Supplemental Information

Table S1. 38 RiPPs discovered by MetaMiner, related to Figure 2. MetaMiner identified 31 known and discovered seven unknown RiPPs in STANDARD, ACTI, BACIL, SPACE, SPONGE, CYANO, HUMAN-iso and HUMAN-CF datasets at 1% FDR threshold for each dataset. All identified RiPPs are linear except for cyclic cyanobactins. The unknown RiPPs are shown in bold. The mode stands for the mode of discovery. The "A" mode stands for RiPPs discovered by the antiSMASH motif search, the "B" mode stands for RiPPs discovered by the BOA search, and the "E" mode stands for RiPPs discovered by exhaustive search (all-ORF mode of MetaMiner). Overall, 11 out of 38 RiPPs are predicted by antiSMASH motif search, 8 by BOA search, and 21 only with exhaustive search.

Dataset	RiPP class	core peptide	mode	p-value	strain	reference
STANDRAD	lanthipeptides	SVAGGGRIDTCPAGGGTSEQTGTCC	EA	$3 \cdot 10^{-47}$	*P. marinus*	prochlorosin (Tang and van der Donk, 2012)
STANDRAD	lanthipeptides	LEAASGGGDTGIQAVLHTAGCYGGTKMCRA	EA	$1 \cdot 10^{-40}$	*P. marinus*	prochlorosin (Tang and van der Donk, 2012)
STANDRAD	lanthipeptides	GAAGGCCITGESPGSAPTNDYKCTKGRGPGGCY	EA	$1 \cdot 10^{-30}$	*P. marinus*	prochlorosin (Tang and van der Donk, 2012)
STANDRAD	lanthipeptides	GVAGGGGGCDGIRITDKQTVADNTIVPCSCFHQ	EA	$1 \cdot 10^{-26}$	*P. marinus*	prochlorosin (Tang and van der Donk, 2012)
STANDRAD	lanthipeptides	VSPQSTIVCVSLRICNWSLRFCPSFKVRCPM	E	$3 \cdot 10^{-24}$	*G. thermodenitrificans*	geobacillin (Garg et al., 2012)
STANDRAD	glycocins	GLGKAQCAALWLQCASGGTIGCGGGAVACQNYRQFCR	EB	$1 \cdot 10^{-17}$	*B. subtilis*	sublancin (Paik et al., 1998)
STANDRAD	lanthipeptides	TTWPCATVGVSVALCPTTKCTSQC	EB	$2 \cdot 10^{-17}$	*B. halodurans*	haloduracin (McClerren et al., 2006)
STANDRAD	lanthipeptides	KGGSGVIHTISHECNMNSWQFVFTCCS	EB	$1 \cdot 10^{-14}$	*L. lactis*	lacticin 481 (Rince et al., 1994)
STANDRAD	lanthipeptides	AVEQRATPATPATPWLIKASYVVSGAGVSFVASYITVN	E	$6 \cdot 10^{-21}$	*B. cereus*	bicereucin alpha (Huo and van der Donk, 2016)
STANDRAD	lanthipeptides	AVEQRATPTLATPLTPHTPYATYVVSGGVVSAISGIFSNNKTCLG	EA	$8 \cdot 10^{-14}$	*B. cereus*	bicereucin beta (Huo and van der Donk, 2016)
STANDRAD	lanthipeptides	NITGAGSTIQCVNTTIGTILSVVFDCCPTSACTPPCRF	E	$5 \cdot 10^{-55}$	*R. flavefaciens*	flavecin A2.d (Zhao and van der Donk, 2016)
STANDRAD	lanthipeptides	NMAGAGSPLTVTITGLIVAATTGFDWCPTGACTYSCRV	E	$9 \cdot 10^{-44}$	*R. flavefaciens*	flavecin A2.b (Zhao and van der Donk, 2016)
STANDRAD	lanthipeptides	TTGGASTVNTVGIHTTYLISKGLQNCPLKPTTILPILPRK	E	$2 \cdot 10^{-39}$	*R. flavefaciens*	flvecin A2.a (Zhao and van der Donk, 2016)
STANDRAD	lanthipeptides	TTVGAASTLPCAEVVVTVTGIIVKATTGFDWCPTGACTHSCRF	E	$2 \cdot 10^{-27}$	*R. flavefaciens*	flavecin A2.g (Zhao and van der Donk, 2016)
STANDRAD	lanthipeptides	TTVGAGSSNDCADLILKITGVVVSATSKFDWCPTGACTTSCRF	E	$3 \cdot 10^{-25}$	*R. flavefaciens*	flavecin A2.c (Zhao and van der Donk, 2016)
STANDRAD	lanthipeptides	KQTIVCTIAQGTVGCLVSYGLGNGGYCCTYTVECSKTCNK	EB	$8 \cdot 10^{-25}$	*R. flavefaciens*	flavecin A1 (Zhao and van der Donk, 2016)
STANDRAD	lanthipeptides	MFDDSVVGAVGYTTYWGILPLVTKNPQICPVSENTVKCRLL	E	$1 \cdot 10^{-17}$	*R. flavefaciens*	flavecin A2.f (Zhao and van der Donk, 2016)
STANDRAD	lanthipeptides	SNVIGGTSSIDCVRLASNTPEGTVNLTVRIEFCPSAACTYSCRL	E	$4 \cdot 10^{-17}$	*R. flavefaciens*	flavecin A2.h (Zhao and van der Donk, 2016)
ACTI	linaridins	ATPAVAQFVIQGSTICLVC	EB	$3 \cdot 10^{-36}$	*S. griseus*	grisemycin (Claesen and Bibb, 2011; Kersten et al., 2011)
ACTI	lanthipeptides	TGSQVSLLVCEYSSLSVVLCTP	EAB	$4 \cdot 10^{-12}$	*S. griseus*	AmfS (Ueda et al., 2002)
ACTI	lanthipeptides	TVTVCSPTGTLCGSCSMGTRGCC	EA	$9 \cdot 10^{-12}$	*S. roseosporous*	SRO-3108 (Kersten et al., 2011)
ACTI	lanthipeptides	ASTVSLLSCISAASVLLCL(-64Da)	EA	$2 \cdot 10^{-18}$	*S. viridochromogenes*	informatipeptin (Mohimani et al., 2014)
ACTI	lanthipeptides	TGSRASLLLCGDSSLSITTCN	EAB	$6 \cdot 10^{-11}$	S. coelicolor	SapB (Kodani et al., 2004)
ACTI	lanthipeptides	**GGGASTVSLLSCVSAGSVILCV**	EA	$5 \cdot 10^{-12}$	*S. cattlya*	**informatipeptin B** (Figure S2)
ACTI	lanthipeptides	**DTGGCSGLCTVLVCTVIVC**	EA	$9 \cdot 10^{-11}$	*Streptomyces* sp. CNT360	**Compound CNT360-1769** (Figure S3)
ACTI	lassopeptide	**LLGRHGNDRLILSKN**(-59Da)	E	$5 \cdot 10^{-31}$	*S. viridochromogenes*	**Compound Svirid-1645** (Figure S4)
BACIL	lanthipeptides	TTPATTSSWTCITAGVTVSASLCPTTKCTSRC	EB	$2 \cdot 10^{-14}$	*B. licheniforis ES-221*	lichenicidin (Begley et al., 2009)
SPACE	lanthipeptides	**DATITTVTVTSTSIWASTVSNHC**(-87Da)	E	$3 \cdot 10^{-25}$	*Bacillus* sp. ISSFR-3F	**Compound Bac-ISS-2196** (Figure S5)
SPONGE	proteusins	TGIGVVVAVVAGAVANTGAGVNQVAGGNINVVGNINVVNANVSVNMNQTT	E	$1 \cdot 10^{-20}$	*T. swinhoei symbiont*	polytheonamide (Freeman et al., 2012; Hamada et al., 2005)
CYANO	cyanobactins	Cyclic(**ISNGYLIP**)	E	$2 \cdot 10^{-17}$	PNG22APR06-1	**Cyanobactin X** (Figure S6)
CYANO	cyanobactins	Cyclic(ISAPPGVTFSFP)	E	$2 \cdot 10^{-22}$	PNG14DEC03	Wewakazole (Figure S7) (Nogle et al., 2003)
HUMAN-iso	phenol-soluble modulins	MAADIISTIGDLVKWIIDTVNK	E	$1 \cdot 10^{-27}$	*S. epidermidis* 1457	delta-toxin (Figure S8) (Nolte and Kapral, 1981)
HUMAN-iso	phenol-soluble modulins	**MENIFNLFIKFFTTILE**	E	$1 \cdot 10^{-32}$	*S. epidermidis* 12228	**PSM-2147** (Figure S14)
HUMAN-CF	AIP	INCDFLL	E	$4 \cdot 10^{-9}$	*S. aureus* VVP0292	AIP-III (Figure S9)(Kalkum et al., 2003)
HUMAN-CF	phenol-soluble modulins	MDFTGVITSIIDLIKTCIQAFG	E	$6 \cdot 10^{-43}$	*S. aureus* VVP0114	Mec-PSM (Figure S10) (Kaito et al., 2011; Queck et al., 2009)
HUMAN-CF	phenol-soluble modulins	MAQDIISTIGDLVKW	E	$6 \cdot 10^{-47}$	*S. aureus* VVP0033	delta-hemolysin (Figure S11) (Fitton et al., 1980)
HUMAN-CF	phenol-soluble modulins	**MAQDIISTIGDLVKR**	E	$1 \cdot 10^{-27}$	*S. aureus* VVP0045	**delta-hemolysin-W-R** (Figure S12)
HUMAN-CF	phenol-soluble modulins	MEGLFNAIKDTVTAAINNDGAKLGT	E	$2 \cdot 10^{-11}$	*S. aureus* VVP0304	PSM-beta (Figure S13) (Wang et al., 2007)

Table S2. 31 known RiPPs identified by MetaMiner, related to Figure 2. In order to evaluate the consistency of MetaMiner identifications with the previous results, we compared the strains/species/samples in which we discovered known RiPPs to the previous reports. The results show that out of 31 known RiPPs, 23 were identified in strains identical to the previous reports, three were identified in strains with 99% or higher 16S rRNA similarity, two were identified in the same species, one was identified in the same genus, and two were identified in the same samples.

RiPP name	Current strain/sample	Original strain/sample	Similarity
Prochlorosin	*Prochlorococcus marinus* MIT9313	*Prochlorococcus marinus* MIT9313 (Tang and van der Donk, 2012)	The same strain
Prochlorosin	*Prochlorococcus marinus* MIT9313	*Prochlorococcus marinus* MIT9313 (Tang and van der Donk, 2012)	The same strain
Prochlorosin	*Prochlorococcus marinus* MIT9313	*Prochlorococcus marinus* MIT9313 (Tang and van der Donk, 2012)	The same strain
Prochlorosin	*Prochlorococcus marinus* MIT9313	*Prochlorococcus marinus* MIT9313 (Tang and van der Donk, 2012)	The same strain
geobacillin	*Geobacillus thermodenitrificans* NG80-2	*Geobacillus thermodenitrificans* NG80-2 (Garg et al., 2012)	The same strain
sublancin	*Bacillus subtilis* 168	*Bacillus subtilis* 168 (Paik et al., 1998)	The same strain
haloduracin	*Bacillus halodurans* C-125	*Bacillus halodurans* C-125 (McClerren et al., 2006)	The same strain
lacticin 481	*Lactococcus lactis* ADRIA 85LO30	*Lactococcus lactis* ADRIA 85LO30 (Rince et al., 1994)	The same strain
bicereucin alpha	*Bacillus cereus* SJ1	*Bacillus cereus* SJ1 (Huo and van der Donk, 2016)	The same strain
bicereucin beta	*Bacillus cereus* SJ1	*Bacillus cereus* SJ1 (Huo and van der Donk, 2016)	The same strain
flavecin A2.d	*Ruminococcus flavefaciens* FD-1	*Ruminococcus flavefaciens* FD-1 (Zhao and van der Donk, 2016)	The same strain
flavecin A2.b	*Ruminococcus flavefaciens* FD-1	*Ruminococcus flavefaciens* FD-1 (Zhao and van der Donk, 2016)	The same strain
flvecin A2.a	*Ruminococcus flavefaciens* FD-1	*Ruminococcus flavefaciens* FD-1 (Zhao and van der Donk, 2016)	The same strain
flavecin A2.g	*Ruminococcus flavefaciens* FD-1	*Ruminococcus flavefaciens* FD-1 (Zhao and van der Donk, 2016)	The same strain
flavecin A2.c	*Ruminococcus flavefaciens* FD-1	*Ruminococcus flavefaciens* FD-1 (Zhao and van der Donk, 2016)	The same strain
flavecin A1	*Ruminococcus flavefaciens* FD-1	*Ruminococcus flavefaciens* FD-1 (Zhao and van der Donk, 2016)	The same strain
flavecin A2.f	*Ruminococcus flavefaciens* FD-1	*Ruminococcus flavefaciens* FD-1 (Zhao and van der Donk, 2016)	The same strain
flavecin A2.h	*Ruminococcus flavefaciens* FD-1	*Ruminococcus flavefaciens* FD-1 (Zhao and van der Donk, 2016)	The same strain
grisemycin	*Streptomyces griseus* IFO 13350	*Streptomyces griseus* IFO 13350(Claesen and Bibb, 2011)	The same strain
AmfS	*Streptomyces griseus* IFO 13350	*Streptomyces griseus* IFO 13350 (Ueda et al., 2002)	The same strain
SRO15-3108	*Streptomyces roseosporus* NRRL 15998	*Streptomyces roseosporus* NRRL 15998 (Kersten et al., 2011)	The same strain
informatipeptin	*Streptomyces viridochromogenes* DSM 40736	*Streptomyces viridochromogenes* DSM 40736 (Mohimani et al., 2014)	The same strain
SapB	*Streptomyces coelicolor* J1501	*Streptomyces coelicolor* J1501 (Kodani et al., 2004)	The same strain
lichenicidin	*Bacillus licheniformis ES-221*	*Bacillus licheniformis* ATCC 14580 (Begley et al., 2009)	The same species
polytheonamide	*Theonella swinhoei* symbionts	*Theonella swinhoei* symbionts (Freeman et al., 2012; Hamada et al., 2005)	The same sample
Wewakazole	PNG14DEC03	PNG14DEC03 (Nogle et al., 2003)	The same sample
delta-toxin	*S. epidermidis* 1457	*Staphylococcus aureus* PG114 (Nolte and Kapral, 1981)	The same genus
AIP-III	*S. aureus* VVP0292	*Staphylococcus aureus* RN8465 (Kalkum et al., 2003)	The same species
Mec-PSM	*S. aureus* VVP0114	*Staphylococcus aureus* NCTC8325(Kaito et al., 2011)	The same species, 99.87% 16S similarity
PSM-beta	*S. aureus* VVP0304	*Staphylococcus aureus* MW2 (Wang et al., 2007)	The same species, 100% 16S similarity
delta-hemolysin	*S. aureus* VVP0033	*Staphylococcus aureus* RN25 (Fitton et al., 1980)	The same species, 99% 16S similarity

Table S3. List of RiPP modifications considered in this paper, related to Figure 1. The list is curated from (Arnison et al., 2013).

RiPP-type	enzyme (Pfam/TIGR)	modification / mass-shift	pos
Lanthipeptides	Lant-dehydration	TS-18, TS-C	All
Lanthipeptides	ElxO	S-15	N-ter
Lanthipeptides	Flavoprotein	TS-64, C-80	C-ter
Lanthipeptides	LtnJ	S-16	all
Lanthipeptides	CinX	D+16	all
Lanthipeptides	Cinorf7	K-18	all
Lanthipeptides	DsbB	C-1	all
Lanthipeptides	MibHS	W+34	all
Lanthipeptides	MibO	P+16, P+32	all
Lanthipeptides	GarO	C+16	all
LAPs	YcaO	CTS-20, T-18	all
LAPs	Methyltransf-31	R+28	N-ter
Lassopeptides	Asn-synthase	DE-18, Cyclization to DE	all
Lassopeptides	DsbA	C-1	all
Linaridins	CypL	TS-18	all
Linaridins	CypM	AIL+28	N-ter
Linaridins	Flavoprotein	C-80	C-ter
Phenol-soluble modulins	AgrB/MreB	M+28/M+44 *	all
Proteusins	Cupin-4	NV+16/N+30	all
Proteusins	MTS	N+14	all
Proteusins	B12-Binding	M+44/IQVT+14	all
Proteusins	DUF4135	D+39	N-ter
Cyanobactins	dehydratase	CTS-18,CTS-20, YFST+68, I+136	all
Glycocins	Glucose-Cys	C+162	all
Glycocins	DsbB	C-1	all
AIP	AgrB	C-18	all

Figure S1. Comparison of performance of MetaMiner in all-ORF and motif-ORF modes, related to Figure 1. Number of peptide-spectrum matches/peptides identified by MetaMiner at different p-value thresholds for the target and decoy databases in the search of ACTI dataset in the all-ORF (a,b) and motif-ORF (c,d) modes. X-axis represents different p-value threshold, and Y-axis represent number of identified peptide-spectrum match or peptides. These distributions are very consistent with the distribution of peptide spectrum match scores/p-values reported in the database search using Dereplicator and Dereplicator+, and also the distributions reported by Fenyo et al. (Fenyo and Beavis, 2003). Here, the distributions are the truncated Gumbel distributions, as we only looked at p-values below 10^{-5}.

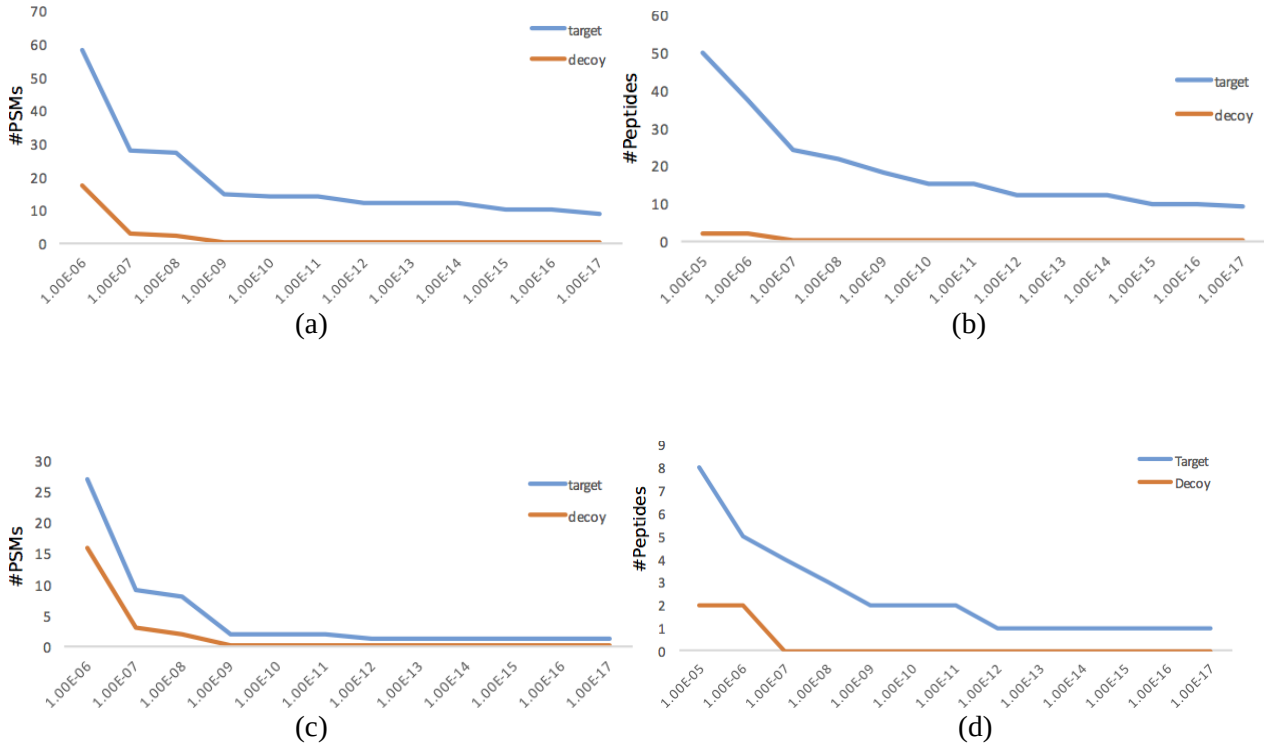
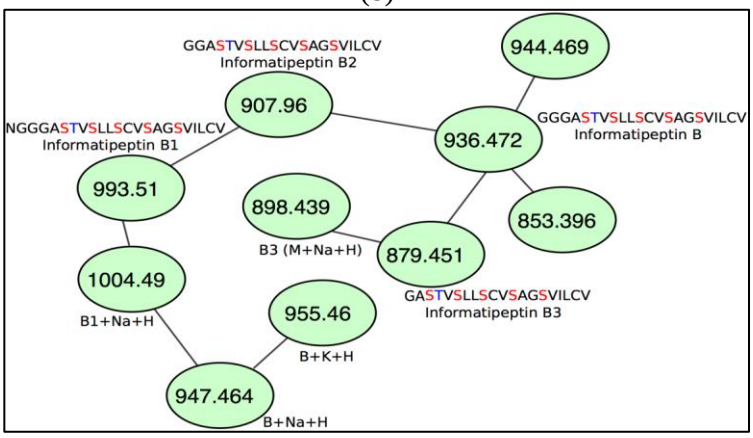


Figure S2. Identification of informatipeptin B, related to Figure 2. (a) Biosynthetic gene cluster of informatipeptin B from *S. cattleya* and the precursor peptide correctly predicted by antiSMASH. (b) The BGC of informatipeptin B has all essential class III lanthipeptide genes, a lanM-like enzyme, a transporter, and a regulator. (c) Spectral network revealed a plethora of compounds similar to informatipeptin B. Masses shown here are in charge +2 state. Dehydrated serines are shown in red and dehydrated threonines are shown in blue (d) Tandem mass spectrum of informatipeptin B (score 7, p-value $5 \cdot 10^{-12}$).

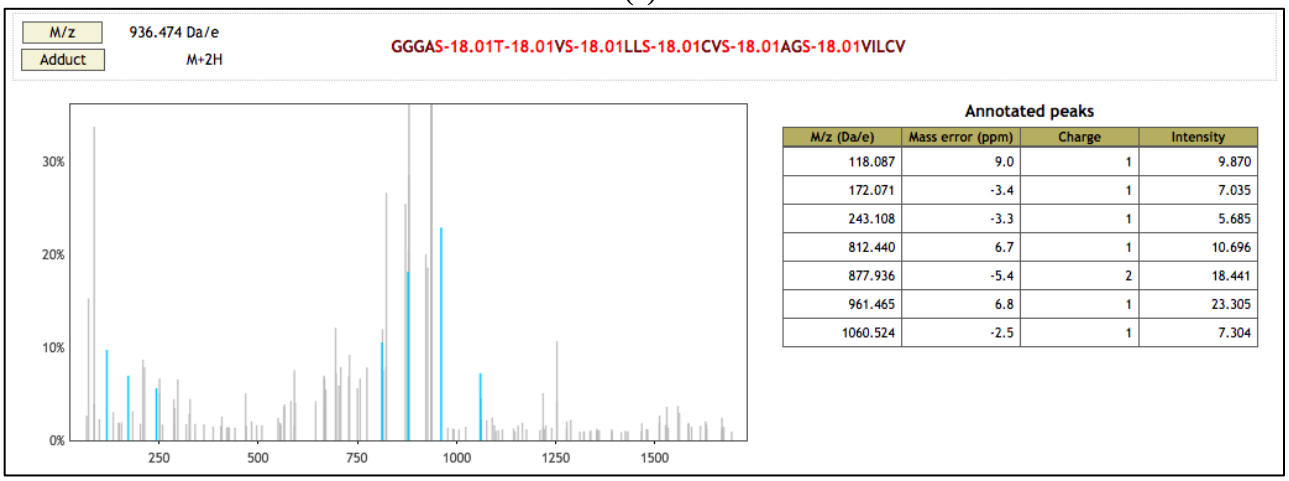
(a)

Gene	Predicted Function	Length
1	glycosyl transferase family 1	468
2	TetR family transcriptional regulator	208
3	helix-turn-helix transcriptional regulator	239
4	ABC transporter ATP-binding protein	591
5	ABC transporter, AmfB	737
6	AmfS protein	37
7	serine/threonine protein kinase	882
8	aldehyde dehydrogenase	488
9	metal dependent hydrolase 1	351
10	hypothetical protein	410

(b)



(c)



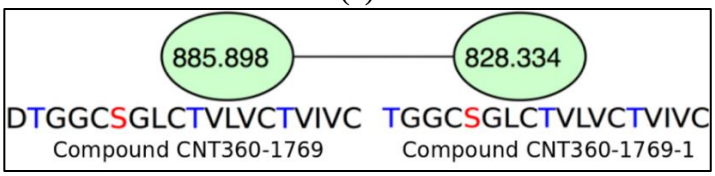
(d)

Figure S3. Identification of Compound CNT360-1769, related to Figure 2. (a) Biosynthetic gene cluster of Compound CNT360-1769 from *Streptomyces* sp. and the precursor peptide as correctly predicted by antiSMASH. (b) The gene cluster of Compound CNT360-1769 has several genes with approximately 60% similarity to genes from BGCs encoding subtilin and other class I lanthipeptide, suggesting that Compound CNT360-1769 is a class I lanthipeptide. (c) Spectral networking revealed an analog of Compound CNT360-1769, called Compound CNT360-1769-1, lacking the N-terminal aspartic acid residue. Dehydrated serines are shown in red and dehydrated threonines are shown in blue (d) Tandem mass spectrum of Compound CNT360-1769 (score 7, p-value $9.8 \cdot 10^{-11}$).

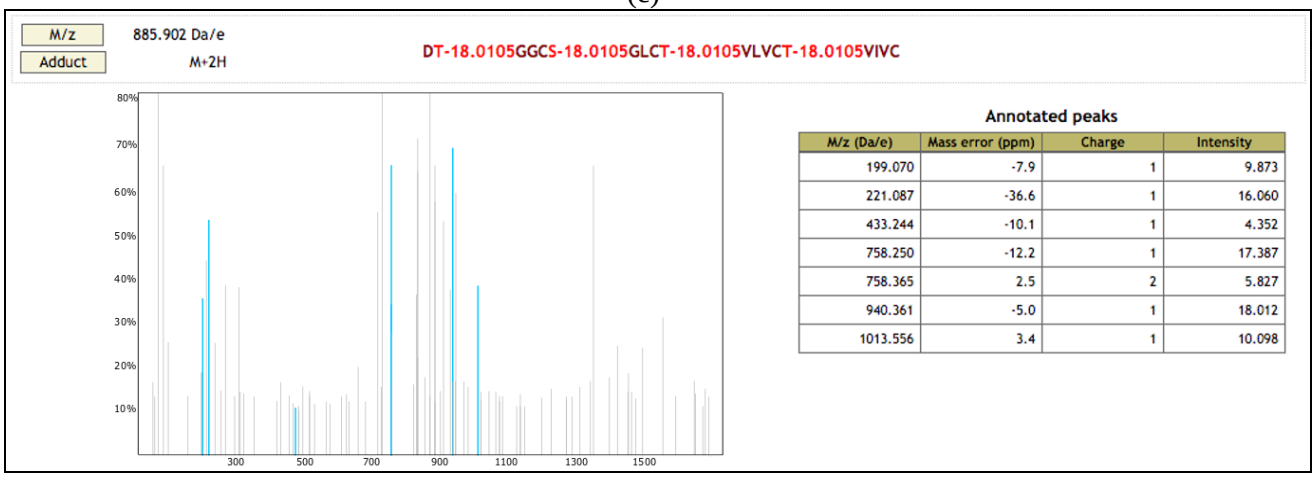
(a)

Gene	Predicted Function	Length
1	TetR family transcriptional regulator	175
2	NADPH: quinone reductase	327
3	hypothetical protein TGGTI 239580	57
4	urea transporter	314
5	subtilin biosynthesis protein spaC	450
6	hypothetical protein K530 51195	53
7	lantibiotic dehydratase	954
8	O-methyltransferase clustered with LanBC	400
9	NAD(P)H dehydrogenase	102
10	oxctanocin A resistance protein	324
11	translation initiation factor IF-2	269
12	transcriptional regulator, PucR family protein	536
13	MerR family transcriptional regulator	131

(b)



(c)



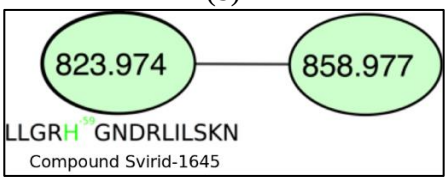
(d)

Figure S4. Identification of Compound Svirid-1645, related to Figure 2. (a) Biosynthetic gene cluster of Compound Svirid-1645 from *S. viridochromogenes*. AntiSMASH failed to predict any precursor peptide for this gene cluster. (b) The biosynthetic gene cluster of Compound Svirid-1645 has several genes with similarity to the biosynthetic gene cluster of lassopeptide microcin J25. (c) Spectral network of Compound Svirid-1645 (d) Tandem mass spectrum of Compound Svirid-1645 (score 14, p-value $5 \cdot 10^{-31}$). MetaMiner discovered Compound Svirid-1645 in blind modification search mode, and assigned a modification of -59Da to the histidine residue, shown in green. While *m/z* 858.97 clusters with Compound Svirid-1645 in the spectral network, MetaMiner does not identify it with a low p-value.

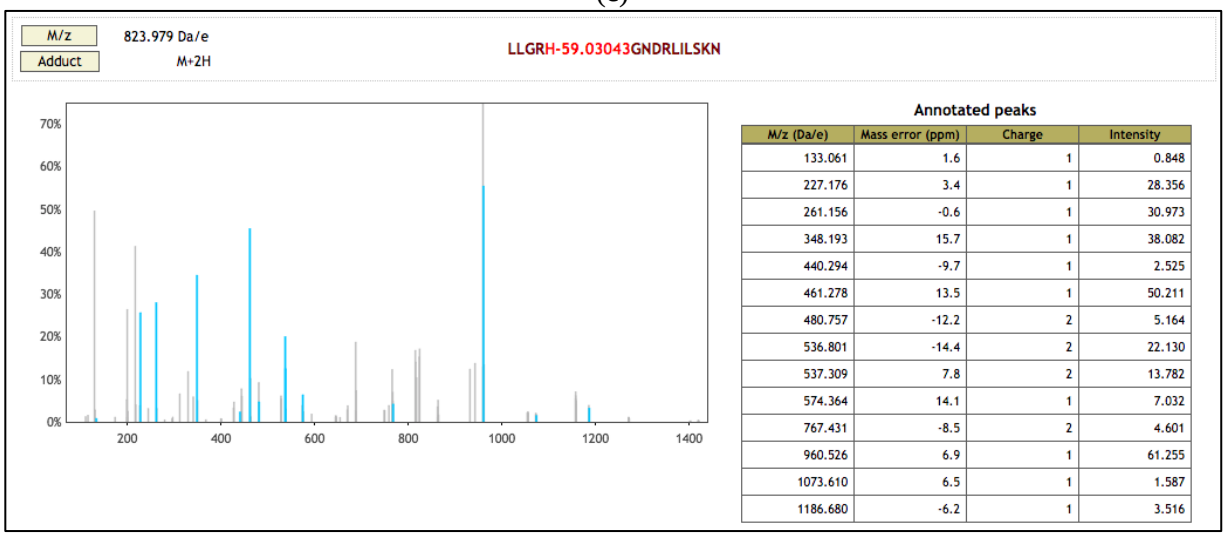
(a)

Gene	Predicted Function	Length
1	Transcriptional regulator	253
2	Endo-1,4-beta-xylanase	1620
3	Dihydrolipyl dehydrogenase	470
4	Putative microcin mature peptide	36
5	Asparagine Synthetase B/C	617
6	PQQ synthetase	84
7	Microcin J25 C	145
8	Lipid A export permease/ATP-binding protein	613
9	Calcium-binding protein	205
10	2-keto-3deoxy-gluconate kinase	331
11	adolase	260
12	3-oxoacyl-[acyl-carrier-protein] reductase	170
13	TetR family transcriptional regulator	218
14	Cytochrome P450 monooxygenase	437
15	Dihydroxy-acid dehydratase	571
16	Regulatory protein	275
17	Alcohol dehydrogenase	336

(b)



(c)



Annotated peaks

M/z (Da/e)	Mass error (ppm)	Charge	Intensity
133.061	1.6	1	0.848
227.176	3.4	1	28.356
261.156	-0.6	1	30.973
348.193	15.7	1	38.082
440.294	-9.7	1	2.525
461.278	13.5	1	50.211
480.757	-12.2	2	5.164
536.801	-14.4	2	22.130
537.309	7.8	2	13.782
574.364	14.1	1	7.032
767.431	-8.5	2	4.601
960.526	6.9	1	61.255
1073.610	6.5	1	1.587
1186.680	-6.2	1	3.516

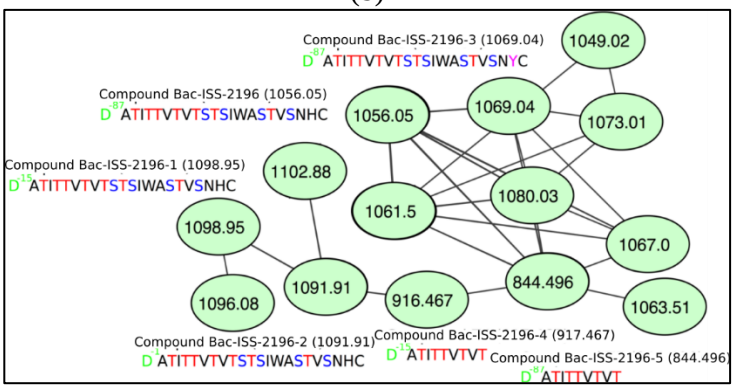
(d)

Figure S5. Identification of Compound Bac-ISS-2196, related to Figure 2. (a) Biosynthetic gene cluster of Compound Bac-ISS-2196 in *Bacillus* sp. ISSFR-3F. AntiSMASH failed to predict any precursor peptide for this gene cluster. The precursor gene is repeated fourteen times, with thirteen being exactly the same and one differing from the rest by a His to Tyr replacement (b) The BGC of Compound Bac-ISS-2196 (c) Spectral networking revealed analogs of Compound Bac-ISS-2196 with various N-terminal and C-terminal modifications. Dehydrated serines are shown in red and dehydrated threonines are shown in blue (d) Tandem mass spectrum of Compound Bac-ISS-2196 (score 13, p-value $3 \cdot 10^{-25}$). MetaMiner discovered Compound Bac-ISS-2196 in blind modification search mode, and assigned -87Da, -15Da, and -1Da modifications to the aspartic acid residue, shown in green residue. Compound Bac-ISS-2196-4 and Compound Bac-ISS-2196-5 are discovered in charge +1, while other compounds are charge +2.

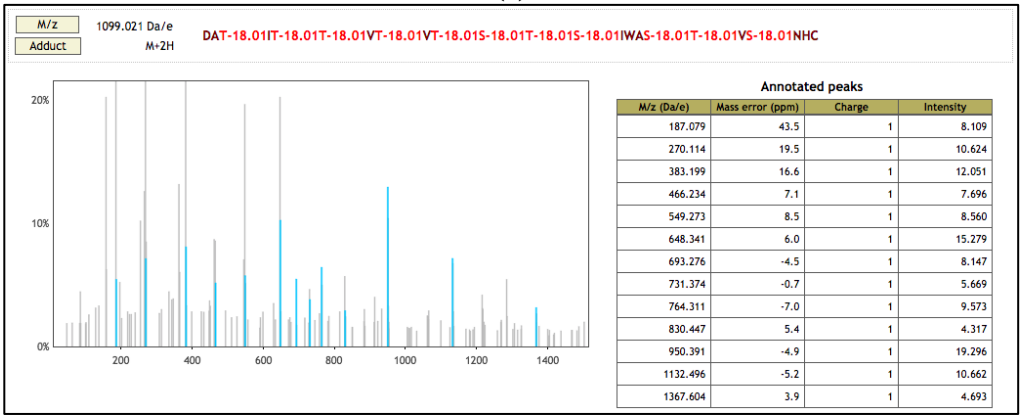
(a)

Gene	Predicted Function	Length
1	transcriptional regulator	117
2	enterotoxin	421
3	multidrug ABC transporter permease	555
4	putative SAM-dependent methyltransferase	277
5	serine/threonine protein kinase	872
6	ABC-2 transporter family protein	197
7	multidrug ABC transporter ATP-binding protein	284
8	peptidase M16	427
9	multidrug ABC transporter permease	201
10	sodium ABC transporter ATP-binding protein	290

(b)



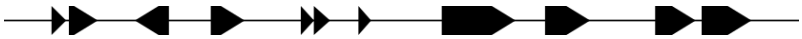
(c)



M/z (Da/e)	Mass error (ppm)	Charge	Intensity
187.079	43.5	1	8.109
270.114	19.5	1	10.624
383.199	16.6	1	12.051
466.234	7.1	1	7.696
549.273	8.5	1	8.560
648.341	6.0	1	15.279
693.276	-4.5	1	8.147
731.374	-0.7	1	5.669
764.311	-7.0	1	9.573
830.447	5.4	1	4.317
950.391	-4.9	1	19.296
1132.496	-5.2	1	10.662
1367.604	3.9	1	4.693

(d)

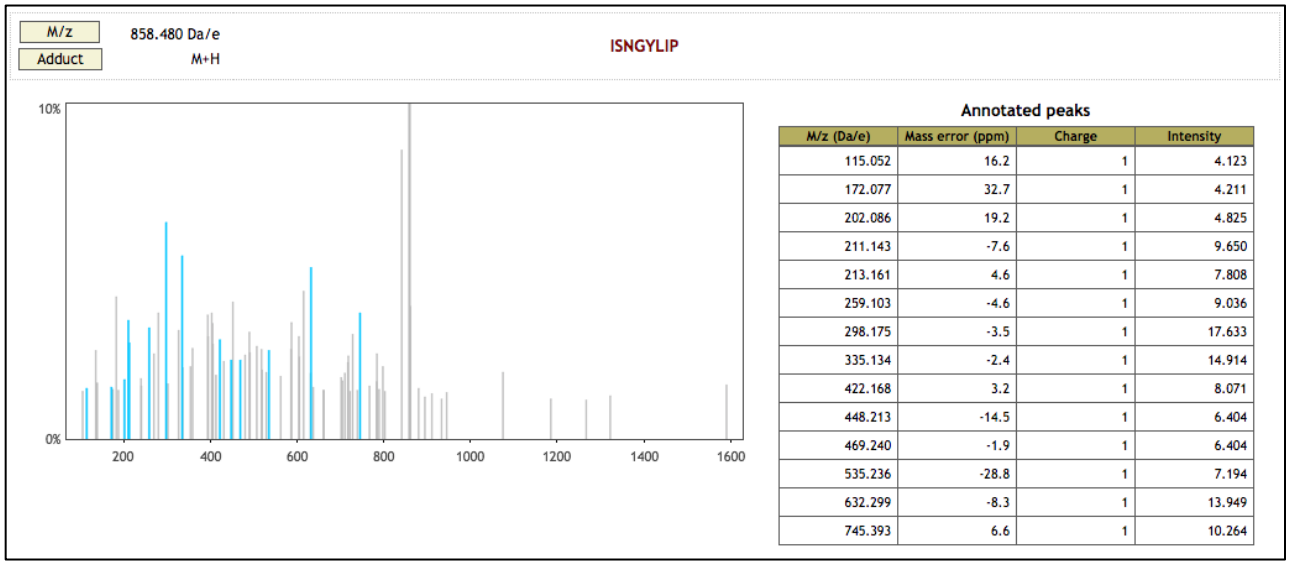
Figure S6. Identification of Cyanobactin X, related to Figure 2. (a) The biosynthetic gene cluster of cyanobactin X (b) Annotation of the genes in the biosynthetic gene cluster of cyanobactin X by DFAST (Tanizawa et al., 2018) and HMMER (Eddy, 2011) (c) Tandem mass spectrum of Cyanobactin X (score 14, p-value $2 \cdot 10^{-17}$).



(a)

Gene	Predicted Function	Length
1	hypothetical protein	62
2	hypothetical protein	122
3	Asparaginase N-terminal	144
4	Domain of unknown function DUF29	143
5	HicA toxin of bacterial toxin-antitoxin	56
6	conserved domain protein	69
7	Cyanobactin X	54
8	Acetyltransferase (GNAT) family5	311
9	protein of unknown function DUF820	190
10	Putative restriction endonuclease	171
11	Putative restriction endonuclease	210

(b)



Annotated peaks

M/z (Da/e)	Mass error (ppm)	Charge	Intensity
115.052	16.2	1	4.123
172.077	32.7	1	4.211
202.086	19.2	1	4.825
211.143	-7.6	1	9.650
213.161	4.6	1	7.808
259.103	-4.6	1	9.036
298.175	-3.5	1	17.633
335.134	-2.4	1	14.914
422.168	3.2	1	8.071
448.213	-14.5	1	6.404
469.240	-1.9	1	6.404
535.236	-28.8	1	7.194
632.299	-8.3	1	13.949
745.393	6.6	1	10.264

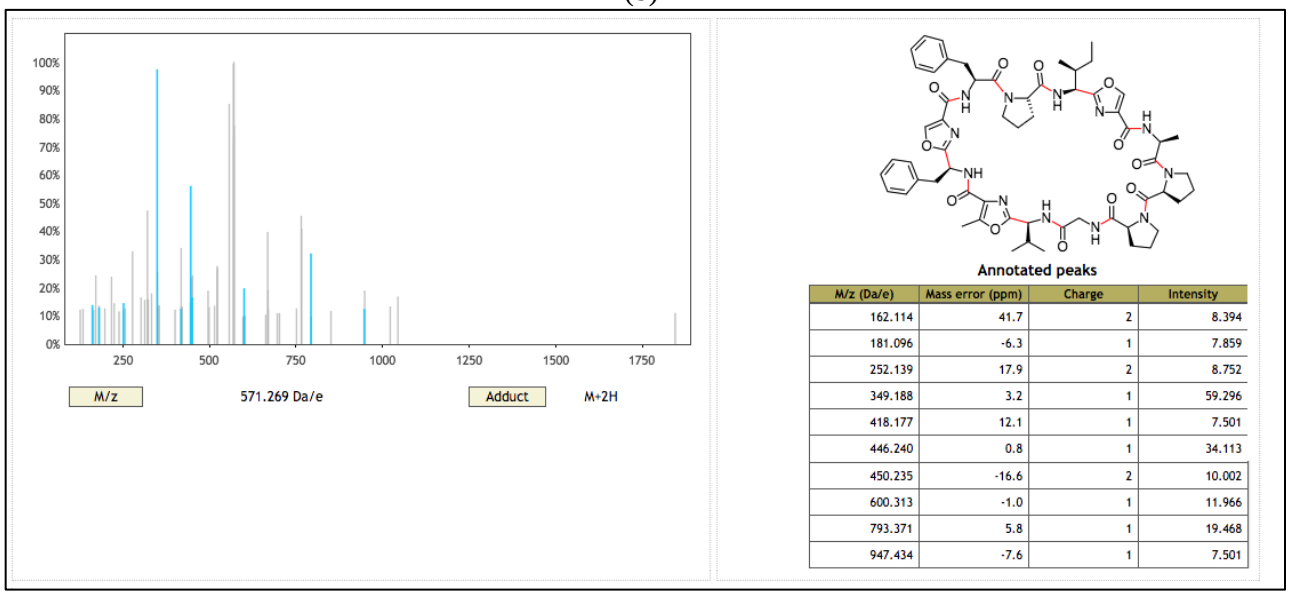
(c)

Figure S7. Identification of wewakazole, related to Figure 2. (a) Biosynthetic gene cluster of wewakazole in the Cyanobacterium PNG26APR06-4. AntiSMASH failed to predict the precursor peptide for this gene cluster by motif search; however, MetaMiner was able to discover this peptide in all-ORF mode. (b) Annotation of the BGC of wewakazole by DFAST (Tanizawa et al., 2018) and HMMER (Eddy, 2011) (c) Tandem mass spectrum of wewakazole (score 10, p-value $2 \cdot 10^{-22}$).

(a)

Gene	Predicted Function	Length
1	hypothetical protein	494
2	peptidase M16 domain protein	439
3	fructosamine kinase	295
4	four helix bundle protein	120
5	sporulation-control protein Spo0M homolog	257
6	PilT protein-like	148
7	hypothetical protein	86
8	methylglyoxal synthase-like protein putative	425
9	peptidase S8 and S53 subtilisin kexin sedolisin	738
10	Wewakazole	68
11	hypothetical protein	65
12	hypothetical protein	68
13	hypothetical protein	66
14	Putative restriction endonuclease	215
15	XisI protein fdxN element excision controlling factor protein	110
16	unknown protein	64
17	hypothetical protein	151
18	YcaO cyclodehydratase ATP-ad Mg2+-binding	827
19	S23 ribosomal protein	133
20	hypothetical protein	202
21	hypothetical protein	105
22	hypothetical protein	111
23	alpha-amylase family protein	554
24	hypothetical protein	545
25	methyltransferase type 11	278

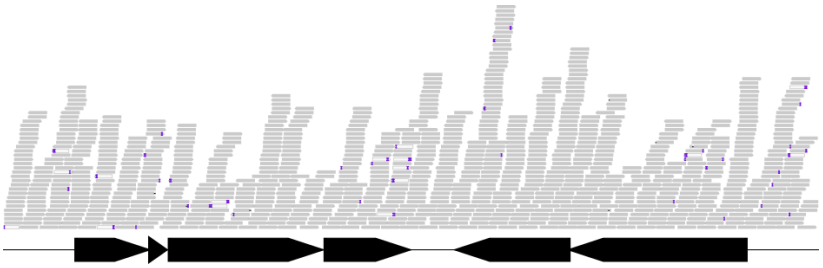
(b)



M/z (Da/e)	Mass error (ppm)	Charge	Intensity
162.114	41.7	2	8.394
181.096	-6.3	1	7.859
252.139	17.9	2	8.752
349.188	3.2	1	59.296
418.177	12.1	1	7.501
446.240	0.8	1	34.113
450.235	-16.6	2	10.002
600.313	-1.0	1	11.966
793.371	5.8	1	19.468
947.434	-7.6	1	7.501

(c)

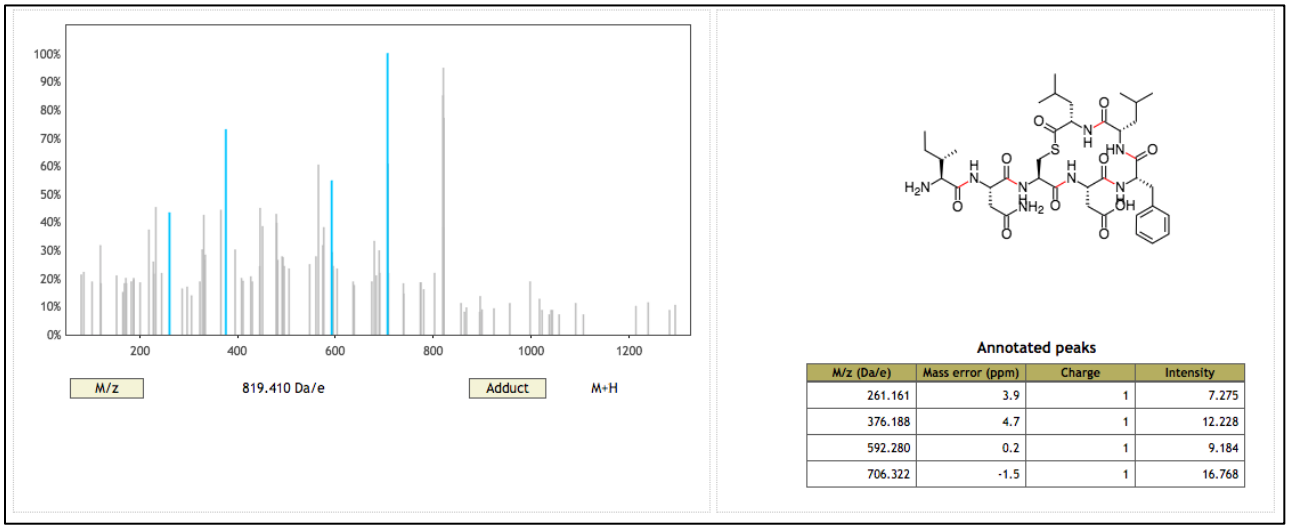
Figure S8. Identification of AIP-III from HUMAN-CF dataset, related to Figure 2. (a) Mapping of the short metagenomics reads from the HUMAN-CF dataset to the biosynthetic gene cluster of AIP-III. Both AntiSMASH and BOA failed to predict the precursor peptide for this gene cluster by motif search; however, MetaMiner was able to discover this peptide in all-ORF mode. (b) Annotation of the BGC of AIP-III by DFAST (Tanizawa et al., 2018) and HMMER (Eddy, 2011). (c) Tandem mass spectrum of AIP-III (score 4, p-value $4 \cdot 10^{-9}$).



(a)

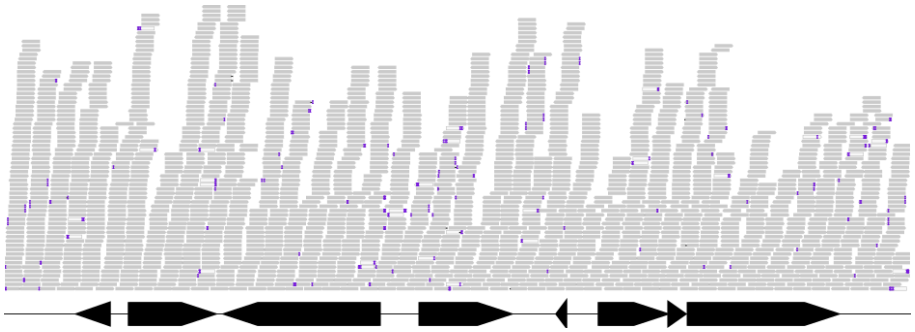
Gene	Predicted Function	Length
1	accessory gene regulator protein B	206
2	AIP-III	47
3	histidine kinase	431
4	accessory gene regulator protein A	239
5	fructokinase	320
6	invertase	498

(b)



(c)

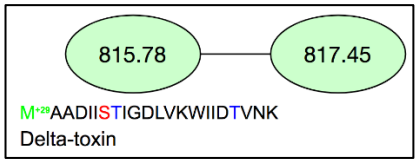
Figure S9. Identification of delta-toxin from HUMAN-iso dataset, related to Figure 2. (a) Mapping of the short metagenomics reads from the HUMAN-iso dataset to the biosynthetic gene cluster of delta-toxin. Both AntiSMASH and BOA failed to predict the precursor peptide for this gene cluster by motif search; however, MetaMiner was able to discover this peptide in all-ORF mode. (b) Annotation of the BGC of delta-toxin. (c) Spectral network of delta-toxin. (d) Tandem mass spectrum of delta-toxin (score 14, p-value $1 \cdot 10^{-27}$).



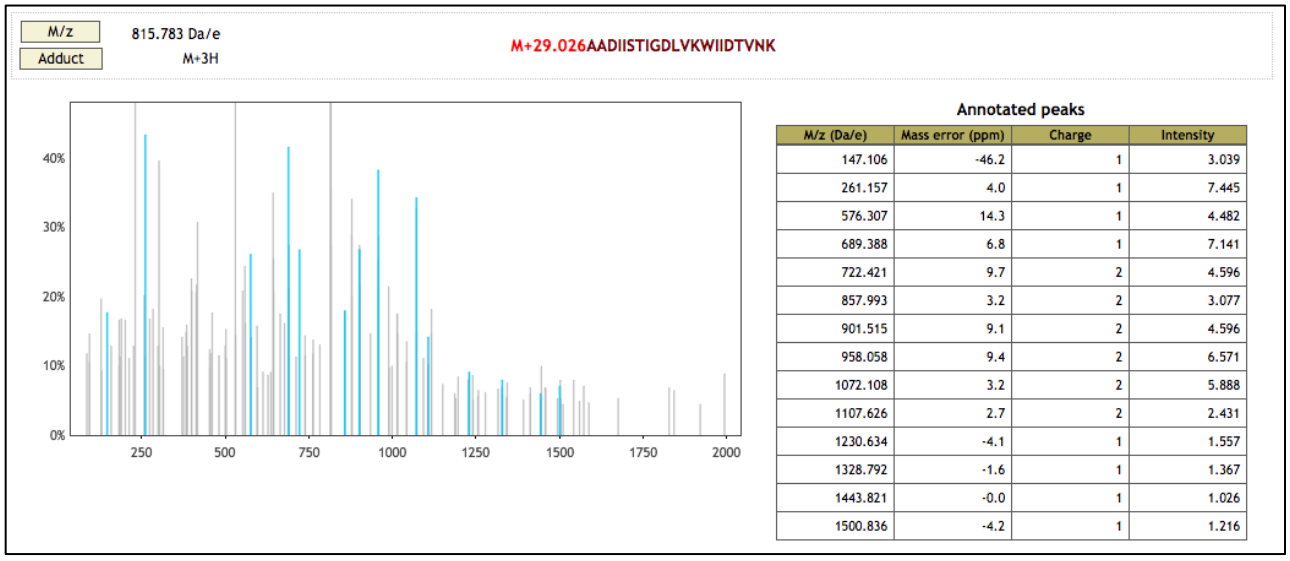
(a)

Gene	Predicted Function	Length
1	10 kDa chaperonin	95
2	CAAX amino protease	247
3	hypothetical protein	445
4	hydrolase	262
5	delta-toxin	25
6	accessory gene regulator protein B	197
7	accessory gene regulator protein D	47
8	histidine kinase	430

(b)



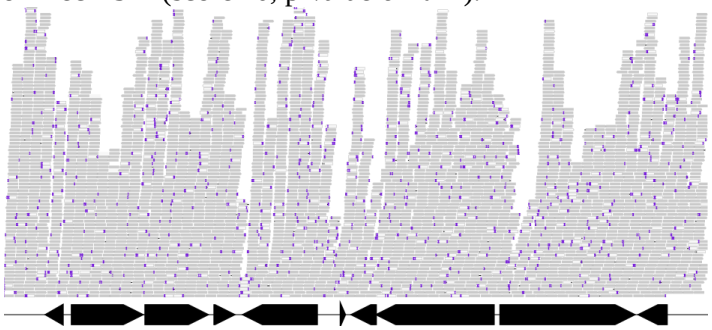
(c)



M/z (Da/e)	Mass error (ppm)	Charge	Intensity
147.106	-46.2	1	3.039
261.157	4.0	1	7.445
576.307	14.3	1	4.482
689.388	6.8	1	7.141
722.421	9.7	2	4.596
857.993	3.2	2	3.077
901.515	9.1	2	4.596
958.058	9.4	2	6.571
1072.108	3.2	2	5.888
1107.626	2.7	2	2.431
1230.634	-4.1	1	1.557
1328.792	-1.6	1	1.367
1443.821	-0.0	1	1.026
1500.836	-4.2	1	1.216

(d)

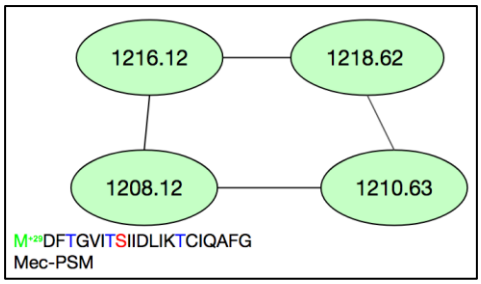
Figure S10. Identification of Mec-PSM from HUMAN-CF dataset, related to Figure 2. (a) Mapping of the short metagenomics reads from the HUMAN-CF dataset to the biosynthetic gene cluster of Mec-PSM. Both AntiSMASH and BOA failed to predict the precursor peptide for this gene cluster by motif search; however, MetaMiner was able to discover this peptide in all-ORF mode. (b) Annotation of the BGC of Mec-PSM by DFAST (Tanizawa et al., 2018) and HMMER (Eddy, 2011). (c) Spectral network of Mec-PSM. (d) Tandem mass spectrum of Mec-PSM (score 20, p-value $6 \cdot 10^{-43}$).



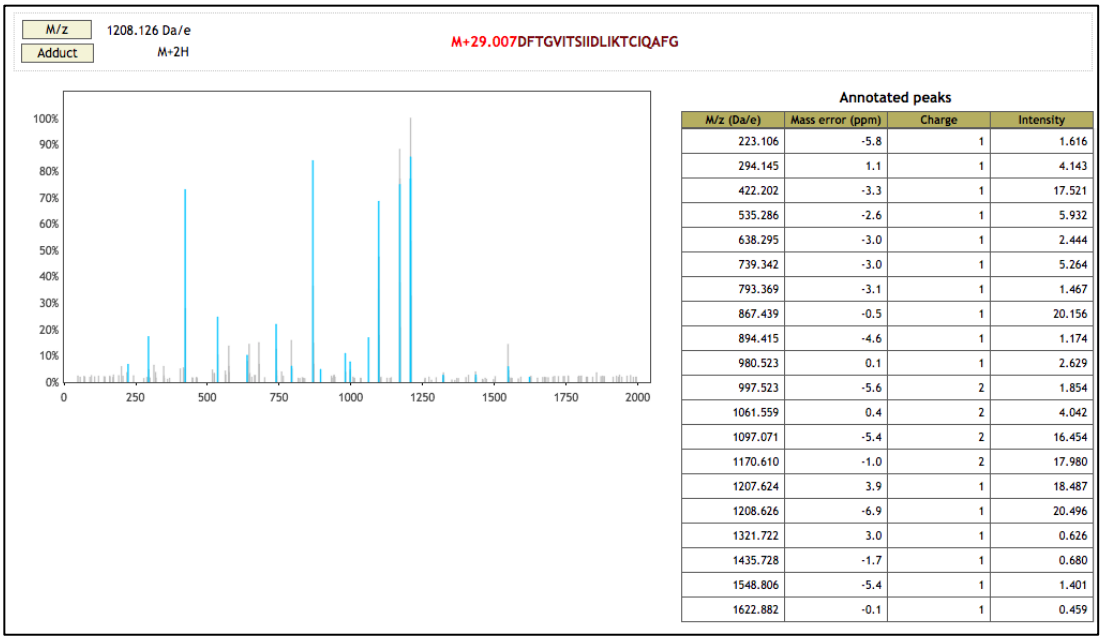
(a)

Gene	Predicted Function	Length
1	Metal-sensitive transcriptional repressor	87
2	dihydroneopterin aldolase	355
3	Metallo-beta-lactamase superfamily	313
4	Rhodanese-like domain	102
5	transcriptional regulator	370
6	Mec-PSM	22
7	penicillinase repressor BlaI	124
8	BlaR1 family beta-lactam sensor/signal transducer	586
9	penicillin-binding protein 3	669
10	MaoC like domain	143

(b)



(c)

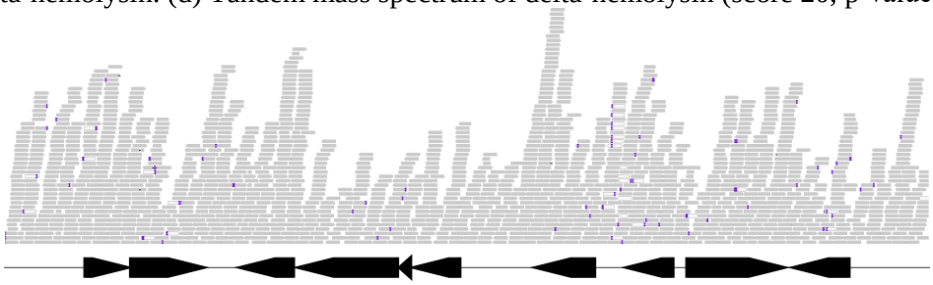


Annotated peaks

M/z (Da/e)	Mass error (ppm)	Charge	Intensity
223.106	-5.8	1	1.616
294.145	1.1	1	4.143
422.202	-3.3	1	17.521
535.286	-2.6	1	5.932
638.295	-3.0	1	2.444
739.342	-3.0	1	5.264
793.369	-3.1	1	1.467
867.439	-0.5	1	20.156
894.415	-4.6	1	1.174
980.523	0.1	1	2.629
997.523	-5.6	2	1.854
1061.559	0.4	2	4.042
1097.071	-5.4	2	16.454
1170.610	-1.0	2	17.980
1207.624	3.9	1	18.487
1208.626	-6.9	1	20.496
1321.722	3.0	1	0.626
1435.728	-1.7	1	0.680
1548.806	-5.4	1	1.401
1622.882	-0.1	1	0.459

(d)

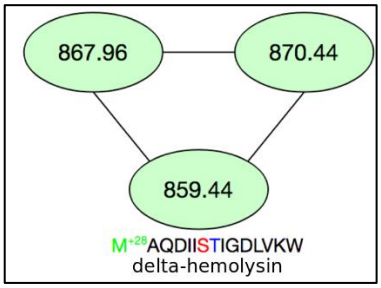
Figure S11. Identification of delta-hemolysin from HUMAN-CF dataset, related to Figure 2. (a) Mapping of the short metagenomics reads from the HUMAN-CF dataset to the biosynthetic gene cluster of delta-hemolysin. Both AntiSMASH and BOA failed to predict the precursor peptide for this gene cluster by motif search; however, MetaMiner was able to discover this peptide in all-ORF mode. (b) Annotation of the BGC of delta-hemolysin by DFAST (Tanizawa et al., 2018) and HMMER (Eddy, 2011). (c) Spectral network of delta-hemolysin. (d) Tandem mass spectrum of delta-hemolysin (score 20, p-value $6 \cdot 10^{-47}$).



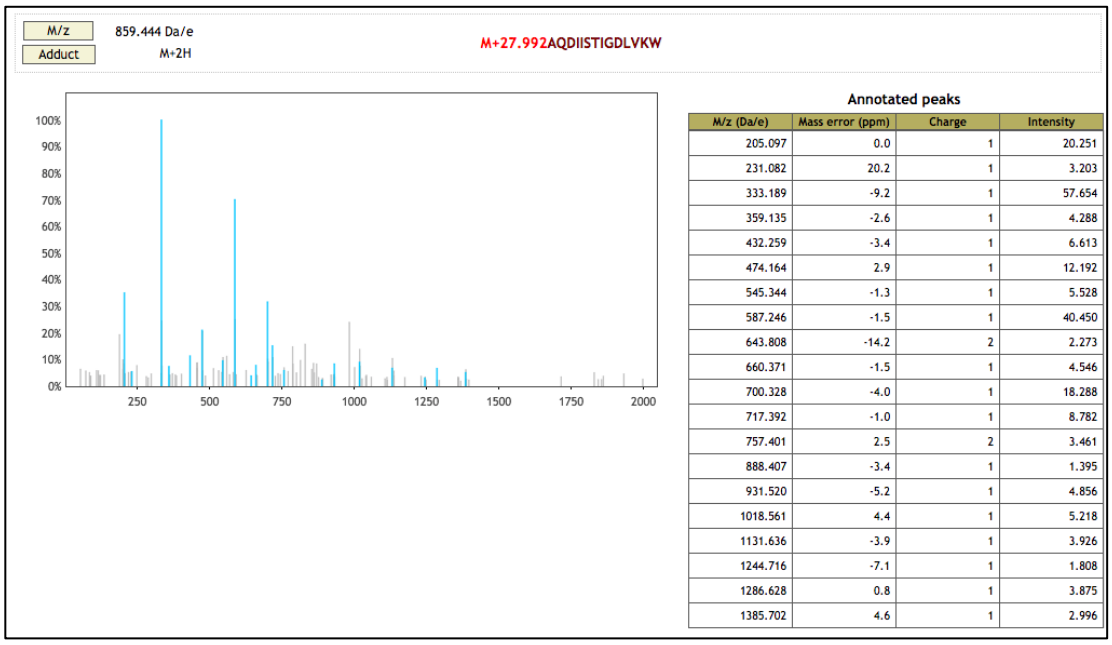
(a)

Gene	Predicted Function	Length
1	fructokinase	320
2	accessory gene regulator protein A	239
3	histidine kinase	431
4	accessory gene regulator protein D	47
5	accessory gene regulator protein B	208
6	PSM-y	29
7	hydrolase in agr operon	262
8	nitroreductase	209
9	hypothetical protein	418
10	CAAX amino protease	248

(b)



(c)

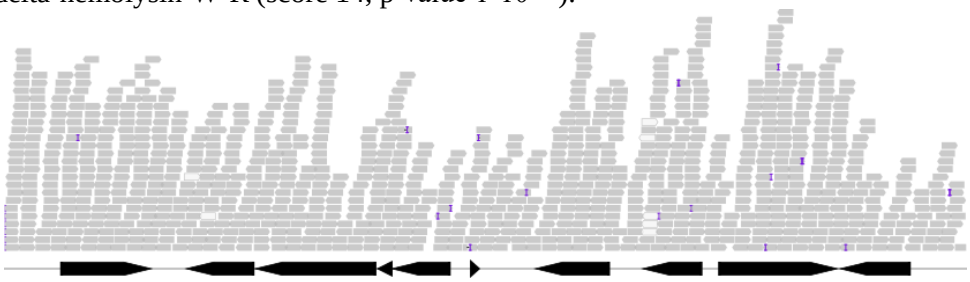


Annotated peaks

M/z (Da/e)	Mass error (ppm)	Charge	Intensity
205.097	0.0	1	20.251
231.082	20.2	1	3.203
333.189	-9.2	1	57.654
359.135	-2.6	1	4.288
432.259	-3.4	1	6.613
474.164	2.9	1	12.192
545.344	-1.3	1	5.528
587.246	-1.5	1	40.450
643.808	-14.2	2	2.273
660.371	-1.5	1	4.546
700.328	-4.0	1	18.288
717.392	-1.0	1	8.782
757.401	2.5	2	3.461
888.407	-3.4	1	1.395
931.520	-5.2	1	4.856
1018.561	4.4	1	5.218
1131.636	-3.9	1	3.926
1244.716	-7.1	1	1.808
1286.628	0.8	1	3.875
1385.702	4.6	1	2.996

(d)

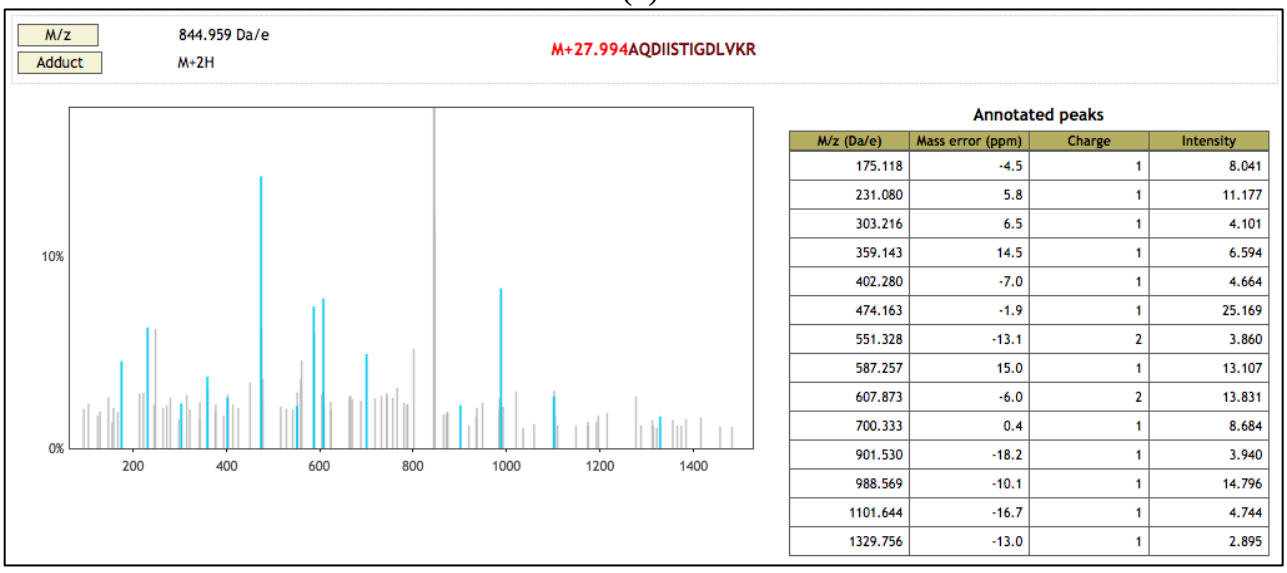
Figure S12. Identification of delta-hemolysin-W-R from HUMAN-CF dataset, related to Figure 2. (a) Mapping of the short metagenomics reads from the HUMAN-CF dataset to the biosynthetic gene cluster of delta-hemolysin-W-R. AntiSMASH failed to predict the precursor peptide for this gene cluster by motif search; however, MetaMiner was able to discover this peptide in all-ORF mode. (b) Annotation of the BGC of delta-hemolysin-W-R by DFAST (Tanizawa et al., 2018) and HMMER (Eddy, 2011). (c) Tandem mass spectrum of delta-hemolysin-W-R (score 14, p-value $1 \cdot 10^{-27}$).



(a)

Gene	Predicted Function	Length
1	fructokinase	320
2	accessory gene regulator protein A	239
3	histidine kinase	428
4	accessory gene regulator protein D	48
5	accessory gene regulator protein B	206
6	PSM-y-WR	27
7	hydrolase in agr operon	262
8	nitroreductase	209
9	hypothetical protein	420
10	CAAX amino protease	248

(b)

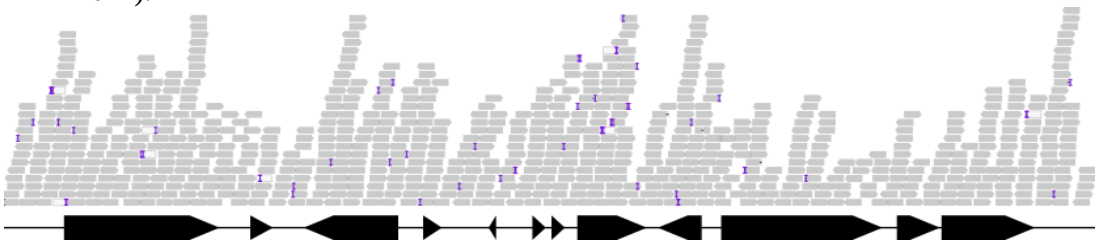


Annotated peaks

M/z (Da/e)	Mass error (ppm)	Charge	Intensity
175.118	-4.5	1	8.041
231.080	5.8	1	11.177
303.216	6.5	1	4.101
359.143	14.5	1	6.594
402.280	-7.0	1	4.664
474.163	-1.9	1	25.169
551.328	-13.1	2	3.860
587.257	15.0	1	13.107
607.873	-6.0	2	13.831
700.333	0.4	1	8.684
901.530	-18.2	1	3.940
988.569	-10.1	1	14.796
1101.644	-16.7	1	4.744
1329.756	-13.0	1	2.895

(c)

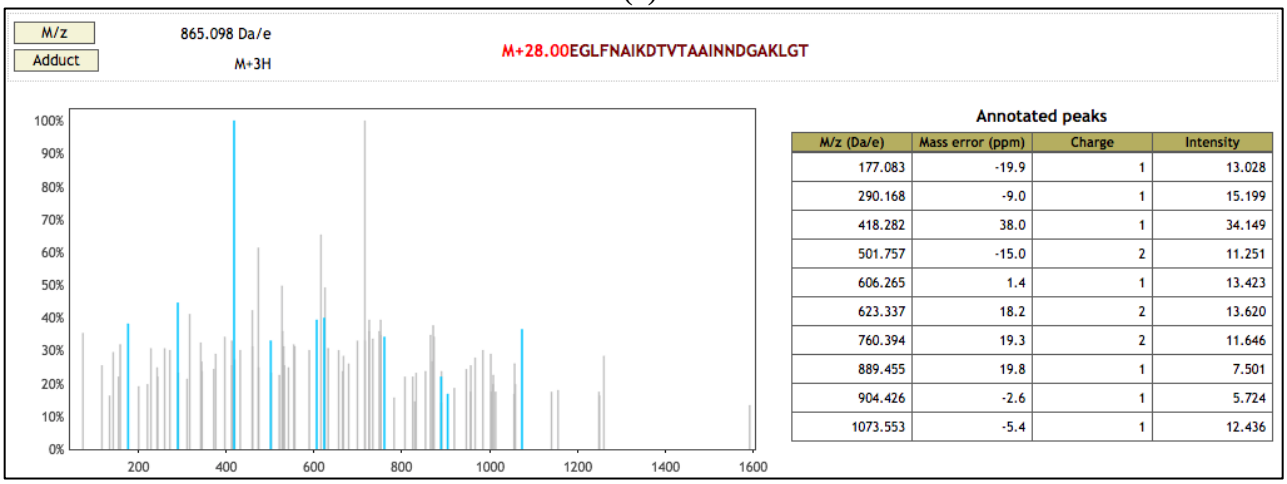
Figure S13. Identification of PSM-beta from HUMAN-CF dataset, related to Figure 2. (a) Mapping of the short metagenomics reads from the HUMAN-CF dataset to the biosynthetic gene cluster of PSM-beta. AntiSMASH failed to predict the precursor peptide for this gene cluster by motif search; however, MetaMiner was able to discover this peptide in all-ORF mode. (b) Annotation of the BGC of PSM-beta by DFAST (Tanizawa et al., 2018) and HMMER (Eddy, 2011). (c) Tandem mass spectrum of PSM-beta (score 10, p-value $2 \cdot 10^{-11}$).



(a)

Gene	Predicted Function	Length
1	C4-dicarboxylate anaerobic carrier	519
2	hypothetical protein	76
3	tellurite resistance protein related permease	316
4	hypothetical protein	63
5	tRNA-Arg	26
6	PSM-beta	45
7	hemolytic protein	45
8	HAD family hydrolase	232
9	putative N-acetyltransferase	147
10	putative cysteine ligase BshC	538
11	cell division protein MraZ	144
12	16S rRNA (cytosine(1402)-N(4))-methyltransferase	312

(b)



Annotated peaks

M/z (Da/e)	Mass error (ppm)	Charge	Intensity
177.083	-19.9	1	13.028
290.168	-9.0	1	15.199
418.282	38.0	1	34.149
501.757	-15.0	2	11.251
606.265	1.4	1	13.423
623.337	18.2	2	13.620
760.394	19.3	2	11.646
889.455	19.8	1	7.501
904.426	-2.6	1	5.724
1073.553	-5.4	1	12.436

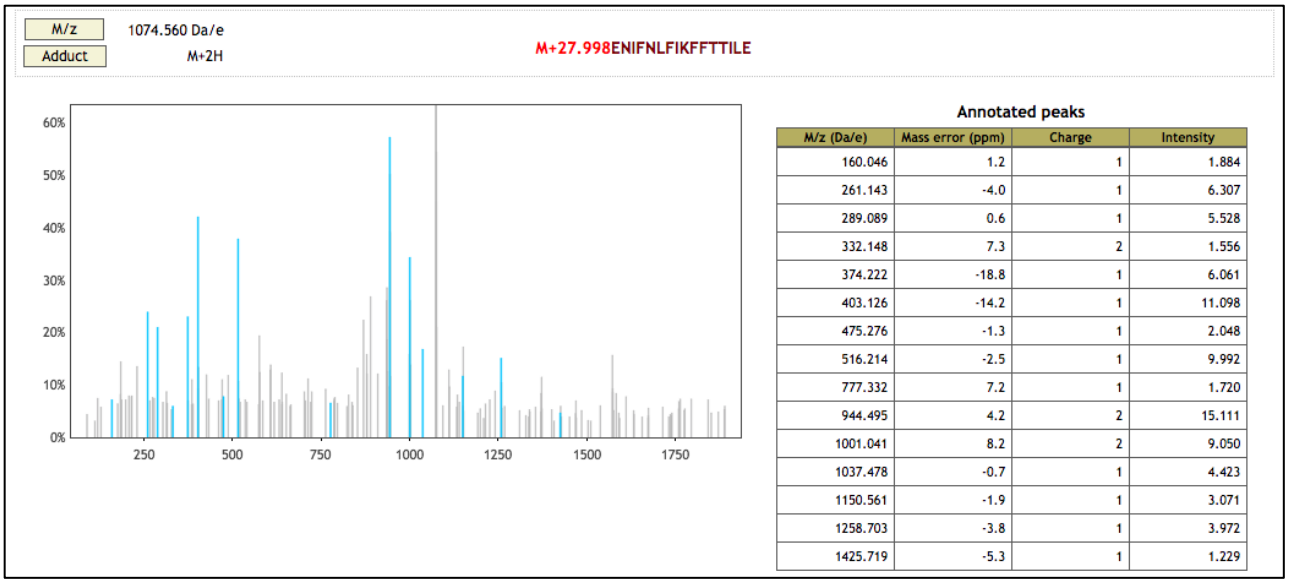
(c)

Figure S14. Identification of PSM-2147 from HUMAN-iso dataset, related to Figure 2. (a) The biosynthetic gene cluster of PSM-2147. AntiSMASH failed to predict the precursor peptide for this gene cluster by motif search; however, MetaMiner was able to discover this peptide in all-ORF mode. (b) Annotation of the BGC of PSM-2147 by DFAST (Tanizawa et al., 2018) and HMMER (Eddy, 2011). (c) Tandem mass spectrum of PSM-2147 (score 15, p-value $1 \cdot 10^{-32}$).

(a)

Gene	Predicted Function	Length
1	CAAX amino protease	247
2	hypothetical protein	482
3	hydrolase	265
4	accessory gene regulator protein B	204
5	PSM-2147	47
6	hypothetical protein	267
7	accessory gene regulator protein A	239
8	fructokinase	320
9	invertase	491

(b)



Annotated peaks

M/z (Da/e)	Mass error (ppm)	Charge	Intensity
160.046	1.2	1	1.884
261.143	-4.0	1	6.307
289.089	0.6	1	5.528
332.148	7.3	2	1.556
374.222	-18.8	1	6.061
403.126	-14.2	1	11.098
475.276	-1.3	1	2.048
516.214	-2.5	1	9.992
777.332	7.2	1	1.720
944.495	4.2	2	15.111
1001.041	8.2	2	9.050
1037.478	-0.7	1	4.423
1150.561	-1.9	1	3.071
1258.703	-3.8	1	3.972
1425.719	-5.3	1	1.229

(c)

Figure S15. High Resolution Mass Spectrum of Wewakazole, related to Figure 2. $[M+Na]^+$ 1163.5282 (calcd for $C_{59}H_{72}N_{12}O_{12}Na$ 1163.5285; Δ= -0.3 ppm).

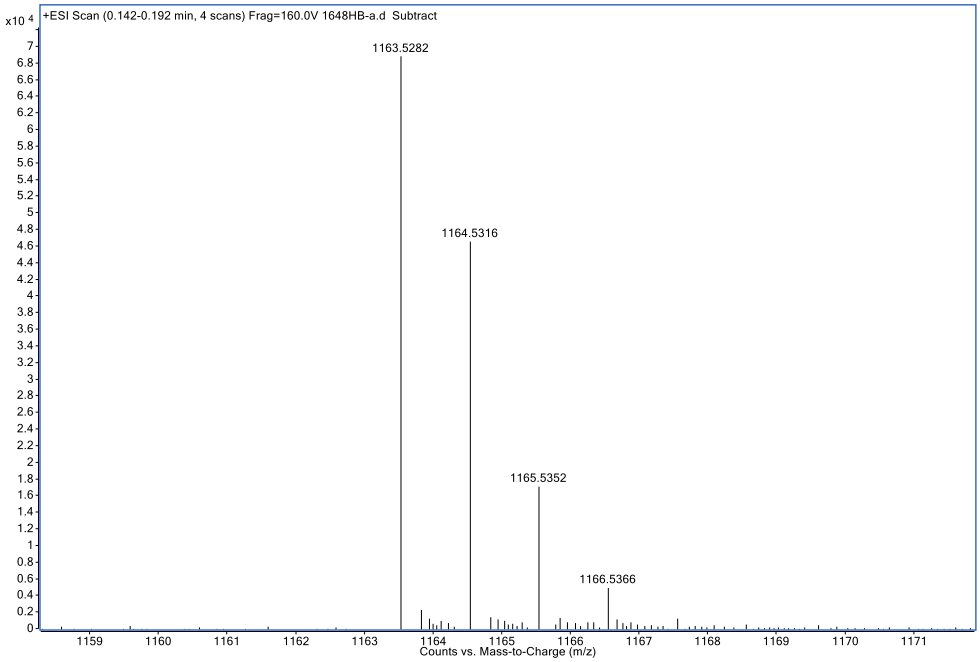


Figure S16. ^{1}H-^{13}C HSQC NMR spectrum of Wewakazole in $CDCl_3$ at 600MHz, related to Figure 2.

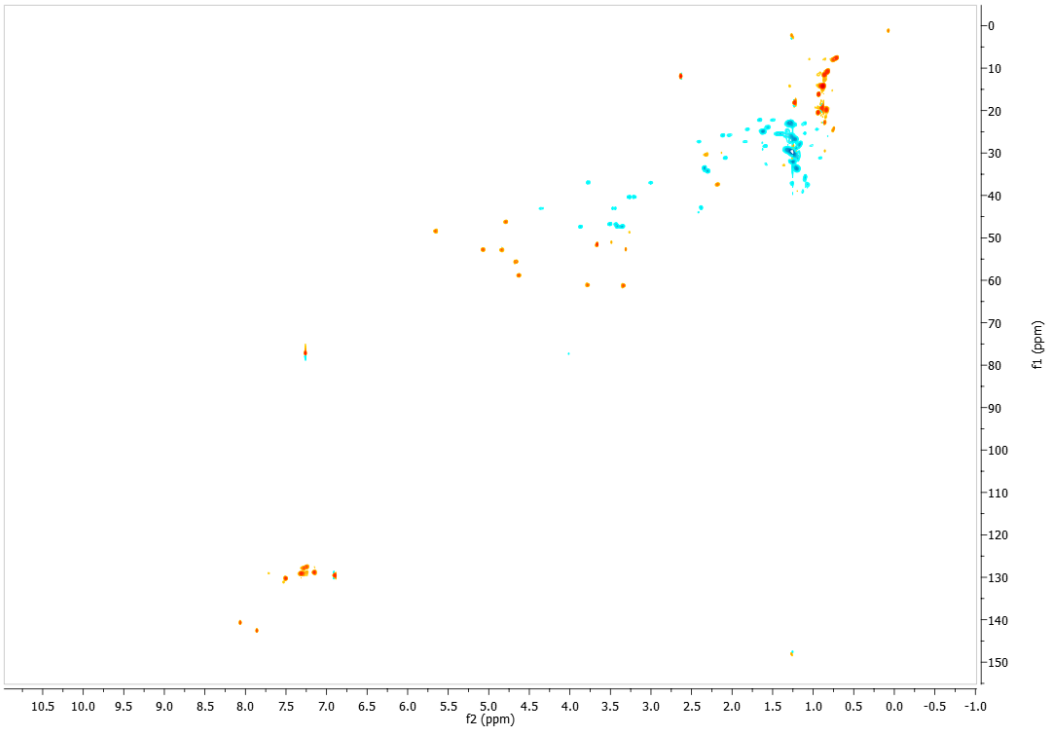


Figure S17. 1H-13C HMBC NMR spectrum of Wewakazole in CDCl3 at 600MHz, related to Figure 2.

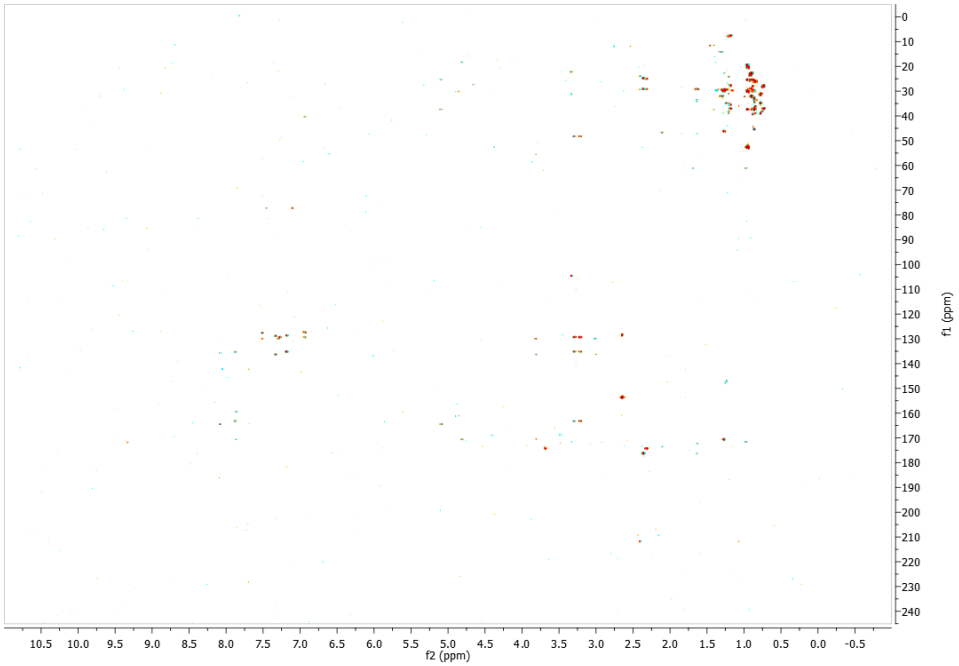


Figure S18. ^{1}H NMR spectrum of Wewakazole in $CDCl_3$ at 500MHz, related to Figure 2.

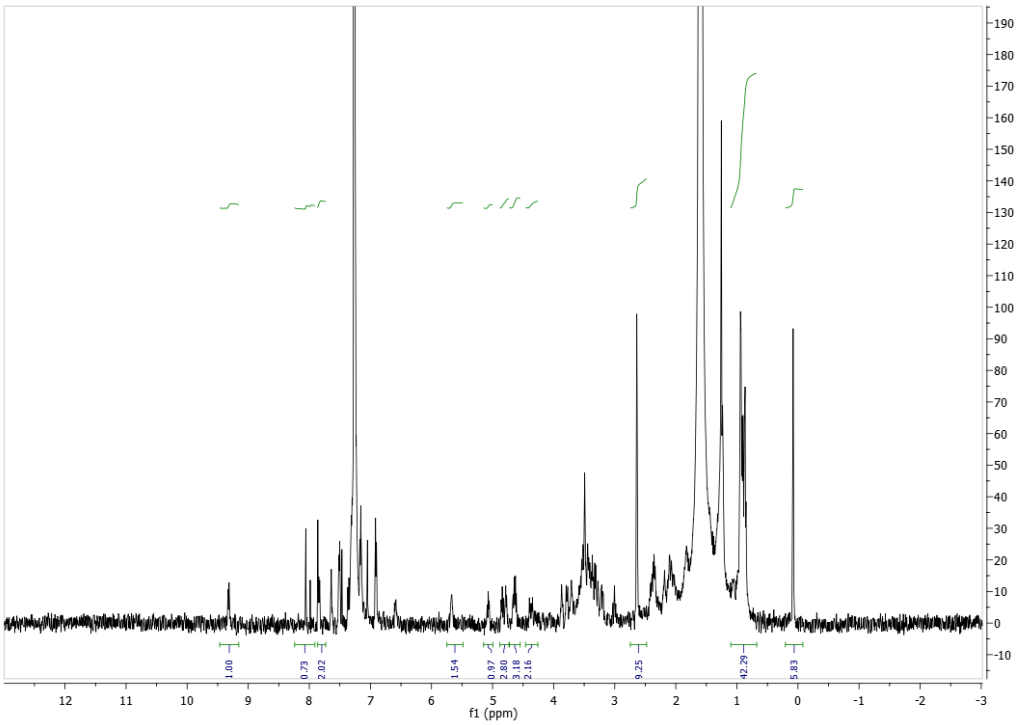


Figure S19. LC-MS/MS spectrum of Wewakazole, related to Figure 2.

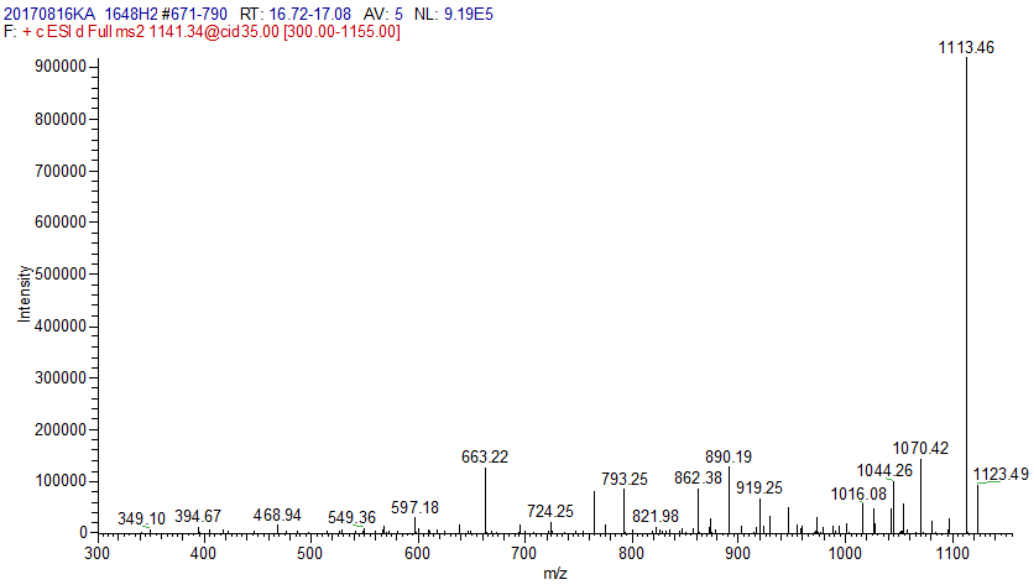


Figure S20. LC-MS chromatogram of Wewakazole, related to Figure 2.

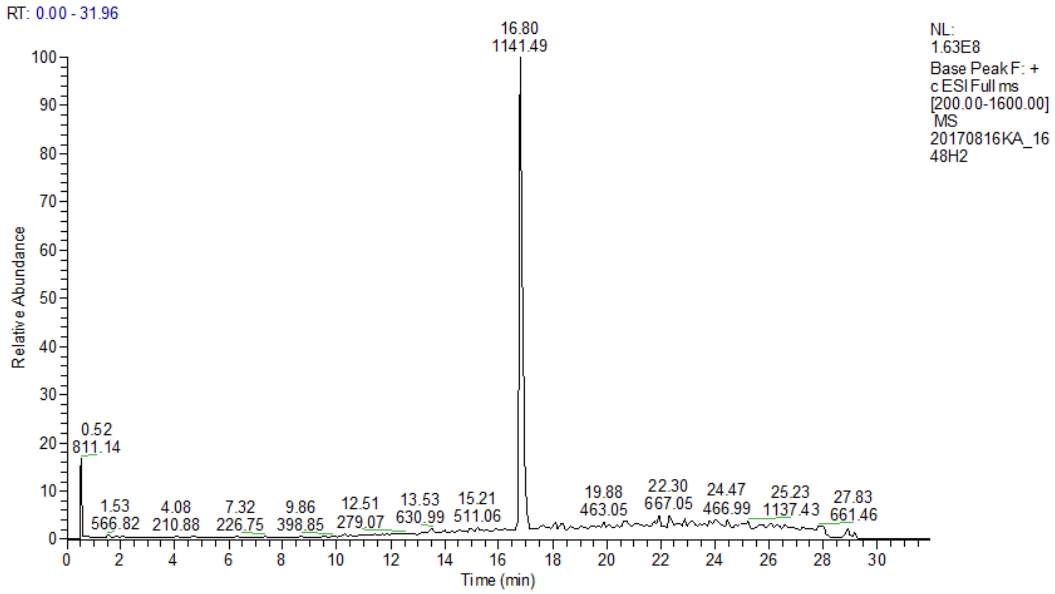


Figure S21. ECCD spectrum of Wewakazole in MeOH, related to Figure 2. Solvent absorptions are seen in the region below 200 nm.