Supplementary material for:

Title: Mapping gene clusters captured in arrayed metagenomic libraries to expand the structural diversity of biomedically relevant natural products

Authors: Jeremy G. Owen, Boojala Vijay B. Reddy, Melinda A. Ternei, Zachary Charlop-Powers, Paula Y. Calle, Jeffrey H. Kim and Sean F. Brady*
Author affiliation: Laboratory of Genetically Encoded Small Molecules, Howard Hughes Medical Institute, The Rockefeller University, 1230 York Avenue, New York, NY 10065.
Corresponding Author: Sean F. Brady
Contact: Laboratory of Genetically Encoded Small Molecules
The Rockefeller University
1230 York Avenue
New York, NY 10065
Phone: 212-327-8280
Fax: 212-327-8281
Email: sbrady@rockefeller.edu

Table of contents

Table S1. 454 sequencing and gene cluster scan data statistics from the analysis of New Mexico soil DNA library KS and AD amplicons.

	Ketosynthase Domain	Adenylation Domain
	NM	**NM**
Raw Reads	155,531	640,101
Clean Reads	150,175	528,308
Cluster Consensus Seq	4,167	16,949
Known Related CCSs	400	626
Related Known Domains	164	117
Related Known Molecules	81	57
Molecules Scanned	192	207
Domains Scanned	1611	975

Table S2. 454 sequencing and gene cluster data statistics for AD amplicons derived from crude eDNA samples.

	NM	NM*	CA	AZ	UT	PA	TZ
Raw Reads	78,780	10,550	13,591	16,363	10,003	15,737	11,099
Clean Reads	74,692	10,086	9,733	12,014	7,438	11,831	8,384
Cluster Consensus Seq	15,371	4,503	5,491	6,659	2,843	2,900	3,654
Known Related CCSs	555	158	401	475	186	152	223
Related Known Domains	107	62	109	114	73	45	54
Related Known Molecules	50	39	53	58	45	34	33
Molecules Scanned	207	207	207	207	207	207	207
Domains Scanned	975	975	975	975	975	975	975

*A subset of New Mexico amplicon reads randomly sampled to yield ~10,000 clean reads. For all other soil samples, 454-sequencing was performed at a depth that was expected to yield approximately 10,000 cleaned reads.

Table S3. Biosynthetic gene cluster hits detected from the screening of KS and A domain sequence tags amplified from the New Mexico mega-library. Gene clusters that were not detected are listed below the table. (HD = number of domains detected from each gene cluster. HS = number of unique amplicons returned from the search. e-max/min = maximum and minimum expectation values returned for hits corresponding to each gene cluster)

Molecule	HD	HS	e-min	e-max	Molecule	HD	HS	e-min	e-max	Molecule	HD	HS	e-min	e-max
A40926	1	6	e-27	e-21	FR901464	1	5	e-36	e-20	Pimaricin	1	1	e-69	e-69
A54145	2	3	e-35	e-25	Friulimicin	4	8	e-80	e-25	Pladienolide	2	2	e-66	e-38
Actinomycin	1	1	e-32	e-32	FriulimicinUC	6	13	e-162	e-20	Polyene	3	7	e-90	e-41
Ajudazol	6	51	e-61	e-22	GE81112	1	4	e-36	e-20	Polymyxin	1	1	e-23	e-23
AlphaLipomycin	2	2	e-72	e-49	Halstoctacosanolide	1	2	e-43	e-41	Pristinamycin	1	1	e-25	e-25
Ambruticin	1	1	e-55	e-55	Indanomycin	2	2	e-67	e-21	Prodigiosin	1	6	e-53	e-27
Amphotericin	2	4	e-81	e-42	JerangolidA	2	3	e-90	e-36	Putisolvin	2	5	e-43	e-26
Ansamitocin	2	4	e-87	e-63	Kalimantacin	3	7	e-46	e-22	Pyoverdine	2	2	e-67	e-26
Anthramycin	1	1	e-87	e-87	Kijanimicin	3	3	e-91	e-37	Pyridomycin	2	28	e-84	e-27
Arthrofactin	3	10	e-42	e-20	Kirromycin	3	7	e-139	e-24	Pyrrolomycin	4	23	e-101	e-22
Aurafuron	1	1	e-34	e-34	Kutznerides	2	2	e-36	e-29	SaframycinMx1	3	52	e-48	e-20
Avermectin	1	1	e-80	e-80	Lactimidomycin	1	6	e-45	e-28	Salinasketal	6	7	e-143	e-104
AvilamycinA	1	4	e-98	e-44	Lankacidin	1	1	e-26	e-26	SanglifehrinA	2	2	e-93	e-78
AzinomycinB	2	3	e-71	e-21	Lankamycin	2	2	e-42	e-25	Skyllamycin	1	3	e-49	e-22
Bafilomycin	1	1	e-54	e-54	Lasalocid	1	1	e-40	e-40	Sorangicin	8	23	e-70	e-20
Bleomycin	5	31	e-51	e-20	Laspartomycin	5	9	e-76	e-29	Spiramycin	3	3	e-82	e-70
CA37Glycopeptide	4	4	e-158	e-97	Leinamycin	2	4	e-49	e-22	Spirangien	3	3	e-74	e-31
Calcimycin	2	2	e-91	e-34	Leupyrrin	1	9	e-59	e-30	Stigmatellin	1	3	e-38	e-22
Calicheamicin	1	14	e-98	e-27	Lysobactin	3	12	e-83	e-22	Streptolydigin	1	2	e-68	e-54
Capreomycin	1	1	e-21	e-21	MacrolactamHSAF	1	2	e-31	e-29	SW163s	3	6	e-113	e-21
CDA	3	28	e-77	e-30	MacrolactamML449	1	1	e-79	e-79	Syringopeptin	2	5	e-50	e-35
Chivosazol	2	2	e-34	e-33	Maduropeptin	2	8	e-92	e-45	Tallysomycin	5	44	e-53	e-20
Chlorothricin	3	4	e-94	e-40	Mannopeptimycin	3	6	e-28	e-21	Tautomycetin	2	2	e-41	e-20
Chondramide	3	31	e-75	e-20	Melithiazol	3	5	e-56	e-41	Teicoplanin	1	6	e-24	e-20
Chondrochloren	3	6	e-55	e-24	Meridamycin	2	3	e-92	e-74	TetrocarcinA	3	6	e-97	e-26
Complestatin	2	7	e-31	e-21	Midecamycin	1	1	e-31	e-31	Tetronasin	3	3	e-41	e-25
ConcanamycinA	2	2	e-60	e-39	Migrastatin	1	1	e-72	e-72	TetronateRK682	1	1	e-22	e-22
ConcanamycinAII	2	3	e-103	e-77	Monensin	1	3	e-104	e-87	Thiocoraline	1	1	e-129	e-129
CorallopyroninA	2	11	e-49	e-20	Mupirocin	1	4	e-55	e-27	ThuggacinCmc	2	3	e-49	e-29
CystothiazoleA	3	8	e-61	e-25	Myxalamid	1	2	e-36	e-33	ThuggacinCmc5	1	1	e-56	e-56
Daptomycin	2	2	e-34	e-34	Myxochelin	1	1	e-37	e-37	ThuggacinSoce	4	16	e-93	e-23
DaptomycinSV	2	2	e-38	e-33	Myxothiazol	2	5	e-62	e-23	ThuggacinSoce895	3	9	e-66	e-20
DisorazoleA1	3	4	e-54	e-24	Myxovirescin	4	11	e-29	e-21	TiacumicinB	3	22	e-114	e-31
DKxanthene534	1	2	e-21	e-20	Nanchangmycin	2	3	e-76	e-54	Tirandamycin	3	9	e-163	e-21
Echinomycin	3	5	e-68	e-30	Nemadectin	2	2	e-82	e-59	TriostinA	3	4	e-67	e-23
Echinosporamicin	1	1	e-79	e-79	Nigericin	2	2	e-105	e-73	Tubulysin	5	68	e-56	e-20
Enduracidin	3	6	e-70	e-32	Nystatin	4	18	e-105	e-38	Tylactone	2	2	e-72	e-46
Epothilone	2	5	e-44	e-24	Oligomycin	2	4	e-91	e-25	Virginiamycin	2	3	e-38	e-24
Erythromycin	1	1	e-51	e-51	Oxazolomycin	2	7	e-40	e-24	WAP8294A2	6	61	e-73	e-20
FD891	2	7	e-118	e-27	Pactamycin	1	1	e-88	e-88	Zorbamycin	1	1	e-23	e-23
Filipin	2	3	e-84	e-77	Paebacillibactin	1	1	e-61	e-61					
FR008	6	8	e-124	e-45	Pikromycin	1	1	e-105	e-105					

Screened but not detected in the NM library: A47934, A500359s, A503083, Aeruginoside126A, Aeruginosin98A, Aflatoxin, Althiomycin, Amicetin, AminoHydroxyBenzoicAcid, Anabaenopeptilide, Anabaenopeptin, AnatoxinA, Ansamycin, Apicidin, Apoptolidin, Ascomycin, Aureothin, Auricin, Bacillaene, BacillaeneII, BacillibactinI, BacillibactinII, BacillibactinIII, BacillibactinIV, BacillomycinD, Bacitracin, Balhimycin, Barbamide, Bassianolide, Beauvericin, Borrelidin, BrasilicardinA, Bryostatin, Burkholdac, CA878Glycopeptide, Cephalosporin, Cereulide, CetoniacytoneA, Chaetoglobosin, ChaetoglobosinA, Chalcomycin, Chloramphenicol, Chloroeremomycin, Cinnabaramide, Clorobiocin, Coelichelin, Cryptophycin, Cyanopeptolin1138, Cyanopeptolin984, Cyclosporin, Cylindrospermopsin, Dapdiamide, Desmethylbassianin, Difficidin, Dihydrochalcomycin, Dkxanthene, DKxanthene544, EicosapantaenoicAcid, Elansolide, Enediyne, EnediyneC1027, Equisetin, EsperamicinEnediyne, FengycinI, FengycinII, FK228, FK506, Fostriecin, Fusaricidin, Fusarin, Geldanamycin, GlidobactinA, Glycopeptidolipid, GramicidinA, GramicidinS, Griseobactin, HCToxin, Hectochlorin, Hedamycin, Herbimycin, Himastatin, Homoanatoxin, Hormaomycin, IndigoidineI, IndigoidineII, Iturin, IturinA, Jamaicamide, Koranimine, lactonamycinZ, Lichenysin, Macbecin, MacrolactamBE14106, Macrolactin, MassetolideA, Megalomicin, Meilingmycin, Microcystin, MicrocystinRR, Microginin, Micropeptin, Microviridin, Muraymycin, Mycinamicin, Mycobactin, Mycolactone, Mycosubtilin, MyxochromideS, Naphthomycin, Napsamycin, Natamycin, Neoaureothin, Neocarzinostatin, NG391, Niddamycin, Nikkomycin, NocathiacinI, Nodularin, Nostocyclopeptide, Nostopeptolide, Nostophycin, Oleandomycin, OrphanLike, Pacidamycin, Pederin, Penicillin, Phoslactomycin, PhoslactomycinB, Plipastatin, Psymberin, Pyochelin, Pyoluteorin, Rapamycin, Reveromycin, Rhizopodin, Rhizoxin, Rifamycin, Rubradirin, Rubrinomycin, SaframycinA, Salinomycin, Salinosporamide, SalinosporamideK, SaquayamycinZ, Simocyclinone, SoraphenA, Spinosad, Streptothricin, SurfactinI, SurfactinII, SyringolinA, Syringomycin, Tautomycin, TegGlycopeptide, Tenellin, ThaxtominA, Tomaymycin, Tyrocidin, Valinomycin, Vanchrobactin, Vancomycin, VEGGlycopeptide, Vicenistatin, Viomycin, Yersiniabactin, ZwittermicinA,

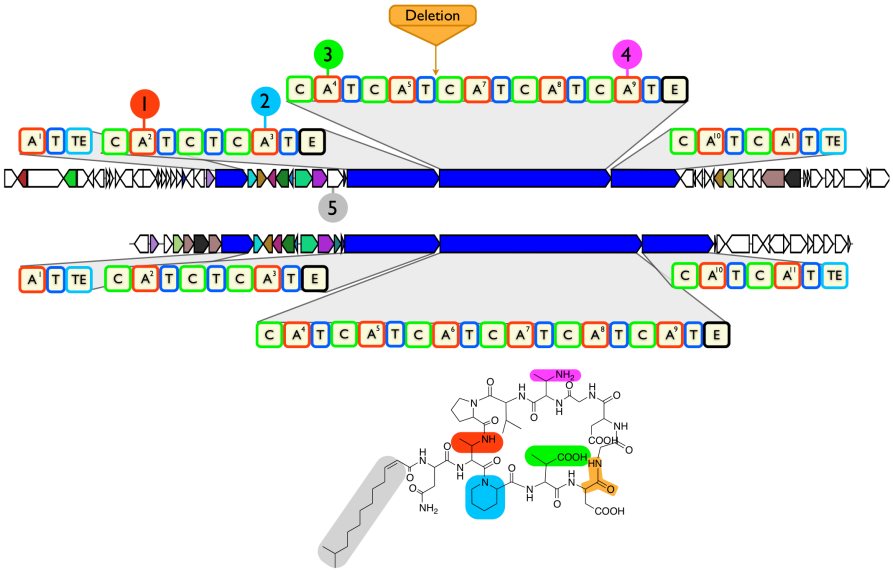
Table S4. Expectation value (E-value) summary for amplicons used to guide the recovery of the 26 eDNA gene clusters described in the manuscript.

Family	Closest Relative	E-value[1]	On Target[2]	New Derivative[3]	Accession Number(s)	Seq. No. (Si Data)
Lipopeptide	Friulimicin	1.0 E-145	Yes	No	KF264537, KF264538	1
Lipopeptide	Friulimicin	4.0 E-47	Yes	Yes	KF264539	2
Lipopeptide	Friulimicin*	2.0 E-128	Yes	No	KF264540, KF264541	3
Lipopeptide	Friulimicin	1.0 E -66	Yes	Yes	KF264542	4
Lipopeptide	CDA	1.0 E-141	Yes	No	KF264543, KF264544	5
Bisintercalator	Thiocoraline	3.0 E-129	Yes	Yes	KF264545	6
Bisintercalator	Thiocoraline	3.0 E-148	Yes	Yes	KF264546	7
Bisintercalator	Thiocoraline	1.0 E-109	Yes	Yes	KF264547	8
Bisintercalator	SW-163C	2.0 E-113	Yes	No	KF264548	9
Streptogramin	Virginamycin	1.0 E-53	Yes	No	KF264549	10
Streptogramin	Virginamycin	1.0 E-50	Yes	No	KF264550	11
Erythromycin	Erythromycin	5.0 E-101	Yes	No	KF264551	12
Rapamycin	Rapamycin	1.0 E-57	Yes	Yes	KF264552	13
Bleomycin	Tallisomycin	2.0 E-39	Yes	Yes	KF264553	14
Glycopeptide	Teicoplanin**	4.0 E-146	Yes	Yes	KF264554	15
Glycopeptide	A47934**	2.0 E-175	Yes	Yes	KF264565	26
Glycopeptide	A47934**	2.0 E-132	Yes	No	KF264555, KF264556	16
Azinomycin	Azinomycin	3.0 E-75	Yes	Yes	KF264557	17
Bisintercalator	Echinomycin	6.0 E-30	No	N/A	KF264558	18
Enduracidin	Enduracidin	4.0 E-70	No	N/A	KF264559	19
Enduracidin	Enduracidin	2.0 E-67	No	N/A	KF264560	20
Enduracidin	Enduracidin	9.0 E-60	No	N/A	KF264561	21
Glycopeptide	Teicoplanin	8.0 E-35	No	N/A	KF264562	22
Glycopeptide	A47934	2.0 E-22	No	N/A	KF264563	23
Safaramycin	Safaramycin	5.0 E-44	No	N/A	KF264564	24
Ajudazol	Ajudazol	3.0 E-59	No	N/A	KF264566	25

[1]E-values are reported for alignments between library sequences and the characterized pathway listed in column 2. [2]Pathways not related to any member of the expected family were considered off target. [3]On target pathways with changes in core biosynthetic function or tailoring enzymes were classified as new derivatives. *Initial amplicon hit was to a previously isolated eDNA cluster predicted to encode friulimicin (1). **Initial amplicon hit was to a previously isolated eDNA cluster predicted to encode a glycopeptide antibiotic (14).

Introduction to Figures S1 - S11: For each recovered eDNA gene cluster, open reading frames were initially delineated using MetaGeneMark (2). The preliminary annotation of open reading frames was conducted using RAST (3) followed by a Blast search and manual analysis to determine the predicted function for each gene. Putative closest relatives for each cluster were then assigned based on a combination of 1. a manual comparison of this data to sequenced gene clusters and 2. the cumulative BLAST score returned by the ClusterBlast function of AntiSMASH (4). The domain organization within each megasynth(et)ase was determined using the NRPS/PKS analysis webserver (5) and AntiSMASH. A-domain binding pockets from characterized and eDNA derived megasynthetases were extracted using NRPS predictor2 and compared manually. Predicted changes in A-domain and AT-domain substrate specificities were assigned based on data from NRPS Predictor2 (6) and the method of Yadav *et al*. (7), respectively, both of which are implemented as part of the AntiSMASH analysis pipeline. Putative function for new tailoring enzymes in eDNA pathways was assigned based on the predicted function of the closest characterized relative identified by Blast in NCBI. The data is presented here for each recovered gene cluster in table, figure and text format. For each gene cluster, NRPS/PKS megasynth(et)ases are colored dark blue and exploded diagrams of megasynth(et)ase domain arrangements are given. Tailoring open reading frames with identical proposed functions have the same color. Grey bars running between eDNA and known pathways indicate equivalent megasynth(et)ases where relative positions are shifted. White open reading frames either represent genes found in only one gene cluster of a pair or, genes not predicted to have a direct role in biosynthesis (e.g. export and regulatory functions).

Figure S1. Lipopeptide eDNA Gene Cluster 1



Closest relative: Friulimicin

Domain arrangement: eDNA pathway contains one fewer module than the known Friulimicin pathway

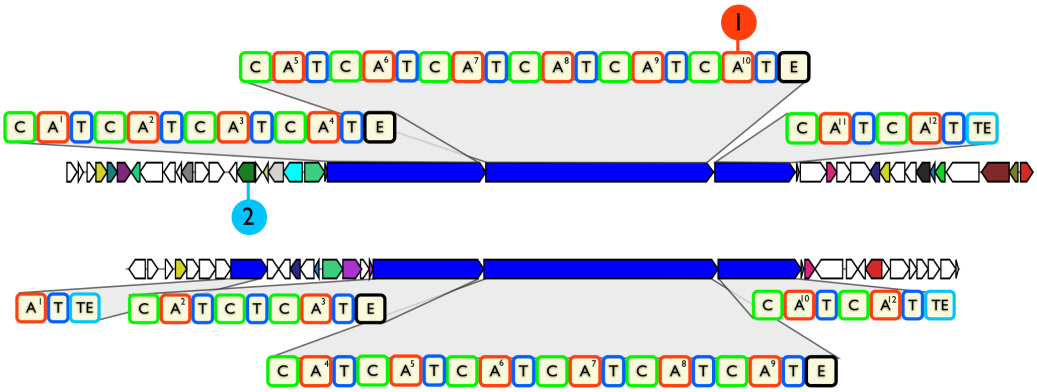
A-domain specificity predictions: Differs at four positions (Gene cluster labels 1-4)

A-domain	Friulimicin Residue	Friulimicin Pocket	eDNA Pocket	eDNA Prediction
1	Asx	DIWQSTADDK	DIWQSTADDK	Asx
2	Dab	DLTKVGDVNK	DLTKMGVVNK	mAsp (3-methyl-asparticacid)
3	Pip	DFQFFGVAVK	DAYWWGGTFK	Val
4	mAsp	DLTKMGVVNK	DTDDMGYVDK	Orn/Lys/Arg
5	Asp	DLTKVGAVNK	DLTKVGAVNK	Asp
6	Gly	DILQLGLVWK	Deletion	N/A
7	Asp	DLTKVGAVNK	DLTKVGAVNK	Asp
8	Gly	DILQLGLVWK	DILQLGLVWK	Gly
9	Dab	DAFFWGEVFK	DLTKMGVVNK	mASP
10	Val	DAYWWGGTFK	DAFWLGGTFK	Val
11	Pro	DVQYVAHVVK	DVQYVGHAIK	Pro

Tailoring Enzymes: Additional acyl-coA dehydrogenase (Gene cluster label 5)

Summary: This gene cluster is predicted to encode a structure related to friulimicin, an antibiotic that inhibits peptidoglycan and teichoic acid synthesis via calcium dependent binding of the carbohydrate carrier undecaprenyl-phosphate (8). The overall cluster architecture and megasynthetase domain arrangement closely resemble that seen in friulimicin however the eDNA gene cluster contains three key changes compared to the friulimicin cluster: 1. it contains one fewer NRPS module, 2. the substrate specificity of four A-domains is predicted to differ from those seen in the friulimicin cluster and 3. it contains an additional acyl-CoA dehydrogenase enzyme that is not seen in the friulimicin cluster. Based on standard biosynthetic logic, this pathway is therefore predicted to encode a novel friulimicin-like structure whose core peptide is one amino acid shorter, differs in side chain composition at four positions and may contain an extra unsaturation in the N-terminal fatty acid.

Figure S2. Lipopeptide eDNA Gene Cluster 2



Closest relative: Friulimicin

Domain arrangement: One additional module (Gene cluster label 1)

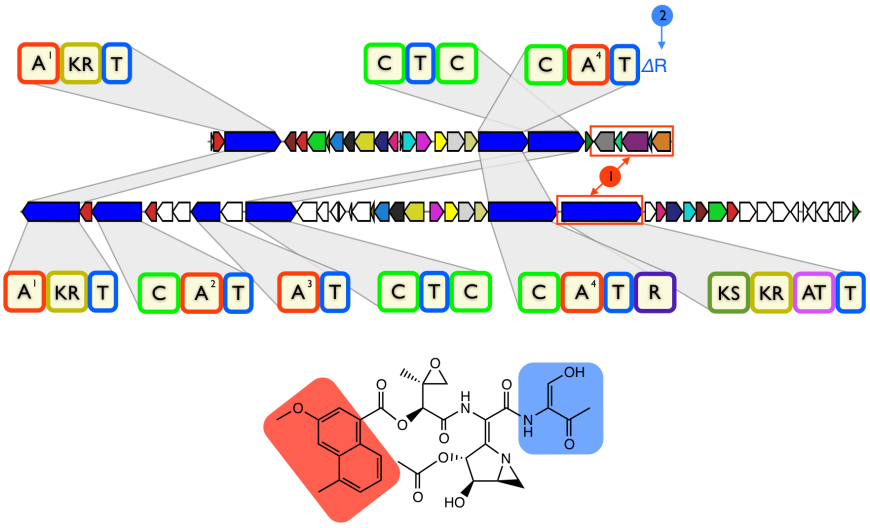
A-domain specificity predictions: Four conserved positions

A-domain	Friulimicin specificity	Friulimicin Pocket	eDNA Pocket	eDNA Prediction
1	Asx	DIWQSTADDK	DLTKVGAVNK	Asp
2	Dab	DLTKVGDVNK	DASTIAAVDK	Tyr
3	Pip	DFQFFGVAVK	DVWHISLVDK	Ser
4	Asp	DLTKMGVVNK	DVLKVGAVNK	Asp
5	Asp	DLTKVGAVNK	DFWNVGMVHK	Thr
6	Gly	DILQLGLVWK	DVWHISLVDK	Ser
7	Asp	DLTKVGDVNK	DMPQFGLVWK	Gly
8	Gly	DILQLGLVWK	DLTKVGEVGK	Asn
9	Dab	DAFFWGEVFK	DAAMCGGVAK	Hydrophobic
10	Val	DAYWWGGTFK	DAYWWGGTFK	Val
11	N/A	N/A	DAFVLAGVAK	Hydrophobic
12	Pro	DVQYVAHVVK	DGQYVSQVMK	Pro

Tailoring Enzymes:. Contains tryptophan halogenase gene (Gene cluster label 2)

Summary: The eDNA gene cluster presented above is most closely related to the friulimicin biosynthetic cluster. Friulimicin belongs to the acidic lipopeptide family of natural products whose members include daptomycin, CDA, laspartomycin and A54145. As is common in this family, the eDNA gene cluster appears to encode an N-terminally acylated peptide with an internal 10 membered macrocycle. The predicted positions of D-amino acids within the encoded structure are also conserved when compared to friulimicin. Beyond this however the predicted structure is predicted to be quite different from known acidic lipopeptides. The eDNA gene cluster contains a total of 12 NRPS modules, compared to the 11 found in the friulimicin cluster. Analysis of A-domains from this gene cluster predicts friulimicin like specificity at only four locations. Four of the remaining nine amino acid binding pockets are perfect, or near perfect matches to A-domains seen the daptomycin and CDA gene clusters. The presence of a tryptophan halogenase suggests that the penultimate A-domain is likely to incorporate a chlorinated tryptophan moiety.

Figure S3. Azinomycin-like eDNA Gene Cluster



Closest relative: Azinomycin

Domain arrangement: Iterative type I PKS is absent and appears to be replaced with a type II PKS system (Gene cluster label 1). One module lacks an NRPS associated reductive domain (Gene cluster label 2).

A-domain specificity predictions: No differences

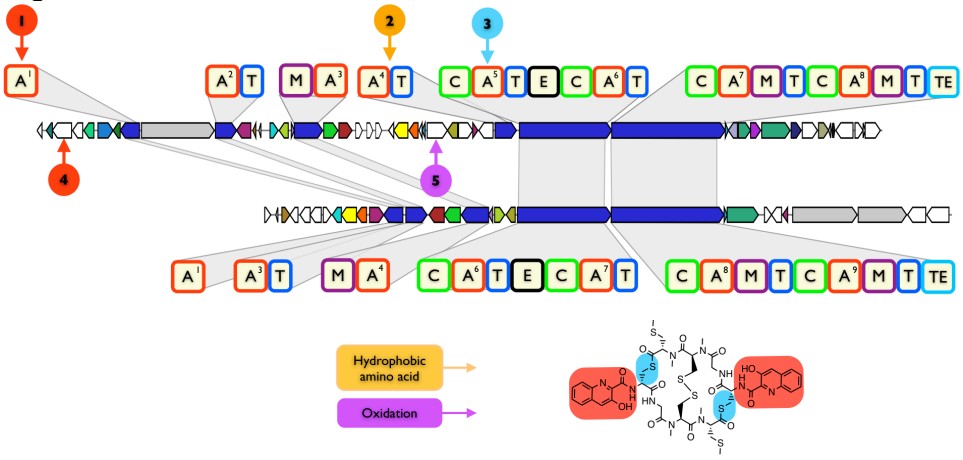
A-domain	Azinomycin Residue	Azinomycin Pocket	eDNA Pocket	eDNA Prediction
1	NAP	GMYWCACSGK	GIYWCACSGK	NAP
2	AKA	DVFDFGGVTK	N/R*	N/R
3	APA	GIHFTGSQIK	N/R	N/R
4	Thr	DFWSVGMVHK	DFWSVGMVHK	Thr

NAP = 5-methyl-napthoic acid, AKA = α-ketoisovaleric acid, APA = aziridino[1, 2a]pyrrolidinyl amino acid
*Region containing this domain not recovered the eDNA library in our study

Tailoring Enzymes: No additional tailoring functionalities

Summary: The partial eDNA gene cluster depicted above is predicted to encode a relative of azinomycin, a DNA damaging agent that causes inter-strand crosslinking via a reactive epoxide functionality. The gene content of the recovered eDNA cluster very closely resembles that of the azinomycin cluster with some key differences that are predicted to result in its encoding for the biosynthesis of a novel azinomycin-like metabolite. The bicyclic aromatic moiety seen in azinomycin is synthesized using an iterative type I PKS enzyme. No type I PKS-like open reading frame is found in the eDNA gene cluster. We propose instead that the type II PKS system found the eDNA cluster serves to construct a related aromatic substructure (Gene cluster label 1) that is used in the biosynthesis of an azinomycin-like metabolite. The majority of the NRPS megasynthetases in the eDNA cluster resembles those seen in the azinomycin cluster with the exception of the missing reductive domain that is believed to act on the terminal threonine residue (Gene cluster label 2).

Figure S4. eDNA Bis-intercalator Gene Cluster 1



Closest relative: Thiocoraline

Domain arrangement: Contains two additional modules (Gene cluster labels 2 and 3).

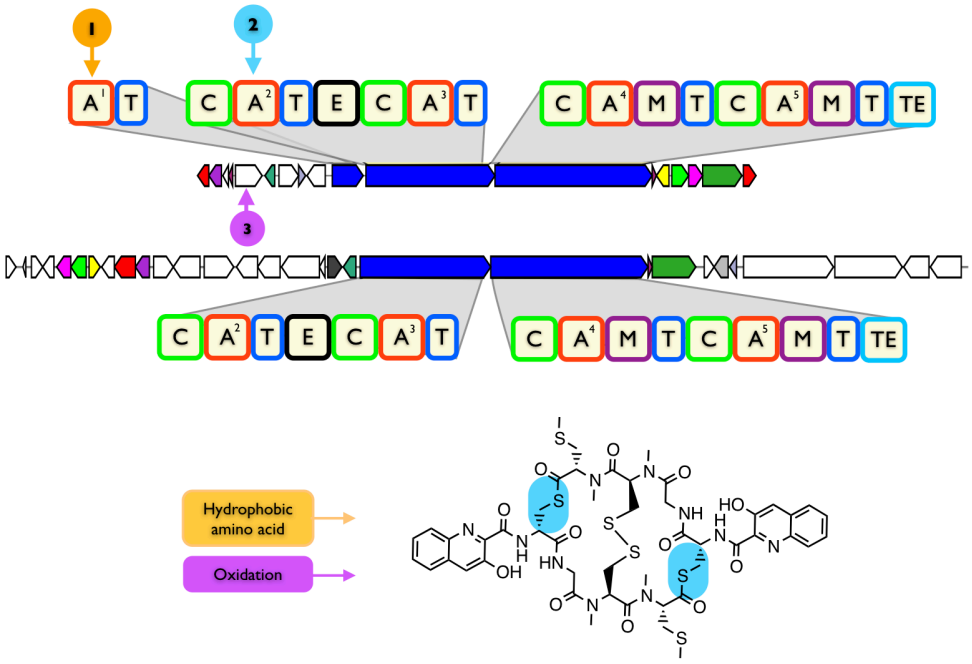
A-domain specificity predictions: Two differences (Gene cluster labels 1 and 4) .

A-domain	Thiocoraline Residue	Thiocoraline Pocket	eDNA Pocket	eDNA Prediction
1	3-HQA	ALPTQGWLTK	PAPSQGWLAK	Aryl-Acid
2	Trp	DAWVMTGVGK	DAWVMTGVGK	Trp
3	Cys	DLYDLSLVWK	DLYDLSLVWK	Cys
4	N/A	N/A	SVVFLGWIVK	Hydrophobic
5	Cys	NAWHMSLPVK	DAWRLGLLDE	Hydrophilic
6	Gly	DILQLGLVWK	DILQLGLIWK	Gly
7	mCys	DLFNFSLVWK	DLFDYSLVWK	mCys
8	dmCys	DAYWWGGTFK	DAYWWGGTFK	dmCys

mCys = N-methyl-Cys, dmCys = N,S-dimethyl-Cys

Summary: The gene cluster presented above is predicted to encode a structure(s) related to thiocoraline, a member of the bis-intercalator family of anti-tumor natural products (10, 11). The core biosynthetic region of this cluster, comprising two large NRPS enzymes, has high sequence identity and near identical architecture to the thiocoraline biosynthetic cluster. Tailoring enzyme functions found in the thiocoraline biosynthetic cluster are also largely conserved in the eDNA cluster. Although the eDNA cluster is very similar to the thiocoraline, there are number of obvious differences predicted in our bioinformatics analysis. 1. The eDNA gene cluster contains an additional single module NRPS enzyme that is predicted to activate a hydrophobic amino acid (Gene cluster label 2). 2. The specificity of two A-domains is predicted to differ from the equivalent positions in the thiocoraline gene cluster (Gene cluster labels 1 and 3). 3. The presence of a tryptophan halogenase suggests a halogenated tryptophan derived bicyclic moiety may replace the tryptophan derived 3-hydroxyquinaldic acid (3HQA) seen in thiocoraline (Gene cluster label 4). This possibility is supported by the altered substrate binding pocket residues seen in the aryl-acid activating A-domain present in the eDNA pathway. In all known bis-intercalator pathways the equivalent A-domain has a highly conserved set of substrate binding pocket residues. The eDNA cluster also contains an additional Baeyer-Villiger monooxygenase (Gene cluster label 5) which is not observed in the known thiocoraline cluster. These differences suggest the eDNA gene cluster is capable of encoding for a previously uncharacterized bis-intercalator natural product.

Figure S5. eDNA Bis-intercalator Gene Cluster 2



Closest relative: Thiocoraline

Domain arrangement: Contains one additional module (Gene cluster label 1)

A-domain specificity predictions: One difference in A-domain specificity (Gene cluster label 2)

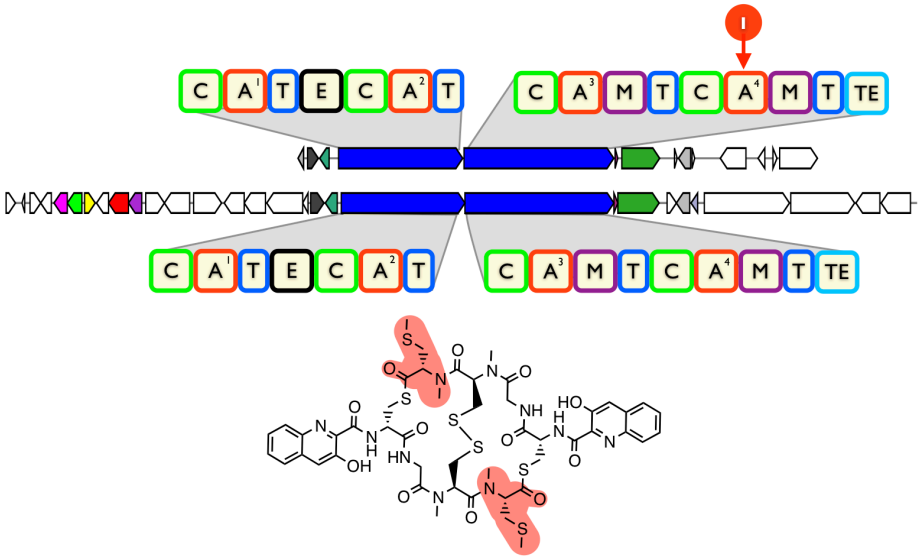
A-domain	Thiocoraline Residue	Thiocoraline Pocket	eDNA Pocket	eDNA Prediction
1	N/A	N/A	NVSFLGWIVK	Hydrophobic
2	Cys	NAWHMSLPVK	AAWHMSLLDK	Hydrophilic
3	Gly	DILQLGLVWK	DILQLGLIWK	Gly
4	mCys	DLFNFSLVWK	DLFDYSLVWK	mCys
5	dmCys	DAYWWGGTFK	DAYWWGGTFK	dmCys

mCys = N-methyl-Cys, dmCys = N,S-dimethyl-Cys

Tailoring Enzymes: new Baeyer-Villiger monooxygenase (Gene cluster label 3)

Summary: The partial gene cluster presented above is predicted to encode a previously uncharacterized structure related to the bis-intercalator thiocoraline. As can be seen in the comparison diagram these two clusters have similar overall gene architecture and gene contents (for the regions recovered from the library). The eDNA cluster contains an additional single module NRPS enzyme (Gene cluster label 1) predicted to activate a hydrophobic residue. In addition, the specificity of one additional A-domain (Gene cluster label 2) is predicted to differ from the equivalent position in the thiocoraline gene cluster. The presence of an open reading frame that is predicted to encode a Baeyer-Villiger monooxygenase (Gene cluster label 3) also represents a potential new tailoring function.

Figure S6. eDNA Bis-intercalator Gene Cluster 3



Closest relative: Thiocoraline

Domain arrangement: Conserved

A-domain specificity predictions: One difference in A-domain specificity (Gene cluster label 1)

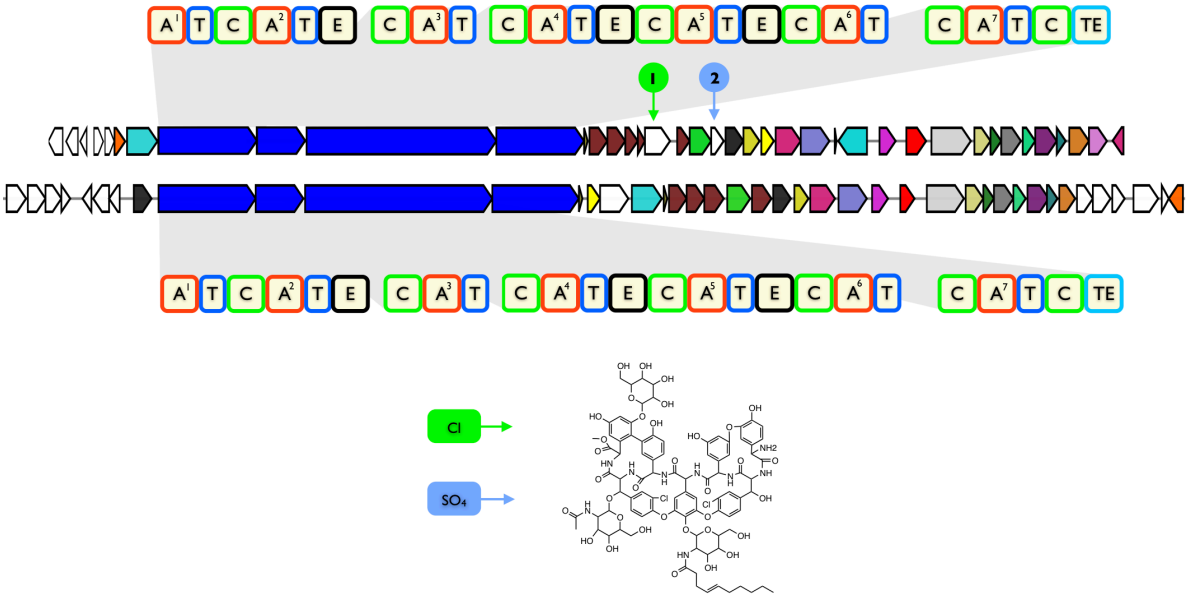
A-domain	Thiocoraline Residue	Thiocoraline Pocket	eDNA Pocket	eDNA Prediction
1	Cys	NAWHMSLPVK	NAWHMSLP**D**K	Cys
2	Gly	DILQLGLVWK	DILQLGL**I**WK	Gly
3	dmCys	DLFNFSLVWK	DLFNFSLVWK	dmCys
4	mCys	DAYWWGGTFK	DAYWWGAAFT	cyProp*

cyProp = cyclopropyl amino acid, mCys = N-methyl-Cys, dmCys = N,S-dimethyl-Cys
*Prediction is based on a manual alignment of all bis-intercalator associated A-domains

Tailoring Enzymes: Conserved

Summary: The partial gene cluster presented above is predicted to encode a structure related to thiocoraline that differs from any of the previously isolated bis-intercalator natural products. As can be seen in the gene cluster comparison diagram, these two clusters have very similar overall architecture and gene contents. The eDNA cluster does however contain a predicted change in A-domain specificity at one location in the NRPS megesynthetases (Gene cluster label 1). The three remaining A-domains specificity predictors show almost perfect matches to the corresponding A-domains seen in the thiocoraline gene cluster.

Figure S7. eDNA Glycopeptide Gene Cluster 1.



Closest relative: Teicoplanin

Domain arrangement: Conserved

A-domain specificity predictions: No differences in equivalent domains

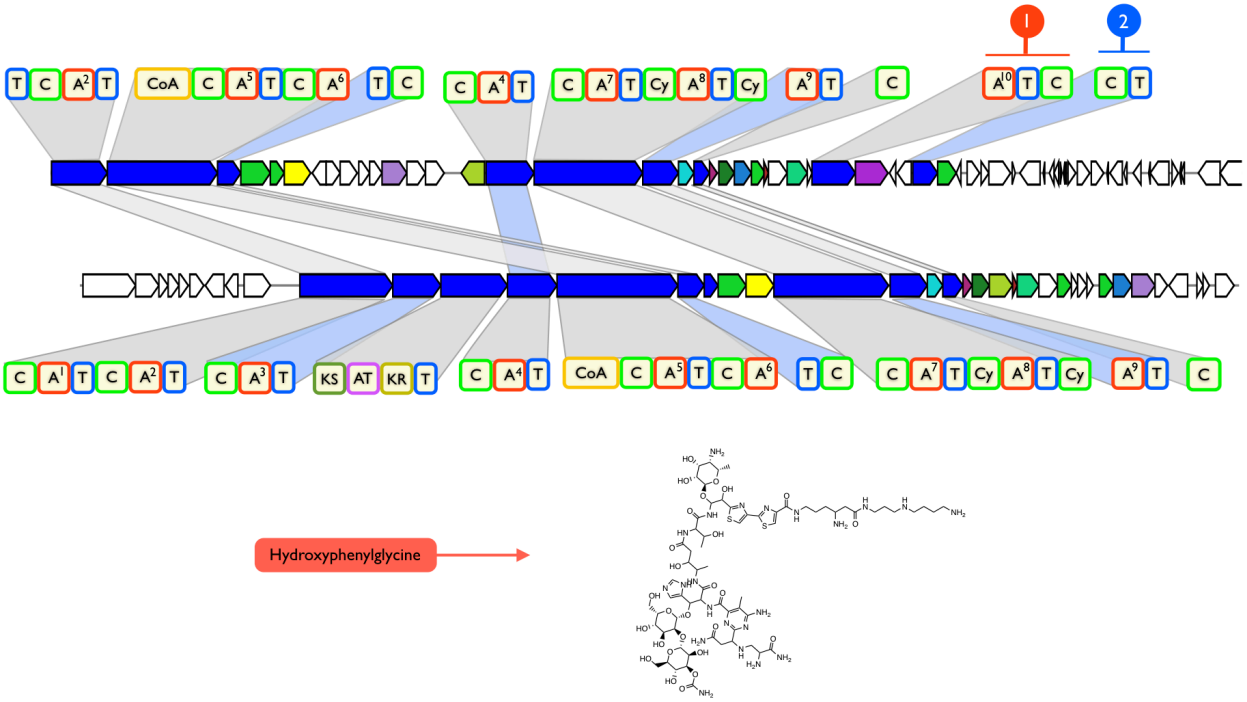
A-domain	Teicoplanin residue	Teicoplanin Pocket	eDNA Pocket	eDNA Prediction
1	Hpg	DAFHLGLLCK	DAFHLGLLCK	Hpg
2	Tyr	DASTVAAVCK	DASTVAAVCK	Tyr
3	dHpg	DAYNLGTLCK	DAYNLGTLCK	dHpg
4	Hpg	DIFHLGLLCK	DIFHLGLLCK	Hpg
5	Hpg	DALHLGLLCK	DALHLGLLCK	Hpg
6	Bht	DASTIAGVCK	DASTVAGVCK	Bht
7	dHpg	DPYHGGTLCK	DPYHGGTLCK	dHpg

hpg = hydroxyphenylglycine, dhpg = hydroxyphenylglycine, bht = β -hydroxytyrosine

Tailoring Enzymes: Additional halogenase (Gene cluster label 1). Additional sulfotransferase (Gene cluster label 2). Missing mannosyltransferase.

Summary: The glycopeptides are a well-characterized family of antibiotics comprising over 200 structural variants. Glycopeptides derive largely from two invariant peptide cores, exemplified by teicoplanin and vancomycin. Structural diversity in this family of compounds is generated primarily by differences in tailoring enzyme function leading to variant patterns of glycosylation, acylation, chlorination, methylation, oxidation and sulfation. In contrast to the previous two examples, where changes in both core and tailoring biosynthetic steps were observed, the teicoplanin-like gene cluster presented above differs from known pathways only in tailoring functionality. The repertoire of tailoring enzymes seen in this cluster includes two halogenases, a sulfotransferase, a glycosyltransferase and an acyl-CoA ligase. This constitutes a combination of tailoring enzymes that has not been observed in any previously characterized glycopeptide cluster and thus would not be predicted to generate any previously characterized glycopeptide antibiotic.

Figure S8. eDNA Bleomycin-like Gene Cluster



Closest relative: Tallisomycin

Domain arrangement: One new module (Gene cluster label 1)

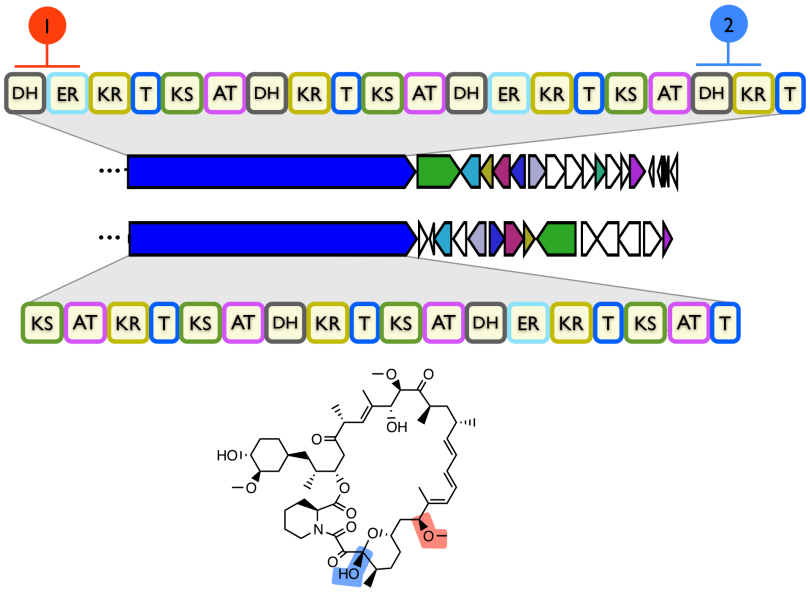
A-domain specificity predictions: No specificity differences in equivalent domains

A-domain	Tallisomycin Residue	Tallisomycin Pocket	eDNA Pocket	eDNA prediction
1	Asn	DLTKVGEVGK	DLTKVGEVGK	Asn
2	His	DSALVAEVWK	DSALIAEVWK	His
3	Ala	DLFNNALTYK	N/R	N/R
4	Thr	DFWGVGMVHK	DFWSVGMVHK	Thr
5	Ser	DVWHVSLVDK	DVWHVSLIDK	Ser
6	Asn	DLTKVGEVGK	N/R	N/R
7	β-Ala	VDWVVSLADK	VDWVISLADK	β-Ala
8	Cys	DLYNMSLIWK	DLYNMSLIWK	Cys
9	Cys	GFYHLGLLWK	DVSHLGLIWK	Cys
10		N/A	DCLHIGLSYK	Hpg

Tailoring Enzymes: N/A

Summary: The partial gene cluster presented above is predicted to encode a structure most closely related to tallysomycin, a member of the bleomycin-like family of antitumor/antibiotics (12, 13). The eDNA gene cluster shows low sequence identity (55 – 65 %) and differs in cluster architecture when compared to the tallysomicin gene cluster, however the vast majority of key biosynthetic elements are conserved between the two gene clusters. The exception to this conservation, are two NRPS enzymes seen in the eDNA cluster that are not observed in the tallisomycin biosynthetic gene cluster. The additional A-domain present in the eDNA gene cluster (Gene cluster label 1) is predicted to activate a hydroxyphenylglycine (HPG).

Figure S9. eDNA Rapamycin-like Gene Cluster



Closest relative: Rapamycin

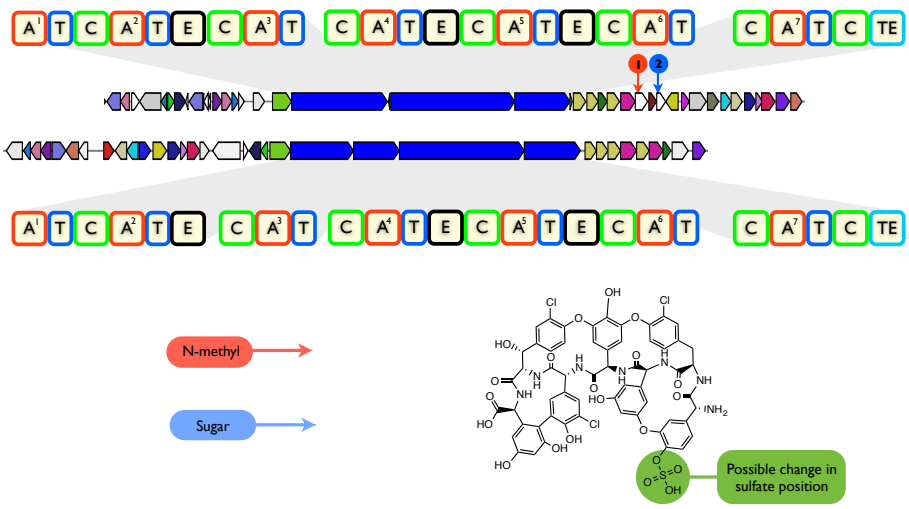
Domain arrangement: Extra DH, KR and ER domains (Gene cluster labels 1 and 2).

A-Domain specificity predictions: N/A

Tailoring Enzymes: New cytochrome P450.

Summary: This partial gene cluster is most closely related to a large fragment of the rapamycin biosynthetic gene cluster. The eDNA clone contains a gene that encodes a PKS megasynthase that is closely related to *rapC* as well as homologs of several key rapamycin associated tailoring enzymes. Among the conserved tailoring enzymes found on this clone are homologs of *rapJ,* the oxidative enzyme believed to install a carbonyl at the C9 position of rapamycin; *rapK,* the oxidative enzyme involved in the biosynthesis of the cyclohexyl starter unit used in rapamycin biosynthesis, and *rapL,* a lysine-cyclodeaminase involved in generation of the pipecolic acid used as an NRPS building block in rapamycin biosynthesis. An additional cytochrome P450 oxidative enzyme that is not closely related to any of the known *rap* biosynthetic genes represents a possible new tailoring function in the eDNA gene cluster. In spite of high sequence identity observed between individual ORFs found on the eDNA clone and the corresponding gene sequences from the rapamycin gene cluster (75 - 85 %) the eDNA derived megasynthase contains a number of key differences compared to RapC. When compared to RapC, the eDNA megasynthase contains additional dehydratase and enoyl reductase domains in one module (Gene cluster label 1) and additional ketoreductase and dehydratase domains in a second module (Gene cluster label 2). The substructure predicted to arise from this RapC-like megasynthase is not observed in the corresponding region of any characterized rapamycin-like natural product (e.g. Rapamycin, FK520, FK506, Meridamycin) suggesting that the eDNA cosmid arises from a gene cluster that encodes for a previously uncharacterized rapamycin-like structure.

Figure S10. eDNA Glycopeptide Gene Cluster 2.



A-domain specificity predictions: No differences in equivalent domains

A-domain	A47934 residue	Teicoplanin Pocket	eDNA Pocket	eDNA Prediction
1	Hpg	DAFHLGLLCK	DAFYQGLVWK	Hydrophobic
2	Tyr	DASTVAAVCK	DASTVAAVCK	Tyr
3	dHpg	DAYNLGTLCK	DVLLVGTIAK	Hydrophobic
4	Hpg	DIFHLGLLCK	DIFHLGLLCK	Hpg
5	Hpg	DALHLGLLCK	DALHLGLLCK	Hpg
6	Bht	DASTIAGVCK	DPYHEGTLCK	Bht
7	dHpg	DPYHGGTLCK	DPYHGGTLCK	dHpg

hpg = hydroxyphenylglycine, dhpg = hydroxyphenylglycine, bht = β-hydroxytyrosine

Closest relative: A47934

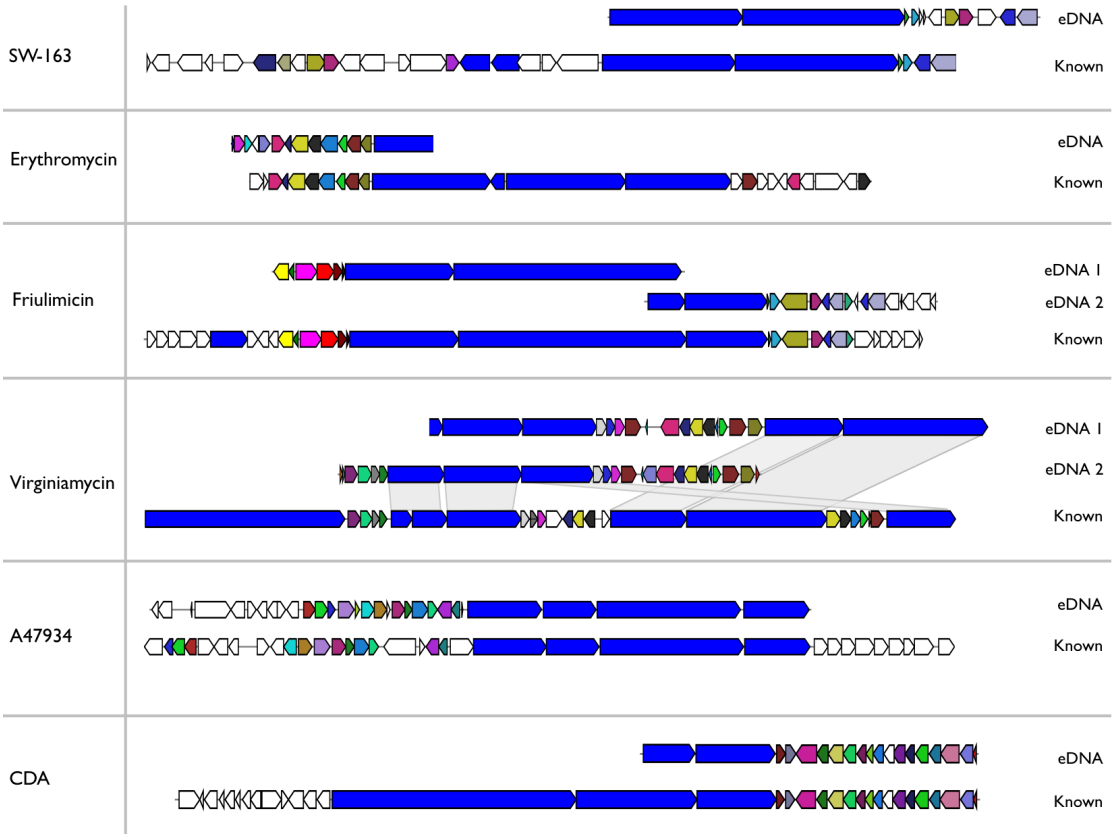
Domain arrangement: Conserved

A-domain specificity predictions: No differences in equivalent domains

Tailoring enzymes: Additional glycosyltransferase and *N*-methyltransferase

Summary: As with the previous glycopeptide pathway (Fig. S7) the pathway presented above contains changes in the collection of encoded tailoring enzymes only. The eDNA pathway contains an additional *N*-methyltransferase and glycosyltransferase when compared to it's closest functional relative (14). In addition to this, it is possible that the regio-specificity of the eDNA-encoded sulfotransferase could differ from that of the A47934 sulfotransferase. The eDNA pathway is therefore predicted to encode an N-methylated, glycosylated derivative of A47934, in which the position of the sulfate group may also be altered.

Figure S11. eDNA gene clusters that are predicted to be functionally identical to characterized gene clusters.

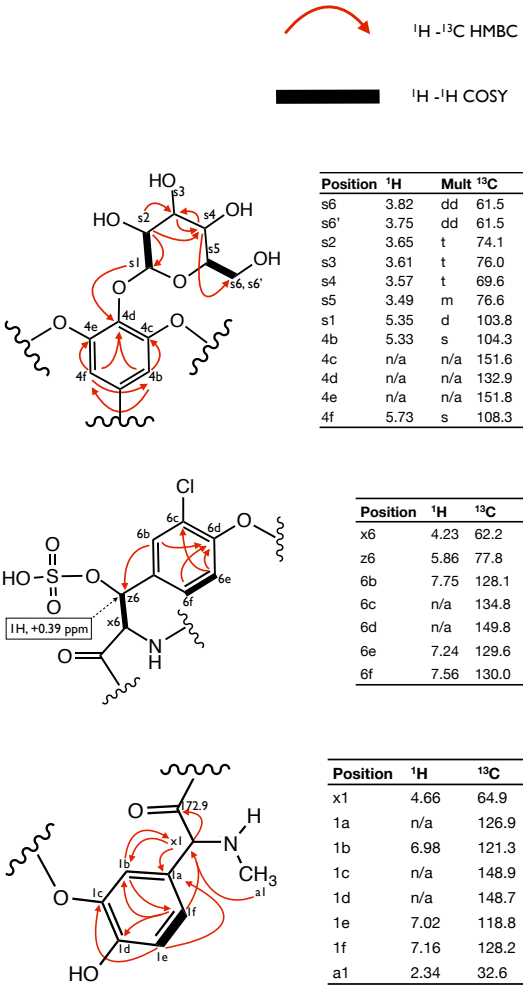


Summary: For each pathway targeted in this study, the initial clone recovered was sequenced and bioinformatically compared to its closest relative cluster. If no significant differences in megasynth(et)ase makeup or finishing enzyme content were observed, the pathway was classified as functionally identical to a known gene cluster and not pursued any further. Each of the partial pathways presented above contains identical domain arrangement and substrate specificity to the indicated characterized biosynthetic system. All tailoring enzymes recovered were predicted to have a biochemical function present in the corresponding characterized pathways. NRPS/PKS encoding ORFS are colored dark blue. Equivalent ORFs are given the same color. Horizontal grey bars indicate equivalent megasynth(et)ase enzymes whose relative positions have shifted. ORFs colored white are those which were not present in partial eDNA pathways recovered or do not have a direct role in biosynthesis (e.g. export and regulatory functions).

a

Position	A47934 ^{1}H	A47934 ^{13}C	Compound 1 ^{1}H	Compound 1 ^{13}C
1b	7.01	121.3	6.98	121.3
1e	**7.6**	**124.2**	**7.02**	**118.8**
1f	**7.38**	**127.3**	**7.16**	**128.2**
2b	7.32	131.8	7.34	132.0
2e	7.21	123.6	7.24	125.2
2f	7.58	130.1	7.58	130.0
3b	6.49	110.3	6.42	110.0
3d	6.62	105.8	6.58	105.8
3f	6.24	104.0	6.19	104.0
4b	5.27	104.0	5.33	104.3
4f	5.68	107.3	5.73	108.3
5b	6.96	134.2	6.92	134.3
5f	7.01	127.1	6.98	126.9
6b	**7.61**	**124.4**	**7.75**	**128.1**
6e	**7.21**	**124.6**	**7.24**	**129.6**
6f	**7.47**	**127.3**	**7.56**	**130.0**
7d	6.52	102.4	6.51	103.3
7f	6.44	107.5	6.51	107.9
a1	N/A	N/A	2.34	32.6
s1	N/A	N/A	5.35	103.8
s2	N/A	N/A	3.65	74.1
s3	N/A	N/A	3.61	76.0
s4	N/A	N/A	3.57	69.6
s5	N/A	N/A	3.49	76.6
s6	N/A	N/A	3.82	61.5
s6'	N/A	N/A	3.75	61.5
x1	**5.4**	**56**	**4.66**	**64.9**
x2	4.94	56.5	4.9	56.4
x3	5.49	58.1	5.47	57.9
x4	5.49	54.9	5.47	56.0
x5	4.35	53.5	4.32	54.2
x6	**4.00**	**63.2**	**4.23**	**62.2**
x7	4.53	N/R	4.49	60.3
z2	3.35	N/R	3.35	36.1
z2'	3.10	N/R	3.13	36.1
z6	**5.47**	**71.6**	**5.86**	**77.8**

b



Position	^{1}H	Mult	^{13}C
s6	3.82	dd	61.5
s6'	3.75	dd	61.5
s2	3.65	t	74.1
s3	3.61	t	76.0
s4	3.57	t	69.6
s5	3.49	m	76.6
s1	5.35	d	103.8
4b	5.33	s	104.3
4c	n/a	n/a	151.6
4d	n/a	n/a	132.9
4e	n/a	n/a	151.8
4f	5.73	s	108.3

Position	^{1}H	^{13}C
x6	4.23	62.2
z6	5.86	77.8
6b	7.75	128.1
6c	n/a	134.8
6d	n/a	149.8
6e	7.24	129.6
6f	7.56	130.0

Position	^{1}H	^{13}C
x1	4.66	64.9
1a	n/a	126.9
1b	6.98	121.3
1c	n/a	148.9
1d	n/a	148.7
1e	7.02	118.8
1f	7.16	128.2
a1	2.34	32.6

c

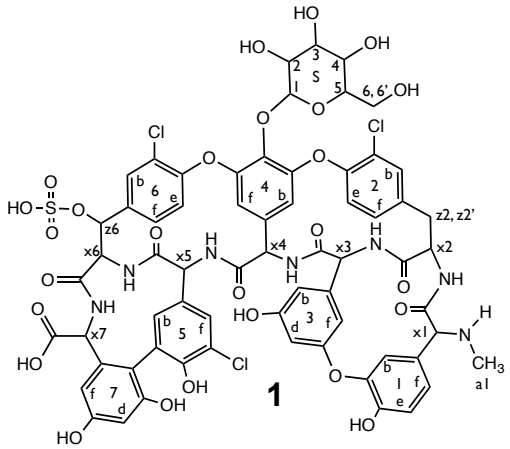


Figure S12. Structure elucidation of compound 1. (a) Comparison of ^{1}H and ^{13}C chemical shift data for A47934 and compound **1**. Bold is used to indicate positions where shifts changed significantly between structures. (b) Key NMR correlations used to define the changes seen in the structure of compound **1**. (c) Structure and numbering scheme for compound **1**.

Structural elucidation of compound 1: High-resolution mass spectrometry predicts a molecular formula of $C_{65}H_{56}Cl_3N_7O_{26}S$ [HR-ESI-MS *m/z* 1488.2134 $[M+H]^+$ (calcd. for $C_{65}H_{57}Cl_3N7O_{26}S$, 1488.2140). for compound **1**. This corresponds, as predicted bioinformatically, to the addition of a methyl, a sulfate and a glucose to desulfo-A47934 (dsA47934) that is natively produced by *S. toyocaensis*:*ΔStaL*. The presence and position of each of these functional groups was assigned based on 1 and 2D NMR data and a comparison of this data to the corresponding data published for A47934 (14) (Fig S12). ^1H and ^{13}C chemical shift data is in good agreement between the two compounds with a few key exceptions (Fig S12a). These exceptions together with key 2D NMR correlations (Fig. S12b) allowed us to define both the identify and position of each new functional group and in turn the structure of compound **1** (Fig. S12).

1. Sulfate - The z6 proton is deshielded by 0.39 ppm compared to A47934, which is consistent with the presence of a sulfate instead of a hydroxyl on carbon z6. As reported in the sulfation of other glycopeptide structures small changes in chemical shifts for positions proximal to z6 (x6, 6f, 6b and 6e) are also observed.

2. Methyl - An additional methyl (^1H s 2.34, ^{13}C 32.6) is seen in the ^1H and ^{13}C spectra of compound 1. The deshielding of this methyl singlet suggested it was attached through either an oxygen or a nitrogen. The presence and position of the new *N*-methyl substituent was established by an HMBC correlation between the new methyl signal and the methine carbon at X1.

3. Glucose - A new collection of carbon and proton chemical shifts seen in 1D spectra are consistent with the presence of a sugar in compound 1 (Fig S12). This sugar is linked to the core dsA47934 structure by an HMBC correlation from the anomeric carbon proton to carbon-4d. The structure of the attached sugar was determined by extensive COSY and HMBC correlations shown in Figure S12 b.

4. Absence of sulfate in dsA47934 compared to A47934 - As expected, protons 1x, 1e and 1f, are shielded by 0.67, 0.63 and 0.27 ppm, respectively due to the absence of the sulfate in the dsA4793 starting material compared to A47934.

HRMS data for compounds 2 and 3.
Compound **2**: $C_{59}H_{46}Cl_3N_7O_{21}S$ [HR-ESI-MS *m/z* 1324.1570 $[M-H]^-$ (calcd. for $C_{59}H_{45}Cl_3N_7O_{21}S$, 1324.1455).
Compound **3**: $C_{59}H_{47}Cl_3N_7O_{18}$ [HR-ESI-MS *m/z* 1246.2002 $[M+H]^+$ (calcd. for $C_{59}H_{47}Cl_3N_7O_{18}$, 1246.2043).

Figure S13: ^{1}H NMR spectrum of compound **1** in 3:1 D_2O:CD_3CN, 300 K

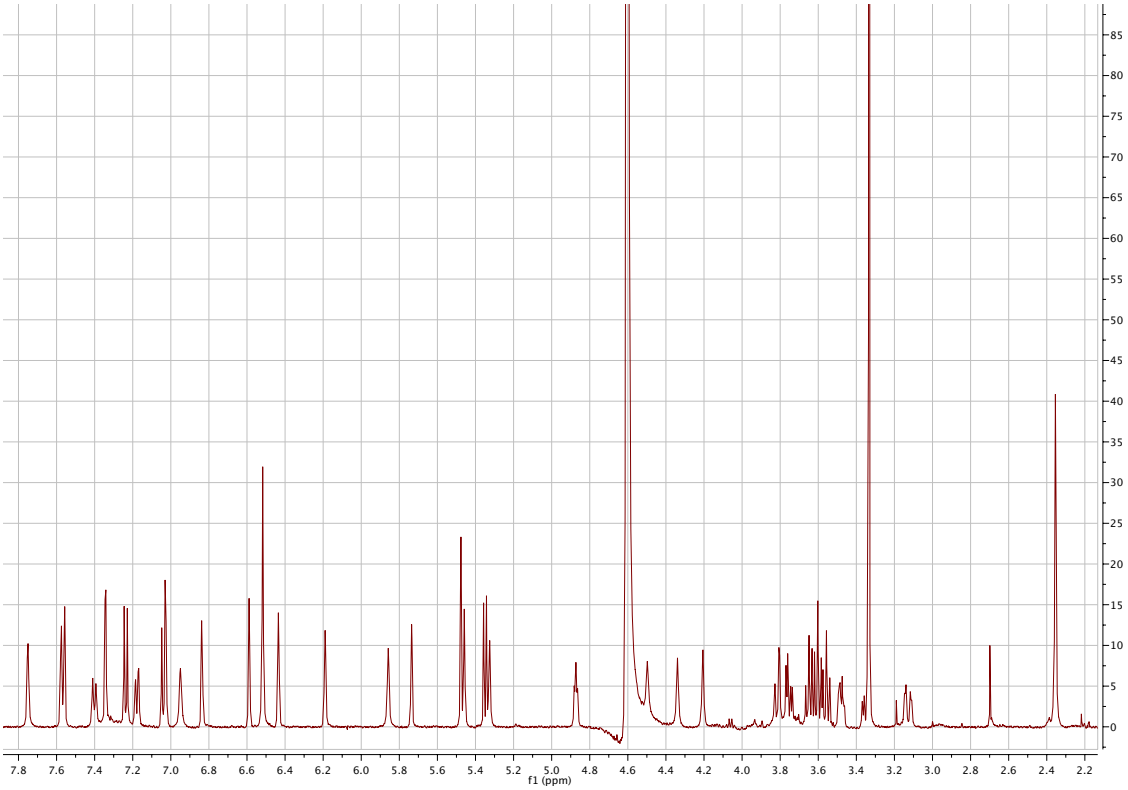


Figure S14: ^{1}H-^{1}H COSY spectrum of compound **1** in 3:1 D_2O:CD_3CN, 300 K

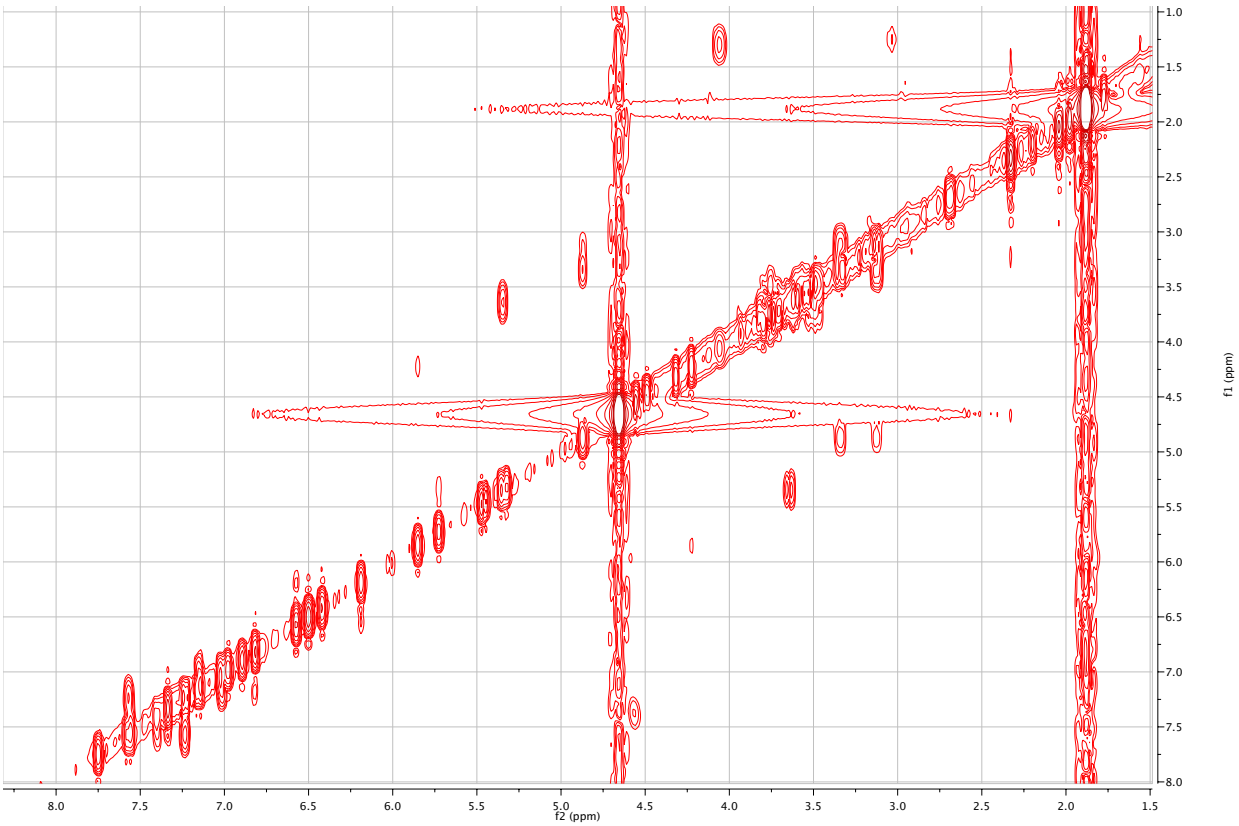


Figure S15: ^{1}H-^{13}C HMBC spectrum of compound **1** in 3:1 D_2O:CD_3CN, 300 K

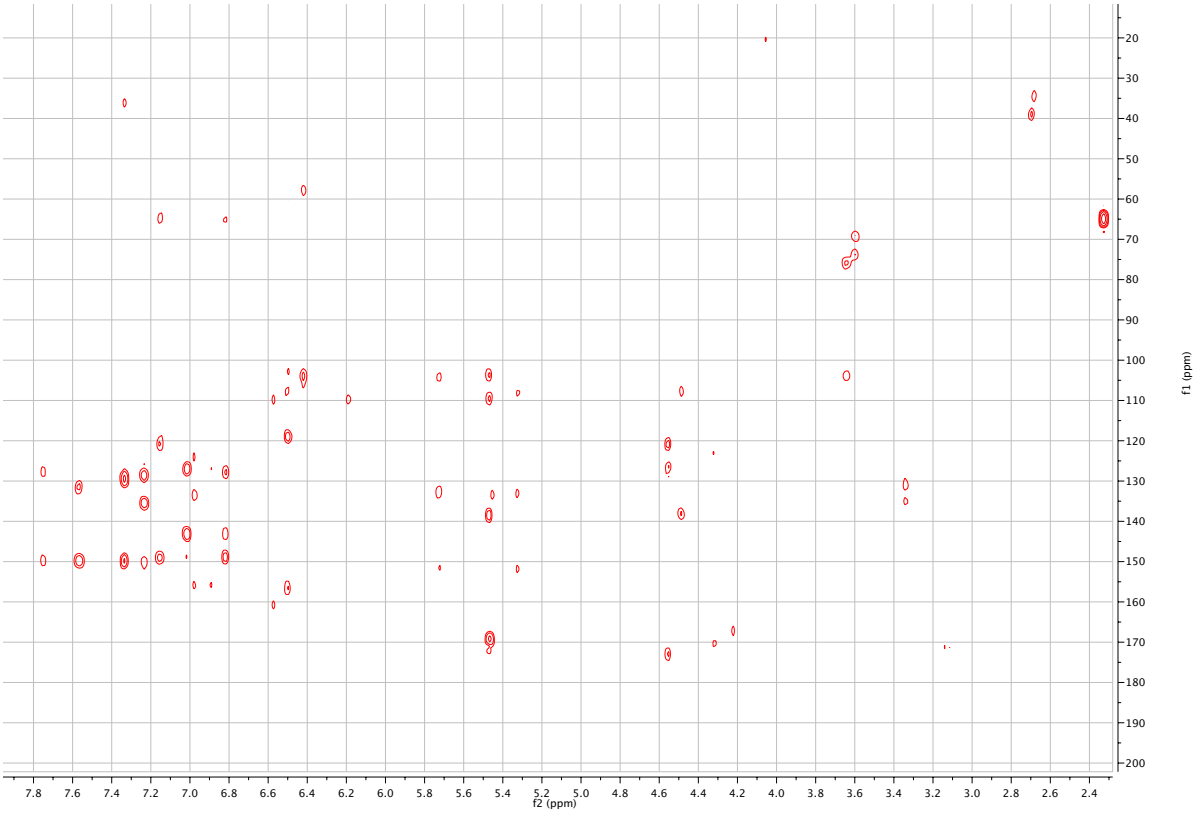


Figure S16: ^{1}H-^{13}C HSQC spectrum of compound **1** in 3:1 D_2O:CD_3CN, 300 K

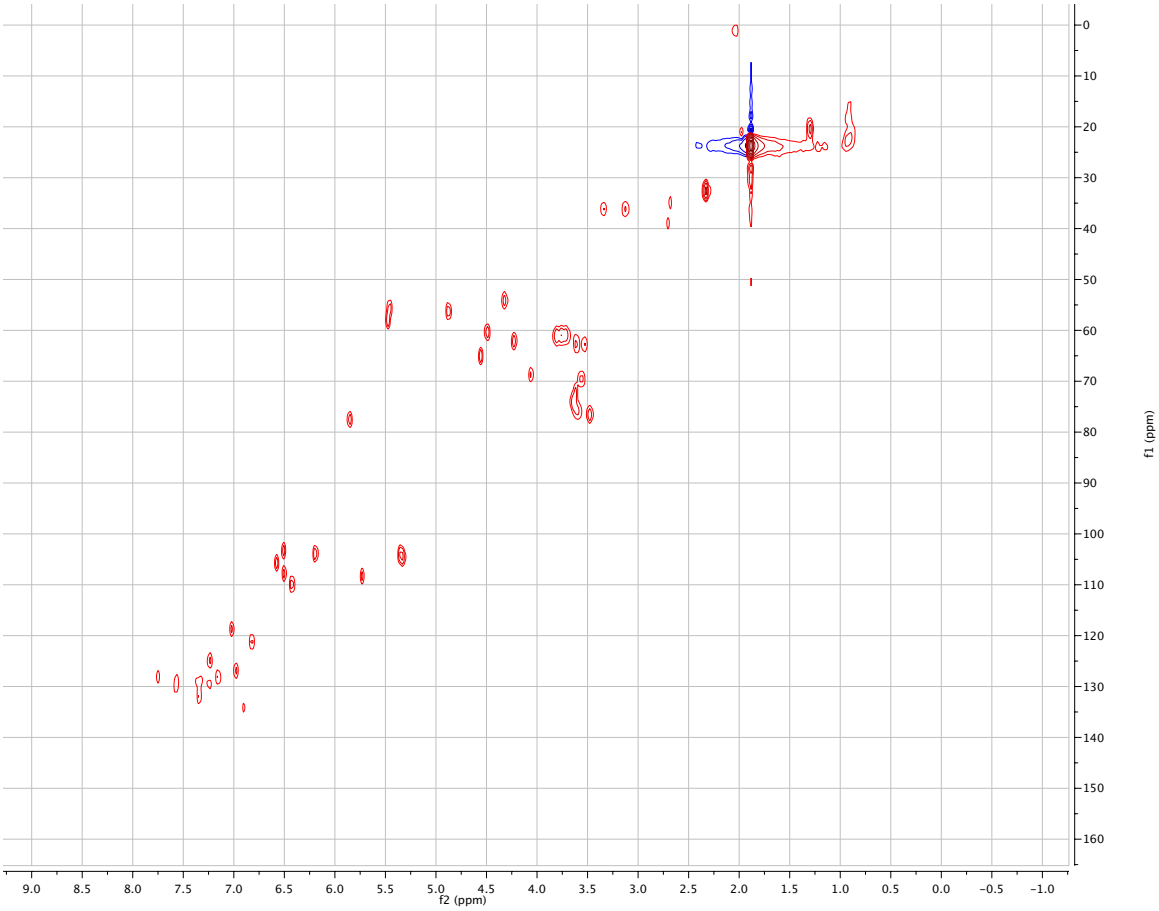
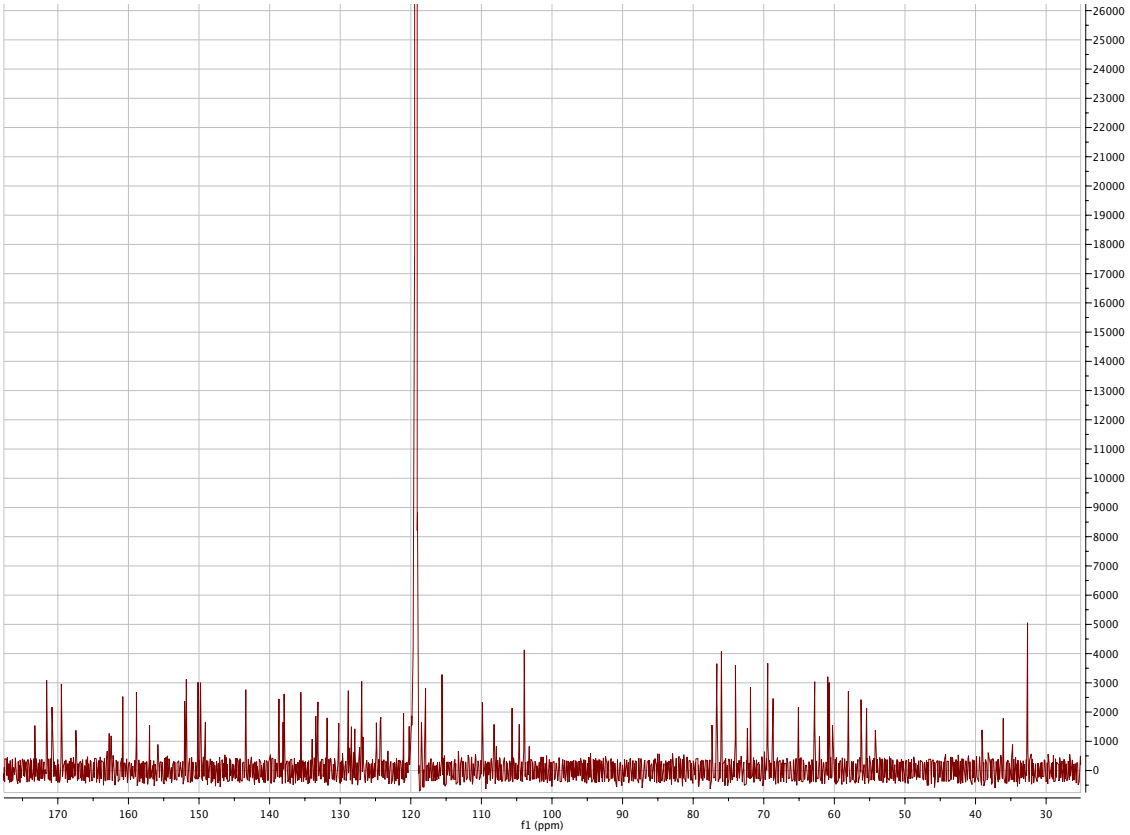


Figure S17: ^{13}C spectrum of compound **1** in 3:1 D_2O:CD_3CN, 300 K



Supplementary methods:

NCBI-NT-A/KS subsets: A/KS-domain sequences were collected from >250 functionally characterized gene clusters representing clinically and biosynthetic interesting natural products (Table S3). Sequenced relatives of these domains were identified in Genbank by blasting these sequences against the NCBI-NT database. Blast hits with expect values of $10e^{-5}$ or lower and Blast alignments of >100 bp were collected to create a database of NCBI-NT A/KS-domain related sequences. The well annotated A and KS-domain sequences from functionally characterized gene clusters of interest were then added to this collection of sequences. All redundant NCBI-NT sequences were removed by clustering at 100% identity. This led to the creation of datasets containing A/KS-domains (NCBI-NT-AD and NCBI-NT-KS, respectively). These contained domains from annotated functionally characterized natural product gene clusters as well as sequenced domain relatives from Genbank and were used in Blast analyses of library and crude soil 454 amplicon sequences.

Recovery of target clones: Specific primers targeting ~200 bp regions from amplicons of interest were designed using the default web setting of Batch Primer3. One µl of an overnight culture of each library pool of interest was used as template in whole cell PCR reactions with the corresponding recovery primers. The resulting amplicons were gel purified and sequenced to confirm the presence of the correct target within the pool. Cultures from library pools containing target sequences were diluted and distributed into 384-well plates such that each well was seeded with ~20 cells in a volume of 100 uL and plates were grown overnight to confluence. The following day pools representing each row within a 384 well plate were screened with recovery primers by whole cell real-time PCR. Individual wells in rows that produced amplicons of correct predicted size were then screened to identify the ~ 20 clone pool containing a target clone. PCR positive wells were then plated on solid medium to obtain single colonies and individual colonies were screened by colony PCR to identify clones with target gene clusters. Typically, between 12-24 clones were recovered in parallel using this process.

***In silico* analysis of recovered gene clusters:** Recovered cosmid clones were pooled and sequenced using 454 pyrosequencing. Reads were quality filtered and assembled into contigs using GS de-novo assembler. Mapping clone end sequencing data onto the pool of assembled contigs using Seqman software identified contigs representing complete cosmid inserts. Overlapping cosmids spanning a single pathway were initially sequenced separately and subsequently assembled into larger contigs using Seqman. Open reading frames were identified using MetaGeneMark (2). Preliminary annotation of open reading frames was conducted using RAST(3) followed by a BLAST search and manual analysis to determine the predicted function for each gene. Putative relatives for each cluster were assigned based on *1*. a manual comparison of this data to sequenced gene clusters and *2*. the cumulative Blast scores returned by the Cluster-BLAST function of AntiSMASH (4). The domain organization within each megasynth(et)ase was determined using the NRPS/PKS analysis webserver (15) and AntiSMASH. A-domain binding pockets from characterized and eDNA derived megasynthetases were extracted using NRPS predictor2 and compared manually. Predicted changes in A-domain and AT-domain substrate specificities were assigned based on data from NRPS Predictor2 (6) and the method of Yadav *et al*. (7), respectively both of which are implemented as part of the AntiSMASH analysis pipeline. Putative functions for new tailoring enzymes in eDNA pathways were assigned based on the predicted function of the closest characterized relative identified by Blast in NCBI.

eSNaPD (environmental Surveyor of Natural Product Diversity) web based tool for analysis, visualization and management of amplicon data: An automated web-based analysis tool eSNaPD was developed in which all amplicon data processing steps were integrated with additional organization, display and data management features. The automated software includes an additional bootstrapping step whereby the Blast analysis is carried out iteratively 10 times. Each time the order of sequence in the database is shuffled. Only known domains that appear as top Blast hits in greater than 50% of the iterations run are reported as hits. This additional boot strapping analysis removes a small number of hits (~2-3%) that arise from clustering anomalies. eSNaPD is designed to perform a completely automated analysis requiring only the uploading of raw 454 sequencing amplicon read files (*.fna*) and if necessary a file of tagged primers used for amplifying A and/or KS domain sequences (text file containing the following primer key: Plate-letter row/column# 454-primer tag-sequence amplification-primer). Two query forms are available for uploading 454 reads. One query form uses data from pooled rows (row.fna) and columns (col.fna) derived from libraries arrayed in 96-well plates and the second query form allows for submission of 454 reads from bulk environmental samples (bulk.fna).

The back-end Perl script analysis software of eSNaPD generates the data files needed for display and for the bulk downloading of hit information as *csv* or *fasta* files. The front-end server side software is written in PHP and provides a web-interface for displaying the analysis output. The final display is an interactive graphical user interphase (GUI) that maps sequence hits to specific positions within a library. It also provides an interactive list of all hits identified in a library or crude environmental sample. The eSNaPD GUI provides buttons for displaying chemical structures of hit relatives, sequences of amplicon hits, Blast alignment data for hits, complete library location information for hits and links to relevant external sites related to each hit. Global dataset statistics are provided for each amplicon collection that is analyzed by eSNaPD. eSNaPD is designed to allow users to manage, store and interact with datasets generated from multiple environmental samples simultaneously. The NCBI-NT-AD and NCBI-NT-KS files can be easily modified to permit screening against domains from additional functionally characterized gene clusters or with environmental amplicons derived from different A/KS-domain degenerate primers or even completely different conserved biosynthetic genes

Construction of a promoter driven eDNA derived tailoring operon for glycopeptide functionalization: The hygromycin resistance cassette and *ermE** promoter region from the vector PIJ10257 (16) were amplified by PCR using the primers PromFw: TACGTTGATCTCTTCGAACGTGCCCCGGACGACGTTGTTCTCCTCGAACACATGGGGCCT CCTGTTCTAGACG and PromRv: CTGCAACCATTCCCCATCCGAGGGCTGGACGAGGCGAAAGGACCTCTGCAGGTTCATGTG CAGCTCCATCAG, where underlined bases indicate cosmid targeting homology regions. 200 ng of the resulting recombinogenic cassette was co-transformed with 200 ng of the target cosmid, by electroporation, into λ/red recombineering proficient cells that had been made competent as previously described (17). Recombinants were selected on LB agar supplemented with 200 µg/mL hygromycin (cassette resistance) and 50 µg/mL apramycin (cosmid resistance). The construct was confirmed by PCR, restriction mapping and sequencing across the cassette integration site.

Heterologous expression of an eDNA cluster to produce new glycopeptide congeners. The promoter driven tailoring construct described above was introduced into *S. toyocaensis ΔstaL* (18) by conjugation using *E. coli* S17-1. Exconjugants were struck onto modified Bennett's agar and incubated for 72 h at 30 °C. Single colonies were then picked into starter

cultures containing 50 ml Streptomyces Vegetative Medium (SVM) in baffled 125 ml flasks. After 72 h growth (30 °C/200 rpm), 0.25 ml of starter culture was used to inoculate production cultures flasks. Each 125 ml baffled production culture flask contained 50 mL of Streptomyces Antibiotic Medium (SAM) supplemented with 1.0 g/L yeast extract. Following 120-160 h growth (30 °C/200 rpm mycelia from production cultures were harvested by centrifugation and extracted once with 2.5% w/v ammonium hydroxide (1 ml/g wet pellet weight) followed by distilled water (1 ml/g wet pellet weight). The combined extracts were adjusted to pH 7.5 and analyzed directly by reversed phased HPLC: C18 (5μM, 4.6 x 150 mm), 20 mM ammonium acetate/acetonitrile, 90:10 to 70:30 over 10 min, 1.5 ml/min.

Purification of compound 1.

Neutralized extracts from 40 production cultures were fractionated by RediSep automated flash chromatography (5.5 g C18 RediSep® Rf column) using a linear gradient of 1:9 methanol:water 3:7 methanol:water over 25 minutes. Individual fractions were then analyzed by LCMS and those containing compound **1** were pooled. Compound **1** was purified from the pooled and dried flash fractions using isocratic reversed phased HPLC: C18 (5μM, 4.6 x 150 mm), 16 mM ammonium acetate/8.5 % Acetonitrile, 5 ml/min.

References:

1. Kim JH*, et al.* (2010) Cloning large natural product gene clusters from the environment: piecing environmental DNA gene clusters back together with TAR. *Biopolymers* 93(9):833-844.
2. Zhu W, Lomsadze A, & Borodovsky M (2010) Ab initio gene identification in metagenomic sequences. *Nucleic Acids Res* 38(12):e132.
3. Aziz RK*, et al.* (2008) The RAST Server: rapid annotations using subsystems technology. *BMC Genomics* 9:75.
4. Medema MH*, et al.* (2011) antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. *Nucleic Acids Res* 39(Web Server issue):W339-346.
5. Bachmann BO & Ravel J (2009) Chapter 8. Methods for in silico prediction of microbial polyketide and nonribosomal peptide biosynthetic pathways from DNA sequence data. *Methods Enzymol* 458:181-217.
6. Rottig M*, et al.* (2011) NRPSpredictor2--a web server for predicting NRPS adenylation domain specificity. *Nucleic Acids Res* 39(Web Server issue):W362-367.
7. Yadav G, Gokhale RS, & Mohanty D (2003) Computational approach for prediction of domain organization and substrate specificity of modular polyketide synthases. *J Mol Biol* 328(2):335-363.
8. Schneider T*, et al.* (2009) The lipopeptide antibiotic Friulimicin B inhibits cell wall biosynthesis through complex formation with bactoprenol phosphate. *Antimicrob Agents Chemother* 53(4):1610-1618.
9. Strieker M & Marahiel MA (2009) The structural diversity of acidic lipopeptide antibiotics. *Chembiochem* 10(4):607-616.
10. Zolova OE, Mady AS, & Garneau-Tsodikova S (2010) Recent developments in bisintercalator natural products. *Biopolymers* 93(9):777-790.
11. Lombo F*, et al.* (2006) Deciphering the biosynthesis pathway of the antitumor thiocoraline from a marine actinomycete and its expression in two streptomyces species. *Chembiochem* 7(2):366-376.
12. Galm U*, et al.* (2011) Comparative analysis of the biosynthetic gene clusters and pathways for three structurally related antitumor antibiotics: bleomycin, tallysomycin, and zorbamycin. *J Nat Prod* 74(3):526-536.
13. Tao M*, et al.* (2007) The tallysomycin biosynthetic gene cluster from Streptoalloteichus hindustanus E465-94 ATCC 31158 unveiling new insights into the biosynthesis of the bleomycin family of antitumor antibiotics. *Mol Biosyst* 3(1):60-74.
14. Banik JJ, Craig JW, Calle PY, & Brady SF (2010) Tailoring enzyme-rich environmental DNA clones: a source of enzymes for generating libraries of unnatural natural products. *J Am Chem Soc* 132(44):15661-15670.
15. Starcevic A*, et al.* (2008) ClustScan: an integrated program package for the semi-automatic annotation of modular biosynthetic gene clusters and in silico prediction of novel chemical structures. *Nucleic Acids Res* 36(21):6882-6892.
16. Hong HJ, Hutchings MI, Hill LM, & Buttner MJ (2005) The role of the novel Fem protein VanK in vancomycin resistance in Streptomyces coelicolor. *J Biol Chem* 280(13):13055-13061.

17. Datsenko KA & Wanner BL (2000) One-step inactivation of chromosomal genes in Escherichia coli K-12 using PCR products. *Proc Natl Acad Sci U S A* 97(12):6640-6645.
18. Lamb SS, Patel T, Koteva KP, & Wright GD (2006) Biosynthesis of sulfated glycopeptide antibiotics by using the sulfotransferase StaL. *Chem Biol* 13(2):171-181.