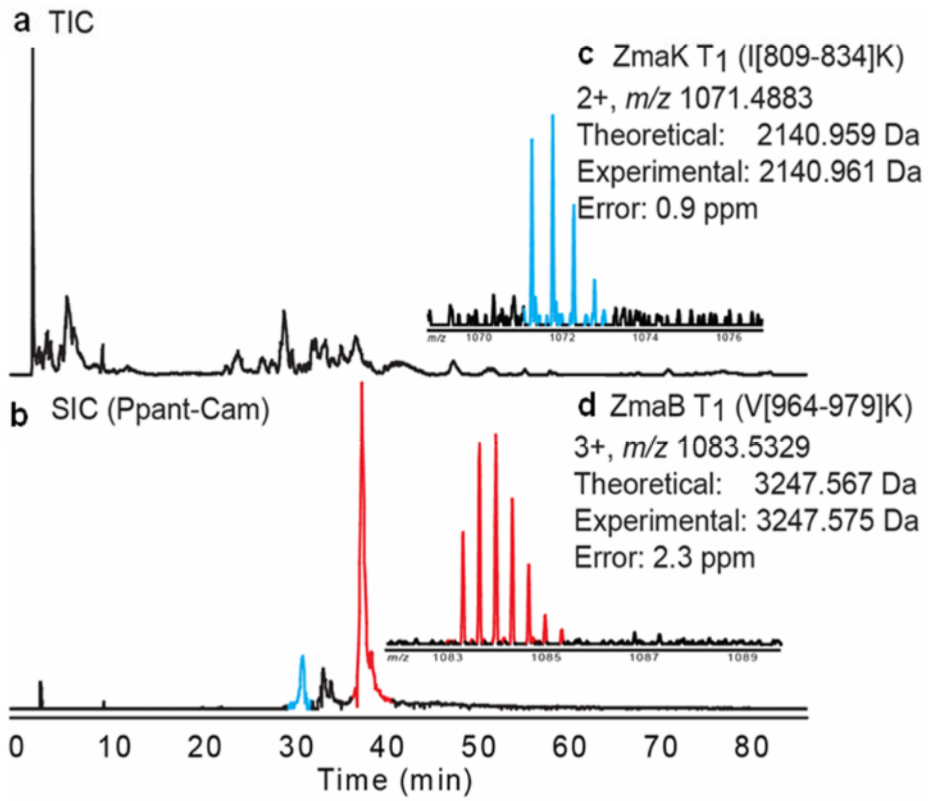
Supplementary Figure 5. Identification of phosphopantetheinylated peptides involved in PTT biosynthesis in *S. viridochromogenes.* **a**, Schematic of NRPS proteins responsible for incorporation of two alanine units into PTT. **b**, Active site peptide from the PhsA T domain identified by MudPIT-like PrISM. **c**, Active site peptide from the PhsB T_1 domain identified by MudPIT-like PrISM. **d**, Bioassay for the presence of PTT against *B. subtilis* ATCC6633. **e**, LC-MS identification of the PTT small molecule.



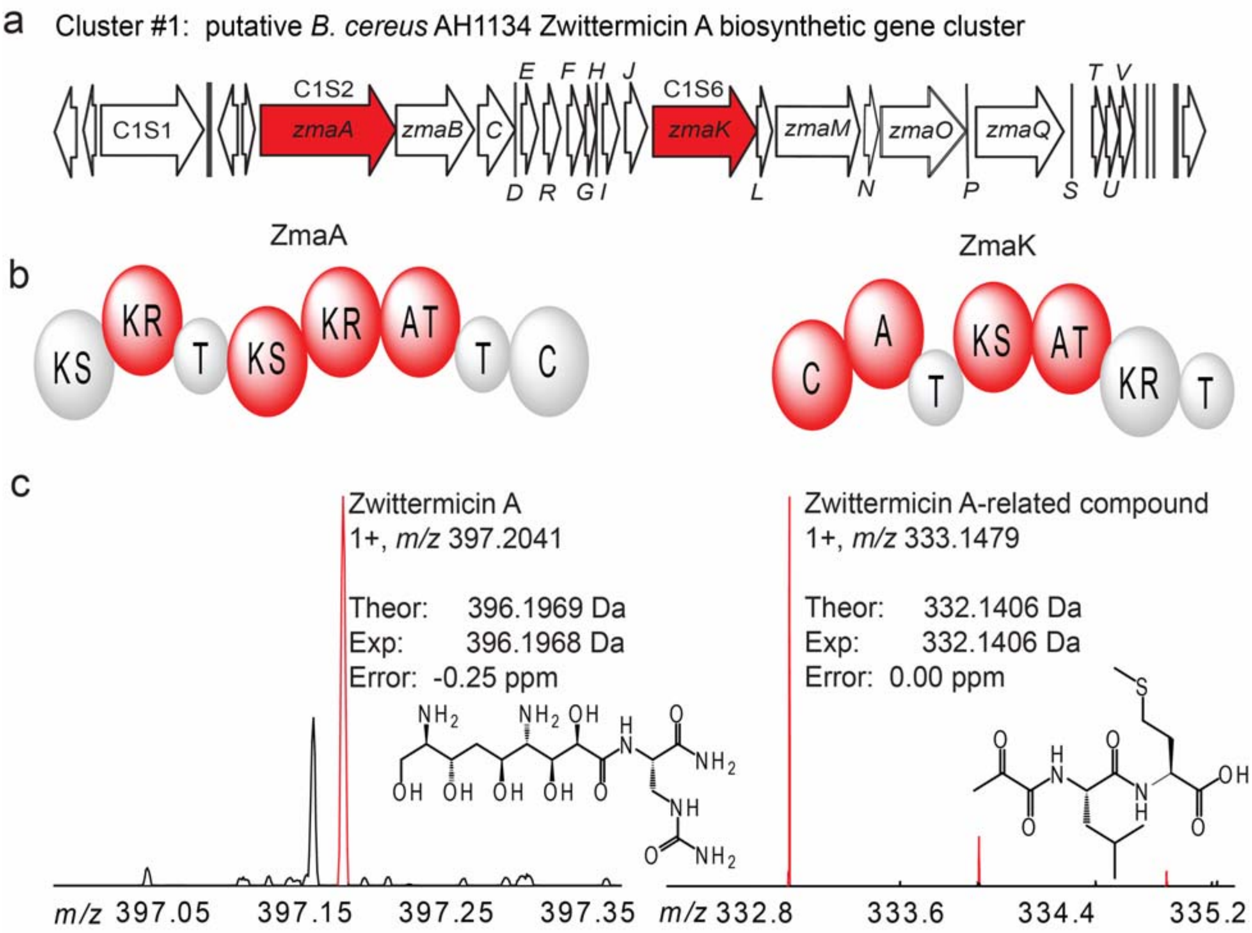
Supplementary Figure 6. Identification of an active site peptide from fatty acid synthase acyl carrier protein (AcpP) in NK2018. **a**, MS^2 on the *holo* peptide identified by the Ppant ejection assay, highlighted by the presence of the *apo* + 80 Da species. **b**, **c**, The *apo* + 80 Da species was isolated and subjected to MS^3, highlighted by the observation of the loss of phosphate to generate the *apo* – 18 Da species. **d**, The *apo* - 18 Da species is isolated and subjected to MS^4. **e**, Sequence tag generated from MS^4 analysis. This peptide corresponds to an active site peptide containing a phosphopantetheinylated serine.



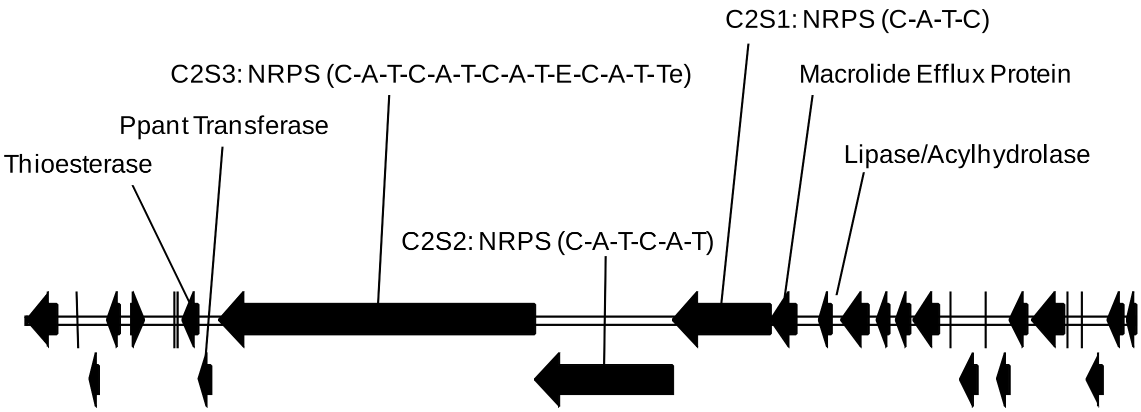
Supplementary Figure 7. SDS-PAGE separation of the soluble NK2018 proteome for in-gel trypsin digestion of HMWPs[17, 18].



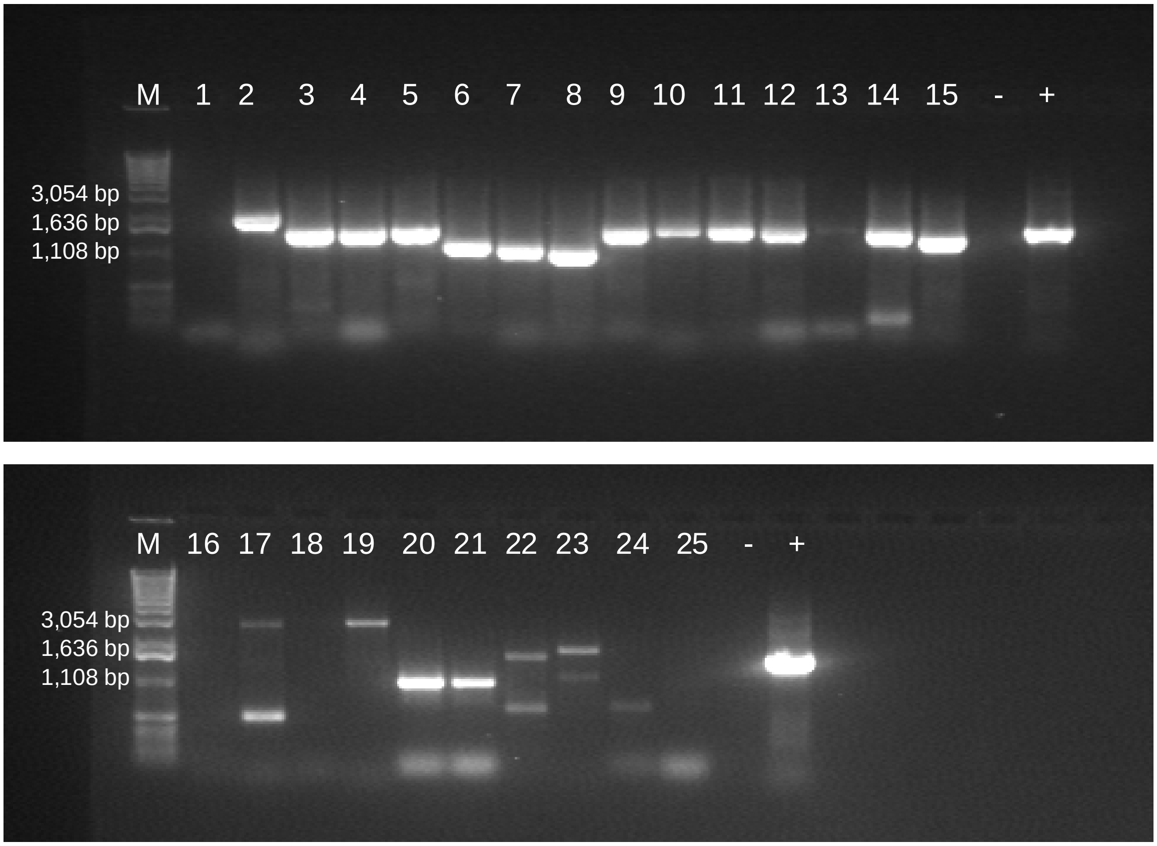
Supplementary Figure 8. Phosphopantetheinylated peptides identified from ZmA biosynthetic proteins during nanoLC-MS of NK2018 HMWPs. **a**, TIC for nanoLC-MS analysis of in-gel digestions. **b**, SIC for the Ppant-Cam product (*m/z* 318.1482) generated by source induced dissociation. **c**, Identification of a phosphopantetheinylated T domain active site peptide from ZmaK. **d**, Identification of a phosphopantetheinylated T domain active site peptide from ZmaB.



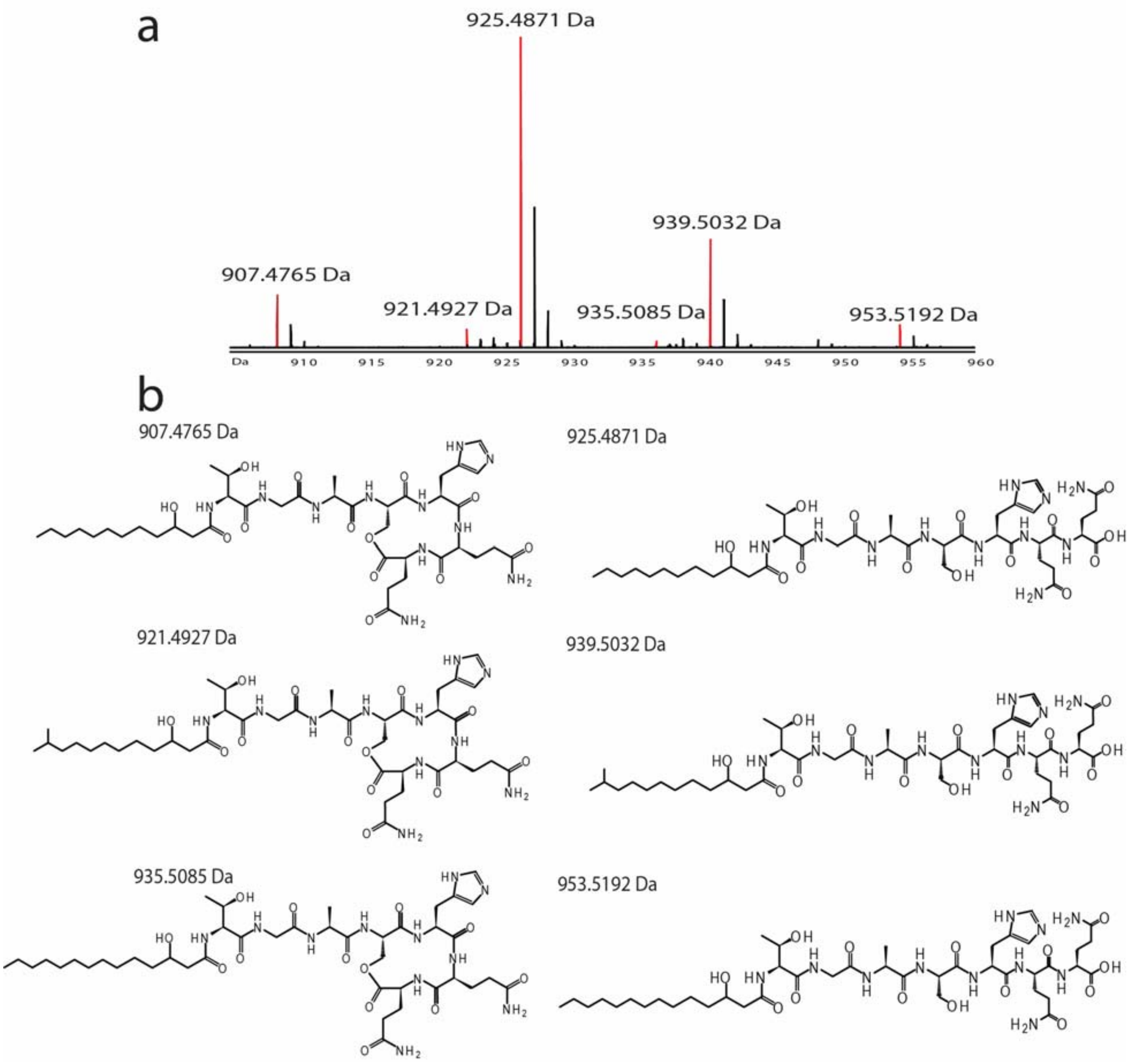
Supplementary Figure 9. Analysis of NK2018 cluster #1: a putative ZmA biosynthetic gene cluster. **a**, Arrangement of the ORFs of the putative ZmA biosynthetic gene cluster from *B. cereus* AH1134. Gene products detected by nanoLC-MS are shown in red. **b**, Domain arrangement of the NRPS/PKS proteins detected by nanoLC-MS analysis. The domains from which peptides were identified are shown in red. **c**, Detection of ZmA and an additional putative product from the ZmA biosynthetic gene cluster by LC-MS. The chemical formula of this species is supported by its unique isotopic distribution.



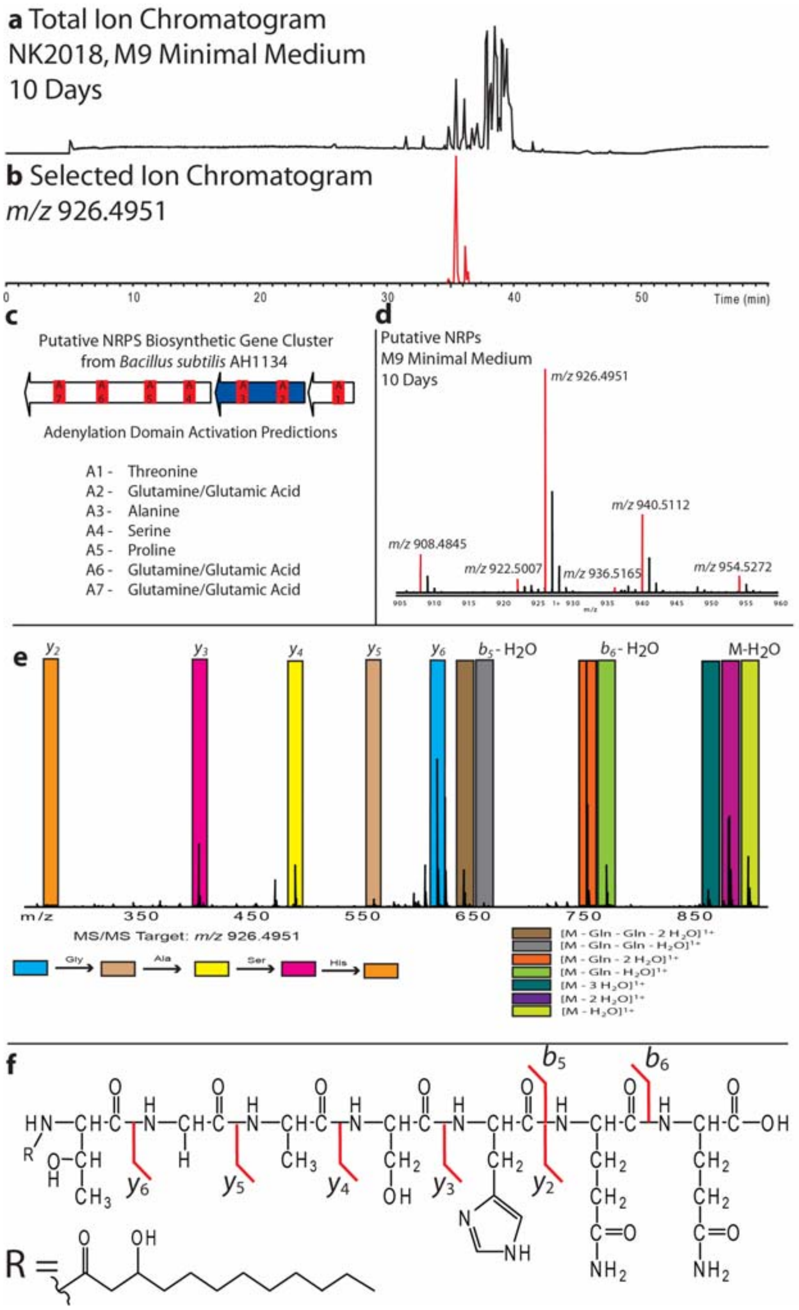
Supplementary Figure 10. Annotation of NK2018 cluster #2. ORF annotations are from BLAST analysis of the corresponding sequence from *B. cereus* AH1134. Shown is the region of the *B. cereus* AH1134 genome containing the NRPS biosynthetic gene cluster.


M 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 - +
3,054 bp
1,636 bp
1,108 bp
M 16 17 18 19 20 21 22 23 24 25 - +
3,054 bp
1,636 bp
1,108 bp

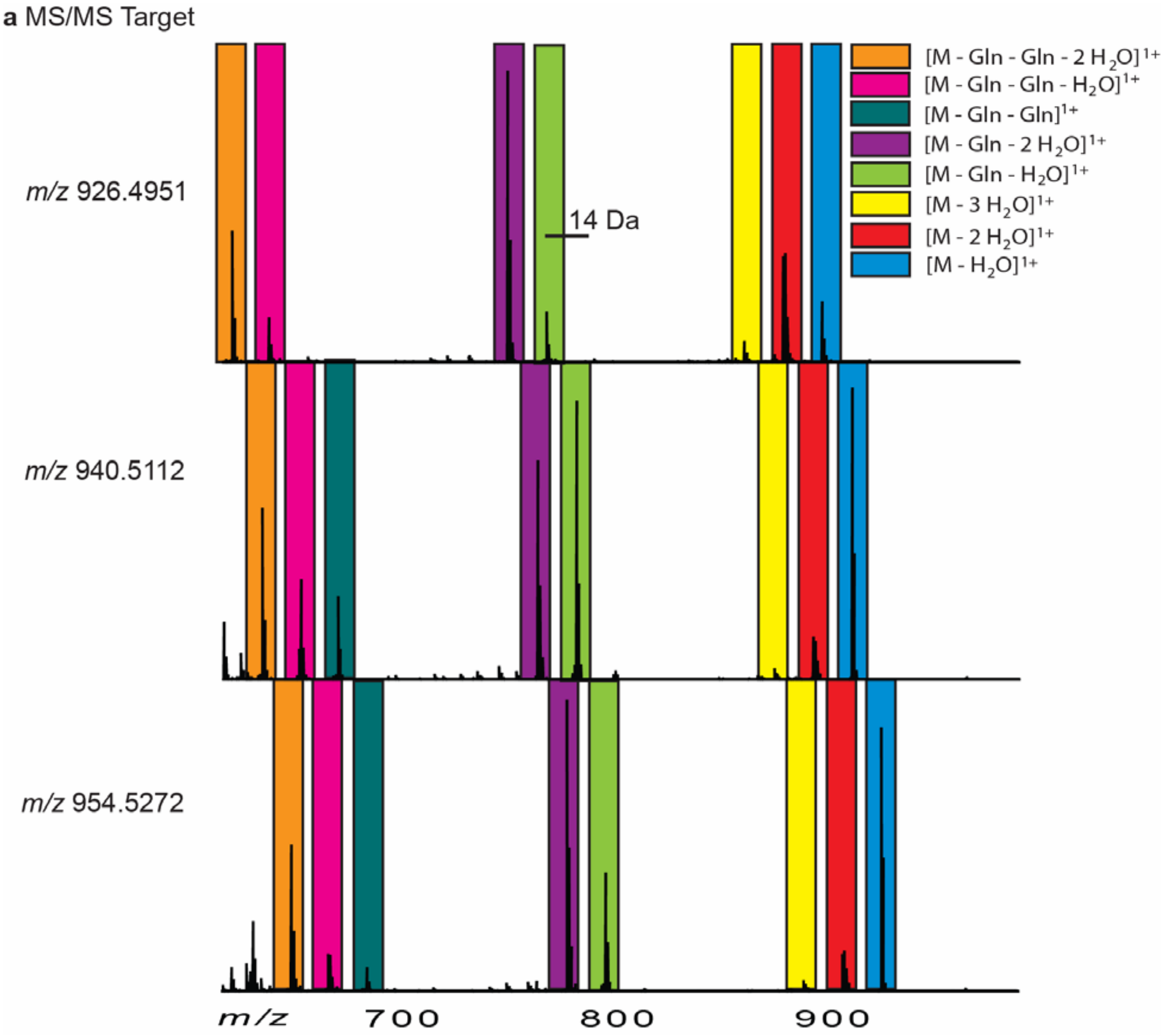
Supplementary Figure 11. Agarose gels of PCR products. Lane identifiers correspond to the PCR number in Supplementary Table 4. M: Marker (Invitrogen 1 kb marker). -: Negative control (omission of primer). +: Positive control (primers for amplification of 16S rDNA, PCR #26)



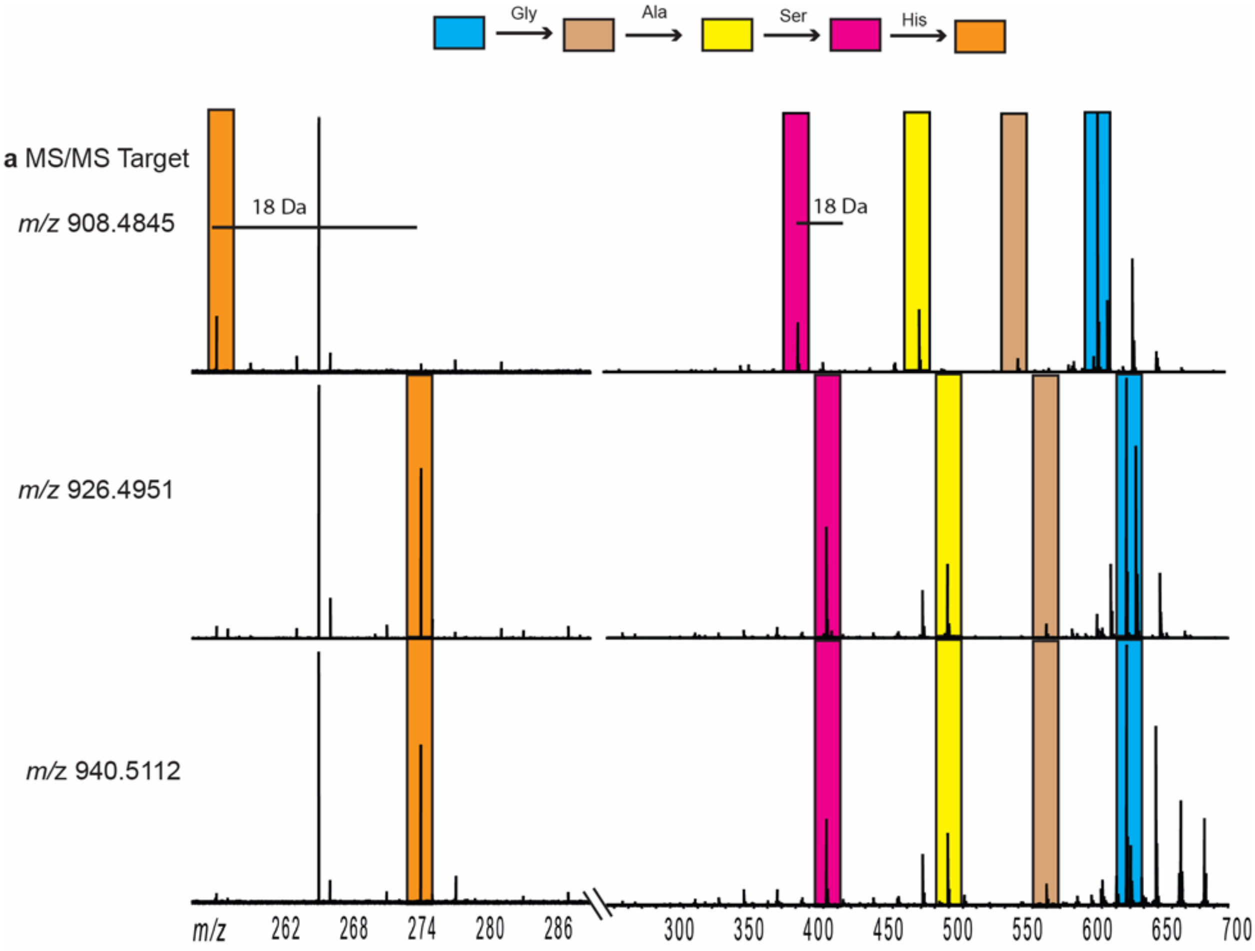
Supplementary Figure 12. Lipoheptapeptides identified from culture supernatant of NK2018 grown in M9 minimal medium and theoretical structures for each mass detected. **a**, Mass spectrum of six lipopeptides produced by NK2018 grown in M9 minimal medium for 10 days. All masses are reported as the neutral monoisotopic mass. **b**, Theoretical structures for the six species observed in Supplementary Figure 12a. The hydroxyl group is placed on the fatty acid based on the precedent of other *Bacillus* lipopeptides[19] and the lactone ring is hypothesized to mimic that of the kurstakins[16].



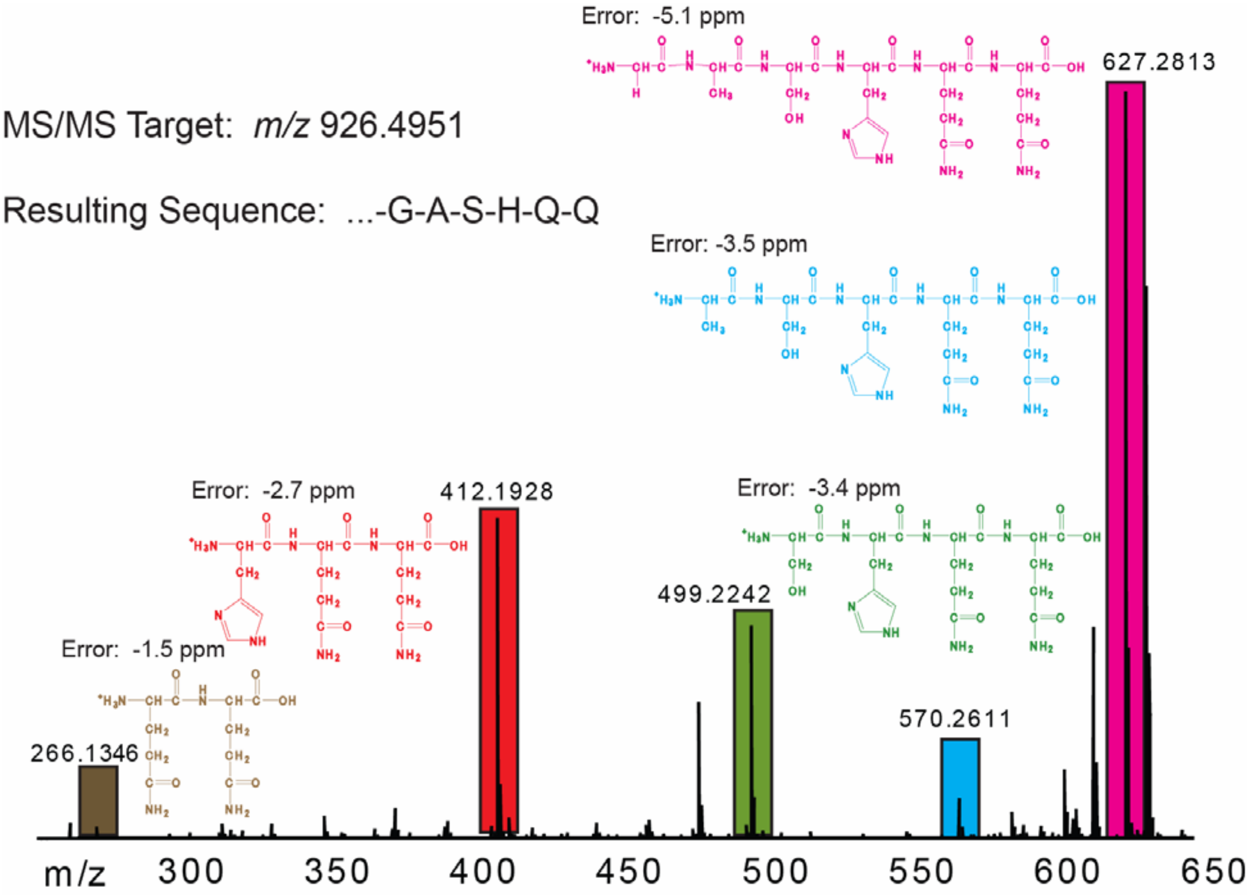
Supplementary Figure 13. Series of nonribosomal peptides identified from NK2018. **a**, Total ion chromatogram from LC-MS analysis of the culture supernatant from NK2018 grown in M9 minimal media for 10 days. **b**, Selected ion chromatogram for the peptide species at *m/z* 926.4951. **c**, Predicted substrates for cluster #2 adenylation domains. **d**, Set of at least six related peptides observed during LC-MS analysis. **e**, Representative MS/MS data of the species at *m/z* 926.4951 with detection of the fragment ions by FTMS. Amino acid mass differences between peaks (in colored boxes) are shown. **f**, Fragmentation map for the proposed lipopeptide structure (*m/z* 926.4951).



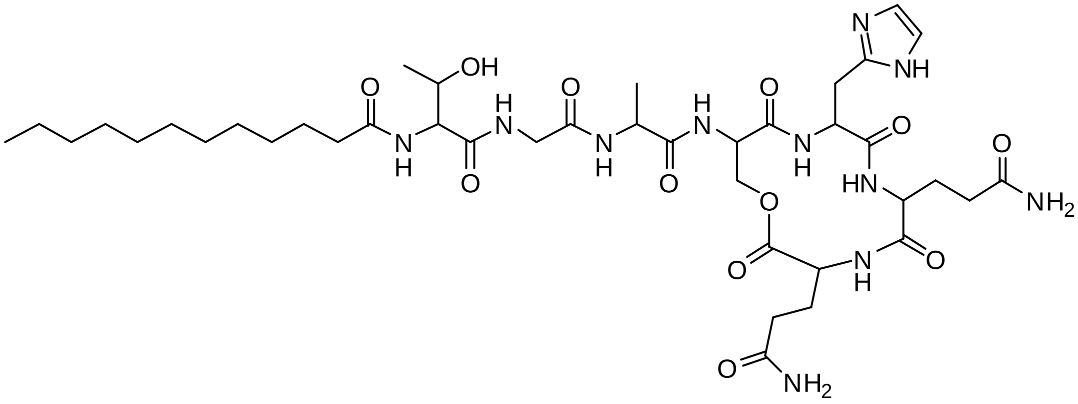
Supplementary Figure 14. Selected MS/MS analysis of the putative nonribosomal peptides identified: localization of the 14 Da mass shifts. **a,** Comparison of a portion of the MS/MS data from three of the peptides of interest, localizing the shift of 14 Da to the high-molecular weight *b*-ions. This corresponds to the 14 Da shifts localizing to the N-terminus of the peptide on the fatty acid chain.



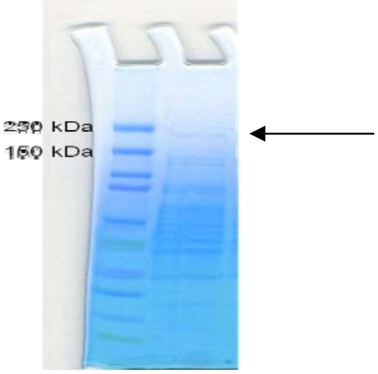
Supplementary Figure 15. Selected MS/MS analysis of the putative nonribosomal peptides identified: localization of the 18 Da mass shifts. **a**, Comparison of a portion of the MS/MS data from three of the peptides of interest, localizing the 18 Da shifts to the detected *y*-ion series. These mass differences are best explained by structures in either lactone ring-open or ring-closed form.



Supplementary Figure 16. Structural assignment of *y*-ions from the MS/MS analysis of the peptide species at *m/z* 926.4951. The largest ion (pink box) most likely results from fragmentation between Thr and Gly, where the mass difference between the ion at *m/z* 627.2813 and the parent ion at *m/z* 926.4951 (299.2138 Da) can be accounted for by the mass of threonine plus a hydroxylated C_{12} fatty acid chain.



Supplementary Figure 17. Representative structure of a kurstakin *Bacillus* lipopeptide[16].



Supplementary Figure 18. Confirmation of HMWP production by NK2018 in M9 minimal medium. Lane 1: Bio-Rad Kaleidoscope Presision Plus Protein Standard. Lane 2: Proteome of NK2018 grown in M9 minimal medium for 3 days at 30°C.

Supplementary Tables

Supplementary Table 1. Peptides identified from GrsA and GrsB in *B. brevis* by nanoLC-MS.

Protein	Sequence[a]	AA Range[b]	-log E[c]
GrsA	WLSDGNIEYLGR	417-428	4
GrsB	DISSLDEEKR	94-103	7
GrsB	AGGAFVPIDPPEYPK	541-554	2
GrsB	DVEQLFDLVKR	685-695	8
GrsB	ENIEVLSFPVAFLK	696-709	4
GrsB	SLPNLEGIVNTAK	955-968	6
GrsB	ESYVAIQPVPEQEYYPVSSQK	1048-1069	8
GrsB	AGGAYLPLDPEYPADR	1579-1594	1
GrsB	YGQLPVGVPGELCVGGDVAR	1831-1852	4
GrsB	WLPDGTIEYLGR	1885-1896	2
GrsB	IEPGEIETLLVK	1908-1919	6
GrsB	LAEIWHNVLGVNK	2015-2027	3
GrsB	LCEINMLSEEEQQR	2513-2526	5
GrsB	VLYDFNGTDATYATNK	2527-2542	5
GrsB	ALLTGGQLIVCPNEVK	2745-2760	6
GrsB	SLPEPDGSISIGTEYVAPR	3035-3053	6
GrsB	TVEQLAQHFIQIVK	3536-3549	3
GrsB	AGGAYVPIDPAYPQER	3661-3676	3

[a] Peptide sequence from *B. brevis* ATCC 9999 with the matching *b* and *y* fragments illustrated as left and right flags, respectively.
[b] Amino acids matched in the parent protein.
[c] OMSSA expectation score for the peptide [20]

Supplementary Table 2. Peptides identified from C1S2 (ZmaA), C1S6 (ZmaK) and C2S2 during nanoLC-MS.

Cluster/Protein	Sequence[a]	AA Range[b]	-log E[c]
C1S2	SEGVYLITGGMGGVGLK	1000-1016	11
C1S2	TFQCVLGLSGLNQVIISSGDLNK	1231-1253	12
C1S2	FPGAQNVNEFWNNIK	1386-1400	7
C1S2	ANVGHLNAASGVAGLIK	1716-1732	8
C1S2	AGVSSFGIGGTNAHIILEEAPK	1780-1801	12
C1S2	SVQQDVSDMYFLFSHITR	2201-2218	10
C1S2	NNNISEWFYTPQWK	2275-2288	8
C1S2	VVTVEPGFAFNK	2327-2338	5
C1S2	IITNGVQQVIGDEELIPEK	2420-2438	9
C1S2	FVQTYEPLQLEQPAK	2499-2513	7
C2S6	FEEAWNYVTR	55-64	2
C2S6	LFWEQYLNELTEQINLSNK	187-205	13
C2S6	NSEDVIFGTTVSGR	256-269	6
C2S6	LVQPNDLLDELETTVADVWK	937-956	16
C2S6	QFWSNLEQGIESIR	1056-1069	8
C2S6	SIASESSNTLEDFQAMLLNEK	1162-1182	12
C2S6	TNIGHLDAAAGVGGFIK	1376-1392	6
C2S6	VNTELIPWKEEVR	1424-1436	3
C2S6	ENPEISLSDTAYTLQK	1495-1510	6
C2S6	TAVVANGIEDAIEK	1519-1532	8
C2S6	TSMTQPLLFTIEYALAK	1614-1630	2
C2S6	MMDEVLDAFEQAVGNIK	1748-1764	11
C2S6	NVICLEIGPGNALSTFVLK	1814-1832	15
C2S2	LTLNTLVQGAWAYLMSR	250-266	8
C2S2	LTDNTSVVDWLR	306-317	7
C2S2	YLTDGNLEFIGR	851-862	3
C2S2	IELGEIEATLEK	874-885	8
C2S2	LVAYVISDGNTEEWR	906-920	9
C2S2	EIEGLIGFFANTLVYR	1344-1359	9
C2S2	LQYILEDAQIK	1611-1621	6
C2S2	MVPIGVVGELYIGGSLAR	1849-1867	13
C2S2	IELGEIEAVLQK	1923-1934	7

[a] Peptide sequence from *B. cereus* AH1134 with the matching *b* and *y* fragments illustrated as left and right flags, respectively.

[b] Amino acids matched in the parent protein.

[c] OMSSA expectation score for the peptide [20].

Supplementary Table 3. Primers designed for PCR analysis of the NK2018 genome.

Name	Sequence
Nac-Xferase F	ATG TTC AAA ATA TAT AAT GGA GTT GA
Nac-Xferase R	TTA AGG GTC ATC TAA CGG TAT T
C1S2-1-F	ATG AAG CAT CAA GAT GAT AGT AAA A
C1S2-1-R	TTA GGA TCT ACA TCC TTA ACT TGT GT
C1S2-2-F	GGA AGG AGT TTA TTT AAT TAC TGG TG
C1S2-2-R	CAC TTG CCA ATA TAA GTT AGG AGA T
C1S2-3-F	ATG AAG CAA ATG TAG AAC CTG A
C1S2-3-R	CTT GTA TTG TCC CTA TTC AAC TCT T
C1S2-4-F	ATT TTT GTG TTC TAA TGT CTT CAA TT
C1S2-4-R	CCT ATA ATA TCT TGT AAA TCC GCA
ZmaR-F	ATG ATT TAT GAA TTG GTA AAA GAA AAG
ZmaR-R	TCA TCT TAA GCT ATC TTC AAC TCT ATC
Alksulf-F	ATG ATT AAT AAA GAA GTA GTT CCG TC
Alksulf-R	TTA GTT AAC TTG GAT TAA CTT TTC AAT
NAcMurLig-F	AGG TAG ACG GAC GTG GTG
NAcMurLig-R	AGC CTT GTT TAC AAT AAT TGC TG
C1S6-1-F	ATG AGG AAA GCA GTA AAG ATT CA
C1S6-1-R	ATC CAA CAT TAC AGC TAC TGG C
C1S6-2-F	CTT CTA ATA CAT TAG AAG ATT TCC AAG
C1S6-2-R	AAA TGT AAT GCA TCT TCA TAA GAC
C1S6-3-F	AAT GCA TTA TGA AAT GCT ATT TT
C1S6-3-R	ATT AAT ACG AAC AAG ATC TAA AGA ACT
C2S2-Deg-F	YTW AAY ACD YTN GTN CAA GGN G
C2S2-Deg-R	GCR AAR AAN CCD ATN AR
C2S2-NonDeg-F	TTA AAT ACA TTA GTA CAA GGA G
C2S2-NonDeg-R	GCA AAA AAT CCA ATT AA
C2S2-NonDeg-F-long	GGA AAT ACA GAA GAA TGG
C2S2-DegF-long	GGN AAY ACN GAR GAR TGG
C2S2-NonDeg-R-long	CCA TTC TTC TGT ATT TCC
C2S2-DegR-long	CCA YTC YTC NGT RTT NCC
C2S2-1-F	TCA TTT TTT TAC CCC TTC CT
C2S2-1-R	GTA TGC ACG AGT GAA GTG AC
C2S2-2-F	CAT TTT GGA TAA GAG GCA ATT T
C2S2-2-R	TTG GCA TTC AGG ATT CCT
C2S2-3-F	CAT CAA ATG AGA ATG GCG T
C2S2-3-R	AGT GAA GTG ATG GAA GTT TAT CTC A
DalaDalaPeptidase-F	TTA GCT TTT TAC TAT ATC AAT CGT ACG
DalaDalaPeptidase-R	GTG AAA GGT ATG TTT TGC AAT AG
C1S6-1-NonDeg-F	CCA TGG AAA GAA GAA GTA
C1S6-1-NonDeg-R	ATC TAA TAC TTC ATC CAT CAT
C1S6-1-Deg-F	CCN TGG AAR GAR GAR GTN
C1S6-1-Deg-R	RTC NAR NAC YTC RTC CAT CAT
C1S2-1-Nondeg-F	GTA AAT GAA TTT TGG AAT AAT
C1S2-1-Nondeg-R	CCA TTG TGG TGT ATA AAA
C1S2-1-Deg-F	GTN AAY GAR TTY TGG AAY AAY

Name	Sequence
C1S2-1-Deg-R	CCA YTG NGG NGT RTA RAA
16SRNA-F	
16SRNA-R	

[a] Standard abbreviations are used: **R**: A or G, **Y**: C or T, **M**: A or C, **K**: G or T, **S**: C or G, **W**: A or T, **H**: A, C or T, **B**: C, G or T, **V** : A, C or G, and **D**: A, G or T, **N**: A, T, C or G.

Rows in white represent primers designed from the *B. cereus* AH1134 genome sequence. Rows in gray represent primers designed from the peptide sequences generated by nanoLC-MS/MS analysis of NK2018. Deg: Degenerate primer. NonDeg: primer designed using the most likely codons as predicted from sequenced *B. cereus* ATCC 10987(http://www.kazusa.or.jp/codon/).

Supplementary Table 4. PCRs completed for analysis of the NK2018 genome.

PCR #	Primer 1	Primer 2	Av. T_m (°C)	Length (bp)
1	Nac-Xferase F	Nac-Xferase R	58.3	900
2	C1S2-1-F	C1S2-1-R	59.7	2000
3	C1S2-2-F	C1S2-2-R	60.5	1500
4	C1S2-3-F	C1S2-3-R	60.2	1500
5	C1S2-4-F	C1S2-4-R	59.8	1500
6	ZmaR-F	ZmaR-R	59.6	1128
7	Alksulf-F	Alksulf-R	58.6	1077
8	NAcMurLig-F	NAcMurLig-R	61.0	1001
9	C1S6-1-F	C1S6-1-R	61.0	1500
10	C1S6-2-F	C1S6-2-R	58.6	1500
11	C1S6-3-F	C1S6-3-R	57.9	1500
12	C2S2-1-F	C2S2-1-R	58.7	1500
13	C2S2-2-F	C2S2-2-R	60.6	1500
14	C2S2-3-F	C2S2-3-R	61.3	1500
15	DalaDalaPeptidase-F	DalaDalaPeptidase-R	59.3	1314
16	C2S2-NonDeg-F	C2S2-NonDeg-R	50.9	3300
17	C2S2-Deg-F	C2S2-Deg-R	50.9	3300
18	C1S2-1-Nondeg-F	C1S2-1-Nondeg-R	52.0	2700
19	C1S2-1-Deg-F	C1S2-1-Deg-R	52.0	2700
20	C1S6-1-NonDeg-F	C1S6-1-NonDeg-R	52.8	1200
21	C1S6-1-Deg-F	C1S6-1-Deg-R	52.8	1200
22	C2S2-NonDeg-F	C2S2-NonDeg-R	51.3	1500
23	C2S2-DegF	C2S2-Deg-R	51.3	1500
24	C2S2-NonDeg-F	C2S2-NonDeg-R	51.3	1800
25	C2S2-Deg-F	C2S2-DegR	51.3	1800
26	16SRNA-F	16SRNA-R		

PCR numbers (PCR#) correspond to the band in the agarose gel (Fig. 2c and Supplementary Fig. 11) labeled with the same number. Rows in gray and white correspond to the same designations as in Supplementary Table 3.

Supplementary Table 5. Summary of PCR results.

Cluster #1			
Primer	Read Length (bp)	Number of Mismatches to AH1134	% Identity to AH1134
2F	640	11	98.3
2R	799	22	97.2
3F	640	10	98.4
3R	720	13	98.2
6F	601	6	99.0
6R	720	10	98.6
19R	719	17	97.6
20F	719	17	97.6
20R	719	17	97.6
21F	293	3	98.9

Cluster #2			
Primer	Read Length (bp)	Number of Mismatches to AH1134	% Identity to AH1134
14F	720	4	99.4
14R	320	1	99.6
15F	560	1	99.8
15R	350	1	99.7
19F	350	1	99.7

“Primer” refers to the PCR in Supplementary Table 3, where F represents that the forward primer was used in that sequencing reaction and R represents that the reverse primer was used in the sequencing reaction.

Supplementary Table 6. Gradient used for strong cation exchange chromatography.

Time (min)	% SCX solvent B
0	0
5	0
43	57 (increase 3% SCX solvent B every 2 min)
47	67 (increase 5% SCX solvent B every 2 min)
49	80
52	100
53	0
60	0

Supplementary Table 7. Gradient used for reverse-phase liquid chromatography.

Time (min)	% RPLC solvent B
0	5
5	5
40	50
45	75
50	95
51	95
52	5
70	5

Supplementary Table 8. HPLC gradient for LC-MS analysis of culture supernatants.

Time (min)	% RPLC solvent B
0	0
10	0
30	20
40	100
44	100
46	0
60	0

Supplementary Table 9. Exact masses and putative empirical formulas for the detected lipoheptapeptides.

m/z	Mass (Da)	Formula
908.4845	907.4765	$C_{40}H_{65}O_{13}N_{11}$
922.5007	921.4927	$C_{41}H_{67}O_{13}N_{11}$
936.5165	935.5085	$C_{42}H_{69}O_{13}N_{11}$
926.4951	925.4871	$C_{40}H_{67}O_{14}N_{11}$
940.5112	939.5032	$C_{41}H_{69}O_{14}N_{11}$
954.5272	953.5192	$C_{42}H_{71}O_{14}N_{11}$

Supplementary Table 10. Information for retrieval of MS data from the Tranche network.

File or File Set	Hash
In-Gel Digest of NK2018-Run1	4KgzmKxdO3VgTTS+qUSoZrx/BQb85RAqdsiXrrRtG2Ly3mtm9S2XKGlGV8eeMBPnoFsMuhxjy1XVqpxSIxStXlALBcoAAAAAAAABOg==
In-Gel Digest of NK2018-Run2	s4ywB1X9jqeexNbUvKDujbBLAsy+snp1nMNXPArzSZKarWziALKxQzNUsZ6ZrKyFCdWP6VWP8oJuDvI9VvqcAfj+OZUAAAAAAAACIA==
Identification of T Domains from PheAT in *E. coli*	oVXf5G6dXMPdXFMKv9IuXbopkRfqvGCs9spFmz9NHLi5xg1i1QsH6G7kvPbUoBqQcU9jc/HEHLnYaMaH8jLItcD7EQkAAAAAAAACWA==
Identification of T Domains from *B. brevis*	qNEsMQXaQ5Wuz00GjYKKNPMY492N1Vz4f9FC+fcx1XuuGd10Tg6q1jJW90BGOHP7rFpcNyVWTzLzdAxjXZCDl3XvHzoAAAAAAAAFEg==
Identification of T domains from *S. viridochromogenes*	DihTnk/HoJueuGEu9ij1hPZke4nyytqmvv6F4pKNPMKQyW7PS6ht4G3brrbratNICzWd7XRCriYr6M716MFJXQJoIccAAAAAAAADXQ==
Detection of Fattyacyl ACP from NK2018	0Mcp2+fNPl63AOLKPvz3x/FINbDVpCyUSAxa6HTkLrBpY8hFsGHhCIcK8hwLvYJXJaviuol+0VqqYFLsAS75C01joxIAAAAAAAASqg==
LCMS Detection of Kurstakins	nEd/ts1nxNyl1qkVrZudqe8VT+3BkDXP6csmDtJCBZgww3HNpdUgA28zAeJVV0kcbS2GTT/WHGklkz4Hx+QoHttKk2IAAAAAAAAHDA==
Offline MSMS of Kurstakins	mTc5ILPIqjJGfRrVkfMdsWdH1OyUtP29JfR/k6CsXM0dUn2Vqjc0XtdD5VwXrtjausVe4+Emo4tjUzNFU61bGUvkJz4AAAAAAABCeg==

Supplementary References

1. Stachelhaus, T.a.M., MA Modular Structure of Peptide Synthetases Revealed by Dissection of the Multifunctional Enzyme GrsA. *Journal of Biological Chemistry* **270**, 6163-6169 (1995).
2. Meluzzi, 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. *Bioorg Med Chem Lett* (2007).
3. Quadri, L.E. et al. Characterization of Sfp, a Bacillus subtilis phosphopantetheinyl transferase for peptidyl carrier protein domains in peptide synthetases. *Biochemistry* **37**, 1585-1595 (1998).
4. Washburn, M.P., Wolters, D. & Yates, J.R., 3rd Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat Biotechnol* **19**, 242-247 (2001).
5. Little, D.P., Speir, J.P, Senko, M.W., O'Conner, P.B., McLafferty, F.W. Infrared multiphoton dissociation of large multiply charged ions for biomolecule sequencing. *Analytical Chemistry* **66**, 2809-2815 (1994).
6. Loo, J., Edmonds, CG, Smith, RD Tandem mass spectrometry of very large molecules: serum albumin sequence information from multiply charged ions from by electrospray ionization. *Analytical Chemistry* **63**, 2488-2499 (1991).
7. Vater, J., Mallow, N., Gerhardt, S., Gadow, A. & Kleinkauf, H. Gramicidin S synthetase. Temperature dependence and thermodynamic parameters of substrate amino acid activation reactions. *Biochemistry* **24**, 2022-2027 (1985).
8. Vater, J. et al. The modular organization of multifunctional peptide synthetases. *J Protein Chem* **16**, 557-564 (1997).
9. Mootz, H.D. & Marahiel, M.A. Biosynthetic systems for nonribosomal peptide antibiotic assembly. *Curr Opin Chem Biol* **1**, 543-551 (1997).
10. Katz, E. & Demain, A.L. The peptide antibiotics of Bacillus: chemistry, biogenesis, and possible functions. *Bacteriol Rev* **41**, 449-474 (1977).
11. Matteo, C.C., Glade, M., Tanaka, A., Piret, J., and Demain, A.L. Microbiological Studies on the Formation of Gramicidin S Synthetases. *Biotechnology and Bioengineering* **17**, 129-142 (1975).
12. Hoerlein, G. Glufosinate (phosphinothricin), a natural amino acid with unexpected herbicidal properties. *Rev Environ Contam Toxicol* **138**, 73-145 (1994).
13. 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).
14. Grammel, N., Schwartz, D., Wohlleben, W. & Keller, U. Phosphinothricin-tripeptide synthetases from Streptomyces viridochromogenes. *Biochemistry* **37**, 1596-1603 (1998).
15. Kevany, B.M., Rasko, D.A. & Thomas, M.G. Characterization of the complete zwittermicin A biosynthesis gene cluster from Bacillus cereus. *Appl Environ Microbiol* **75**, 1144-1155 (2009).
16. 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).
17. Mortz, E., Vorm, O., Mann, M. & Roepstorff, P. Identification of proteins in polyacrylamide gels by mass spectrometric peptide mapping combined with database search. *Biol Mass Spectrom* **23**, 249-261 (1994).
18. Patterson, S.D. & Aebersold, R. Mass spectrometric approaches for the identification of gel-separated proteins. *Electrophoresis* **16**, 1791-1814 (1995).

19. Peypoux, F., Bonmatin, J.M. & Wallach, J. Recent trends in the biochemistry of surfactin. *Appl Microbiol Biotechnol* **51**, 553-563 (1999).

20. Geer, L.Y. et al. Open mass spectrometry search algorithm. *J Proteome Res* **3**, 958-964 (2004).