UNIVERSITY OF CALIFORNIA, SAN DIEGO

Mass spectrometry-guided genome mining of peptidic and glycosylated microbial natural products

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy

in

Marine Biology

by

Roland David Kersten

Committee in charge:

Bradley S. Moore, Chair
Pieter C. Dorrestein, Co-Chair
William H. Gerwick
Eric E. Allen
Paul R. Jensen
Nuno Bandeira

2013

The Dissertation of Roland David Kersten is approved and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____

_____

Co-Chair

_____

Chair

University of California, San Diego

2013

For my parents

“Das ist der Weisheit letzter Schluss:
Nur der verdient sich Freiheit wie das Leben,
Der täglich sie erobern muss."
Johann Wolfgang von Goethe, Faust II

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

Abbreviation	Abbreviation description
2,3-DH, 2,3DH	2,3-dehydratase
2-AT	2-aminotransferase
3,4-DH, 3,4DH	3,4-dehydratase
3,4-IM, 3,4IM	3,4-isomerase
3-AT	3-aminotransferase
3-KR, 3KR	3-ketoreductase
4,6-DH, 4,6DH	4,6-dehydratase
4-AT	4-aminotransferase
4-KR, 4KR	4-ketoreductase
A	adenylation domain
aa	amino acid
ACP	acyl carrier protein
AcT	acetyltransferase
Ala	alanine
ARO	aromatase
Asn	asparagine
Asp	aspartic acid
At	aminotransferase
ATP	adenosine triphosphate
BCSDB	Bacterial Carbohydrate Structure Data Base
BLAST	Basic Local Alignment Search Tool
C	condensation domain
C3-MT	C3-methyltransferase
C5-MT	C5-methyltransferase
CarbT	carbamoyltransferase
CID	collision induced dissociation
CoA	coenzyme A
Cy	heterocyclization domain
CYC	cyclase
Cys	cysteine
DH	dehydratase
Dha	dehydroalanine
Dhb	dehydrobutyrine
DHB	2,5-dihydroxybenzoic acid
Dhg	dehydrogenase
DNA	deoxyribonucleic acid

Abbreviation	Abbreviation description
DQF-COSY	double quantum filter-correlation spectroscopy
E	epimerase
EIC	extracted ion chromatogram
ER	enoylreductase
ESI	electrospray ionization
FAS	fatty acid synthase
FT-ICR	Fourier-transform ion cyclotron resonance
FTMS	Fourier-transform mass spectrometry
FuPyIM	fucopyranose isomerase
Glu	glutamate
Gly	glycine
GNP	glycosylated natural product
GT	glycosyltransferase
HMBC	heteronuclear multiple-bond correlation spectroscopy
HPLC	high performance liquid chromatography
IMS	imaging mass spectrometry
ISP2	International Streptomyces Project medium 2
KR	ketoreductase
KS	ketosynthase
Kyn	kynurenine
Lan	lanthionine
LanC	lanthionine synthetase C
LanKC	lanthionine synthetase KC
LanL	lanthionine synthetase L
LanM	lanthionine synthetase M
LC	liquid chromatography
LC-MS	liquid chromatography-mass spectrometry
LTQ	linear trap quadrupole
MALDI	matrix assisted laser desorption ionization
MeGlu	methylglutamate
MeLan	methyllanthionine
mRNA	messenger ribonucleic acid
MS/MS	tandem mass spectrometry
MS^n	tandem mass spectrometry
MS	mass spectrometry
N,N-MT	N,N-dimethyltransferase
NCBI	National Center for Biotechnology Information
N-ET	N-ethyltransferase
NLC	neutral loss chromatogram

Abbreviation	Abbreviation description
NMR	nuclear magnetic resonance
N-MT	N-methyltransferase
NOESY	nuclear Overhauser effect spectroscopy
N-Ox	N-oxidase
NRP	nonribosomal peptide
NRPS	nonribosomal peptide synthetase
NT	nucleotidylyltransferase
O-MT	O-methyltransferase
Orn	ornithine
Ox	oxidoreductase domain
oxDA	oxidative deaminase
OxRed	oxidoreductase
Phe	phenylalanine
PKS	polyketide synthase
PNP	peptide natural product
PPant	phosphopantetheine
PrISM	Proteomic Investigation of Secondary Metabolism
PTFE	polytetrafluoroethylene
PyT	pyrrolyltransferase
RiPP	ribosomally synthesized and posttranslationally modified peptide
RNA	ribonucleic acid
Ser	serine
T	thiolation domain
TDP-Glc	thiamine diphosphate-glucose
TE	thioesterase
TFA	trifluoroacetic acid
Thr	threonine
TOF	time-of-flight mass analyzer
Trp	tryptophan
UV	ultraviolet (light)

LIST OF SYMBOLS

Symbol	Symbol discription
ΔM	full width at half maximum signal intensity
°C	degree Celsius
a	acceleration
Å	Angstrom
B	magnetic field strength
Da	Dalton
E	electric field strength
e	electron charge
eV	electron volt
F	force (on charged particle)
g	gram
h	hour
L	liter
m	mass
M	mass signal
m(calc)	calculated monoisotopic mass
m(obs)	observed monoisotopic mass
m/z	mass-to-charge
MHz	megahertz
mL	milliliter
mm	millimeter
q	charge
Q	quadrupole mass analyzer
RF	radiofrequency
rpm	revolutions per minute
s	second
UV	ultraviolet (light)
v	velocity
V	volt
µL	microliter
µm	micrometer

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

Many thanks to my advisors, Pieter Dorrestein and Brad Moore, for support, mentorship and guidance during my graduate studies in their labs. I thank Pieter for the opportunity he gave me in 2007 to join his lab as an undergraduate and to recommend me as a jointly mentored student for the SIO graduate studies program. He has been an amazing and motivating teacher of mass spectrometry, natural product chemistry and academic work, in general. I thank Brad for the opportunity to join his lab and the SIO graduate school. He has been an amazing and motivating teacher of natural product biosynthesis and a great guide towards academic life.

Many thanks to former and present members of both the Dorrestein and Moore lab – too many to mention here - for a great work atmosphere and a worthwhile graduate studies time.

Thanks to all collaborators on all research projects during my graduate studies for productive interactions.

Thanks and love to all my friends at SIO that make it a special place. First year forever! Thanks and love to Jenan for shared time, making me a better person and cats. Thanks and love to my family and friends in Germany for support, especially during the first years abroad.

Chapter 2, in full, is a reprint of the material as it appears in 'A mass spectrometry-guided genome mining approach for natural product peptidogenomics', Kersten, R.D., Yang, Y.L., Xu, Y., Cimermancic, P., Nam, S.J., Fenical, W., Fischbach, M.A., Moore, B.S., Dorrestein, P.C. *Nature Chemical Biology*, 2011, 7, 794-802. The dissertation author was the primary investigator and author of this paper.

Chapter 3, in full, is a reprint of the material as it appears in 'Bacterial biosynthesis and maturation of the didemnin anti-cancer agents', Xu, Y., Kersten, R.D., Nam, S.J., Lu, L., Al-Suwailem, A.M., Zheng, H., Fenical, W., Dorrestein, P.C., Moore, B.S., Qian, P.Y. *Journal of the American Chemical Society*, 2012, 134, 8625-8632. The dissertation author was one of two equally contributing primary investigators and authors of this paper.

Chapter 4, in full, is currently being prepared for submission for publication of the material. Kersten, R.D., Ziemert, N., Crüsemann, M., Duggan, B.M., Jensen, P.R., Dorrestein, P.C., Moore, B.S. The dissertation author was the primary investigator and author of this material.

Chapter 5, in full, has been accepted for publication. It is shown as it may appear in *Chembiochem*, Kersten, R.D., Lane, A.L., Nett, M., Richter, T.S.K., Duggan, B.M., Dorrestein, P.C., Moore, B.S. *Chembiochem*, 2013, DOI: 10.1002/cbic.2001300147. The dissertation author was the primary investigator and author of this paper.

VITA

2008	Diplom (German 'Master of Science'), Free University, Berlin, Germany
2008-2009	Research associate, University of California, San Diego
2009-2013	Research assistant, University of California, San Diego
2013	Doctor of Philosophy, University of California, San Diego

PUBLICATIONS

1. **Kersten, R.D.**, Ziemert, N., Crüsemann, M., Duggan, B.M., Jensen, P.R., Dorrestein, P.C., Moore, B.S. Glycogenomics, a mass spectrometry-guided genome mining method to connect biosynthetic genes to glycosylated natural products. *In preparation.*
2. **Kersten, R.D.**, Lane, A.L., Nett, M., Richter, T.K.S., Duggan, B.M., Dorrestein, P.C., Moore, B.S. Bioactivity-guided genome mining identifies the lomaiviticin biosynthetic gene cluster in *Salinispora tropica*. *Chembiochem.* DOI: 10.1002/cbic.2001300147 (2013).
3. Ross, A.C., Xu, Y., Lu, L., **Kersten, R.D.**, Shao, Z., Al-Suwailem, A.M., Dorrestein, P.C., Qian, P.Y., Moore, B.S. Biosynthetic Multitasking Facilitates Thalassospiramide Structural Diversity in Marine Bacteria. *J. Am. Chem. Soc.* **135**, 1155–1162 (2013).
4. Roberts, A.A., Schultz, A.W., **Kersten, R.D.**, Dorrestein, P.C., Moore, B.S. Iron acquisition in the marine actinomycete genus *Salinispora* is controlled by the desferrioxamine family of siderophores. *FEMS Microbiol. Lett.* **335**, 95-103 (2012).
5. Watrous, J., Roach, P., Alexandrov, T., Heath, B.S., Yang, J.Y., **Kersten, R.D.**, van der Voort, M., Pogliano, K., Gross, H., Raaijmakers, J.M., Moore, B.S., Laskin, J., Bandeira, N., Dorrestein, P.C. Mass spectral molecular networking of living microbial colonies. *Proc. Natl. Acad. Sci. U. S. A.* 109, E1743-1752 (2012).
6. Xu, Y.*, **Kersten, R.D.***, Nam, S.J., Lu, L., Al-Suwailem, A.M., Zheng, H., Fenical, W., Dorrestein, P.C., Moore, B.S., Qian, P.Y. Bacterial biosynthesis and maturation of the didemnin anti-cancer agents. *J. Am. Chem. Soc.* **134**, 8625-32 (2012). (* - contributed equally to this work)
7. Gonzalez, D.J., Okumura, C.Y., Hollands, A., **Kersten, R.**, Akong-Moore, K., Pence, M.A., Malone, C.L., Derieux, J., Moore, B.S., Horswill, A.R., Dixon, J.E., Dorrestein, P.C., Nizet, V. Novel phenol-soluble modulin derivatives in community-associated methicillin-resistant *Staphylococcus aureus* identified through imaging mass spectrometry. *J. Biol. Chem.* **287**, 13889-13898 (2012).
8. Liu, W.T., **Kersten, R.D.**, Yang, Y.L., Moore, B.S., Dorrestein, P.C. Imaging mass spectrometry and genome mining via short sequence tagging identified the anti-infective agent arylomycin in *Streptomyces roseosporus. J. Am. Chem. Soc.* **133**, 18010-18013 (2011).
9. **Kersten, R.D.**, Yang, Y.L., Xu, Y., Cimermancic, P., Nam, S.J., Fenical, W., Fischbach, M.A., Moore, B.S., Dorrestein, P.C. A mass spectrometry-guided genome mining approach for natural product peptidogenomics. *Nat. Chem. Biol.* **7**, 794-802 (2011).
10. Yang, Y.L., Xu, Y., **Kersten, R.D.**, Liu, W.T., Meehan, M.J., Moore, B.S., Bandeira, N., Dorrestein, P.C. Connecting chemotypes and phenotypes of cultured marine microbial assemblages by imaging mass spectrometry. *Angew. Chem. Int. Ed. Engl.* 50, 5839-42 (2011).
11. Meier, J.L., Patel, A.D., Niessen, S., Meehan, M., **Kersten, R.**, Yang, J.Y., Rothmann, M., Cravatt, B.F., Dorrestein, P.C., Burkart, M.D., Bafna, V. Practical 4'-phosphopantetheine active site discovery from proteomic samples. *J. Proteome Res.* **10**, 320-319 (2011).

12. **Kersten, R.D.**, Dorrestein, P.C. Metalloenzymes: Natural product nitrosation. *Nat. Chem. Biol.* **6**, 636-7 (2010).
13. **Kersten, R.D.**, Meehan, M.J., Dorrestein, P.C. Applications of Modern Mass Spectrometry Techniques in *Natural Products Chemistry. Comprehensive Natural Products II*, Ch.9, 389-456 (2010).
14. **Kersten, R.D.**, Dorrestein, P.C. Secondary metabolomics: natural products mass spectrometry goes global. *ACS Chem. Biol.* **4**, 599-601 (2009).

ABSTRACT OF THE DISSERTATION

Mass spectrometry-guided genome mining of peptidic and glycosylated microbial natural products

by

Roland David Kersten

Doctor of Philosophy in Marine Biology

University of California, San Diego, 2013

Professor Bradley S. Moore, Chair
Professor Pieter C. Dorrestein, Co-Chair

Scientific progress in organic synthesis, biochemistry and biology and cures to many infectious diseases and cancer rely on discovery of microbial natural products and their biosynthetic pathways. 'Omics' approaches such as genome mining have opened new opportunities for natural product discovery within the last decade as ~90% of pathways in microbial genomes are uncharacterized in their products. Genome mining for natural product discovery can be defined as the connection of a natural product (chemotype) with its biosynthetic genes (genotype) by applied biosynthetic knowledge. Traditional genome mining approaches are *in silico*-guided approaches in which the isolation of a new natural product is guided by bioinformatic predictions from a target cryptic gene cluster. The problem of *in silico*-guided genome mining in natural product discovery is its low-throughput rate as only one pathway is characterized per experiment.

In this dissertation, mass spectrometry (MS)-guided genome mining approaches are introduced which rapidly connect a natural product with its biosynthetic genes by matching *de novo* tandem MS structures of biosynthetic building blocks such as amino acids and sugars to metabolite structures predicted from microbial genomes. As MS-guided genome mining starts at the chemotype level by e.g. liquid chromatography-tandem mass spectrometry analysis of a microbial extract and subsequently connects putative natural products with their gene clusters, it has the potential for automation.

In Chapter 2, peptidogenomics is introduced as a MS-guided genome mining approach for characterization of ribosomal and nonribosomal peptide chemotypes and their corresponding genotypes. Peptidogenomics characterized ten new peptide chemo- and genotypes from *Streptomyces* cultures including lanthipeptides, lassopeptides, linaridins and lipopeptides.

In Chapter 3, a combination of imaging mass spectrometry, tandem MS and genome mining characterized the biosynthetic pathway of the didemnin anti-cancer agents in the marine α-proteobacterium *Tistrella mobilis*.

In Chapter 4, glycogenomics is introduced as a MS-guided genome mining approach to connect chemo- and genotypes of glycosylated natural products. Glycogenomics enabled the discovery of putative arenimycin B, a glycosylated aromatic polyketide from the marine actinobacterium *Salinispora arenicola* and its biosynthetic pathway.

In Chapter 5, bioactivity-guided genome mining combined with genetic knockouts and glycogenomics characterized the biosynthetic gene cluster of the lomaiviticins anti-cancer agents in the marine actinobacterium *Salinispora tropica*.

Chapter 1 – Introduction

1.1 Microbial natural products

Natural products, or so-called secondary metabolites, are small, structurally diverse chemicals produced by cells for many phenotypic purposes, e.g. cell-cell communication, self-regulation or adaptation. The term 'secondary metabolites' indicates that these chemicals are not essential for the general cell survival like primary metabolites. Due to their diversity in structure, biosynthesis and activity, natural products from microbial sources have been a great source of innovation for chemists, biochemists, biologists and pharmacists. Organic total synthesis has been driven by the synthetic challenges set by complex natural products [1]. Our understanding of enzymatic catalysis has been fueled by studies of secondary metabolic pathways [2]. Natural products are involved in many ecological processes such as symbiotic interactions [3]. In modern medicine, 49% of current drugs and 70% of antimicrobials are natural product-derived [4]. The wide scientific significance of natural products is based on their role as a part of the central dogma of life, i.e. the information transfer between DNA, RNA, proteins and metabolites [5]. In microbial genomes, the genes involved in the biosynthesis, transcriptional regulation, self-resistance and transport of a natural product are organized on the genomic DNA level as a gene cluster (genotype). The biosynthetic gene products assemble a natural product (chemotype) from simple chemical building blocks which are unmodified or modified primary metabolites such as sugars, amino acids, fatty acids, isoprene units and acetate-derived units. Many natural products are finally exported from the cell to cause a bioactivity (phenotype), e.g. in the interaction with another cell (Figure 1) [6]. This section is a brief introduction to structural and biosynthetic principles of microbial natural products which are relevant to this dissertation.

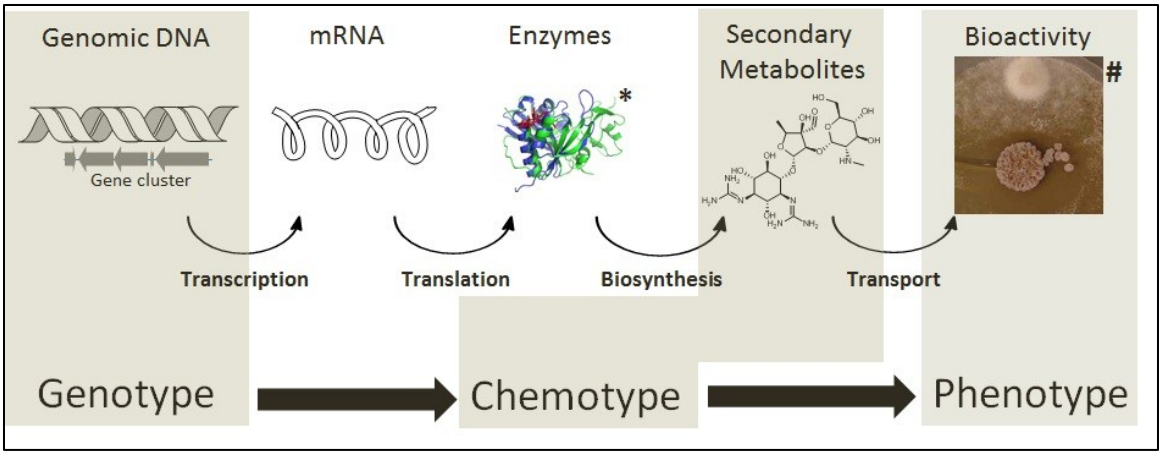


Figure 1: Natural products in the dogma of life, i.e. the information transfer from genomic DNA to RNA to enzymes to metabolites. The information for the biosynthesis of a microbial natural product is encoded in a gene cluster (genotype) which is translated into proteins that catalyze the biosynthesis and secretion of the secondary metabolite (chemotype) to direct its bioactivity (phenotype). * - PDB 3ec4, # - adapted from [3].

1.1.1 Peptide natural products (PNPs)

Peptide natural products (PNPs) are ubiquitious chemicals in all life forms where they have diverse biological functions in development, protection and communication [7]. Microbial peptide natural products are an important source of pharmaceuticals [8] and food preservatives [9]. There are two general strategies of PNP biosynthesis in nature: a ribosomal pathway and a non-ribosomal pathway [10]. PNPs produced by the ribosomal pathway are called ribosomally synthesized and posttranslationally modified peptides (RiPPs) [11], PNPs produced by the nonribosomal pathway are called nonribosomal peptides (NRPs). PNPs are chains of amino acid monomers which can include proteinogenic and non-proteinogenic amino acids. A common modification of PNPs is macrocyclization for increased stability and specific conformation [12].

1.1.1.1 Ribosomal peptides (RiPPs)

Ribosomally synthesized and posttranslationally modified peptides (RiPPs) encompass a rapidly expanding group of natural products [11]. Multiple classes of RiPPs of prokaryotic origin have been characterized through their biosynthetic pathways, which entail diverse posttranslational modification strategies to yield lanthipeptides [13], thiopeptides [14], cyanobactins [15], lasso peptides [16], linaridins [17] and other microcins [18]. Consequently, traditional RiPP classification systems based on bioactivity, producer and structure have shifted

toward a new classification based largely on biosynthesis [19]. In RiPP biosynthesis, the peptide sequence is encoded by a precursor gene directly translated by the ribosome to consist of leader peptide and core peptide regions (Figure 2). The leader peptide serves as a scaffold and contains recognition sites for processing enzymes that introduce posttranslational modifications of the RiPP biosynthetic machinery, whereas the core peptide constitutes the primary sequence of the produced peptide natural product that is modified. Posttranslational modification of the core peptide by biosynthetic enzymes can often be extensive and can provide a wealth of structural diversity rendering these peptides, at first glance, unrecognizable as ribosomally synthesized molecular entities [11].

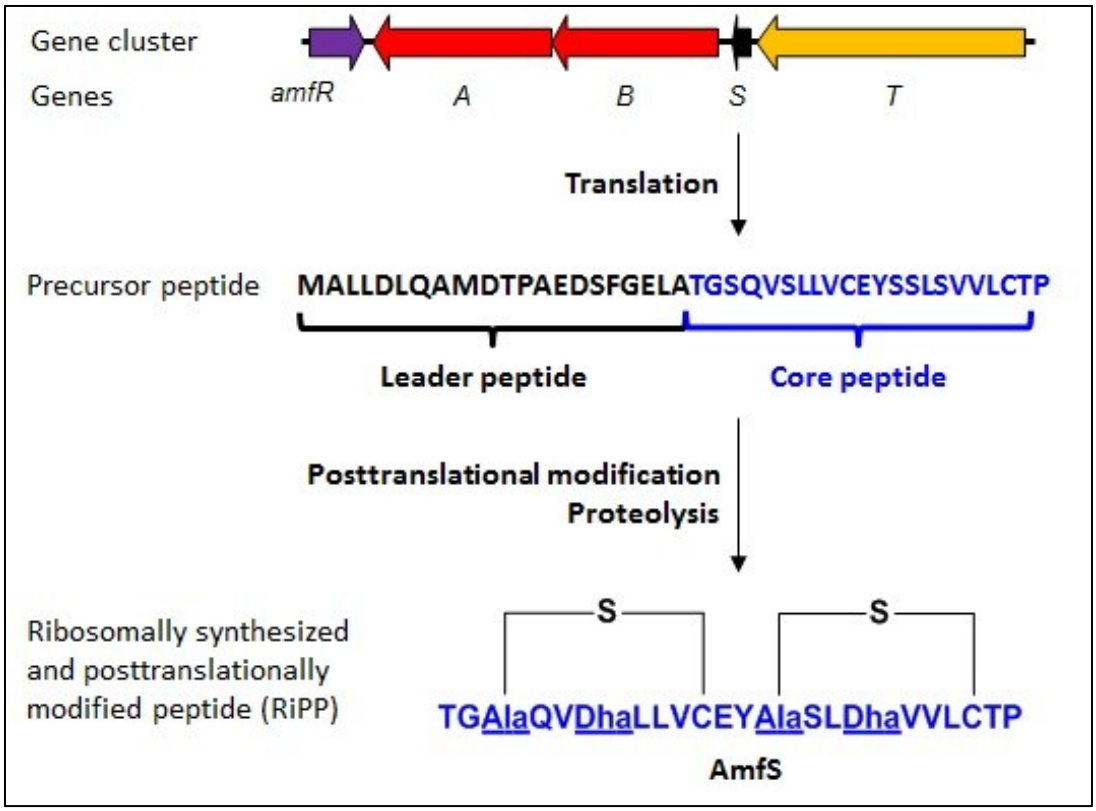


Figure 2: Ribosomal peptide biosynthesis exemplified by the AmfS pathway [20].

1.1.1.1.1 Lanthipeptides

Lanthipeptides are defined by lanthionine (Lan) and methyllanthionine (MeLan) posttranslational modifications. A lanthionine consists of two alanine residues which are connected by a thioether between their β-carbons [11]. A methyllanthionine has an additional

methyl group at one of the β-carbons. A lanthionine is biosynthetically derived from serine and cysteine residues, while methyllanthionine is derived from threonine and cysteine [11,13]. In the first step of Lan/MeLan biosynthesis, serine and threonine residues are dehydrated to dehydroalanine (Dha) and dehydrobutyrine (Dhb), respectively. The only dehydration mechanism characterized to date is phosphorylation-elimination [21]. In the second step, the cysteine thiol attacks Dha/Dhb as a thiolate in a Michael-type addition to form a thioether bond. Protonation of the resulting enolate species yields Lan/MeLan (Figure 3A).

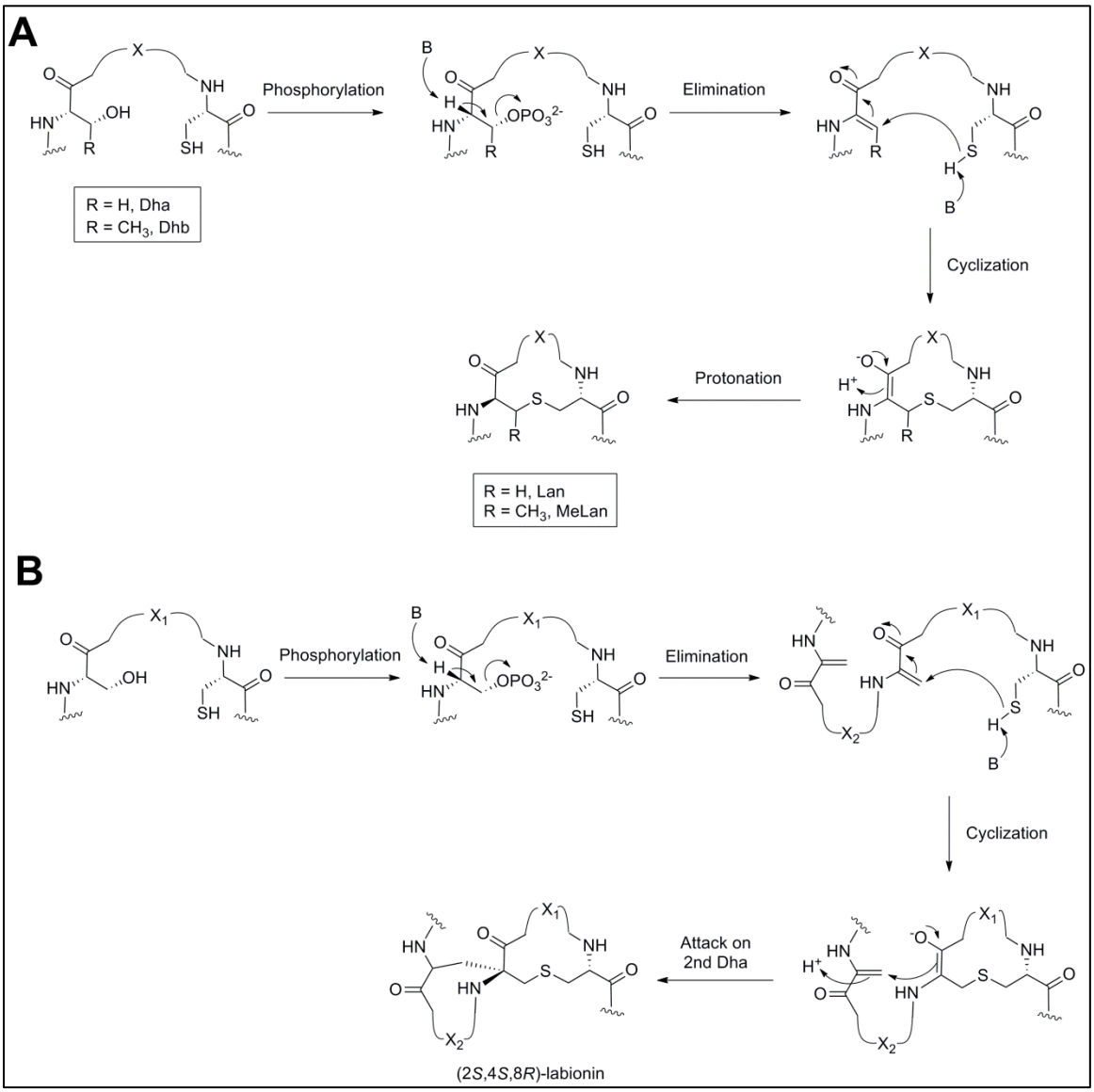


Figure 3: Biosynthesis of (methyl)lanthionine (**A**) and labionin (**B**) motifs in lanthipeptides.

To date, there are four different classes of lanthipeptides which are differentiated by their biosynthetic enzymes forming Lan/MeLan (Figure 4). In class I lanthipeptide pathways, a LanB dehydratase catalyzes the Ser/Thr dehydration [22] whereas a LanC cyclase catalyzes the subsequent thioether formation (Figure 4A) [23]. A prominent class I lanthipeptide is nisin which is one of the first known lanthipeptides and RiPP in general [13]. It is produced by *Lactococcus lactis* and used as a food preservative [9]. In class II, III and IV lanthipeptide pathways, the dehydration and cyclization reactions forming Lan/MeLan motifs are biosynthesized by bifunctional lanthionine synthetases. Class II lanthionine synthetases (LanM) consist of a unique dehydratase domain and a LanC-homologous cyclase domain (Figure 4B) [24]. Class II lanthipeptide gene clusters can yield one or multiple products via one LanM, e.g. lacticin 481 [24] or the prochlorosins [25], respectively. The one pathway-multiple product principle has been observed in several other RiPP pathways, especially from small genome (<3 Mb) organisms such as the microalgae *Prochlorococcus* which indicates that RiPP metabolism might be secondary metabolic adaptation to small genome space. In addition, a class II lanthipeptide pathway can produce two synergistic lanthipeptides from distinct precursor peptides via two precursor-specific LanM enzymes (Figure 4C). An example involves lacticin 3147 A1 and A2 from *Lactococcus lactis* that interact synergistically to promote their antimicrobial activity [26] (Figure 4C). Class III and IV lanthionine synthetases (LanKC and LanL, respectively) comprise a kinase-domain which phosphorylates Ser/Thr, a N-terminal lyase-domain which forms Dha/Dhb from phoshoSer/phoshpoThr via elimination and a C-terminal cyclase-domain which forms the Lan/MeLan bonds (Figure 4D,E) [27,28]. The LanKC and LanL cyclase-domains differ in absence and presence of zinc-binding residues in the active site, respectively [29]. The class III lanthionine synthetase can also catalyze formation of an additional carbon-carbon crosslink between the initial Dha/Dhb enolate with a second Dha to yield a so-called labionin modification (Figure 3B). Labionin was first observed in the labyrinthopeptins (Figure 4D) [30].

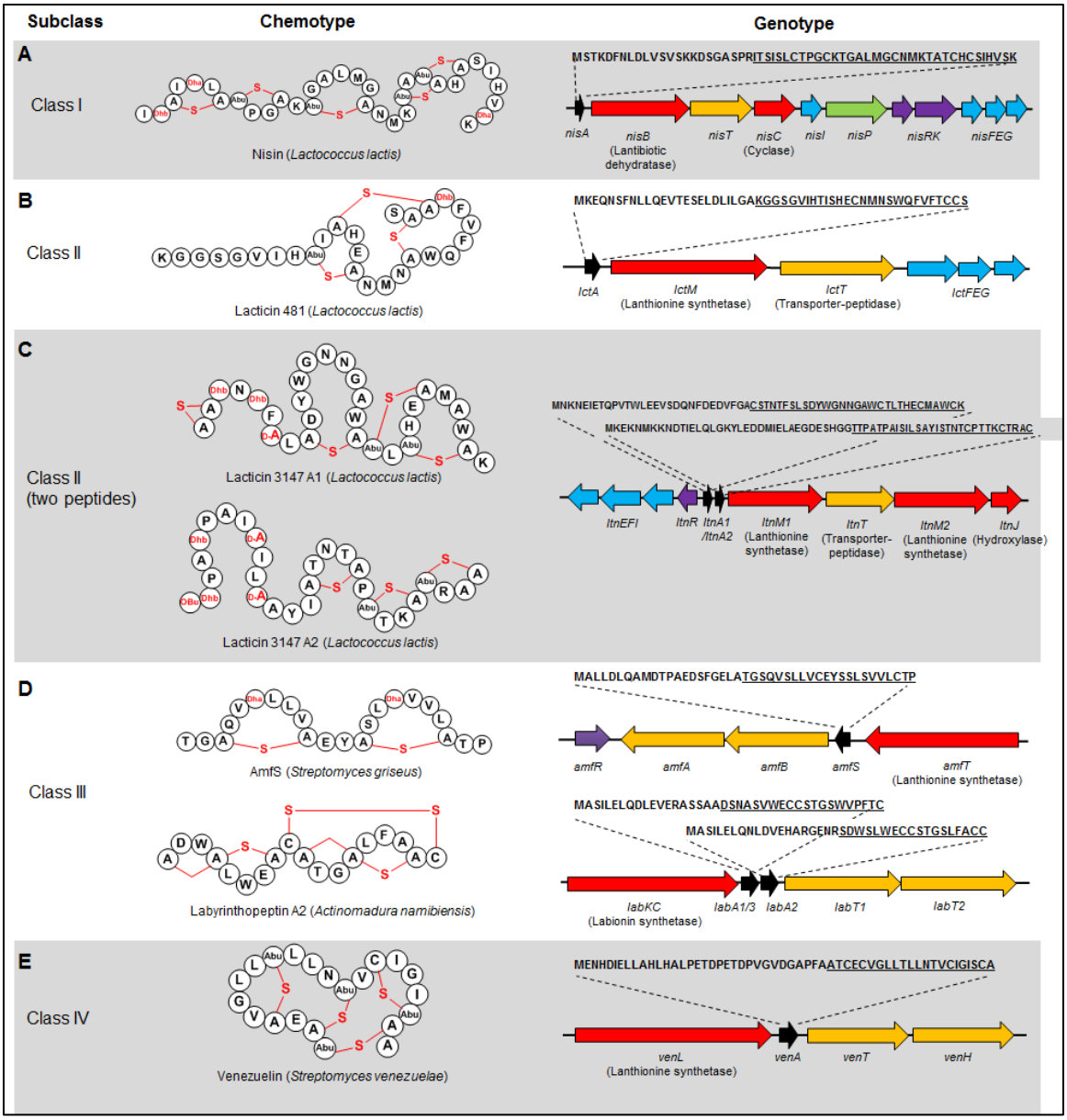


Figure 4: Chemo- and genotypes of lanthipeptide classes (I-IV) based on lanthionine biosynthetic machinery.

1.1.1.1.2 Lassopeptides

Lassopeptides are defined by an N-terminal macrolactam of the N-terminal amino group and the side chain carboxyl group of an Asp/Glu at the 8/9 position. Lassopeptides form lasso-like structures as the C-terminal chain is wrapped through the N-terminal macrocycle and held in position by sterically hindered residues (e.g. Trp, Phe) in the C-terminus [31]. Their lasso-

conformation causes their high stability against denaturing, proteases and temperature. Several lassopeptides have antimicrobial activities and act as enzyme inhibitors or receptor antagonists [32]. Three classes of lassopeptides have been defined based on the presence of disulfide bonds. Class I lassopeptides have two conserved disulfide bonds (Cys1→Cys13, Cys7→Cys19) and a conserved N-terminal macrolactam (Cys1→Asp9) [11,33]. The only known class I lassopeptide pathway is described in this work (Chapter 2) [19] (Figure 5A). Class II lassopeptides have no disulfide bonds and a Gly at position 1. The first characterized lassopeptide was class II lassopeptide microcin J25 from *Escherichia coli* AY25 [34]. Class III lassopeptides have one disulfide bond and a Gly at position 1 (Figure 5B). The only known member of this class is BI-32169 from *Streptomyces* sp. DSM 14996) [35]. No class III lassopeptide gene cluster has been characterized to date (Figure 5C).

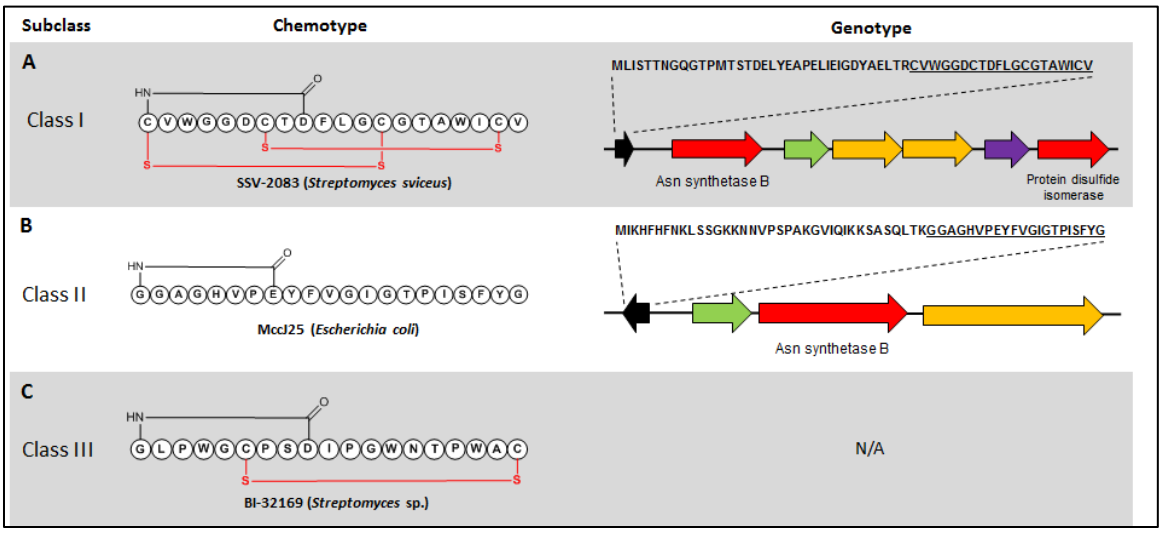


Figure 5: Lassopeptide classes based on presence and number of disulfide bonds.

Lassopeptide pathways share two common enzymes for formation of the N-terminal macrolactam: an ATP-dependent cysteine protease and a macrolactam synthetase. In the first step of the macrolactam biosynthesis, the ATP-dependent cysteine protease cleaves the precursor peptide at a cleavage site that is mainly determined by eight C-terminal residues including a threonine located two positions C-terminal of the cleavage site [37]. In the next step, the macrolactam synthetase (Asn synthetase B-homolog) is adenylating an Asp/Glu side chain at

the position 8/9. Subsequently, the macrolactam synthetase catalyzes the nucleophilic attack of the N-terminal amino group on the Asp/Glu-adenylate to yield the lassopeptide macrolactam modification [11,37] (Figure 6).

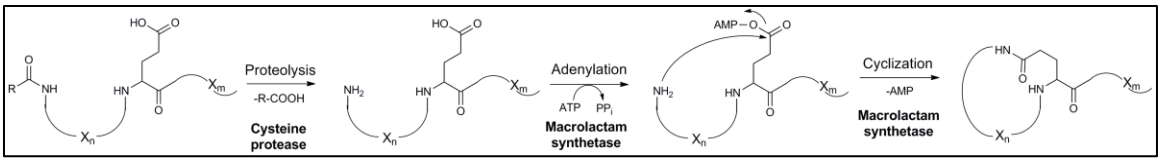


Figure 6: Proposed mechanism for macrolactam formation during lassopeptide biosynthesis

1.1.1.1.3 Linaridins

Linaridins were introduced as a RiPP class in 2011 by the characterization of the cypemycin gene cluster [17] (Figure 7A). Linaridins are defined by a C-terminal aminovinyl cysteine which is also known from lanthipeptides such as epidermin [11]. In both RiPP classes, the aminovinyl cysteine is generated by a Michael-type addition of an oxidative decarboxylated C-terminal cysteine onto a Dha [17,38]. In lanthipeptide pathways, the Dha is generated by dehydration of a Ser, whereas in linaridin pathways, the Dha is formed by dethiolation of a Cys (Figure 7B) [11,17].

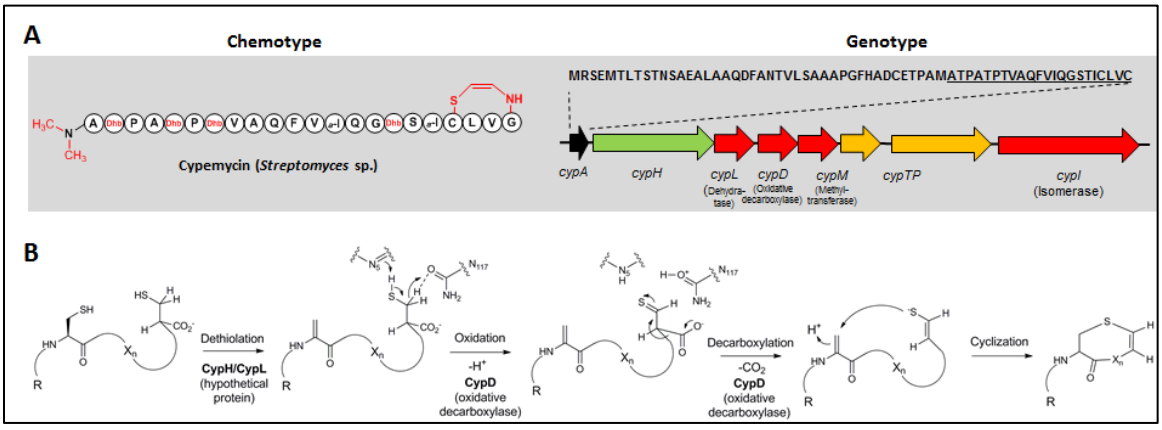


Figure 7: Linaridins. (A) Linaridin chemo- and genotype of founding member cypemycin. (B) Proposed biosynthesis of aminovinyl cysteine in linaridins.

1.1.1.2 Nonribosomal peptides (NRPs)

In nonribosomal peptides, the peptide chain is biosynthesized by large, multimodular nonribosomal peptide synthetases (NRPS, Figure 8) with individual functional domains. A

standard NRPS module consists minimally of two enzymatic domains that catalyze adenylation, condensation and thiolation reactions and a carrier domain. The adenylation domain (A) usually selects the amino acid substrate, activates it with ATP to the aminoacyladenylate and loads it onto the thiolation domain (T) via a thioester bond. The loaded T domain then serves as a noncatalytic carrier for substrates and intermediates during NRP assembly. The condensation domain (C) catalyzes the peptide bond formation of NRP substrate or intermediate loaded on an upstream T domain with the new amino acid loaded on the T domain of the module. Based on this domain interaction, the NRP chain gets extended in an assembly-line fashion from module to module to be finally released by a terminal thioesterase (TE) or reductase domain [39]. Two principles of NRP biosynthesis have enabled the prediction of NRP structures from NRPS genes:

(1) The amino acid substrate that gets incorporated in a specific module can be predicted from the nucleotide sequence of the A domain via the so-called 'Stachelhaus' or '10-letter substrate specificity' code [40,41]. This code consists of ten amino acid (aa) residues in the A domain sequence that form the substrate binding pocket. Many substrate specificity codes of A domains with known substrates have been characterized and implemented in NRPS substrate prediction tools [42-44].

(2) The NRP sequence can be predicted from the NRPS gene sequences based on the sequence of A domains in the corresponding NRPS assembly line. This principle is called colinearity rule.

The striking structural feature of NRPs is the large diversity of >300 non-proteinogenic amino acid building blocks [19,45] which extends the chemical space for this compound class beyond the twenty standard proteinogenic amino acids. Non-proteinogenic NRP building blocks can be formed before, during or after nonribosomal peptide assembly. Before NRP assembly, a non-proteinogenic amino acid can be formed by proteins encoded in the NRP gene cluster. Alternatively, amino acid residues may be derived as a primary metabolite and then loaded on the NRPS by a specific A domain [46-48]. During NRP assembly, a non-proteinogenic amino acid can be formed by incorporation of a proteinogenic or non-proteinogenic amino acid which is

further modified within the NRPS modules, e.g. N-methylation or heterocyclization [49]. After NRP assembly, a non-proteinogenic amino acid can be formed by modification enzymes from the gene cluster that transform incorporated amino acids in the released NRP intermediate [39].

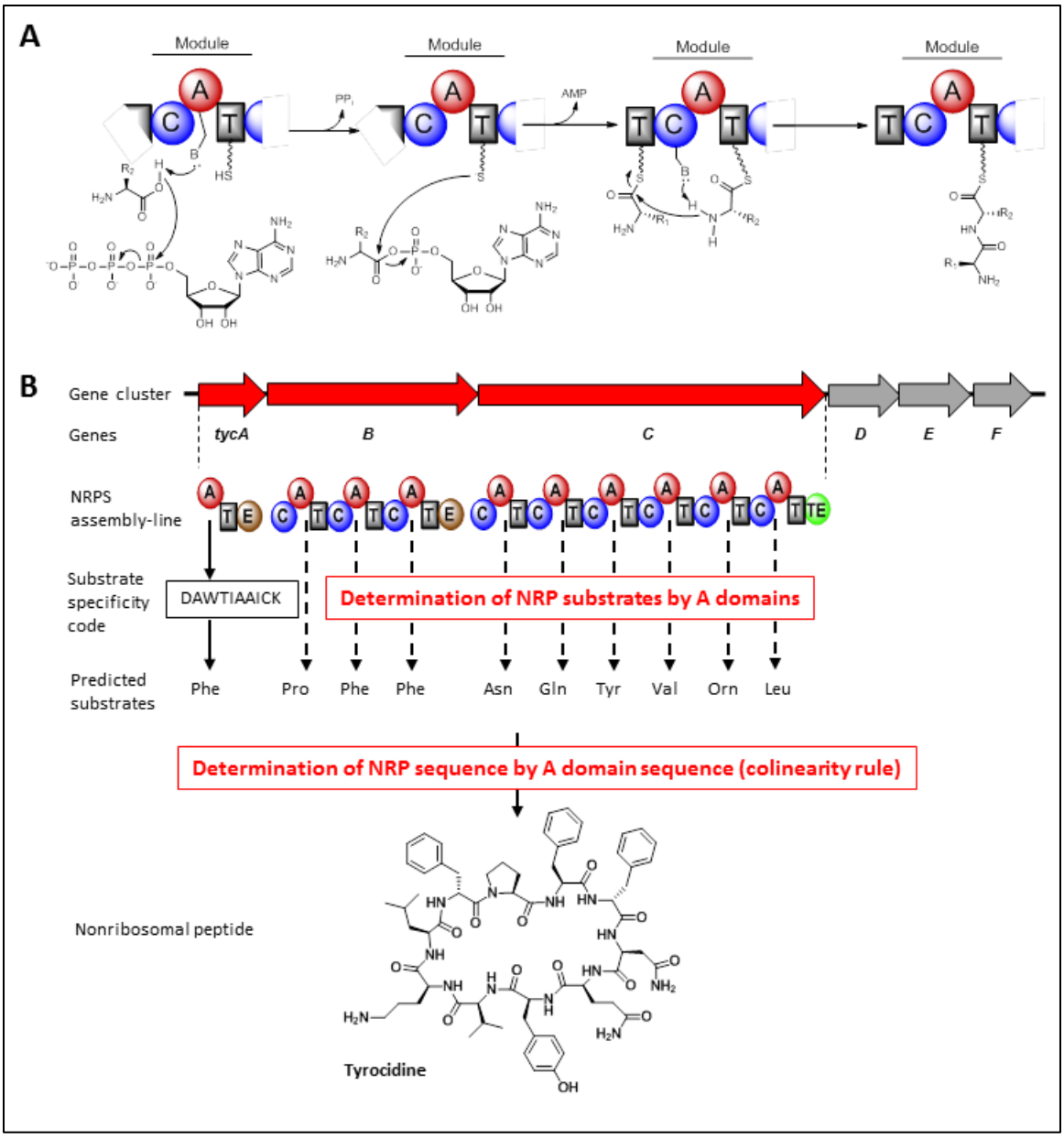


Figure 8: Nonribosomal peptide biosynthesis exemplified by the tyrocidine pathway [50]. NRPS genes (red) get translated into a multidomain NRPS assembly-line in which the adenylation domains (A) select substrates based on a substrate specificity code. The sequence of A domains in the NRPS assembly-line determines the nonribosomal peptide sequence.

An important class of nonribosomal peptides includes the lipopeptides [51]. Lipopeptides contain a N-terminal acyl chain that usually gets incorporated into the nonribosomal peptide via a starter condensation domain in the loading NRPS module. The acyl chains are fatty acid synthase (FAS)-derived and can be further modified before incorporation. Many lipopeptides are macrocyclic which can be differentiated into branched cyclic or head-to-tail cyclic peptides [12]. In branched cyclic peptides, the C-terminal carboxyl group of the NRP forms a macrolactone or a macrolactam bond with a nucleophile in the peptide side chains, e.g. the hydroxyl group of a threonine [8]. In head-to-tail cyclic lipopeptides, the C-terminus forms a bond with a nucleophile at the β-position of the acyl chain, e.g. a macrolactam via iturinic acid in iturin A from *Bacillus subtilis* [52] or a macrolactone via a β-hydroxy fatty acid in surfactin from *Bacillus subtilis* [53]. Macrocyclization is catalyzed by terminal thioesterase domains of NRPS assembly-lines [12].

A prominent example of microbial lipopeptides is daptomycin produced by *Streptomyces roseosporus* (Figure 9) [8,54]. Daptomycin, also known as Cubicin®, is an antibiotic used for treatment of certain Gram-positive bacterial infections. Daptomycin is a branched-cyclic, N-decanoyl tridecapeptide with a decapeptide macrolactone formed by the C-terminus and the Thr^4-side chain hydroxyl group. Several non-proteinogenic amino acids are incorporated in daptomycin which are *D*-Asn^2, Orn^6, *D*-Ala^8, *D*-Ser^{11}, (2*S*,3*R*)-methylglutamate ($MeGlu^{12}$) and kynurenine (Kyn^{13}). Orn^6, $MeGlu^{12}$ and Kyn^{13} are formed in pre-assembly-line steps and are then incorporated via the corresponding A domains. *D*-Asn^2, *D*-Ala^8, and *D*-Ser^{11} are formed from the respective proteinogenic precursors by epimerization domains within the NRPS assembly-line. FAS-derived decanoic acid is incorporated by acyl ligase DptE, acyl-carrier domain DptF and the starter C domain of module 1 of NRPS DptA. Herein, DptE activates decanoic acid by adenylation and loads the acyl chain onto DptF. Subsequently, DptA starter C1 domain catalyzes the condensation of decanoic acid with Trp^1 to yield the *N*-acyl terminus of daptomycin. Macrocyclization via nucleophilic attack of the Thr^4-hydroxyl group on the C-terminal carboxyl group is catalyzed by the DptD TE of the terminal module [54].

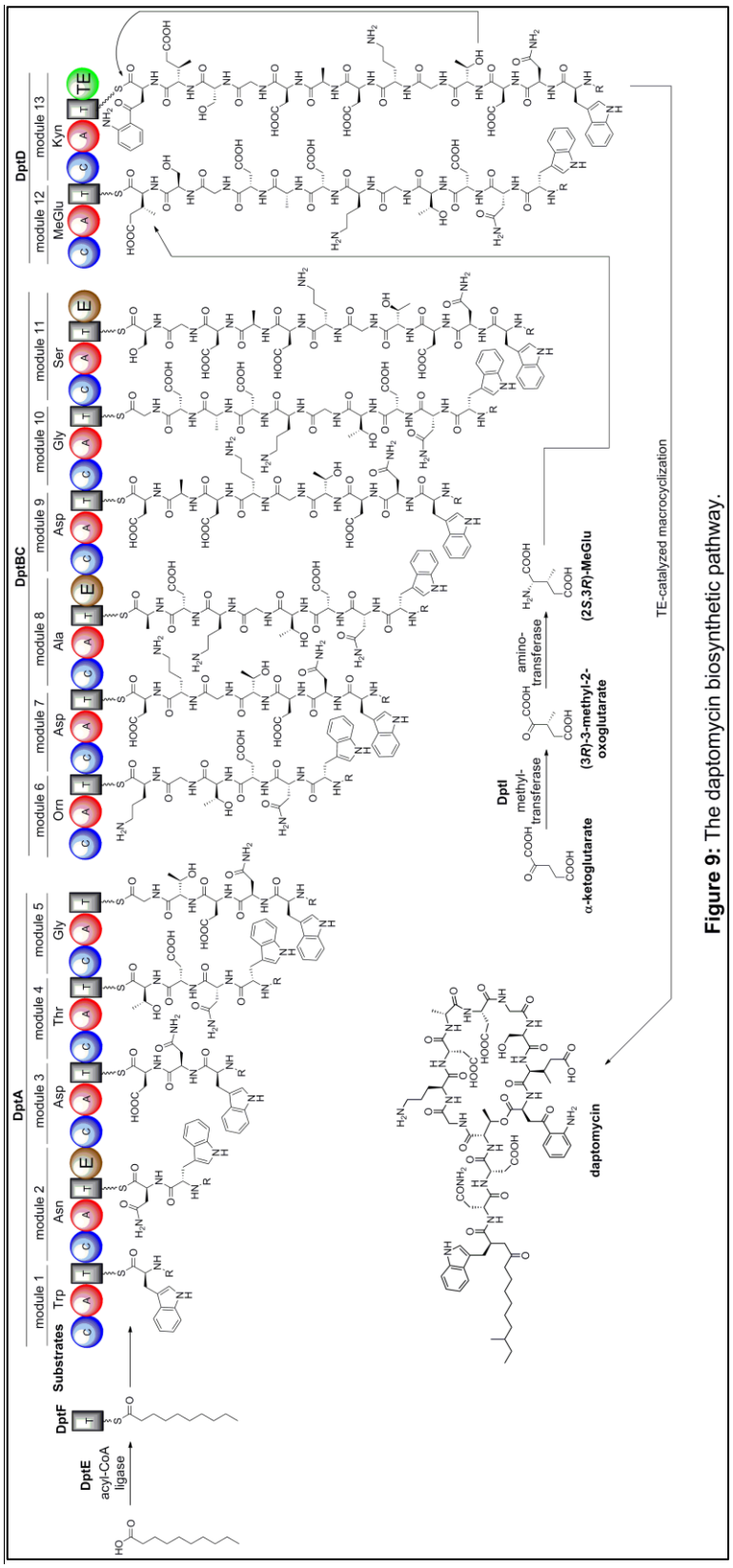


Figure 9: The daptomycin biosynthetic pathway.

1.1.2 Polyketides

Polyketides are a major biosynthetic class of structurally diverse natural products with many prominent examples of pharmaceuticals such as the macrolide antibiotic erythromycin A [55] and the anthracycline anticancer drug doxorubicin [56]. Polyketides are biosynthesized from acetate units (C_2) which are assembled to poly-β-keto carbon chains via Claisen condensations by polyketide synthases (PKS) [39]. Polyketide synthases are divided into type I, II and III classes. Type I PKSs are large, multimodular proteins with individual functional domains including acyl carrier domains (ACP) for polyketide intermediate and substrate attachment during biosynthesis. The PKS ACP domain is analogous to the NRPS T domain. Type I PKS can function non-iteratively or iteratively, i.e. enzymatic domains function once or more than once, respectively, during the assembly line biosynthesis. Type II PKSs consist of a complex of individual monofunctional proteins including an ACP for attachment of polyketide biosynthetic intermediates and substrates. Type III PKSs, or chalcone synthase-like PKSs, are homodimeric proteins that use coenzyme A (CoA) esters rather than ACPs for polyketide biosynthesis. Type I PKS are found in bacteria and fungi, type II PKS only in bacteria and type III PKS in bacteria, fungi and plants [39,59]. Products of all PKS types were isolated from marine invertebrates, however their biosynthetic origin is in question and might be based on bacterial symbionts.

The biosynthetic building blocks of polyketides are distinguished into starter and extender units based on the incorporation step in polyketide biosynthesis [57]. Common starter units for type I and II PKS are acetyl-CoA or propionyl-CoA. Common extender units for type I PKS are malonyl-CoA and methylmalonyl-CoA whereas the only extender unit for type II PKS is malonyl-CoA [57,58].

Polyketide starter and extender units are loaded by acyltransferases (AT) on acyl carrier proteins (ACP) as thioesters via a phosphopantetheine cofactor on the ACP active site serine. Chain elongation in polyketide biosynthesis occurs by Claisen condensation catalyzed by ketosynthases (KS) to yield a C2 extension of the polyketide chain per biosynthetic cycle [39,59] (Figure 10).

Three groups of polyketides of relevance to this dissertation are introduced briefly in their structure and biosynthesis in the following section: macrolides (type I PKS), anthracyclines (type II PKS) and non-ribosomal peptide-polyketide hybrids (mixed NRPS-PKS).

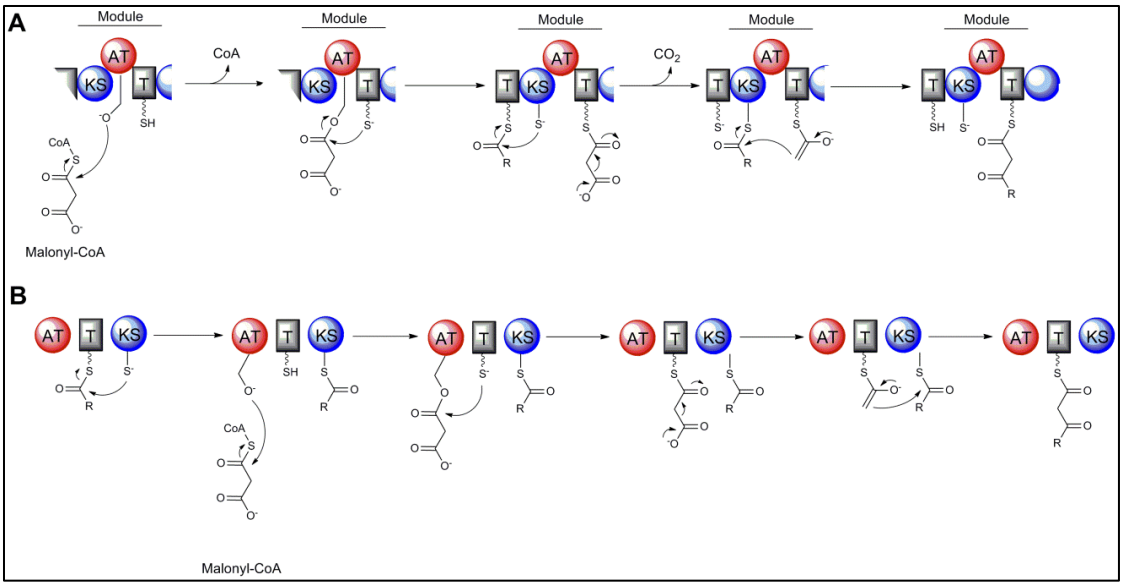


Figure 10: Polyketide biosynthesis. (A) Polyketide biosynthesis by type I PKS. (B) Polyketide biosynthesis by type II PKS.

1.1.2.1 Macrolide polyketides

Macrolides are a large group of polyketide natural products that are characterized by polyketide macrolactams and polyketide macrolactones. They are biosynthesized by type I PKS. Type I PKSs are large multimodular protein assembly-lines in which one module of functional domains is responsible for the incorporation of one C2 unit per extension cycle. A minimal type I PKS module consists of an ACP domain, a KS domain and an AT domain. The AT domain determines and loads the substrate on the ACP domain. The KS domain catalyzes the Claisen condensation of the polyketide intermediate from a downstream ACP with the new extender unit to yield a C2-extended polyketide chain. An extender unit can be reduced after chain elongation – in analogy to fatty acid biosynthesis – by a ketoreductase (KR) domain, by a dehydratase (DH) domain and by an enoylreductase (ER) domain. In macrolide biosynthesis, the reduction of each extender unit is determined by the set of reducing domains in the following module (Figure 11)

[39,59]. In addition, extender units can also be methylated by methyltransferase (MT) domains in type I PKS. The starter unit is selected by the loading module which consists usually of a starter AT domain and an ACP. The macrocyclization is catalyzed by a thioesterase in the terminal module [12]. The off-loaded macrolide can be further modified by e.g. hydroxylation, methylation or glycosylation by post-PKS modification enzymes to yield the final macrolide polyketide [39]. As with NRPS-derived peptide products, a putative polyketide structure can be predicted from its associated modular type I PKS encoding genes. Herein, substrates can be predicted from acyltransferase domain sequences, and the reducing domains in each module enable the prediction of the reduction state of the β-keto chain at each β-carbon. The colinearity rule, i.e. the sequence of modules in the assembly line determines the sequence of building blocks, also applies for type I PKS, although numerous deviations exist [59].

The biosynthetic pathway of the macrolide drug erythromycin A from *Saccharopolyspora erythraea* is shown as an example for type I PKS biosynthesis in Figure 11 [60]. Erythromycin A contains a 14-membered macrolactone that is biosynthesized from a propionate starter unit and six methylmalonate extender units. The PKS consists of seven modules on three PKS proteins and forms the intermediate 6-deoxyerythronolide B via thioesterase-mediated macrocyclization. 6-deoxyerythronolide B is further modified by two α-hydroxylations and glycosylations to ultimately yield erythromycin A.

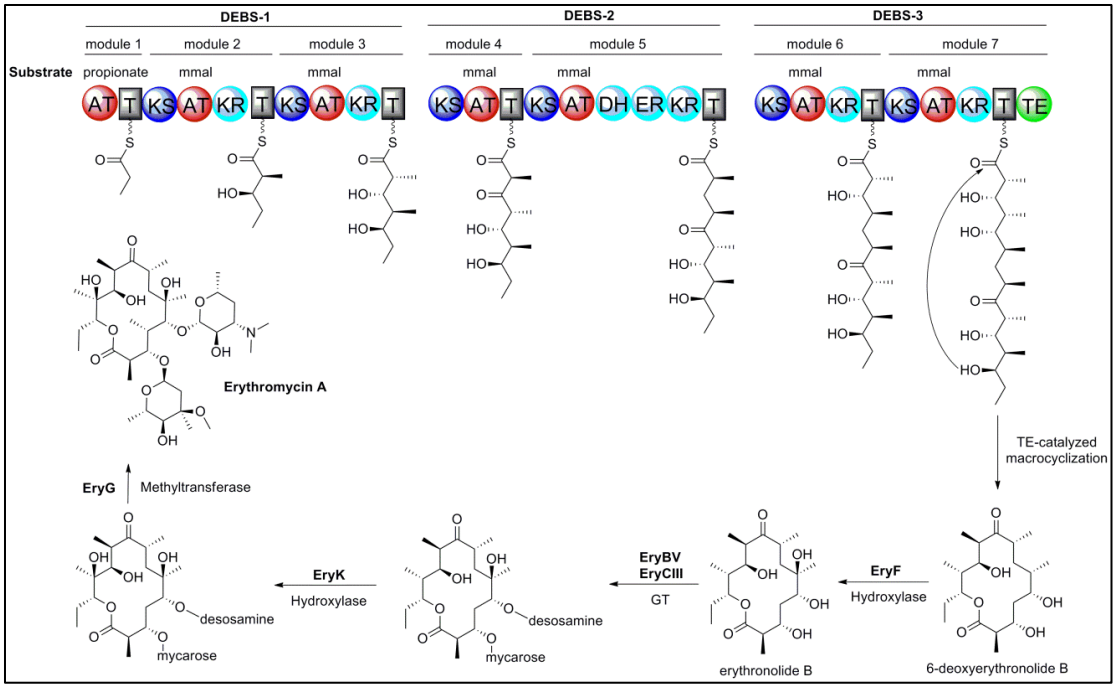


Figure 11: Type I PKS biosynthesis exemplified by the macrolide erythromycin A pathway.

1.1.2.2 Anthracycline polyketides

Anthracyclines are pharmaceutically important aromatic polyketides such as aclacinomycin A, doxorubicin and daunomycin [56]. Anthracyclines are characterized by a tetracyclic anthracyclinone core structure which is usually glycosylated. The anthracyclinone core is biosynthesized by type II PKS. A minimal type II PKS consists of the KS α/β heterodimer and an ACP and it uses the fatty acid synthase malonyl-CoA AT (MCAT, FabD). These domains interact iteratively in chain elongation to yield a poly-β-keto chain from several malonyl-CoA molecules. In addition, a starter PKS consisting of a starter KS, a starter AT and a starter ACP can often enable the incorporation of a different starter unit than acetyl-CoA such as propionyl-CoA. In addition, ketoreductases (KR), cyclases (CYC) and aromatases (ARO) are responsible for the formation of the cyclic structure from the reactive poly-β-keto intermediate [56,59].

An exemplified type II PKS biosynthesis of the anti-cancer drug aclacinomycin A from *Streptomyces galilaeus* is shown in Figure 12 [56,61,62]. Aclacinomycin A belongs to the anthracycline aromatic polyketides. It is formed from one propionate starter unit which is incorporated into the polyketide by a separate starter AT and starter KS. After an iterative

extension of the polyketide chain with 9 malonates by the type II PKS, a ketoreductase, an aromatase and a cyclase catalyze the formation of an aromatic tricyclic, ACP-released intermediate which is further modified by hydroxylations, methylation, cyclization and triglycosylation to yield aclacinomycin A [56].

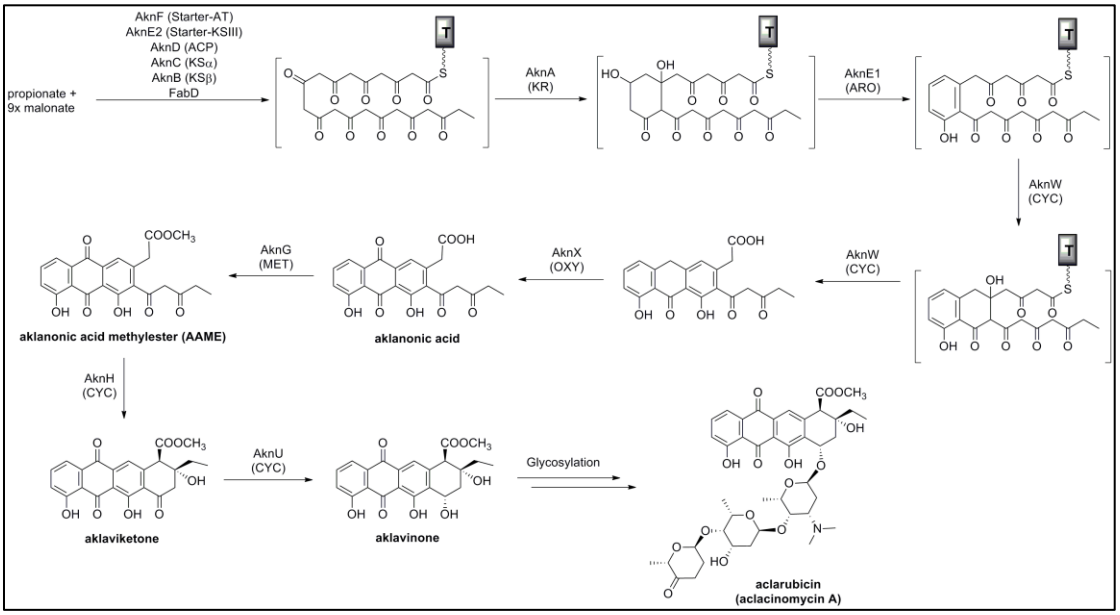


Figure 12: Type II PKS biosynthesis exemplified by the anthracycline pathway.

1.1.2.3 Hybrid nonribosomal peptide-polyketides

Due to their modular organization and shared covalent attachment of intermediates to carrier domains via thioesters, modules of NRPS and modules of type I PKS can be mixed in large NRPS-PKS proteins to yield nonribosomal peptide-polyketide hybrids.

An exemplified mixed NRPS-PKS biosynthesis of the anti-cancer agent epothilone from *Sorangium cellulosum* is shown in Figure 13 [63]. Epothilone is a 16-membered macrolactone with a thiazole group. The epothilone pathway consists of nine type I PKS modules and one NRPS module. The NRPS A domain in module 2 loads cysteine which is further cyclized to thiazoline by a heterocyclization (Cy) domain and oxidized to thiazole by an oxidoreductase (Ox) domain. The heterocyclizing condensation domain in module 2 and the KS domain in module 3 constitute a switch from PKS to NRPS biosynthesis and vice versa, respectively.

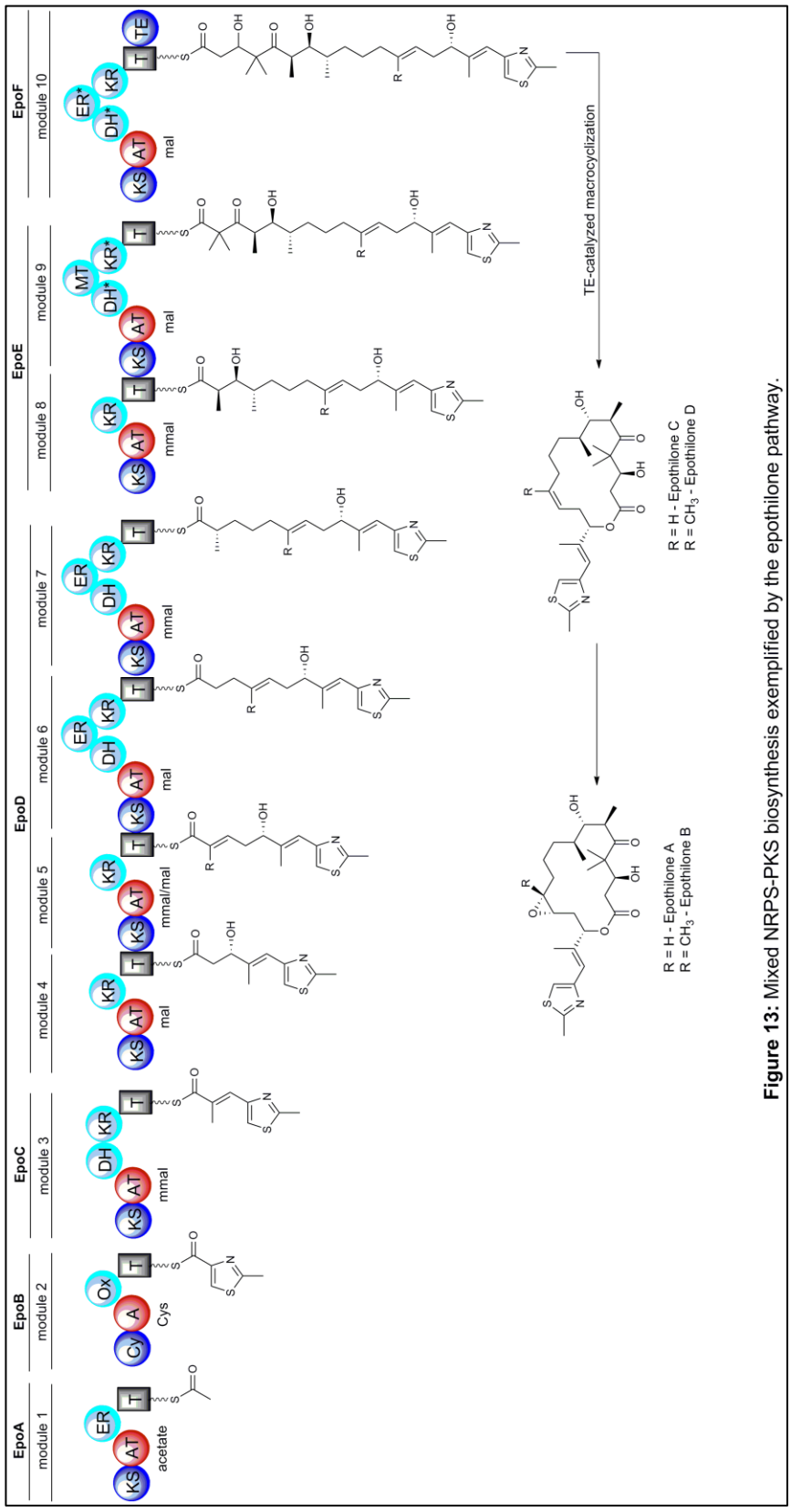


Figure 13: Mixed NRPS-PKS biosynthesis exemplified by the epothilone pathway.

1.1.3 Glycosylated natural products

Glycosylated natural products (GNPs) produced by microbes comprise many compounds with therapeutic and agrochemical applications such as the antibiotic erythromycin [55] and the insecticide avermectin [64], respectively. A glycosylated natural product consists of an aglycone and one or multiple glycosyl units (Figure 14) [65] which often directly mediate the bioactivity of the compound [66]. In microbial genomes, the genes for biosynthesis and attachment of these glycosyl groups are usually clustered with the biosynthetic genes of the aglycone (Figure 15).

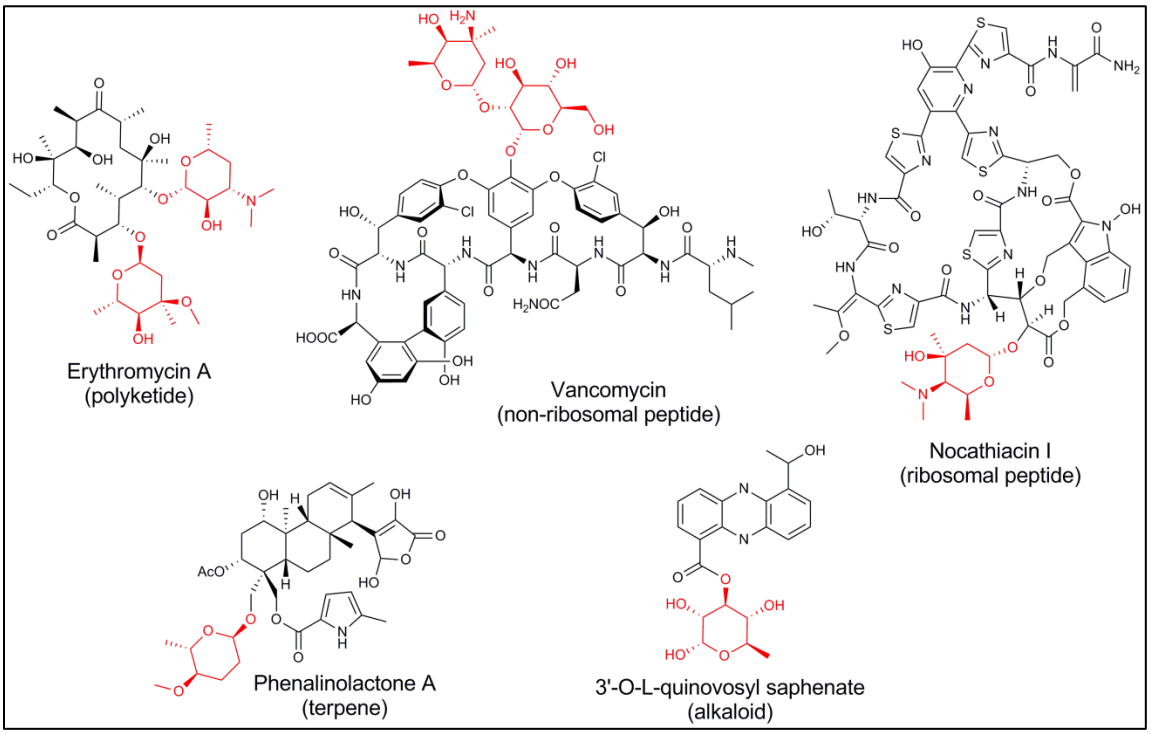


Figure 14: Structure of glycosylated natural products. Selected GNP structures exemplify the diverse biosynthetic origin of the aglycone (black) and the diversity in glycosyl groups (red).

Glycosylated natural products are a structurally very diverse class of natural products in terms of the aglycone, i.e. the non-sugar protion of the molecule and the glycosyl groups. Aglycone diversity is based on the fact that GNPs are found in almost all major biosynthetic classes of natural products (Figure 14), e.g. non-ribosomal [67] and ribosomal peptides [68], polyketides [69], terpenes [70] and alkaloids [71]. Glycosylation diversity arises through sugar monomers and sugar attachment. There are over 100 different sugars found in microbial GNPs

where the majority are deoxysugars [65]. These sugar monomers can be attached to the aglycone or to each other by *C*-, *N*-, *N*-*O*-, *O*- and *S*-glycosidic bonds, where *O*-glycosidic bonds are the most common. Furthermore, glycosidic bonds occur from various sugar ring positions [65].

The genes involved in deoxysugar glycosylation of an aglycone can be distinguished into common glycosylation genes, which are found in all deoxysugar pathways, and specific glycosylation genes, which catalyze sugar-specific modifications to yield different sugar monomers. The common genes encode a nucleotidylyltransferase (NT), which activates glucose-1-phosphate as TDP-glucose, a 4,6-dehydratase (4,6-DH), which subsequently forms the common deoxysugar intermediate TDP-4-keto-6-deoxy-α-glucose, and a glycosyltransferase (GT), which attaches the final sugar monomer to the aglycone or a glycosyl group (Figure 15). Specific glycosylation genes are dehydratases, deoxygenases, dehydrogenases, methyltransferases, aminotransferases and many more that modify the common deoxysugar intermediate to yield the large diversity of sugar monomers in microbial GNPs [72,73]. Phylogenetically, the highest GNP sugar diversity is found in actinobacteria [74]. Among these bacteria, the families of *Micromonosporaceae*, *Pseudonocardiaceae*, *Streptomycetaceae* and *Thermomonosporaceae* have the highest genetic potential to produce GNPs (Table 1, Chapter 4).

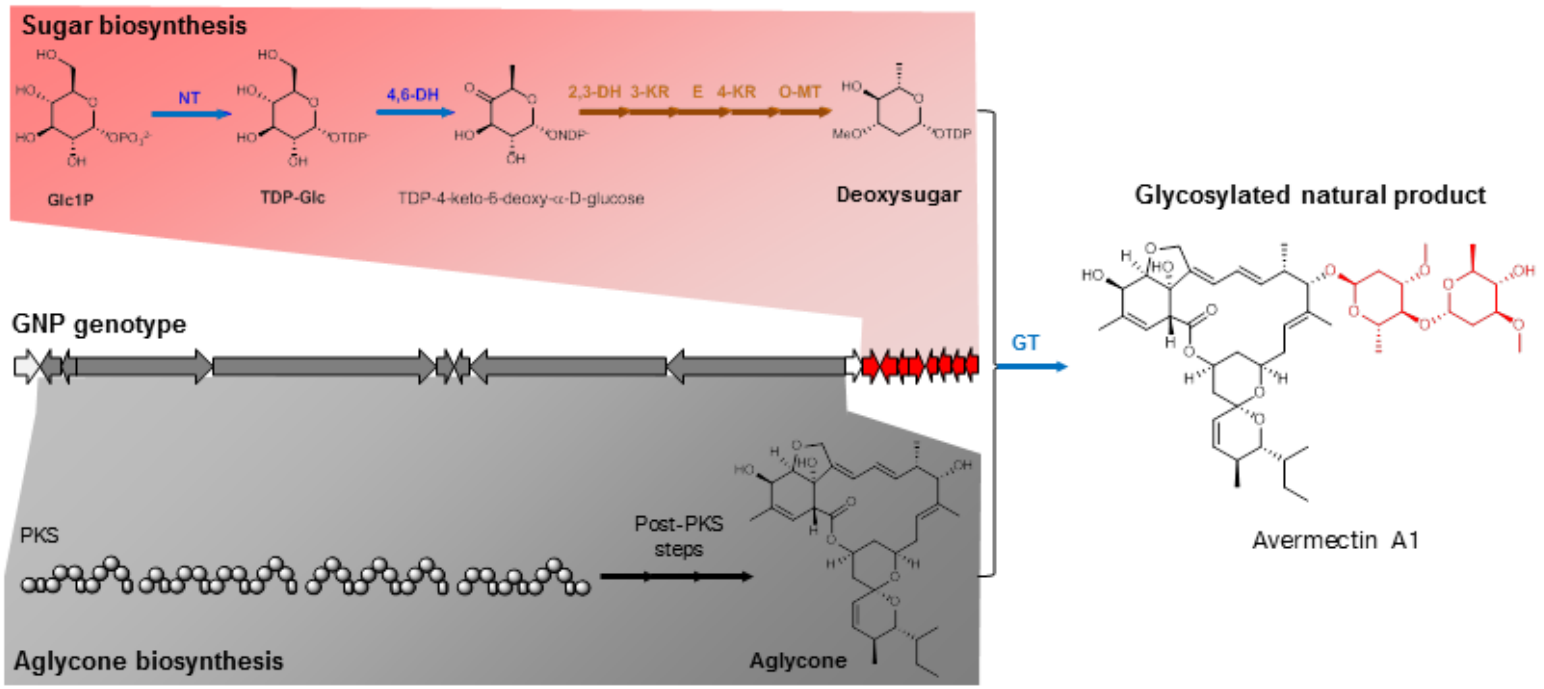


Figure 15: Biosynthesis of glycosylated natural products. The simplified genetic organization of the avermectin biosynthetic pathway shows that it can be differentiated into aglycone biosynthetic genes (grey) and sugar biosynthetic genes (red). In a deoxysugar pathway, there are common biosynthetic genes (NT – nucleotidylyltransferase, 4,6-DH – 4,6-dehydratase, GT – glycosyltransferase, blue arrows) and specific biosynthetic genes (here: 2,3-DH, 2,3-dehydratase, 3-KR – 3-ketoreductase, 5-E – 5-epimerase, 4-KR – 4-ketoreductase, O3-MT – O3-methyltransferase, brown arrows). Sugar groups are indicated in red. Abbreviations: PKS – polyketide syhthetase, Glc1P – D-glucose-1-phosphate, TDP-Glc, 1-TDP-D-glucose.

1.2 Mass spectrometric analysis of natural products

Mass spectrometry (MS) is an important technique in the analysis of microbial natural products because of its high sensitivity, its easy implementation into automated processes such as metabolomic platforms, and its capability for *de novo* structure elucidation and metabolite identification by tandem MS. In this section, the relevant concepts of mass spectrometers and tandem mass spectrometry experiments applied in this work are introduced.

A mass spectrometer is an analytical tool to determine the mass of a given ionized analyte in a sample through measurement of its mass-to-charge (m/z) ratio displayed in a mass spectrum. A mass spectrometer consists of three general parts: an ion source, a mass analyzer, and a detector [75].

In the ion source, analytes are ionized into the gas-phase for further mass spectrometric analysis. Two common ionization techniques in the MS analysis of natural products in liquid and solid samples are electrospray ionization (ESI) [76,77] and matrix-assisted laser desorption ionization (MALDI) [78], respectively. ESI and MALDI are so-called “soft” ionization techniques which are characterized by a low in-source fragmentation of analytes during their conversion into the gas-phase and, thus, are preferred for MS analysis of biomolecules.

Electrospray ionization is an atmospheric pressure ionization technique in the presence of a strong electrostatic field and a heated, inert drying gas such as nitrogen (Figure 16). It is applied for liquid samples and, thus, enables a combination of liquid chromatography with mass spectrometry (LC-MS). ESI consists of four steps: ion formation, nebulization, desolvation, and ion evaporation. Ion formation in ESI can occur before nebulization in solution depending on the analyte and ESI solution or it can occur after nebulization at the charged surface of spray droplets for analytes that don’t ionize in solution. During nebulization, the ESI solution with the analyte is sprayed through a nebulizer needle into the ESI source. A strong electrostatic field is applied between the needle and the capillary of the MS inlet causing a Taylor cone at the nebulizer tip and charged droplet formation as analyte ions of one polarity migrate to the droplet surface. Subsequently, a counter-current drying gas evaporates the droplet solvent. The droplet surface

decreases and forces like surface charges closer together until Coulomb repulsion surpasses surface tension ('Rayleigh limit') and the droplet explodes ('Coulomb fission') [79]. This produces smaller charged droplets which undergo the same desolvation process again. At a certain charge density, ion evaporation occurs, i.e. analyte ions are directly ejected from the droplet surface towards the MS inlet capillary [79]. This ion evaporation mechanism is most likely the major mechanism for electrospray ionization of natural product-like analytes [80]. Standard ESI solutions contain an organic solvent with low heat of vaporization and low surface tension, e.g. methanol or acetonitrile. In addition, they usually contain an acid or a base for increased analyte ionization in solution. A characteristic of ESI is the occurrence of multiple charging of analytes, which enables the detection of larger molecules, such as peptides and proteins, in a small mass range. ESI is applied for MS analysis of liquid samples and, thus, enables a combination of liquid chromatography with mass spectrometry (LC-MS). It is a common technique for proteomics and protein MS analysis [81].

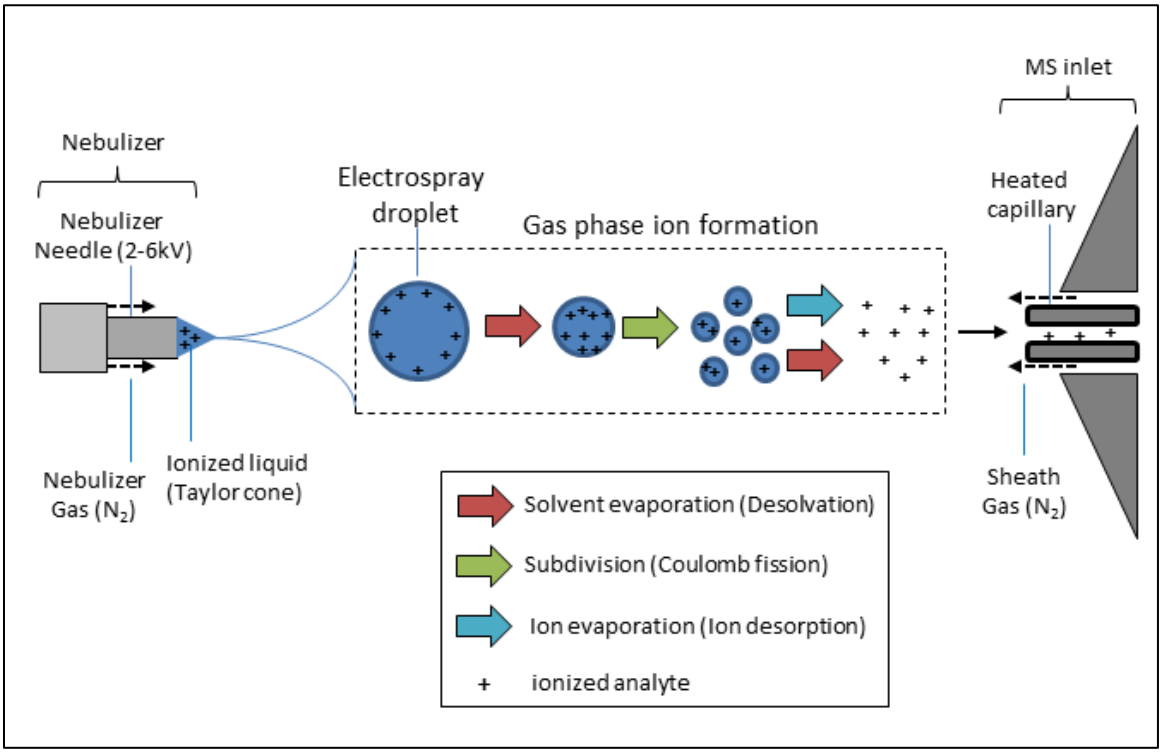


Figure 16: The concept of electrospray ionization in mass spectrometry.

Matrix assisted laser desorption ionization is commonly a vacuum ionization technique which relies on the interaction of an analyte with MALDI matrix. For MALDI-MS analysis, an analyte is co-crystallized with excess of MALDI matrix on a target plate [82]. The matrix is a UV-absorbing, weak organic acid such as 3,5-dimethoxy-4-hydroxycinnamic acid (sinapic acid), α-cyano-4-hydroxycinnamic acid or 2,5-dihydroxybenzoic acid (DHB) [82]. The matrix has two purposes during ionization: transition of the analyte to the gas-phase (desorption) and ionization (Figure 17). During MALDI, laser radiation is absorbed by the matrix molecules causing the vaporization of the matrix-analyte mixture. The matrix also acts as a proton donor or acceptor and, thus, ionizes the analyte for MS analysis in either positive or negative ion mode. The ionized analyte molecules – and the matrix ions – are subsequently accelerated towards the mass analyzer via an electrode. MALDI causes less multiple charging than ESI and results in background MS signals of matrix molecules in the lower mass range of the MALDI-mass spectrum [82]. A MALDI application relevant to this work is MALDI-imaging mass spectrometry [83-85] (see 1.2.1.3).

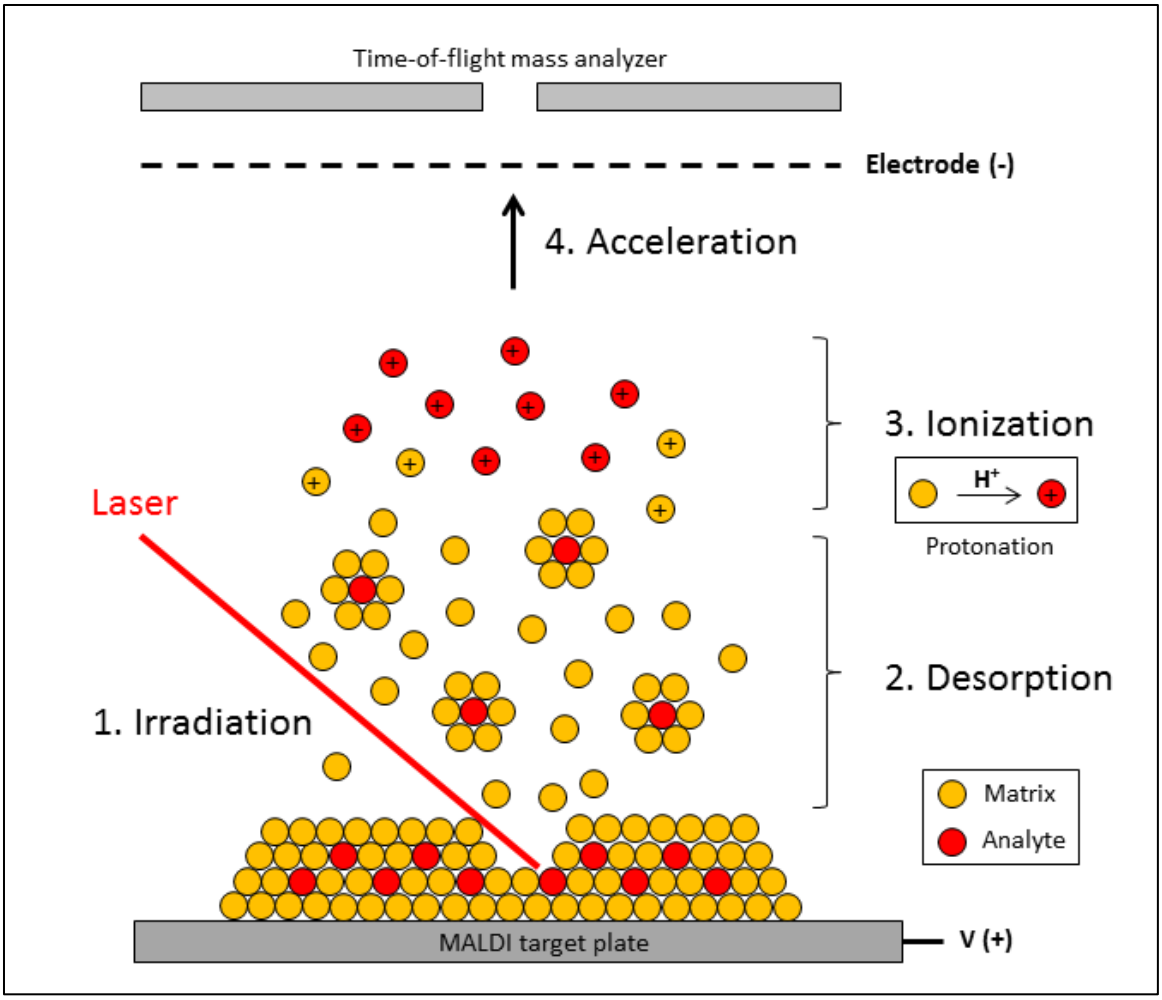


Figure 17: Matrix-assisted laser desorption ionization in mass spectrometry.

Mass analyzers measure gas phase ions generated in the ion source in terms of their mass-to-charge ratio (m/z). The addition of charge allows the analyte to be affected by electric fields during motion in the mass analyzer. Generally, the motion of a charged ion through electric and magnetic fields in vacuum is governed by two laws:

(a) $F = q(E + v \times B)$ (Lorentz force law, F – force on charged particle, q – charge, E – electric field strength, v – velocity, B – magnetic field strength)

(b) $F = ma$ (Newton's 2nd law of motion, F – force on charged particle, m – mass, a – acceleration),

which give the differential equation:

$(m/q)a = E + v \times B$

that determines the motion of a charged ion in space and time, e.g. in a mass analyzer, in terms of m/z (z = q/e, z – charge state, e – electron charge). Thus, the measurement of an ion's mass can be accomplished by ion separation in space or in time in a mass analyzer. A mass analyzer is defined in its performance by several characteristics: mass accuracy, mass resolution, mass range, and scan speed. Mass accuracy is the ability with which the analyzer can accurately provide m/z information. It is defined as the ratio of the measurement error and the calculated mass in ppm:

$$\Delta m[ppm] = \frac{m(calc) - m(obs)}{m(calc)}$$

Resolution is the ability of a mass analyzer to distinguish between ions of different mass-to-charge ratios. It is defined as the ratio of a mass signal and the full width at half maximum (M/ΔM). A higher mass resolution increases to some extend the mass accuracy of a mass analyzer. The mass range is the m/z range of the mass analyzer. The scan rate is the rate at which the analyzer scans over a particular mass range to acquire a mass spectrum [75].

Four types of mass analyzers were used in this work and are introduced as follows. The quadrupole mass analyzer consists of four parallel cyclindrical rods that create a radiofrequency (RF) quadrupole field by applying RF voltages to opposing rod spirals [75]. The quadrupole mass analyzer (Q) acts as a mass filter as only ions in a certain mass range pass through the quadrupole at certain RF rod voltages while the rest of the ions collide with the rods. This enables scanning a mass range by continuously varying the applied RF voltages. Quadrupole mass analyzers are often coupled with ESI as they tolerate high pressures. They have the advantages of a relatively large mass range (up to 4000 m/z) and are low-cost instruments.

The time-of-flight mass analyzer (TOF) is based on the measurement of the flight time of a given analyte ion through the mass analyzer after an acceleration event [75]. As all ions receive the same kinetic energy entering the mass analyzer, the flight time of an ion only depends on its mass-to-charge ratio. Thus, ions with a smaller mass will reach the detector or end of the mass analyzer faster. TOF mass analyzers often have a reflectron, which is an electrostatic mirror at the end of the flight tube [75]. Its purpose is to redirect ions back to the detector at the entrance

end of the flight tube. This increases flight time and decreases the temporal distribution of ions in a TOF mass analyzer and, thus, leads to higher mass resolution. Time-of-flight mass analyzers are often coupled to MALDI sources (MALDI-TOF MS) [75].

The linear ion trap mass analyzer (LTQ) consists of a quadrupole mass analyzer with two endcap electrodes generating a two-dimensional radiofrequency field that can trap ions inside the quadrupole field [75]. The endcap electrode RF field is generally scanned to excite and eject ions through holes in one endcap electrode to the detector over a certain m/z-range. One advantage of the ion trap is that ions a specific m/z can be isolated in order to do tandem MS analysis. Linear ion traps are often coupled with ESI sources. A disadvantage of ion traps are “the one third rule” which states that the maximum ratio between precursor m/z and the lowest trapped fragment ion is 1/3.

Fourier transform-ion cyclotron resonance mass analyzers (FT-ICR MS or FTMS) are based on the measurement of the cyclotron movement of ions in a magnetic field. In a FTMS cell, ions orbit in a strong magnetic field (e.g. 7T) and ultra-high vacuum (~10^{-10} Torr) [86]. A pulsed radio-frequency signal excites the ions and forces them to a larger orbit to produce a detectable image current. The Fourier-transformed current yields the component frequencies of each ion in the FTMS cell corresponding to their m/z values. FTMS instruments have very high mass accuracy and mass resolution which enables MS analysis of large proteins such as NRPS and PKS [81]. The “one third rule” also applies for FT-ICR mass analyzers.

The last part of a mass spectrometer is a detector which transforms separated ions reaching the detector into a current signal that can be processed and displayed as a mass spectrum [75]. The most common MS detector is the electron multiplier which consists of a row of dynodes of increasing potential. When an ion hits the first dynode of the electron multiplier detector, a number of electrons corresponding to the kinetic energy of the ion are emitted from the dynode towards the dynode cascade that amplifies the number of emitted electrons per dynode to ultimately generate a 10^6-enhanced signal. Another detector type is the charge detector which is used in FT-ICR mass spectrometers. An ion is detected by its movement along

two parallel charge detector plates which generates a corresponding image current for further FTMS data processing.

1.2.1 Tandem mass spectrometry

Tandem mass spectrometry (MS^n or MS/MS) implies a sequence of mass analysis events, which are usually separated in space or time by a specific gas-phase fragmentation event of analyte ions [75,87]. A common tandem mass spectrometric experiment is collision-induced dissociation (CID). For CID, precursor ions (also called 'parent' ions) are selected and accelerated to collide with an inert collision gas such as nitrogen or helium. Upon collision of an ion and a collision gas molecule, kinetic energy is converted into internal energy of the ion which results in bond breakage of the precursor ion into molecular fragments including charged and neutral fragments. Tandem MS fragment ions (also called 'daughter' ions) can then be mass analyzed and detected as a tandem MS spectrum of the precursor ion (Figure 18). Tandem MS analysis of a precursor ion and its fragment ions can be realized in space, e.g. in a Q-TOF MS, or in time, e.g. in an ion trap. In a Q-TOF, the precursor ion analysis occurs in the quadrupole mass analyzer, the fragmentation occurs in a collision cell quadrupole and the fragment ion analysis occurs in a TOF mass analyzer (Figure 18). In an ion trap, the precursor ion gets separated from other ions first, a collision gas gets injected into the trap and fragment ions are then analyzed in the same trap. The MS/MS fragments of a specified precursor ion can yield *de novo* structural information of an analyte or can identify the analyte by comparison of its tandem MS spectrum to a tandem mass spectral library [88].

CID fragmentation of natural products occurs preferentially at polar bonds, i.e. carbon-heteroatom bonds such as peptide bonds, glycosidic bonds or ester bonds [89]. This general MS/MS fragmentation behavior enables the prediction of diagnostic tandem MS fragments of e.g. a peptide or a glycosylated natural product. Due to the relevance to this work, CID fragmentation of peptides and glycosides is discussed in the following sections.

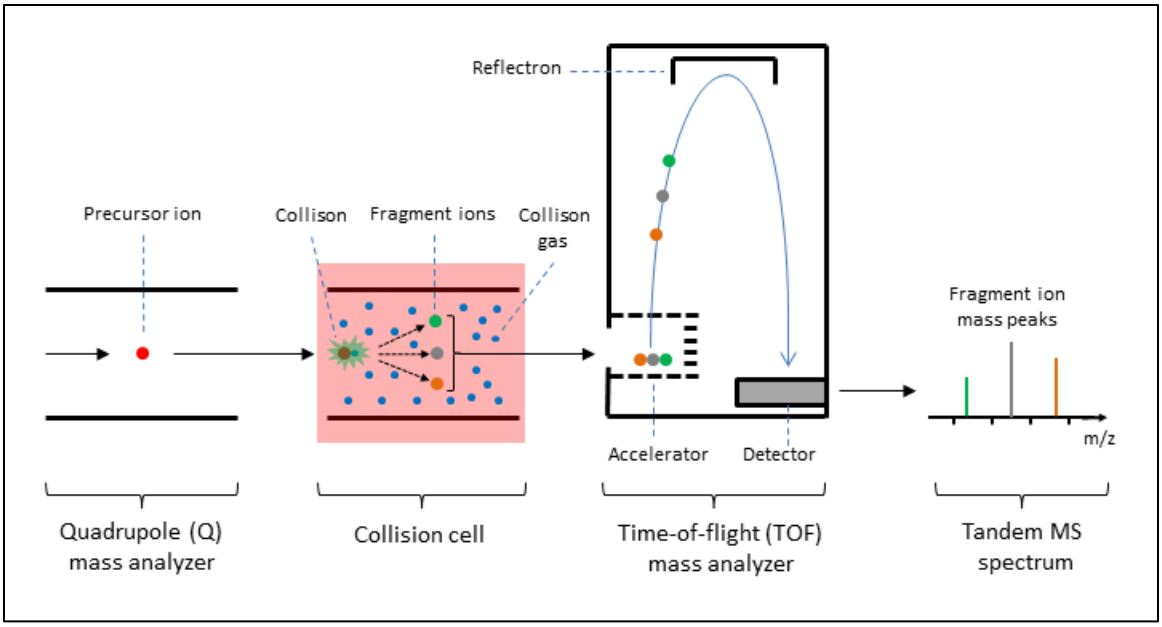


Figure 18: Tandem mass spectrometric analysis via collision induced dissociation in a Q-TOF MS.

1.2.1.1 Tandem mass spectrometric analysis of peptides

Tandem mass spectrometry is a common method to gain sequence and structural information of peptides [90]. Peptide fragmentation in CID experiments is defined in nomenclature of observed fragments as shown in Figure 19A [91] and can be explained mechanistically by the 'mobile proton' model [92]. In singly charged peptides, a proton can be sequestered by the N-terminal amino group, the amide bond oxygen or nitrogen and amino acid side chains. During CID, these peptide ions fragment preferentially in the sequence b-y ion pathway (Figure 19B). It starts with the mobilization of the C-terminal proton of the peptide precursor to an amide bond nitrogen. Next, the oxygen of the N-terminally adjacent amide bond attacks the carbonyl carbon of the protonated amide bond to yield a protonated oxazolone derivative and a C-terminal amine fragment [93]. The extra proton can end up on the b- or the y-ion depending on their proton affinities [92]. This predictable peptide fragmentation enabled the development of proteomic approaches for protein identification by comparing observed peptide MS/MS data to predicted protein database spectra [94,95]. Based on the CID b-y fragmentation pathway, peptide

sequences can also be determined *de novo*, i.e. without database comparison, from peptide MS/MS spectra by identification of amino acid-specific mass shifts in the b- and y-ion series [96].

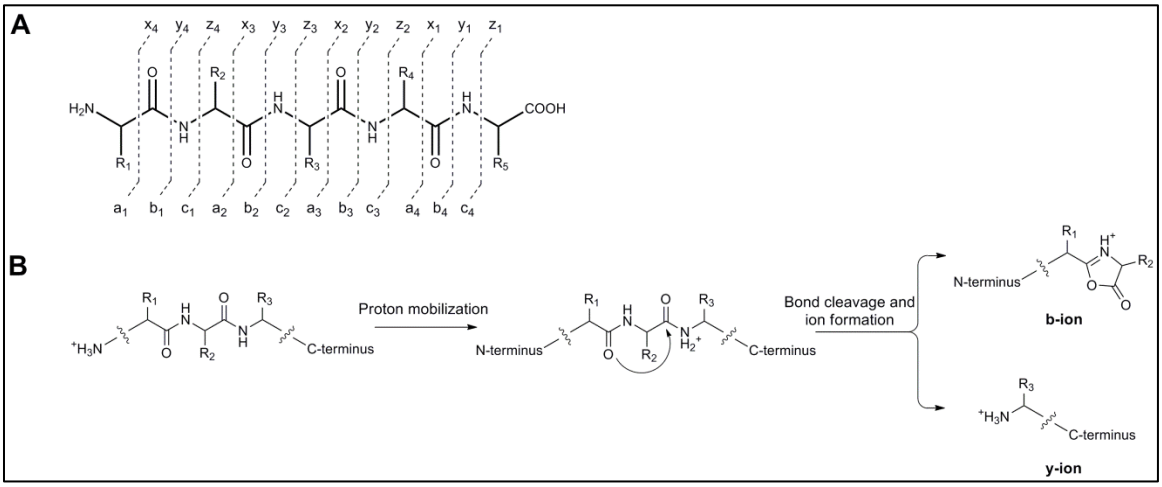


Figure 19: Tandem mass spectrometric analysis of peptides. A – Peptide fragmentation nomenclature. B – The b-y ion fragmentation pathway upon CID.

1.2.1.2 Tandem mass spectrometric analysis of glycosylated natural products

Tandem mass spectrometry is also a common method to gain structural information of oligosaccharides such as glycans [97]. For example, oligosaccharides can be sequenced by MS^n based on the cleavage of *O*-/*N*-glycosidic bonds in low-energy collision induced dissociation (CID) [94]. The same fragmentation of *O*-/*N*-glycosidic bonds has been observed in GNPs, such as erythromycin [98], thus enabling a similar fragmentation nomenclature for GNPs (Figure 20A). *O*-/*N*-glycosyl groups in GNPs are preferred leaving groups from the parent ion because of the mentioned lability of carbon-heteroatom bonds in the gas phase [89]. The predictable fragmentation of glycosyl groups in GNPs yields B/C-sugar fragments in the low-m/z region of the MS^n spectrum and Y/Z-aglycone fragments in the higher m/z region. Both fragmentation footprints correspond to specific sugar losses from the glycosylated natural product (Figure 20B) and, thus, can reveal these biosynthetic building blocks in MS^n experiments. For tandem MS analysis of GNP, the 'one third rule' of ion traps has to be considered as it can prevent detection of low-m/z B/C-sugar fragments by these mass analyzers (Figure 20C).

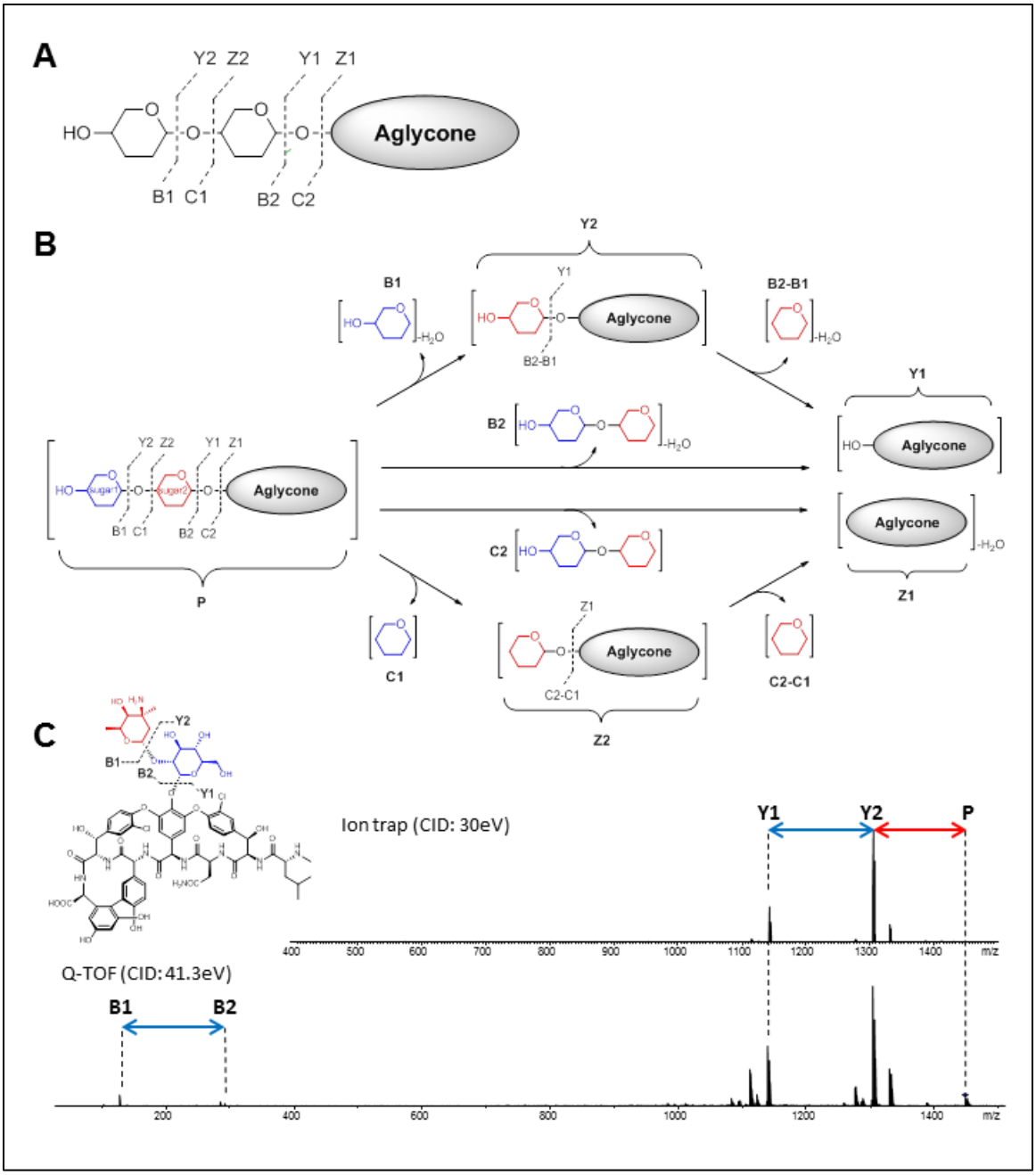


Figure 20: Tandem mass spectrometric analysis of glycosylated natural products. A – GNP fragmentation nomenclature. B – GNP fragmentation mechanisms upon CID. C – Sugar footprints in tandem MS spectra using different tandem MS instruments.

1.2.2 MALDI-imaging mass spectrometry

Imaging mass spectrometry (IMS) is a method to characterize and visualize the spatial distribution of analytes on a sample surface [83-85]. MALDI-TOF MS is commonly used for IMS approaches and has recently been applied to study the production of microbial natural products.

Briefly, a microbial colony on a thin agar plate is placed on a MALDI target plate, coated with MALDI matrix, dried to an even sample surface and analyzed with a MALDI-TOF MS by a MALDI imaging software. The MALDI-IMS data can be displayed as a spatial distribution of a target mass signal intensity on the sample surface. This enables e.g. the detection of secreted natural products from a microbial colony at a sampled timepoint [99].

1.3 Natural product discovery approaches

Discovery of natural products for pharmaceutical use has been traditionally guided by bioactivity, e.g. an inhibitory phenotype of a purified natural product in an antimicrobial bioassay. A classic example is the discovery of penicillin through its antibiotic activity by Alexander Fleming in 1928 [100]. Bioactivity-guided strategies have provided a wealth of antibiotics and anti-cancer agents. However these strategies were slow and expensive and they didn't meet the needs of industrial high-throughput screening technologies that only synthetic chemicals could provide e.g. via combinatorial synthesis [101,102]. This made natural products in the last decades a less attractive source of new drug candidates for pharmaceutical companies [103]. Within the last decade, new strategies for natural product discovery have been introduced based on advances in our understanding of natural product biosynthesis and advances in 'omics' methodologies such as genome [104], proteome [105] or metabolome analysis [106] (Figure 21).

1.3.1 'Omics'-approaches for natural product discovery

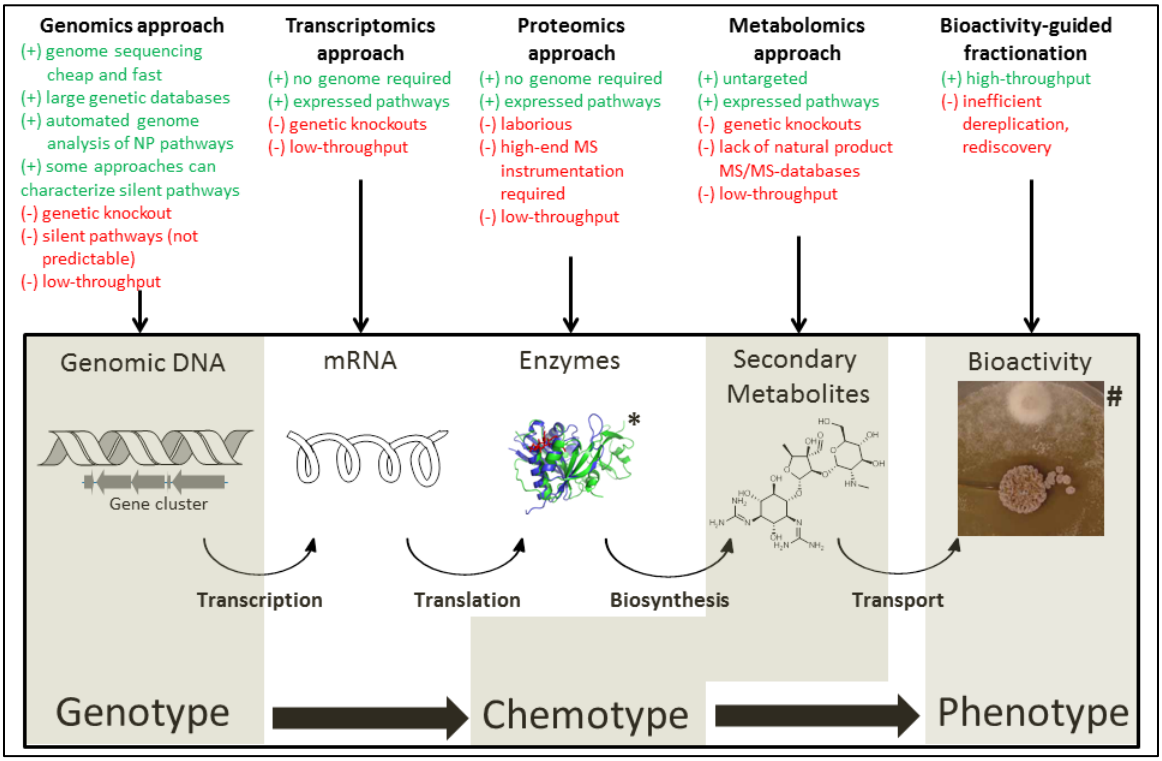


Figure 21: 'Omics'-approaches for natural product discovery.

1.3.1.1 Metabolomics approach

Untargeted metabolomics approaches enable the characterization of many phenotype-associated metabolites by LC-MS-based comparison of metabolomes. In natural product discovery, these metabolomic approaches led to the characterization of new metabolites associated with the pyochelin gene cluster (*pch*) in *Pseudomonas aeruginosa* [106]. The comparative metabolic profiling of a wild-type and a *pch* mutant identified the upregulation of several new pathway side-products [106].

An advantage of untargeted metabolomics is the characterization of expressed pathway products. The disadvantage is that metabolomics databases still lack reference spectra for many secondary metabolites under various fragmentation conditions [107]. This prevents a quick confirmation of pathway-related products. In addition, in the given example, a genome or at least gene cluster sequence, had to be present for genetic knockout of the target pathway. A recently applied networking algorithm can cluster LC-MS/MS metabolomic datasets with reference spectra

to identify new derivatives of reference metabolites [108] promising an extended search-space of small molecule structures with current and future tandem MS database references.

1.3.1.2 Proteomics approach

A proteomic approach for natural product discovery was introduced in 2009 by the Kelleher group [105]. In their PrISM approach (Proteomic Investigation of Secondary Metabolism), large proteins were isolated from bacterial cultures and analyzed by LC-MS/MS for the identification of large NRPS or PKS proteins by PPant tandem MS fragments [109]. From the proteomic data, identified NRPS or PKS active sites were characterized in amino acid sequence and the corresponding gene cluster was sequenced with degenerate primers. A target NRPS gene cluster sequence led to the MS-based discovery of a new *Bacillus* lipopeptide based on bioinformatic prediction of the pathway product.

The advantage of the PrISM approach is that no genome sequence is necessary for the identification of a new natural product and that expressed pathways are targeted. The first point can be an advantage for the investigation of organisms with large, complex genomes, e.g. dinoflagellates [110]. However, the approach requires expensive FTMS instrumentation, is laborious and includes challenging proteomic data analysis. In addition, follow-up studies of the PrISM approach used genome sequencing to assist the proteomics-based discovery of lipopeptide natural products [111]. Finally, the PrISM strategy only targets assembly line biosynthetic pathways and is thus limited to nonribosomal peptides and polyketides.

1.3.1.3 Transcriptomics approach

A recent study used transcriptome-screening for characterization of natural products from a microbial culture [112]. Herein, mRNA was isolated from certain cultivation time points and screened with degenerate PCR primers of genes involved in secondary metabolic pathways. Several transcribed genes of a mixed NRPS-PKS pathway were identified and linked to products via genetic knockouts. The genotype-chemotype connection was finally established by sequencing of the corresponding gene cluster from genomic DNA and comparative metabolic profiling of the wild-type and the pathway mutant.

The transcriptome discovery approach has the advantage that no genome sequence is required and that targeted pathways are transcribed under tested cultivation conditions. The disadvantage is that genetic knockouts are necessary to establish a chemotype-genotype connection with further gene cluster sequencing for verification.

1.3.1.4 Genomics approach or "Genome mining"

Through genome sequencing, researchers realized that microbial genomes comprise ~90% of biosynthetic gene clusters with unknown natural products [113]. These so-called cryptic gene clusters represent a large untapped source of new natural product chemistry and bioactivity as genome sequence information in databases is growing fast [114]. Based on the predicted metabolic resource, genomics-guided discovery strategies, generally termed as genome mining strategies, have been developed to characterize the molecules connected to cryptic pathways [104]. Genome mining in natural product discovery can be loosely defined as the connection of an unknown natural product structure with its biosynthetic genes by applied biosynthetic knowledge. Traditional genome mining approaches are *in silico*-guided approaches. They start with the prediction of a biosynthetic gene cluster and of properties of its product such as its building blocks. Subsequently, the unknown natural product is isolated by experiments guided by these predictions [104]. Genome mining has advantages over other 'omics'-approaches in data acquisition, data storage and data processing. The characterization of a whole microbial genome by DNA sequencing is easier and cheaper than characterization of a microbial proteome or metabolome. Genetic databases such as the NCBI Genbank [115] surpass protein and metabolomic databases in size and applicability for natural product discovery by the corresponding 'omics'-approaches [107]. Genome sequences can be analyzed automatically by prediction softwares such as NP.searcher [42], AntiSMASH [44] or NaPDoS [116] for the presence of secondary metabolic genotypes. These advantages make genome mining a promising approach for rapid characterization of new natural products. In the following section, traditional *in silico*-guided genome mining approaches are described.

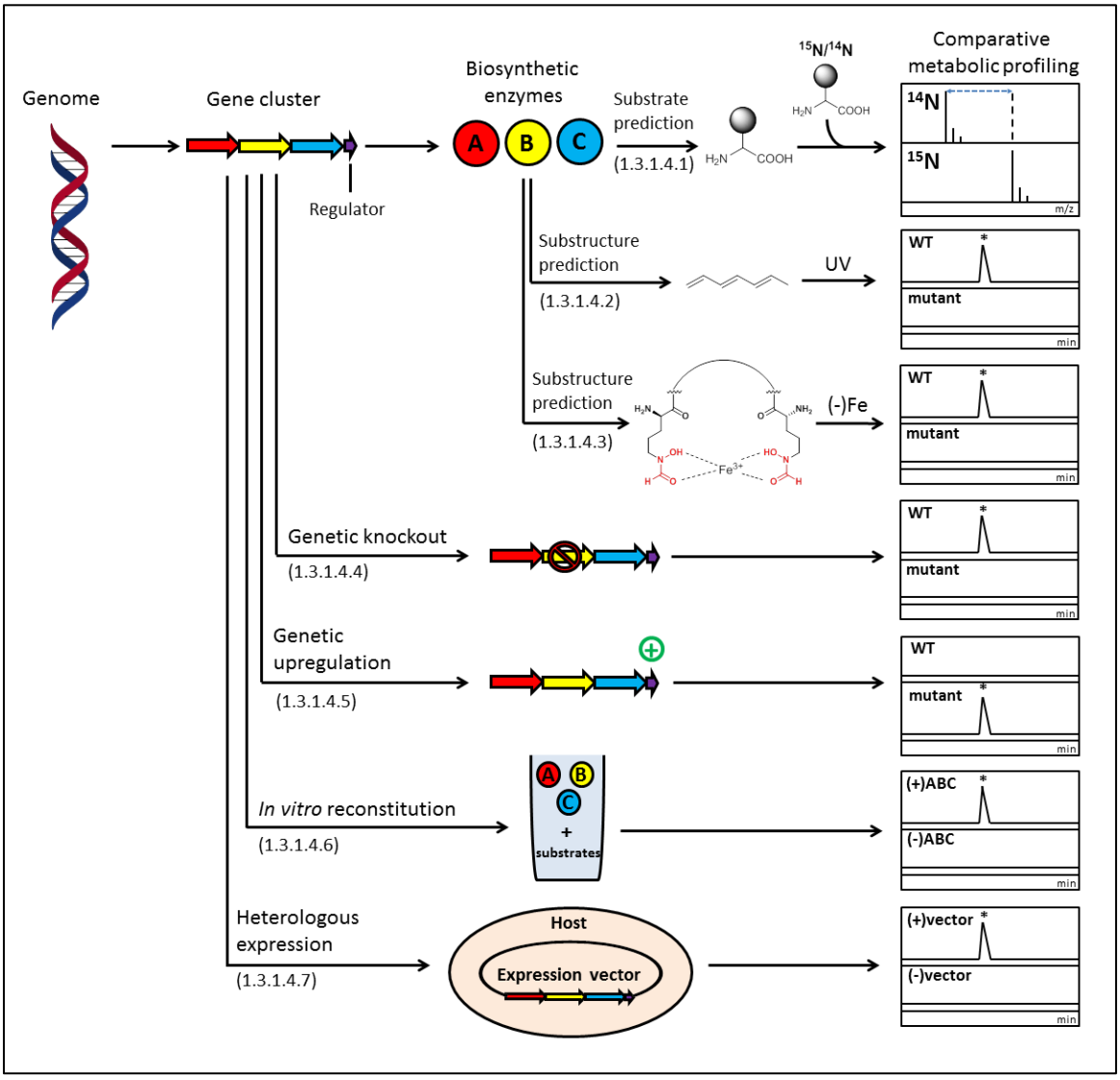


Figure 22: *In silico*-guided genome mining approaches.

1.3.1.4.1 Biosynthetic substrate-guided genome mining

A common strategy in *in silico*-guided genome mining is the prediction of biosynthetic substrates of a cryptic pathway to isolate its biosynthetic product. For example, in a genomisotopic approach [117], Gross *et al.* predicted leucine as a substrate specific for an orphan NRPS pathway (*ofa*) in *Pseudomonas fluorescens* Pf-5 based on determination of A domain substrate specificities in all NRPS gene clusters in the corresponding genome. Subsequently, ^{15}N-labeled leucine was fed to a *Pseudomonas fluorescens* culture to isolate a

putative *ofa* product by ^{1}H-^{15}N-HMBC signals in corresponding fractions. The target compound was a new bioactive lipopeptide, orfamide A, which was confirmed as the *ofa* product by genetic knockout of the pathway.

1.3.1.4.2 Physico-chemical-guided genome mining

Another strategy in *in silico*-guided genome mining is the prediction of physico-chemical properties of a cryptic pathway product such as UV absorbance. Nguyen *et al.* targeted an orphan *trans*-AT PKS pathway of *Burkholderia thailandensis* for natural product discovery and predicted a putative product with a conjugated double bond system from the PKS assembly line [118]. As polyenes have a characteristic UV-absorbance, *Burkholderia thailandensis* cultures were screened by UV-HPLC-MS and compounds with polyene-like absorbance, thailandamide A and B, were isolated as the cryptic pathway products.

1.3.1.4.3 Bioactivity-guided genome mining

In silico-guided genome mining can also start with the prediction of functional properties, e.g. bioactivity, of a cryptic pathway product. For example, in the first genome mining study, Lautru *et al.* targeted an orphan NRPS gene cluster (*cch*) from *Streptomyces coelicolor* M145 [119]. The NRPS genes enabled a prediction of *N*-formylhydroxyornithine substrates which contain siderophore-like functionalities. Based on the prediction of a putative siderophore function, wild-type and *cch* mutant *S. coelicolor* M145 were cultivated under iron-limited conditions and compared in their metabolic profiles by LC-MS. The comparative metabolic profiling revealed the iron-chelating *cch* product, coelichelin, a nonribosomal tetrapeptide.

1.3.1.4.4 Genetic knockout and comparative metabolic profiling

Cryptic pathways that don't enable a substrate prediction from their biosynthetic genes can be characterized in their product by genetic knockout of an essential biosynthetic gene and LCMS-based comparative metabolic profiling of the wild-type and mutant strains. An example for this strategy is the identification of the germicidins as the products of a cryptic type III PKS gene cluster in *Streptomyces coelicolor* [120].

1.3.1.4.5 Genetic upregulation and comparative metabolic profiling

A problem of *in silico*-guided genome mining is that a targeted pathway can be 'silent', i.e. not expressed or low-expressed under investigated cultivation conditions. Several genome mining studies have applied genetic upregulation of cryptic pathways and isolated the corresponding natural products by comparative metabolic profiling of the regulator mutant and the wild-type strain [117]. A recent example is the characterization of the 51-membered glycosylated macrolide stambomycins from a large cryptic type I PKS gene cluster in *Streptomyces ambofaciens* by constitutive expression of a pathway activator gene [121].

1.3.1.4.6 *In vitro* reconstitution

Another strategy to circumvent silent gene clusters in genome mining studies is an *in vitro* reconstitution approach of the biosynthetic enzymes in a cryptic pathway. An example is the heterologous expression of two cryptic sesquiterpene synthase genes, *sscg_02150* and *sscg_03688*, from *Streptomyces clavuligerus* in *Escherichia coli* and the subsequent *in vitro* reconstitution of the recombinant enzymes with substrate farnesyl diphosphate yielding the cyclic sesquiterpene products, (-)-δ-cadinene and (+)-T-muurolol, respectively [122].

1.3.1.4.7 Heterologous pathway expression

A third effective *in silico*-guided genomics strategy to capture natural products of silent cryptic gene clusters is heterologous expression of whole pathways in a host system. Recent examples of this strategy involve the characterization of cryptic ribosomal peptide pathways predicted by bioinformatic identification of precursor genes [123] or biosynthetic genes [124]. Maksimov *et al.* could clone and heterologously express a cryptic lassopeptide gene cluster from *Asticcacaulis excentricus* in *Escherichia coli* to yield new lassopeptide products. Garg *et al.* cloned and heterologously expressed a class I lanthipeptide gene cluster from *Geobacillus thermodenitrificans* in *E. coli* to yield a nisin-like lantibiotic.

1.4 Problem and aim of dissertation

The problem of traditional *in silico*-guided genome mining approaches is that one pathway is targeted per experiment (Figure 22). In addition, most highlighted genomics approaches involve a laborious genetic inactivation step of a target pathway. *In silico*-guided approaches are effective but are not matching the pace with which genomes, i.e. cryptic gene clusters, are sequenced nowadays.

The aim of this dissertation was to introduce mass spectrometry-guided genome mining approaches that connect unknown structures of major microbial natural product classes more rapidly with their biosynthetic genes with a potential for automation. The proposed approaches are methodologically based on the advantages of mass spectrometry (MS) in analysis of natural products, i.e. detection of low quantities of compounds, easy integration into automated platforms and *de novo* structure elucidation by tandem MS experiments. The work is focuses on two large classes of natural products, peptide natural products (PNPs) and glycosylated natural products (GNPs) which are targeted with two complementary approaches, peptidogenomics (Chapter 2) and glycogenomics (Chapter 4), respectively.

The innovation of the proposed work is the rapid connection of an unknown chemotype with its biosynthetic genes (genotype) by matching *de novo* MS^n substructures such as amino acid sequences or sugar mass shifts to corresponding structures that can be predicted from biosynthetic gene clusters in genome sequences. Herein, structure elucidation of peptide natural products and glycosylated natural products by tandem MS will be combined with genotype analysis based on current biosynthetic knowledge of peptidic and glycosylated natural products. These rapid chemotype-to-genotype connections can then guide the characterization of new natural product chemistry, biosynthesis and bioactivity for all researchers interested in natural products.

The approaches, peptidogenomics and glycogenomics, are further tested for discovery of new PNP and GNP chemo- and genotypes from marine bacteria (Chapters 3 and 5).

References

1. Seiple, I.B., Su, S., Young, I.S., Lewis, C.A., Yamaguchi, J., Baran, P.S. Total synthesis of palau'amine. *Angew. Chem. Int. Ed. Engl*. **49**, 1095-1098 (2010).

2. Eustáquio, A.S., Pojer, F., Noel, J.P., Moore, B.S. Discovery and characterization of a marine bacterial SAM-dependent chlorinase. *Nat. Chem. Biol.* **4**, 69-74 (2008).

3. Oh, D.C., Poulsen, M., Currie, C.R., Clardy, J. Dentigerumycin: a bacterial mediator of an ant-fungus symbiosis. *Nat. Chem. Biol.* **5**, 391-393 (2009).

4. Koehn, F.E., Carter, G.T. The evolving role of natural products in drug discovery. *Nat. Rev. Drug Discov.* **4**, 206-220 (2005).

5. Gerwick, W., Dorrestein, P.C., Moore, B.S. *personal communication.*

6. Phelan, V.V., Liu, W.T., Pogliano, K., Dorrestein, P.C. Microbial metabolic exchange--the chemotype-to-phenotype link. *Nat. Chem. Biol*. **8**, 26-35 (2011).

7. Daffre, S., Bulet, P., Spisni, A. Ehret-Sabatier, L., Rodrigues, E.G., Travassos, L.R. Bioactive natural peptides. in *Studies in Natural Products Chemistry*, 1st edn., Vol. 35 (ed. Rahman, A.U.) 597–691 (Elsevier, 2008).

8. Micklefield, J. Daptomycin structure and mechanism of action revealed. *Chem. Biol.* **11**, 887-888 (2004).

9. Hansen, J.N. Nisin as a model food preservative. *Crit. Rev. Food Sci. Nutr.* **34**, 69-93 (1994).

10. Nolan, E.M., Walsh, C.T. How nature morphs peptide scaffolds into antibiotics. *Chembiochem* **10**, 34-53 (2009).

11. Arnison, P.G., Bibb, M.J., Bierbaum, G., Bowers, A.A., Bugni, T.S., Bulaj, G., Camarero, J.A., Campopiano, D.J., Challis, G.L., Clardy, J., Cotter, P.D., Craik, D.J., Dawson, M., Dittmann, E., Donadio, S., Dorrestein, P.C., Entian, K.D., Fischbach, M.A., Garavelli, J.S., Göransson, U., Gruber, C.W., Haft, D.H., Hemscheidt, T.K., Hertweck, C., Hill, C., Horswill, A.R., Jaspars, M., Kelly, W.L., Klinman, J.P., Kuipers, O.P., Link, A.J., Liu, W., Marahiel, M.A., Mitchell, D.A., Moll, G.N., Moore, B.S., Müller, R., Nair, S.K., Nes, I.F., Norris, G.E., Olivera, B.M., Onaka, H., Patchett, M.L., Piel, J., Reaney, M.J., Rebuffat, S., Ross, R.P., Sahl, H.G., Schmidt, E.W., Selsted, M.E., Severinov, K., Shen, B., Sivonen, K., Smith, L., Stein, T., Süssmuth, R.D., Tagg, J.R., Tang, G.L., Truman, A.W., Vederas, J.C., Walsh, C.T., Walton, J.D., Wenzel, S.C., Willey, J.M., van der Donk, W.A. Ribosomally synthesized and post-translationally modified peptide natural products: overview and recommendations for a universal nomenclature*. Nat. Prod. Rep.* **30**, 108-160 (2013).

12. Kopp, F., Marahiel, M.A. Macrocyclization strategies in polyketide and nonribosomal peptide biosynthesis. *Nat. Prod. Rep.* **24**, 735-749 (2007).

13. Willey, J.M., van der Donk, W.A. Lantibiotics: peptides of diverse structure and function. *Annu. Rev. Microbiol.* **61**, 477-501 (2007).

14. Li, C., Kelly, W.L. Recent advances in thiopeptide antibiotic biosynthesis. *Nat. Prod. Rep.* **27**, 153-164 (2010).

15. Donia, M.S., Ravel, J., Schmidt, E.W. A global assembly line for cyanobactins. *Nat. Chem. Biol,* **4**, 341-343 (2008).

16. Duquesne, S., Destoumieux-Garzón, D., Zirah, S., Goulard, C., Peduzzi, J., Rebuffat, S. Two enzymes catalyze the maturation of a lasso peptide in Escherichia coli. *Chem. Biol.* **14**, 793-803 (2007).

17. Claesen, J., Bibb, M. Genome mining and genetic analysis of cypemycin biosynthesis reveal an unusual class of posttranslationally modified peptides. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 16297-16302 (2010).

18. Cotter, P.D., Hill, C., Ross, R.P. Bacteriocins: developing innate immunity for food. *Nat. Rev. Microbiol.* **3**, 777-788 (2005).

19. Kersten, R.D., Yang, Y.L., Xu, Y., Cimermancic, P., Nam, S.J., Fenical, W., Fischbach, M.A., Moore, B.S., Dorrestein, P.C. A mass spectrometry-guided genome mining approach for natural product peptidogenomics. *Nat. Chem. Biol.* **7**, 794-802 (2011).

20. Ueda, K., Oinuma, K., Ikeda, G., Hosono, K., Ohnishi, Y., Horinouchi, S., Beppu, T. AmfS, an extracellular peptidic morphogen in Streptomyces griseus. *J. Bacteriol.* **184**, 1488-1492 (2002).

21. Chatterjee, C., Miller, L.M., Leung, Y.L., Xie, L., Yi, M., Kelleher, N.L., van der Donk, W.A. Lacticin 481 synthetase phosphorylates its substrate during lantibiotic production. *J. Am. Chem. Soc.* **127**, 15332-15333 (2005).

22. Karakas, Sen. A., Narbad, A., Horn, N., Dodd, H.M., Parr, A.J., Colquhoun, I., Gasson, M.J. Post-translational modification of nisin. The involvement of NisB in the dehydration process. *Eur. J. Biochem.* **261**, 524-532 (1999).

23. Li, B., Yu, J.P., Brunzelle, J.S., Moll, G.N., van der Donk, W.A., Nair, S.K. Structure and mechanism of the lantibiotic cyclase involved in nisin biosynthesis. *Science* **311**, 1464-1467 (2006).

24. Xie, L., Miller, L.M., Chatterjee, C., Averin, O., Kelleher, N.L., van der Donk, W.A. Lacticin 481: in vitro reconstitution of lantibiotic synthetase activity. *Science* **303**, 679-681 (2004).

25. Li, B., Sher, D., Kelly, L., Shi, Y., Huang, K., Knerr, P.J., Joewono, I., Rusch, D., Chisholm, S.W., van der Donk, W.A. Catalytic promiscuity in the biosynthesis of cyclic peptide secondary metabolites in planktonic marine cyanobacteria. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 10430-10435 (2010).

26. Morgan, S.M., O'Connor, P.M., Cotter, P.D., Ross, R.P., Hill, C. Sequential actions of the two component peptides of the lantibiotic lacticin 3147 explain its antimicrobial activity at nanomolar concentrations. *Antimicrob. Agents Chemother.* **49**, 2606-2611 (2005).

27. Müller, W.M., Schmiederer, T., Ensle, P., Süssmuth, R.D. In vitro biosynthesis of the prepeptide of type-III lantibiotic labyrinthopeptin A2 including formation of a C-C bond as a post-translational modification. *Angew. Chem. Int. Ed. Engl.* **49**, 2436-2440 (2010).

28. Goto, Y., Li, B., Claesen, J., Shi, Y., Bibb, M.J., van der Donk, W.A. Discovery of unique lanthionine synthetases reveals new mechanistic and evolutionary insights. *PLoS Biol.* **8**, e1000339 (2010).

29. Kodani, S., Hudson, M.E., Durrant, M.C., Buttner, M.J., Nodwell, J.R., Willey, J.M. The SapB morphogen is a lantibiotic-like peptide derived from the product of the developmental gene ramS in Streptomyces coelicolor. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 11448-11453 (2004).

30. Meindl, K., Schmiederer, T., Schneider, K., Reicke, A., Butz, D., Keller, S., Gühring, H., Vértesy, L., Wink, J., Hoffmann, H., Brönstrup, M., Sheldrick, G.M., Süssmuth, R.D. Labyrinthopeptins: a new class of carbacyclic lantibiotics. *Angew. Chem. Int. Ed. Engl.* **49**, 1151-1154 (2010).

31. Maksimov, M.O., Pan, S.J., Link, A.J. Lasso peptides: structure, function, biosynthesis, and engineering. *Nat. Prod. Rep.* **29**, 996-1006 (2012).

32. Rebuffat, S., Blond, A., Destoumieux-Garzón, D., Goulard, C., Peduzzi, J. Microcin J25, from the macrocyclic to the lasso structure: implications for biosynthetic, evolutionary and biotechnological perspectives. *Curr. Protein Pept. Sci.* **5**, 383-391 (2004).

33. Lin, P.F., Samanta, H., Bechtold, C.M., Deminie, C.A., Patick, A.K., Alam, M., Riccardi, K., Rose, R.E., White, R.J., Colonno, R.J. Characterization of siamycin I, a human immunodeficiency virus fusion inhibitor. *Antimicrob. Agents Chemother.* **40**, 133-138 (1996).

34. Salomón, R.A., Farías, R.N. Microcin 25, a novel antimicrobial peptide produced by Escherichia coli. *J. Bacteriol.* **174**, 7428-7435 (1992).

35. Knappe, T.A., Linne, U., Xie, X., Marahiel, M.A. The glucagon receptor antagonist BI-32169 constitutes a new class of lasso peptides. *FEBS Lett.* **584**, 785-789 (2010).

36. Yan, K.P., Li, Y., Zirah, S., Goulard, C., Knappe, T.A., Marahiel, M.A., Rebuffat, S. Dissecting the maturation steps of the lasso peptide microcin J25 in vitro. *Chembiochem* **13**, 1046-1052 (2012)

37. Pan, S.J., Rajniak, J., Maksimov, M.O., Link, A.J. The role of a conserved threonine residue in the leader peptide of lasso peptide precursors. *Chem. Commun. (Camb)* **48**, 1880-1882 (2012).

38. Kempter, C., Dr. Kupke, T., Kaiser, D., Metzger, J.W., Jung, G. Thioenols from Peptidyl Cysteines: Oxidative Decarboxylation of a ^{13}C-Labeled Substrate. *Angew. Chem. Int. Ed. Engl.* **35**, 2104–2107 (1996).

39. Fischbach, M.A., Walsh, C.T. Assembly-line enzymology for polyketide and nonribosomal Peptide antibiotics: logic, machinery, and mechanisms. *Chem. Rev.* **106**, 3468-3496 (2006).

40. Stachelhaus, T., Mootz, H.D., Marahiel, M.A. The specificity-conferring code of adenylation domains in nonribosomal peptide synthetases. *Chem. Biol.* **6**, 493-505 (1999).

41. Challis, G.L., Ravel, J., Townsend, C.A. Predictive, structure-based model of amino acid recognition by nonribosomal peptide synthetase adenylation domains. *Chem. Biol.* **7**, 211-224 (2000).

42. Li, M.H., Ung, P.M., Zajkowski, J., Garneau-Tsodikova, S., Sherman, D.H. Automated genome mining for natural products. *BMC Bioinformatics* **10**, 185 (2009).

43. Röttig, M., Medema, M.H., Blin, K., Weber, T., Rausch, C., Kohlbacher, O. NRPSpredictor2--a web server for predicting NRPS adenylation domain specificity. *Nucleic Acids Res*. **39**, W362-367 (2011).

44. Medema M.H., Blin, K., Cimermancic, P., de Jager, V., Zakrzewski, P., Fischbach, M.A., Weber, T., Takano, E., Breitling, R. antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. *Nucleic Acids Res*. **39**, W339-346 (2011).

45. Caboche, S., Pupin, M., Leclère, V., Fontaine, A., Jacques, P., Kucherov, G. NORINE: a database of nonribosomal peptides. *Nucleic Acids Res.* **36**, D326-331 (2008).

46. Yin, X., McPhail, K.L., Kim, K.J., Zabriskie, T.M. Formation of the nonproteinogenic amino acid 2*S*,3*R*-capreomycidine by VioD from the viomycin biosynthesis pathway. *Chembiochem* **5**, 1278-1281 (2004).

47. Ju, J., Ozanick, S.G., Shen, B., Thomas, M.G. Conversion of (2*S*)-arginine to (2*S*,3*R*)-capreomycidine by VioC and VioD from the viomycin biosynthetic pathway of Streptomyces sp. strain ATCC11861. *Chembiochem* **5**, 1281-1285 (2004).

48. Felnagle, E.A., Podevels, A.M., Barkei, J.J., Thomas, M.G. Mechanistically distinct nonribosomal peptide synthetases assemble the structurally related antibiotics viomycin and capreomycin. *Chembiochem* **12**, 1859-1867 (2011).

49. Hur, G.H., Vickery, C.R., Burkart, M.D. Explorations of catalytic domains in non-ribosomal peptide synthetase enzymology. *Nat. Prod. Rep.* **29**, 1074-98 (2012).

50. Mootz, H.D., Marahiel, M.A. The tyrocidine biosynthesis operon of Bacillus brevis: complete nucleotide sequence and biochemical characterization of functional internal adenylation domains. *J. Bacteriol*. **179**, 6843-6850 (1997).

51. Strieker, M., Marahiel, M.A. The structural diversity of acidic lipopeptide antibiotics. *Chembiochem* **10**, 607-616 (2009).

52. Isogai, I., Takayama, S., Murakoshi, S., Suzuki, A. Structure of β-amino acids in antibiotics iturin A. *Tetrahedron Lett.* **23**, 3065–3068 (1982).

53. Kakinuma, A., Ouchida, A., Shima, T., Sugino, H., Isono, M., Tamura, G., Arima, K. Conformation of the structure of surfactin by mass spectrometry. *Agric. Biol. Chem.* **33**, 1669–1671 (1969).

54. Robbel, L., Marahiel, M.A. Daptomycin, a bacterial lipopeptide synthesized by a nonribosomal machinery. *J. Biol. Chem.* **285**, 27501-27508 (2010).

55. Staunton, J., Weissman, K.J. Polyketide biosynthesis: a millennium review. *Nat. Prod. Rep.* **18**, 380–416 (2001).

56. Fujii, I., Ebizuka, Y. Anthracycline Biosynthesis in Streptomyces galilaeus. *Chem. Rev.* **97**, 2511-2524 (1997).

57. Moore, B.S., Hertweck, C. Biosynthesis and attachment of novel bacterial polyketide synthase starter units. *Nat. Prod. Rep*. **19**, 70-99 (2002).

58. Wilson, M.C., Moore, B.S. Beyond ethylmalonyl-CoA: the functional role of crotonyl-CoA carboxylase/reductase homologs in expanding polyketide diversity. *Nat Prod Rep*. **29**, 72-86 (2012).

59. Dewick, J. *Medicinal Natural Products: A Biosynthetic Approach*. 3rd edition, John Wiley & Sons (2009).

60. Cortes, J., Haydock, S.F., Roberts, G.A., Bevitt, D.J., Leadlay, P.F. An unusually large multifunctional polypeptide in the erythromycin-producing polyketide synthase of Saccharopolyspora erythraea. *Nature* **348**, 176-178 (1990).

61. Räty, K., Kantola, J., Hautala, A., Hakala, J., Ylihonko, K., Mäntsälä, P. Cloning and characterization of Streptomyces galilaeus aclacinomycins polyketide synthase (PKS) cluster. *Gene* **293**, 115-122 (2002).

62. Räty, K., Kunnari, T., Hakala, J., Mäntsälä, P., Ylihonko, K. A gene cluster from Streptomyces galilaeus involved in glycosylation of aclarubicin. *Mol. Gen. Genet.* **264**, 164-172 (2000).

63. Müller, R. Biosynthesis and Heterologous Production of Epothilones. In *"The Epothilones—An Outstanding Family of Anti-Tumor Agents"* (J. Mulzer, ed.). Springer (2009).

64. Ikeda, H., Nonomiya, T., Usami, M., Ohta, T., Omura, S. Organization of the biosynthetic gene cluster for the polyketide anthelmintic macrolide avermectin in Streptomyces avermitilis. *Proc. Natl. Acad. Sci. U. S. A.* **96**, 9509-9514 (1999).

65. Thibodeaux, C.J., Melançon, C.E. 3rd, Liu, H.W. Natural-product sugar biosynthesis and enzymatic glycodiversification. *Angew. Chem. Int. Ed. Engl*. **47**, 9814-9859 (2008).

66. La Ferla, B., Airoldi, C., Zona, C., Orsato, A., Cardona, F., Merlo, S., Sironi, E., D'Orazio, G., Nicotra, F. Natural glycoconjugates with antitumor activity. *Nat. Prod. Rep.* **28**, 630-648 (2011).

67. Hubbard, B.K., Walsh, C.T. Vancomycin assembly: nature's way. *Angew. Chem. Int. Ed. Engl.* **42**, 730-765 (2003).

68. Ding, Y., Yu, Y., Pan, H., Guo, H., Li, Y., Liu, W. Moving posttranslational modifications forward to biosynthesize the glycosylated thiopeptide nocathiacin I in Nocardia sp. ATCC 202099. *Mol. Biosyst.* **6**, 1180-1185 (2010).

69. Ahlert, J., Shepard, E., Lomovskaya, N., Zazopoulos, E., Staffa, A., Bachmann, B.O., Huang, K., Fonstein, L., Czisny, A., Whitwam, R.E., Farnet, C.M., Thorson, J.S. The calicheamicin gene cluster and its iterative type I enediyne PKS. *Science* **297**, 1173-1176 (2002).

70. Gebhardt, K., Meyer, S.W., Schinko, J., Bringmann, G., Zeeck, A., Fiedler, H.P. Phenalinolactones A-D, terpenoglycoside antibiotics from Streptomyces sp. Tü 6071. *J. Antibiot. (Tokyo)* **64**, 229-232 (2011).

71. Pathirana, C., Jensen, P.R., Dwight, R., Fenical, W. Rare phenazine L-quinovose esters from a marine actinomycete. *J. Org. Chem*. **57**, 740-742 (1992).

72. Singh, S., Phillips, G.N. Jr., Thorson, J.S. The structural biology of enzymes involved in natural product glycosylation. *Nat. Prod. Rep*. **29**, 1201-1237 (2012).

73. Rohr, J. Modifying Oxidation and Glycosylation Events in Biosyntheses of Natural Product Anticancer Drugs – *Challenges for Combinatorial Biosynthesis, Functional Molecules from Natural Sources*, S. Wrigley, RSC Publishing, Cambridge, 161-183 (2011).

74. Chen, F., Lin, L., Wang, L., Tan, Y., Zhou, H., Wang, Y., He, W. Distribution of dTDP-glucose-4,6-dehydratase gene and diversity of potential glycosylated natural products in marine sediment-derived bacteria. *Appl. Microbiol. Biotechnol.* **90**, 1347-1359 (2011).

75. Siuzdak, G. *The Expanding Role of Mass Spectrometry in Biotechnology*; MCC Press: San Diego (2003).

76. Fenn, J.B., Mann, M., Meng, C.K., Wong, S.F., Whitehouse, C.M. Electrospray ionization for mass spectrometry of large biomolecules. *Science* **246**, 64–71 (1989).

77. Ho, C.S., Chan, M.H.M., Cheung, R.C.K., Law, L.K., Lit, L.C.W., Ng, K.F., Suen, M.W.M., Tai, H.L. Electrospray Ionisation Mass Spectrometry: Principles and Clinical Applications. *Clin. Biochem. Rev.* **24**, 3–12 (2003).

78. Karas, M., Bachmann, D., Hillenkamp, F. Influence of the Wavelength in High-Irradiance Ultraviolet Laser Desorption Mass Spectrometry of Organic Molecules. *Anal. Chem.* **57,** 2935–2939 (1985).

79. Wilm, M. Principles of electrospray ionization. *Mol. Cell. Proteomics* **10**, M111.009407 (2011).

80. Nguyen, S., Fenn, J.B. Gas-phase ions of solute species from charged droplets of solutions. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 1111-1117 (2007).

81. Dorrestein, P.C., Kelleher, N.L. Dissecting non-ribosomal and polyketide biosynthetic machineries using electrospray ionization Fourier-Transform mass spectrometry. *Nat. Prod. Rep.* **23**, 893-918 (2006).

82. Hillenkamp, F. and Karas, M. (2007) The MALDI Process and Method, in *MALDI MS: A Practical Guide to Instrumentation, Methods and Applications* (eds F. Hillenkamp and J. Peter-Katalinić), Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, Germany.

83. Caprioli, R.M., Farmer, T.B., Gile, J. Molecular imaging of biological samples: localization of peptides and proteins using MALDI-TOF MS. *Anal. Chem*. **69**, 4751-4760 (1997).

84. Watrous, J.D., Dorrestein, P.C. Imaging mass spectrometry in microbiology. *Nat. Rev. Microbiol*. **9**, 683-694 (2011).

85. Spengler, B. (2007) Microprobing and Imaging MALDI for Biomarker Detection, in *MALDI MS: A Practical Guide to Instrumentation, Methods and Applications* (eds F. Hillenkamp and J. Peter-Katalinić), Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, Germany.

86. Marshall, A.G., Hendrickson, C.L., Jackson, G.S. Fourier transform ion cyclotron resonance mass spectrometry: A primer. *Mass Spectrom. Rev*., **17**, 1–35 (1998).

87. McLafferty, F.W. Tandem mass spectrometry. *Science* **214**, 280-287 (1981).

88. Zhu, Z.J., Schultz, A.W., Wang, J., Johnson, C.H., Yannone, S.M., Patti, G.J., Siuzdak, G. Liquid chromatography quadrupole time-of-flight mass spectrometry characterization of metabolites guided by the METLIN database. *Nat. Protoc*. **8**, 451-460 (2013).

89. Gräfe, U., Heinze, S., Schlegel, B., Härtl, A. Disclosure of new and recurrent microbial metabolites by mass spectrometric methods. *J. Ind. Microbiol. Biotechnol.* **27**, 136-143 (2001).

90. Ng, J., Bandeira, N., Liu, W.T., Ghassemian, M., Simmons, T.L., Gerwick, W.H., Linington, R., Dorrestein, P.C., Pevzner, P.A. Dereplication and de novo sequencing of nonribosomal peptides. *Nat. Methods* **6**, 596–599 (2009).

91. Roepstorff, P., Fohlmann, J. Proposal for a common nomenclature for sequence ions in mass spectra of peptides. *Biomed. Mass Spectrom*. **11**, 601 (1984).

92. Paizs, B., Suhai, S. Fragmentation pathways of protonated peptides. *Mass Spectrom. Rev.* **24**, 508-548 (2005).

93. Polce, M.J., Ren, D., Wesdemiotis, C. Dissociation of the peptide bond in protonated peptides. *J. Mass. Spectrom*. **35**, 1391–1398 (2000).

94. Perkins, D.N., Pappin, D.J., Creasy, D.M., Cottrell, J.S. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20**, 3551–3567 (1999).

95. Duncan, M.W., Aebersold, R., Caprioli, R.M. The pros and cons of peptide-centric proteomics. *Nat. Biotechnol.* **28**, 659–664 (2010).

96. Dancik, V., Addona, T.A., Clauser, K.R., Vath, J.E., Pevzner, P.A. De novo peptide sequencing via tandem mass spectrometry. *J. Com. Biol.* **6**, 327-342 (1999).

97. An, H.J., Lebrilla, C.B. Structure elucidation of native N- and O-linked glycans by tandem mass spectrometry (tutorial). *Mass Spectrom. Rev.* **30**, 560-578 (2011).

98. Gates, P.J., Kearney, G.C., Jones, R., Leadlay, P.F., Staunton, J. Structural elucidation studies of erythromycins by electrospray tandem mass spectrometry. *Rapid Commun. Mass Spectrom.* **13**, 242-246 (1999).

99. Yang, Y.L., Xu, Y., Straight, P., Dorrestein, P.C. Translating metabolic exchange with imaging mass spectrometry. *Nat. Chem. Biol.* **5**, 885–887 (2009).

100. Fleming, A. On the antibacterial action of cultures of a penicillium, with special reference to their use in the isolation of B. influenzae*. Br. J. Exp. Pathol.* **10**, 226-236 (1929).

101. Fischbach, M.A., Walsh, C.T. Antibiotics for emerging pathogens. *Science* **325**, 1089-1093 (2009).

102. Grushkin, D. Natural products emergent. *Nat. Med*. **19**, 390-392 (2013).

103. Von Nussbaum, F., Brands, M., Hinzen, B., Weigand, S., Häbich, D. Antibacterial natural products in medicinal chemistry--exodus or revival? *Angew. Chem. Int. Ed. Engl.* **45**, 5072-5129 (2006).

104. Zerikly, M., Challis, G.L. Strategies for the discovery of new natural products by genome mining. *Chembiochem*. **4**, 625-633 (2009).

105. Bumpus, S.B., Evans, B.S., Thomas, P.M., Ntai, .I, Kelleher, N.L. A proteomics approach to discovering natural products and their biosynthetic pathways. *Nat. Biotechnol*. **10**, 951-956 (2006).

106. Vinayavekhin, N., Saghatelian, A. Regulation of Alkyl-dihydrothiazole-carboxylates (ATCs) by Iron and the Pyochelin Gene Cluster in Pseudomonas aeruginosa. *ACS Chem. Biol.* **8**, 617–623 (2009).

107. Smith, C.A., O'Maille, G., Want, E.J., Qin, C., Trauger, S.A., Brandon, T.R., Custodio, D.E., Abagyan, R., Siuzdak, G. METLIN: a metabolite mass spectral database. *Ther. Drug Monit.* **27**, 747-751 (2005).

108. Watrous, J., Roach, P., Alexandrov, T., Heath, B.S., Yang, J.Y., Kersten, R.D., van der Voort, M., Pogliano, K., Gross, H., Raaijmakers, J.M., Moore, B.S., Laskin, J., Bandeira, N., Dorrestein, P.C. Mass spectral molecular networking of living microbial colonies. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 1743-1752 (2012).

109. Dorrestein, P.C., Bumpus, S.B., Calderone, C.T., Garneau-Tsodikova, S., Aron, Z.D., Straight, P.D., Kolter, R., Walsh, C.T., Kelleher, N.L. Facile detection of acyl and peptidyl intermediates on thiotemplate carrier domains via phosphopantetheinyl elimination reactions during tandem mass spectrometry. *Biochemistry* **45**, 12756-12766 (2006).

110. Hou, Y., Lin, S. Distinct gene number-genome size relationships for eukaryotes and non-eukaryotes: gene content estimation for dinoflagellate genomes. *PLoS One* **4**, e6978 (2009).

111. Evans, B.S., Ntai, I., Chen, Y., Robinson, S.J., Kelleher, N.L. Proteomics-based discovery of koranimine, a cyclic imine natural product. *J. Am. Chem. Soc*. **133**, 7316-7319 (2011).

112. Qu, X., Lei, C., Liu, W. Transcriptome Mining of Active Biosynthetic Pathways and Their Associated Products in Streptomyces flaveolus. *Angew. Chem. Int. Ed.* **123**, 9825-9828 (2011).

113. Bentley, S.D., Chater, K.F., Cerdeño-Tárraga, A.M., Challis, G.L., Thomson, N.R., James, K.D., Harris, D.E., Quail, M.A., Kieser, H., Harper, D., Bateman, A., Brown, S., Chandra, G., Chen, C.W., Collins, M., Cronin, A., Fraser, A., Goble, A., Hidalgo, J., Hornsby, T., Howarth, S., Huang, C.H., Kieser, T., Larke, L., Murphy, L., Oliver, K., O'Neil, S., Rabbinowitsch, E., Rajandream, M.A., Rutherford, K., Rutter, S., Seeger, K., Saunders, D., Sharp, S., Squares, R., Squares, S., Taylor, K., Warren, T., Wietzorrek, A., Woodward, J., Barrell, B.G., Parkhill, J., Hopwood, D.A. Complete genome sequence of the model actinomycete Streptomyces coelicolor A3(2). *Nature* **417**, 141-147 (2002).

114. Pagani, I., Liolios, K., Jansson, J., Chen, I. M. A., Smirnova, T., Nosrat, B., Kyrpides, N. C. The Genomes OnLine Database (GOLD) v. 4: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res*. **40**, D571-D579 (2012).

115. Benson, D.A., Karsch-Mizrachi, I., Clark, K., Lipman, D.J., Ostell, J., Sayers, E.W. GenBank. *Nucleic Acids Res.* **40**, D48-D53 (2012).

116. Ziemert, N., Podell, S., Penn, K., Badger, J.H., Allen, E., Jensen, P.R. The natural product domain seeker NaPDoS: a phylogeny based bioinformatic tool to classify secondary metabolite gene diversity. *PloS One* **7**, e34064 (2012).

117. Gross, H., Stockwell, V.O., Henkels, M.D., Nowak-Thompson, B., Loper, J.E., Gerwick, W.H. The genomisotopic approach: a systematic method to isolate products of orphan biosynthetic gene clusters. *Chem. Biol*. **14**, 53-63 (2007).

118. Nguyen, T., Ishida, K., Jenke-Kodama, H., Dittmann, E., Gurgui, C., Hochmuth, T., Taudien, S., Platzer, M., Hertweck, C., Piel, J. Exploiting the mosaic structure of trans-acyltransferase polyketide synthases for natural product discovery and pathway dissection. *Nat. Biotechnol*. **26**, 225-233 (2008).

119. Lautru, S., Deeth, R.J., Bailey, L.M., Challis, G.L. Discovery of a new peptide natural product by Streptomyces coelicolor genome mining. *Nat. Chem. Biol*. **5**, 265-269 (2005).

120. Song, L., Barona-Gomez, F., Corre, C., Xiang, L., Udwary, D.W., Austin, M.B., Noel, J.P., Moore, B.S., Challis, G.L. Type III polyketide synthase beta-ketoacyl-ACP starter unit and ethylmalonyl-CoA extender unit selectivity discovered by Streptomyces coelicolor genome mining. *J. Am. Chem. Soc.* **128**, 14754-14755 (2006).

121. Laureti, L., Song, L., Huang, S., Corre, C., Leblond, P., Challis, G.L., Aigle, B. Identification of a bioactive 51-membered macrolide complex by activation of a silent polyketide synthase in Streptomyces ambofaciens. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 6258-6263 (2011).

122. Hu, Y., Chou, W.K., Hopson, R., Cane, D.E. Genome Mining in Streptomyces clavuligerus: Expression and Biochemical Characterization of Two New Cryptic Sesquiterpene Synthases. *Chem. Biol*. **18**, 32-37 (2011).

123. Maksimov, M.O., Pelczer, I., Link, A.J. Precursor-centric genome-mining approach for lasso peptide discovery., *Proc. Natl. Acad. Sci. U. S. A.* **109**, 15223-15228 (2012).

124. Garg, N., Tang, W., Goto, Y., Nair, S.K., van der Donk, W.A. Lantibiotics from Geobacillus thermodenitrificans. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 5241-5246 (2012).

Chapter 2 - A mass spectrometry–guided genome mining approach for natural product peptidogenomics

ARTICLE
PUBLISHED ONLINE: 9 OCTOBER 2011 | DOI: 10.1038/NCHEMBIO.684
nature chemical biology

A mass spectrometry–guided genome mining approach for natural product peptidogenomics

Roland D Kersten[1], Yu-Liang Yang[2], Yuquan Xu[2], Peter Cimermancic[3], Sang-Jip Nam[1], William Fenical[1,2], Michael A Fischbach[3], Bradley S Moore[1,2]* & Pieter C Dorrestein[1,2,4]*

Peptide natural products show broad biological properties and are commonly produced by orthogonal ribosomal and nonribosomal pathways in prokaryotes and eukaryotes. To harvest this large and diverse resource of bioactive molecules, we introduce here natural product peptidogenomics (NPP), a new MS-guided genome-mining method that connects the chemotypes of peptide natural products to their biosynthetic gene clusters by iteratively matching *de novo* tandem MS (MS^n) structures to genomics-based structures following biosynthetic logic. In this study, we show that NPP enabled the rapid characterization of over ten chemically diverse ribosomal and nonribosomal peptide natural products of previously unidentified composition from *Streptomycete* bacteria as a proof of concept to begin automating the genome-mining process. We show the identification of lantipeptides, lasso peptides, linardins, formylated peptides and lipopeptides, many of which are from well-characterized model *Streptomycetes*, highlighting the power of NPP in the discovery of new peptide natural products from even intensely studied organisms.

Peptide natural products (PNPs) are ubiquitous chemicals found in all life forms, where they have diverse biological functions in development, protection and communication[1]. Nature has evolved two orthogonal biosynthetic pathways to these highly modified peptides involving ribosomal and nonribosomal processes[2]. Although nonribosomal peptides have limited distribution, being restricted mainly to microorganisms with large genomes[3], ribosomally synthesized and post-translationally modified peptides seem to have a much broader distribution throughout nature, including being present in humans[4,5]. The enormous diversity and distribution of PNPs and their associated biological functions, however, are only now being fully realized because of time-consuming discovery options. We report here a new MS–guided genome mining method that quickly connects the chemotypes of expressed PNPs to their biosynthetic pathways, thereby enabling the rapid identification of transcriptionally active PNP biosynthetic gene clusters and the classification of their associated products in a streamlined discovery platform.

Among PNPs, ribosomally synthesized peptides encompass a rapidly expanding group of natural products[6]. Multiple classes of ribosomal peptide natural products (RNPs) of prokaryotic origin have been characterized through their biosynthetic pathways, which entail diverse post-translational modification strategies to yield lantipeptides[7], thiopeptides[8], cyanobactins[9], lasso peptides[10] and other microcins[11]. Consequently, traditional RNP classification systems based on bioactivity, producer and structure[11,12] have shifted toward a new classification based largely on biosynthesis (**Supplementary Results** and **Supplementary Table 1**). In RNP biosynthesis, the peptide sequence is encoded by a precursor gene directly translated by the ribosome to consist of leader peptide and core peptide regions[13]. The leader peptide serves as a scaffold and contains recognition sites for processing enzymes that introduce post-translational modifications of the RNP biosynthetic machinery, whereas the core peptide constitutes the primary sequence of the produced peptide natural product that is modified. Post-translational modification of the core peptide by biosynthetic enzymes can often be extensive and can provide a wealth of structural diversity rendering these peptides, at first glance, unrecognizable as ribosomally synthesized molecular entities[6] (**Fig. 1**). Nonribosomal peptides are conversely synthesized by multifunctional assembly line proteins that instead code for their amino acid precursors through an adenylating enzyme that selects and transfers its substrates to carrier proteins to facilitate peptide synthesis by the nonribosomal peptide synthetase (NRPS) machinery[14]. This process can capture a much wider array of substrates beyond the 20 proteinogenic amino acid building blocks, which limit input into RNPs, to produce notable examples such as the clinical agents vancomycin, daptomycin and cyclosporin[2] (**Fig. 1**).

To estimate the extent of PNP chemical diversity in bacteria, we systematically queried the Joint Genome Institute (JGI) database of 1,035 completed genomes as of September 2010 for RNP and NRPS pathways. Searching for gene clusters harboring characteristic RNP biosynthetic Pfam (protein family) domains[15], we estimate that at least 71% of the deposited bacterial genomes contain biosynthetic features that support common RNP classes (**Supplementary Table 3**). We identified 1,966 candidate RNP gene clusters, 637 of which have two or more of the nine Pfam domains found most frequently in RNP gene clusters (**Supplementary Table 2**). In comparison, 69% of the genomes we searched contained NRPS Pfam domains and 53% had hybrid NRPS-PKS biosynthetic features (**Supplementary Table 3**). Because the training set for our algorithm contained only 24 known RNP gene clusters, our estimate of RNPs is not comprehensive. Nonetheless, this analysis shows that the genetic capacity to produce RNPs is common in most microbial phyla and that RNPs represent one of the most underappreciated classes of bioactive molecules.

Given the sheer volume of predicted bacterial PNPs in publicly available genome strains, we set out to develop a method that takes advantage of recent technological advances in MS and genomics

[1]Center for Marine Biotechnology and Biomedicine, Scripps Institution of Oceanography, University of California at San Diego, La Jolla, California, USA. [2]Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California at San Diego, La Jolla, California, USA. [3]Department of Bioengineering and Therapeutic Sciences and California Institute of Quantitative Biosciences, University of California, San Francisco, California, USA. [4]Department of Pharmacology and Department of Chemistry and Biochemistry, University of California at San Diego, La Jolla, California, USA. *e-mail: pdorrestein@ucsd.edu or bsmoore@ucsd.edu.

NATURE CHEMICAL BIOLOGY DOI: 10.1038/NCHEMBIO.684 ARTICLE

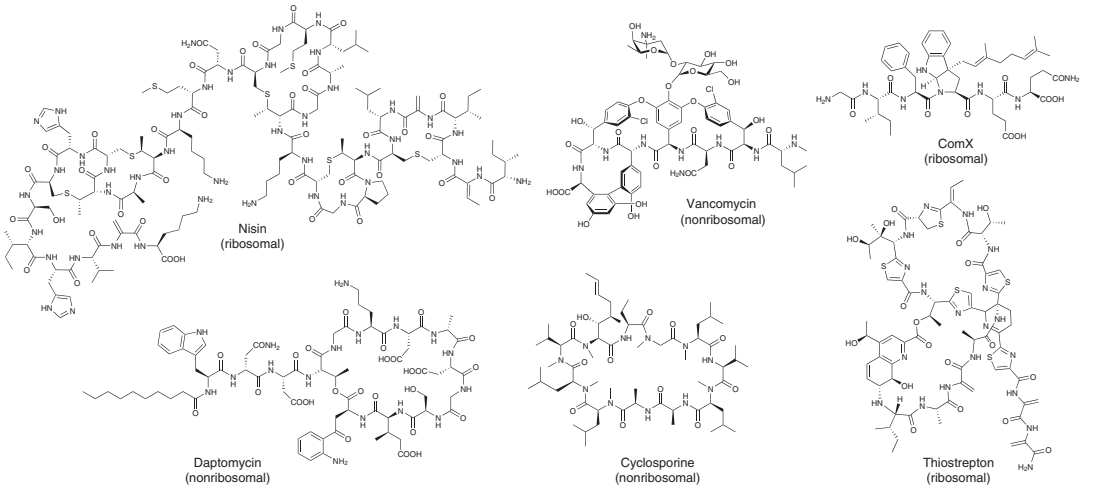


Figure 1 | Structural diversity of peptide natural products.

to streamline the discovery process. The recent development of genome mining has transformed natural product discovery by allowing the targeting of new chemical entities predicted by bioinformatics[16]. In the case of RNPs, a produced peptide structure can be directly linked to the corresponding biosynthetic genes by identifying the core peptide sequence in the translated genome sequence[4]. Furthermore, large portions of NRPs often readily correlate to the predicted amino acid specificity found on their associated modular synthetases[17] (**Supplementary Fig. 1**). This connection of PNP chemotype to genotype has been accomplished in numerous genome-mining studies[8–10,17–20]. One of the major limitations with these approaches is that they only characterize one molecule at a time or require extensive genetic manipulations[21]. With an increasing number of available genome sequences, there is a growing need for new genome-mining methods that can readily connect expressed natural products (chemotype) with their gene clusters (genotype) and that have the potential for automation.

MS is an important technique in the analysis of peptide natural products because of its high sensitivity, its easy implementation into automated processes such as metabolomic or proteomic platforms and its capability for *de novo* peptide structure elucidation by tandem MS[22]. Peptides fragment in MS^n experiments, for example, collision-induced dissociation (CID), in a common way to yield fragment ions in the MS^n spectrum that differ in mass by the amino acid monomers of the corresponding peptide sequence and, thus, enable *de novo* peptide sequencing. MS^n is used in proteomic workflows to identify proteins by connecting peptide MS^n data to protein sequence databases. One approach to link a proteolytic peptide to its database gene uses short *de novo* sequence tags for the database search[23]. However, automated *de novo* sequencing makes errors in one in every four amino acids, and this error rate is enhanced when post-translational modifications (PTMs) are included. In addition, database proteomic tools still struggle to connect modified RNPs with their precursor genes in genomic databases because of scoring functions, which have difficulty in recognizing many PTMs per peptide[24]. The scoring allows for a specific percentage of false-positive rates (usually 1–5%) without any further confirmation of a spectrum-peptide match (**Supplementary Table 4**). Finally, there are no tools that connect MS^n data of nonribosomally synthesized peptides to the corresponding NRPS genes. Given the advantage of MS to automatically acquire data of partial peptide structures from small amounts of material, it could enable a more rapid connection of peptidic natural products with their biosynthetic genes if MS^n data processing is effectively combined with genome mining of RNP and NRP biosynthetic pathways.

In this study, we establish the concept of MS–guided genome mining for peptide natural products called natural product peptidogenomics (NPP). We first highlight proof-of-concept experiments in which NPP characterizes the ribosomal lantipeptide AmfS from *Streptomyces griseus* IFO 13350, the nonribosomal lipopeptide stendomycin I from *Streptomyces hygroscopicus* ATCC 53653 and their corresponding biosynthetic gene clusters. In all, we show that NPP can be applied to characterize many PNP chemotypes and genotypes by introducing 14 new streptomycete PNPs in a very effective genome mining approach.

RESULTS

The NPP concept

NPP is an easy to implement and unbiased, MS-based, chemotype-to-genotype genome mining approach to rapidly characterize ribosomal and nonribosomal peptide natural products and their biosynthetic gene clusters from sequenced organisms (**Fig. 2**). In short, NPP aims to match a series of mass shifts obtained from an MS^n spectrum of a putative PNP to the genes that are responsible for its production. The NPP genome-mining workflow has several iteration steps, which ensure that a match of peptide MS^n data to a genomics-derived peptide structure makes sense biosynthetically. In this way, NPP takes advantage of the enormous wealth of knowledge of PNP biosynthesis gained over the past decade[2].

In practice, the NPP workflow starts with a MALDI-TOF MS analysis of the organism or extract in order to detect unknown masses. We targeted the mass range of 1,500–5,000 Da, as most masses in this window are not described in microbes at the chemical level, and thus they provided an opportunity to apply the NPP approach. However, there is no inherent limitation in size in the NPP approach as long as the MS^n data becomes a unique identifier for a biosynthetic pathway. MALDI-TOF MS analysis of crude butanol extracts or MALDI imaging of agar cultures ensure that the compounds are actively expressed and are captured on semi-solid media. Though not necessary for the PNP discovery process, MALDI imaging links secreted metabolites directly to the morphology of microbial colonies[25] and, thus, decreases potential media or extraction artifacts. Putative peptides are subsequently enriched using a MS-guided isolation using

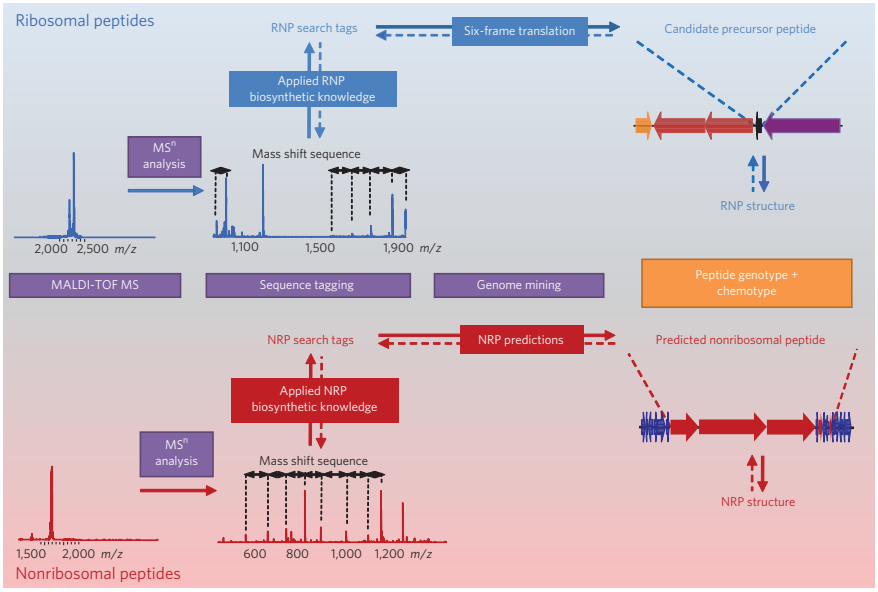


Figure 2 | General workflow of natural product peptidogenomics. NPP can be applied to characterize both ribosomal and nonribosomal peptide natural products in their genotype and chemotype from genome-sequenced organisms. Two proof-of-concept NPP experiments are outlined: ribosomal peptides (RNPs) or nonribosomal peptides (NRPs) and their respective biosynthetic gene cluster can be characterized from a *Streptomyces* extract by MALDI-TOF mass spectrometry (MS) detection, MS^n sequence tagging and PNP genome mining. The iterative approach in matching MS^n data to the genomics-derived peptide structures is shown with dashed arrows. See **Figures 3** and **4** for a detailed NPP analysis of ribosomal and nonribosomal peptides.

size-exclusion chromatography followed by enrichment and desalting steps. An enriched putative PNP is then analyzed in a sequence-tagging step by MS^n. In general, NPP sequence tagging is the formation of *de novo* sequence tags that are searchable in the genome-mining query space of PNPs (**Fig. 2**). This includes the generation of an amino acid sequence tag from a mass shift sequence in an MS^n spectrum and the subsequent processing of the MS^n sequence tag into search tags. The mass shifts define the candidate amino acid residues from all possible monomers that could be encoded in an RNP-based precursor gene or that could be loaded by a corresponding NRPS. This processing of MS^n mass shifts to genome-mining monomers considers PTMs, nonribosomal substrates, fragmentation gas-phase behavior and chemical modifications of amino acid residues during purification and MS analysis. NPP-based RNP genome mining interrogates the six-frame translation of the genome for candidate precursor peptides that comprise any of the search tags. As there may be multiple matches to a search tag that is 5–10 amino acids long, the correct RNP precursor gene is identified by applied biosynthetic knowledge in which the search tag should associate with the C-terminal half of a <100-amino-acid–long open reading frame (ORF) that clusters with RNP biosynthetic genes. NPP-based NRP genome mining, on the other hand, queries all predicted nonribosomal peptides of the target genome for the search tags.

The effectiveness of NPP in connecting PNP structures with biosynthetic genes lies in its iterative approach in matching MS^n-based structures to genomics-based candidate structures following biosynthetic logic, as each search tag match has to be confirmed in mass, sequence and biosynthetic signatures with the MS^n analysis (**Fig. 2**). We showed this effectiveness in a comparison of the NPP approach to current proteomic approaches in identifying precursor genes in RNP genome mining. None of the standard proteomic platforms such as Mascot[26] or InsPecT[23] could identify any of the NPP-characterized RNPs in a search with variable common RNP PTMs or in blind or unrestricted searches designed to find unknown PTMs (**Supplementary Table 4**). InsPecT was able to characterize two of the RNPs after predefining NPP-dissected PTMs in the analysis for each peptide.

NPP characterization of ribosomal peptide AmfS

As a proof of concept of the NPP workflow for RNPs, we targeted the known ribosomal peptide AmfS from *S. griseus* IFO 13350 because this is a well-characterized lantipeptide with four PTMs[27] (**Fig. 3**). MALDI imaging of *S. griseus* and MALDI-TOF MS analysis of an extract resulted in the detection of a secreted mass of 2,212 Da. We then subjected the peptide to CID fragmentation. In the MS^2 spectrum, we assigned the charge states of sequential fragment ions and identified the mass shift sequence 99-99-113-69-101 (**Fig. 3**). We matched the mass shifts to all likely candidate amino acids to yield sequence tags by first substituting with proteinogenic amino acids where possible (**Supplementary Table 5**). We then substituted nonproteinogenic masses with all possible RNP monomers arising from known PTMs. We substituted the shift of 69 Da to the nonproteinogenic amino acid dehydroalanine (Dha) (**Fig. 3** and **Supplementary Table 6**). Dha is a candidate amino acid for ribosomal peptides because dehydrated serine and threonine or dethiolated cysteine are commonly observed in PNP MS^n spectra either as a post-translational modification[28] or as an MS^n gas-phase rearrangement (**Supplementary Fig. 3**). From the resulting sequence tag VVI(L)S(C)T, we created a list of all possible search tags in both sequence directions to give eight putative PNP sequence tags for a search against the *S. griseus* genome sequence (**Fig. 3**)[29]. Of the millions of possible peptide sequences based on a six-frame translation, we identified just one candidate 43-amino-acid–long precursor by the search tag VVLCT (**Fig. 3**). This result fulfilled the

NATURE CHEMICAL BIOLOGY DOI: 10.1038/NCHEMBIO.684 ARTICLE

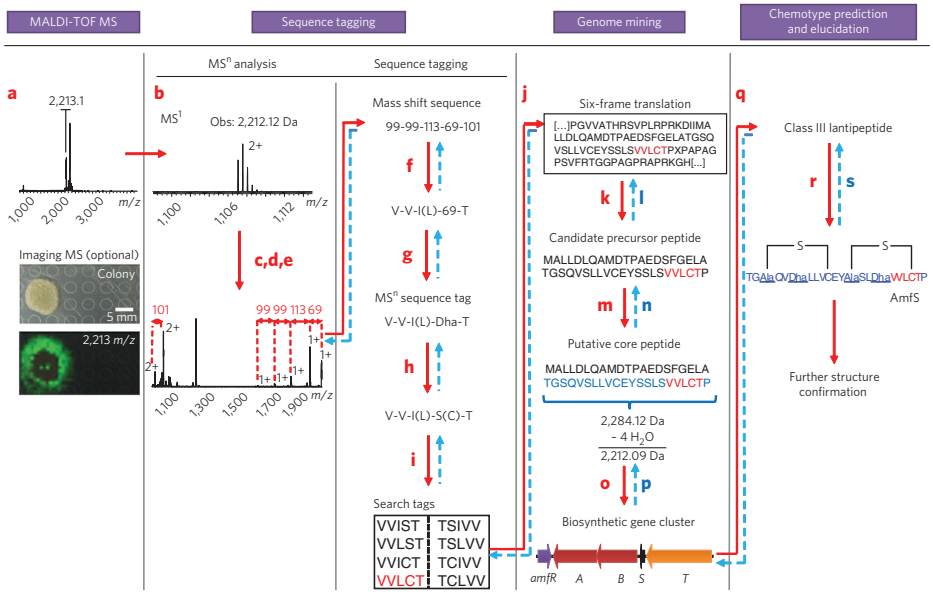


Figure 3 | Peptidogenomic connection of a RNP chemotype with its biosynthetic genes in the characterization of the class III lantipeptide AmfS from *S. griseus* IFO 13350. Analysis was carried out through sequence tagging and genome mining. Iterative aspects in connecting MS^n data of the peptide chemotype to the genotype are highlighted in blue and with the dashed arrows. The steps are as follows: (**a**) detection of putative peptide mass signals by MALDI-TOF MS or imaging MS, (**b**) determination of molecular weight, (**c**) MS^n fragmentation (CID), (**d**) assignment of charge states, (**e**) identification of mass shifts, (**f**) substitution of proteinogenic mass shifts (**Supplementary Table 5**), (**g**) substitution of nonproteinogenic mass shifts with putative RNP monomers (**Supplementary Table 6**), (**h**) MS^n sequence-tag processing of putative biosynthetic or MS gas-phase modifications, (**i**) MS^n sequence-tag processing of sequence tag direction, (**j**) search in six-frame translation of the target genome, (**k**) identification of candidate precursor peptide through RNP biosynthetic rationale, (**l**) verification of precursor peptide sequence, (**m**) prediction of core peptide sequence based on observed mass and putative PTM mass shifts, (**n**) verification of core peptide sequence and mass, (**o**) prediction of biosynthetic gene cluster, (**p**) verification of putative PTMs, (**q**) RNP classification, (**r**) structure prediction based on RNP class and MS^n data and (**s**) structure verification by MS^n data. Dha, dehydroalanine; Obs, observed.

RNP biosynthetic requirement of the search tag being located in the C-terminal half of a <100-amino-acid–long gene product. Next, we compared a predicted core peptide sequence in its calculated mass to the observed mass in the MS^1 spectrum considering putative PTMs such as, for example, dehydrations. The calculated mass of the 22-amino-acid–long core peptide ^{22}T-^{43}P differed by four putative PTM dehydrations from the observed mass (**Fig. 3**), which is in agreement with the formation of two Dha and two lanthionine bridges in AmfS[30]. In addition, the predicted core peptide sequence could be further verified at this step by comparison to the MS^n data. Subsequent BLAST analysis of the neighboring genes identified the remainder of the AmfS biosynthetic gene cluster[27] and, therefore, further verified the connection of RNP chemotype and genotype. Based on the gene cluster components, in particular the AmfS core peptide and the PTM-introducing enzymes AmfA and AmfB, we could characterize the analyzed peptide as a class III lantipeptide from known RNP biosynthetic gene clusters (**Supplementary Table 1**). Finally, we could verify an AmfS structure based on the given core peptide sequence, the MS^n data and the knowledge about AmfS-like lantipeptide PTMs[30] (**Fig. 3**, **Supplementary Fig. 2** and **Supplementary Table 1**). The proof-of-concept characterization of AmfS and its gene cluster highlights the effectiveness of the NPP workflow by its iterative utilization of MS^n data and genetic data to enable a peptidogenomic connection of a PNP chemotype with its genotype.

NPP characterization of nonribosomal lipopeptides

With a minor adjustment to the NPP workflow, we can also discover NRPs (**Fig. 4**). We exemplified this approach with a set of lipopeptides detected by MALDI imaging from a colony of *S. hygroscopicus* ATCC 53653 (**Fig. 4**). MS^n analysis of SHY-1628 yielded the sequence fragment 99-99-83-83-71-113-99-57-115 (**Fig. 4** and **Supplementary Fig. 4b**), which we first processed into the RNP workflow because most of its mass shifts corresponded to proteinogenic amino acids through the sequence tag V-V-83-83-A-I(L)-V-G. We substituted the 83-Da masses with Dhb, whose biosynthetic precursor is threonine (**Supplementary Table 6**). Although we queried the six-frame translation of *S. hygroscopicus* with the search tags VVTTAI(L)VG, we detected no precursor peptides based on the described biosynthetic requirements of RNPs. The inability to identify a precursor peptide from a long sequence tag suggested that the SHY-1628–based peptides could instead be a set of nonribosomal peptides. To explore this scenario, we revised the original sequence tag to include NRP-specific, nonproteinogenic residues for NRP genome mining (**Supplementary Table 7**). Hence, the 83-Da mass shifts could correspond to Dhb, *N*Me-Dha or homoserine lactone (HseL). We excluded HseL because of its common C-terminal NRP location. Dhb and *N*Me-Dha most likely derived biosynthetically from threonine and serine, respectively, during or after NRP assembly because of the enamine instability of putative Dha and Dhb monomers[31]. In addition, a Dhb mass shift could also derive from a MS gas-phase–induced ring-opening elimination of a threonine-macrolactone bond with the NRP C terminus[32]. We evaluated the VVT(S)T(S)AI(L)VG sequence tags (**Fig. 4**) against NRP sequences predicted by NP.searcher[33] and antiSMASH[34] algorithms that predict NRPS gene clusters and their NRP products from the genome supercontig. This analysis matched the reduced and full eight-amino-acid

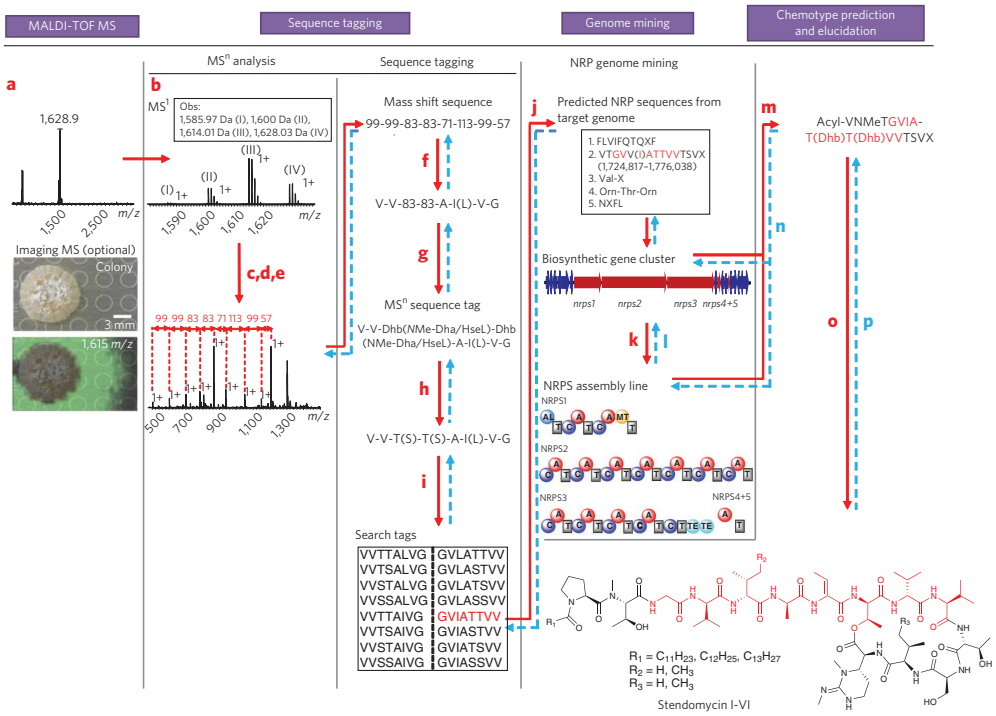


Figure 4 | Peptidogenomic connection of a NRP chemotype with its biosynthetic genes in the characterization of the lipopeptide stendomycin complex from *S. hygroscopicus* ATCC 53653. Analysis was carried out through sequence tagging and genome mining. Iterative aspects in connecting MS^n data of the peptide chemotype to the genotype are highlighted in blue and with the dashed arrows. The steps are as follows: (**a**) detection of putative peptide mass signals by MALDI-TOF MS or imaging MS, (**b**) determination of molecular weight, (**c**) MS^n fragmentation (CID), (**d**) assignment of charge states, (**e**) identification of mass shifts, (**f**) substitution of proteinogenic mass shifts (**Supplementary Table 5**), (**g**) substitution of nonproteinogenic mass shifts with putative NRP monomers (**Supplementary Table 7**), (**h**) MS^n sequence-tag processing of putative biosynthetic and MS gas-phase modifications, (**i**) MS^n sequence-tag processing of sequence tag direction, (**j**) search of predicted NRP sequences from the target genome with all search tags (NP.searcher or antiSMASH), (**k**) biosynthetic gene cluster analysis, (**l**) verification of predicted NRPS assembly line, (**m**) NRP structure prediction, (**n**) verification of predicted NRP structure, (**o**) full structure elucidation based on MS^n and NMR data and (**p**) verification of NRP structure. Dha, dehydroalanine; dhb, dehydrobutyrine; hseL, homoserine lactone; orn, ornithine; Obs, observed.

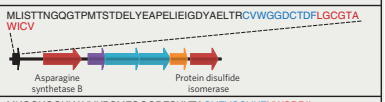
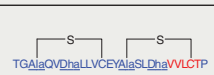
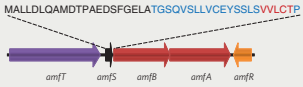
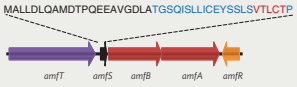
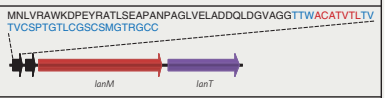
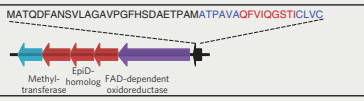
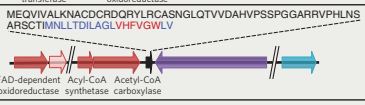
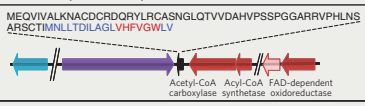
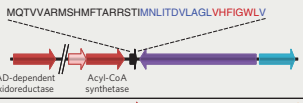
sequence tag to one candidate NRP sequence out of the five predicted NRPS sequences in the *S. hygroscopicus* genome (**Fig. 4** and **Supplementary Fig. 5**).

Again, through an iterative process, we inspected the corresponding gene cluster and found it to contain an N-terminal acyl ligase domain associated with lipopeptide biosynthesis, which is in full agreement with the observed 14-Da separation of the parent ions characteristic of lipopeptides. Further MS^n (**Supplementary Fig. 7b** and **Supplementary Tables 8–12**) and nuclear magnetic resonance (NMR) analysis (**Supplementary Fig. 6** and **Supplementary Table 13**) identified the lipopeptides as members of the stendomycin antibiotic family of lipotetradecapeptides that contain a seven-membered macrolactone and a total of seven modifications[35]. Aside from stendomycin I, which was originally characterized in *Streptomycetes endus*[35], we characterized five new stendomycin analogs (II–VI) that differed in the acyl chain and in valine or isoleucine substitutions at positions 5 and 13 for the first time in *S. hygroscopicus* ATCC 53653. The biosynthetic features of the identified gene cluster matched the structure of stendomycin I in the NRPS substrates and modifications (**Supplementary Figs. 5,8**). Thus, as we found for RNPs, the iteration between the MS^n analysis and genome mining enabled the fast and reliable connection of an NRP chemotype and genotype (**Fig. 4**). For example, we detected a low-resolution mass shift of 115 Da in the MS^2 spectrum of stendomycin I (**Supplementary Fig. 4b**) that was first assigned to aspartic acid. However, the corresponding module of the putative stendomycin NRPS instead predicts *N*Me-threonine (also a 115-Da shift) at this position and, thus, the mass shift in the MS^n spectrum could be explained. This example illustrates that in NRP sequence tagging, modifications such as N-methylations of proteinogenic masses and even nonproteinogenic masses should be considered if the first iterative round of NRP genome mining misses the assignment of the tag.

We also successfully applied the NPP method to other NRPS-derived molecules such as the structurally diverse calcium-dependent antibiotic[36], surfactin[37], plipastatin[37], pyoverdine[38] and daptomycin[36], and in each case, we identified the correct gene cluster (data not shown). Recently, the NPP workflow enabled the discovery of the arylomycin gene cluster with a sequence tag of just two amino acids[39]. This highlights the point that with NRPS-derived molecules, minimal sequence information can be sufficient to find a match in an NRP database of less than ten predicted NRP sequences per genome despite there being >526 known NRP monomers[40] because of the iterative nature of using biosynthetic knowledge in the workflow. To complete the structure analysis, additional analytical methods such

NATURE CHEMICAL BIOLOGY DOI: 10.1038/NCHEMBIO.684

Table 1 | NPP characterization of nine new RNPs and their associated gene clusters from seven genome-sequenced *Streptomyces* strains

Observed PNP	Class	Chemotype	Genotype
SSV-2083	Class I lasso peptide		
SRO15-2005	Class II lasso peptide		
SRO15-2212	Class III lantipeptide		
SAL-2242	Class III lantipeptide		
SRO15-3108	Class II lantipeptide	TTWACATVTLTVTVCSP TGTLCGSCSMGTRGCC (Core peptide – $9H_2O$)	
SGR-1832	Linaridin		
SLI-2138	N-formylated peptide		
SCO-2138	N-formylated peptide		
SWA-2138	N-formylated peptide		

Precursor peptide | Post-translationally modifying enzyme | Transporter | Protease | MreB | Regulator

Shown is a summary of the diverse RNP chemotypes and genotypes characterized by NPP in this study. Detailed analyses are described in **Supplementary Results.**

as NMR and Marfey's analysis are needed to complement the wealth of tandem MS and biosynthetic information, as has been done with stendomycins. The characterization of five stendomycin derivatives and their biosynthetic gene cluster in *S. hygroscopicus* shows that the NPP workflow can be readily accommodated to additionally discover modified NRPs.

NPP characterization of new RNP chemotypes and genotypes

Next, we interrogated several sequenced *Streptomycetes* to explore the practicality of NPP in the identification of other uncharacterized RNPs. From seven *Streptomyces* strains, we identified multiple previously uncharacterized RNPs and their gene clusters using the NPP approach (**Table 1**). The first unknown RNP and its gene cluster that we characterized by NPP was a class I lasso peptide, SSV-2083, from *Streptomyces sviceus* ATCC 20983 (**Table 1** and **Supplementary Fig. 9**). The discovery and isolation of secreted SSV-2083 from sporulating colonies was guided by MALDI imaging and MALDI-TOF MS of the ion at 2,084 m/z. An MS^n analysis of the unmodified compound provided no sequence information (**Supplementary Fig. 9b**). One of the main experimental challenges in the generation of the sequence tag is that many of these molecules are constrained by disulfide or thioether linkages, and therefore they provide poor to no fragmentation data (**Supplementary Fig. 10**). In such cases, samples are reductively dethiolated with $NaBH_4$ and $NiCl_2$ treatment[41] and resubjected to tandem MS to reveal longer sequence tags for PNP genome mining. Deconstrained SSV-2083 yielded a ten-amino-acid MS^n sequence tag that we identified in the six-frame translation of the *S. sviceus* genome in a 56-amino-acid candidate precursor peptide. This observation enabled the identification of the SSV-2083 biosynthetic gene cluster containing conserved lasso peptide biosynthetic genes as well as a new protein disulfide-isomerase–encoding gene (**Supplementary Fig. 9c**). Alignment with known class I lasso peptides in combination with tandem MS data (**Supplementary Fig. 9d**) enabled the prediction of the candidate SSV-2083 structure (**Table 1**), and these results represent the first class I lasso peptide gene cluster[42].

NPP characterization of new RNP classes from *Streptomyces*

Our NPP analysis also resulted in the discovery of two new RNP classes and their genetic origins from well-scrutinized

Streptomycetes, namely SGR-1832 (**Table 1** and **Supplementary Fig. 11**) from *S. griseus* IFO 13350 and SCO-2138 (**Table 1** and **Supplementary Fig. 12**) from *Streptomyces coelicolor* A3(2)[43]. Based on the gene cluster and the MS fragmentation data, we determined SGR-1832 to be a linear 19-residue peptide with an N-terminal *N*,*N*-dimethylalanine, two dehydrobutyrines and a rare C-terminal aminovinylcysteine (AviCys) residue. These unusual post-translational modifications are reminiscent of those seen in cypemycin, a related AviCys-containing linaridin from *Streptomyces* sp. OH-4156, whose biosynthesis was recently revealed by genome mining[18]. Peptide SCO-2138, detected only in organic extracts, is also a previously unidentified 19-amino-acid RNP from *S. coelicolor* A3(2) that produces a number of other peptide natural products[44]. The corresponding gene neighborhood containing a conserved unknown protein, a protease and a rod-shape–determining protein[45] is also found in other *Streptomyces* genomes (**Supplementary Fig. 12c**,**d**). Accordingly, we isolated and characterized two SCO-2138 homologs using NPP from *Streptomyces lividans* TK24 (SLI-2138, which is identical to SCO-2138; **Table 1** and **Supplementary Fig. 12**) and *S.* sp. E14 (SWA-2138, which is isomeric to SCO-2138; **Table 1** and **Supplementary Fig. 12**). These RNPs have a 28-Da N-terminal modification, which we confirmed by Fourier transform MS (FTMS^n) to be an *N*-formyl unit (**Supplementary Fig. 12e**). The SGR-1832 and SCO-2138 peptides represented undiscovered classes of RNPs at the time of this analysis and showcase that new RNP classes can be discovered by the NPP method.

Characterization of multiple PNPs in one NPP experiment

NPP analysis of the daptomycin-producing bacterium *Streptomyces roseosporus* NRRL 15998 (ref. 46) enabled the identification of three new RNPs and their gene clusters in a single NPP experiment (**Supplementary Fig. 13**). SRO15-2005 (**Table 1** and **Supplementary Fig. 14**) is a class II lasso peptide; SRO15-2212 (**Table 1** and **Supplementary Fig. 15**) is identical to the class III lantipeptide AmfS, which was previously uncharacterized in this strain; and SRO15-3108 (**Table 1** and **Supplementary Fig. 16**) is a class II lantipeptide that undergoes nine dehydrations during maturation. The detection of these three RNPs and their corresponding gene clusters in one NPP experiment shows the potential of NPP as a high-throughput discovery methodology.

DISCUSSION

In this work, we introduce NPP as a chemotype-to-genotype genome-mining approach for the characterization of ribosomal and nonribosomal peptide natural products and their respective biosynthetic gene clusters by identifying 14 peptides from well-known genome-sequenced *Streptomycetes*. In contrast to global metabolomic[47] and peptidomic[24] strategies, NPP is a targeted approach in which MALDI imaging or MALDI-TOF MS analysis of organic extracts is defined by a preselection of ions that are putative peptide natural products of expressed biosynthetic pathways. The innovation of NPP in efficiently linking these putative peptides to their gene clusters is firmly grounded in the connection of *de novo* MS^n peptide sequence tags of modified peptides to precursor peptides or to predicted NRPS products by applying biosynthetic knowledge and iterative steps between MS^n analysis and PNP genome mining for confirmation of putative chemotype-genotype matches. Because peptides are often structurally constrained, the generation of an MS^n sequence tag is facilitated by structural deconstraining the peptide before MS^n analysis. This yields simpler peptide structures and, thus, higher quality sequence tags, as in the case of the class I lasso peptide SSV-2083 (**Supplementary Fig. 9b**). Deconstraining also aids in the elucidation of post-translational modifications, such as the AviCys group of linaridin SGR-1832. In MS^n sequence-tag processing, the approach takes advantage of the degeneration of residues in the MS^n sequence tag by mass, reactions in the mass spectrometer, biosynthesis or sequence directionality to ensure that the resulting search tags can be found in genomics-derived peptide sequences (**Figs. 3** and **4**). In PNP genome mining, the sequence tags are searched against a query space that is different for RNPs and NRPs. In RNP genome mining, the query space is the six-frame translation of the target genome and, thus, is large. The sequence tag for effective genome mining of a precursor peptide in this large query space should be at least five amino acids; otherwise too many candidate precursor peptides will be obtained to be further differentiated based on RNP biosynthetic requirements. Several characterized precursor peptides that we identified in this study were not previously annotated in the NCBI database[48] (peptides SCO-2138 and SGR-1832). We found these peptides only in the six-frame translations of the *S. coelicolor* and *S. griseus* genome supercontigs. Although the drawback of an extended database providing more candidate precursor peptides for a certain sequence tag is a potential concern, this larger protein inference problem (as it is known in global proteomics[24]) is effectively solved in NPP by the iterative matching of the candidate precursor peptides in mass, sequence and biosynthetic signatures to the MS^n data.

MS^n sequence tag processing and the iterative MS^n and genomics analysis make the NPP *de novo* sequencing approach more effective in identifying precursor genes in RNP genome mining than current proteomic approaches. Neither Mascot[26] nor InsPecT[22] could identify any of the NPP-characterized RNPs in searches for unknown PTMs (**Supplementary Table 4**). InsPecT, which also relies on *de novo* sequence tagging, was able to characterize just two of the RNPs (SCO-2138 and SLI-2138), but only after we predefined NPP-characterized PTMs in the analysis. This is about what one would expect, as proteomic tools typically annotate 5–15% of the collected data, although in rare cases this percentage can be higher. The main reason that these programs do not work for these peptides is because their scoring functions have been designed to work for protease-cleaved, water-soluble peptides. Proteomic programs require specific scoring functions for specific PTMs (for example, specific for trypsin-cleaved ubiquitination tags or specific for phosphorylation) and simply have not been developed for RNP-based PTMs.

We further showed that NRPs are readily incorporated in the NPP workflow, as in the case of stendomycins (**Fig. 4**). Even though >50% of all amino acids in NRPs are L- or D-proteinogenic amino acids[40], mass shift sequences obtained from an MS^n spectrum define the candidate monomers to be used for the generation of all possible sequences that are to be compared to the predicted sequences based on the amino acid specificity of the adenylation domains using programs such as NRPSpredictor2 (ref. 49). In NRP genome mining, the query space consists of NRP megasynthetases predicted from the target genome by NP.searcher or antiSMASH and, thus, is relatively small, as most microbial genomes contain less than ten NRPS gene clusters. Consequently, short sequence tags of just two amino acids can be sufficient to correlate the NRP to its cognate NRPS gene cluster[39]. In the case of stendomycin, even though we ultimately applied the 8-amino-acid tag GVIATTVV, we could have functionally operated with and would have obtained the similar results with just a two-amino-acid tag such as VV, VI, TT, IA, AT or GV, as only one of the five *S. hygroscopicus* NRPS gene sets was appropriate in size and sequence. NRP sequences often contain modified and/or nonproteinogenic amino acid residues that can be addressed by including all appropriate nonproteinogenic monomers to a mass shift sequence and by considering their corresponding biosynthetic machineries during genome mining (**Supplementary Table 7**).

Because NPP is a MS-guided approach, it is ultimately dependent on generating quality sequence tags. The challenge in NPP characterization of peptides <500-Da or four-amino-acids long or less is in applying a limited sequence tag for genome mining rather than for dealing with matrix background in the low m/z region during peptide detection by MALDI-TOF MS. The analysis of putative peptides in the mass range <1,500 m/z will also increase the discovery

NATURE CHEMICAL BIOLOGY DOI: 10.1038/NCHEMBIO.684 ARTICLE

of PNPs, and in particular, of NRPs. NRPs with curated gene clusters in the NORINE database (which contains nonribosomal peptides) have an average mass of ~950 Da and eight monomers (**Supplementary Fig. 1**), whereas RNPs usually have a higher molecular weight, and NPP is appropriate for all such peptides. NPP, however, in its current implementation, is challenged by NRPs with multiple heterocycles, such as thiopeptides[8], and hybrid NRPS-PKS products with major polyketide portions. This will remain a challenge until the fragmentation rules are established. Another NPP restraint is the bioinformatics predictability of PNP sequences from inadequate genomic data in which poor sequence or annotation quality result in misassigned precursor and NRPS genes. Better genome assembly, improved gene annotation (especially of small ORFs), increased understanding of gas-phase fragmentation behaviors and deeper knowledge of NRPS substrate specificity codes will further empower the tools described in this work.

In conclusion, NPP is a new, MS-based genome-mining platform to guide the discovery of new ribosomal and nonribosomal peptides. This approach enables streamlined screening of peptide chemotypes from multiple organisms and facilitates expanded studies on their isolation, complete structure elucidation, biological evaluation and pathway engineering that leads to an increased appreciation for the understanding of the biological roles and therapeutic potential of peptide natural products. With further automatization of the NPP workflow such as training for offset functions of complex peptides, better understanding of MS fragmentation behaviors and the expansion to smaller masses and additional organisms, NPP has the potential to open up new research directions in the (bio)chemistry of peptide natural products.

METHODS

MALDI imaging of *Streptomyces* colonies. *Streptomyces* strains were grown on solid ISP2 medium (1 l of medium contained 4 g yeast, 10 g malt extract, 4 g dextrose and 20 g agar at pH 7) for 4–10 d at 28 °C until sporulation. *Streptomyces* spores from one plate were suspended in 1 ml sterile water and glycerol (3:1) and stored at –80 °C after inoculation. Thin-layer ISP2 agar plates of sporulating *Streptomyces* colonies were prepared as described elsewhere[25]. The applied matrix was a universal MALDI matrix (Sigma-Aldrich). MALDI imaging of *Streptomyces* samples on a Bruker MSP 96 anchor plate was performed on a Microflex Bruker Daltonics mass spectrometer outfitted with Compass 1.2 software suite (which consists of flexImaging 2.0, flexControl 3.0 and flexAnalysis 3.0). Target plate calibration was done as described elsewhere[25]. The sample was run in positive reflectron mode, with 800-µm laser intervals in XY. After the target-plate calibration was complete, the AutoXecute command was used to analyze the samples. The flexControl method we used had settings as previously described[25] with detection parameters adjusted as follows: mass range of 800–4,200 m/z and detector gain, reflector of 3.7–8.1. Mass calibration was accomplished using a peptide standard mix (Bruker Daltonics) as an external standard. After data acquisition, the data were analyzed using the flexImaging software. The resulting mass spectrum was analyzed manually for mass signals >1,500 m/z. Putative peptide mass signals >1,500 m/z were assigned with individual colors for a display of the distribution of the mass signal in the image.

Mass spectrometry analysis and sequence tagging. Peptide extraction, enrichment and preparation for MS analysis are described in the **Supplementary Methods**. Prepared peptide samples were injected for MS analysis by a nanomate-electrospray ionization robot (Advion) for consecutive electrospray into the MS inlet of a LTQ 6.4T Fourier transform ion cyclon resonance (FT-ICR) mass spectrometer (Thermo Finnigan). MS and MS^n data were acquired in the positive ion mode. FTMS data were acquired in 400–2,000 m/z scans. Selected peptide mass signals were manually isolated and fragmented by CID. MS^n data was collected either in ion trap or FT detection mode. All data were analyzed using QualBrowser, which is part of the Xcalibur LTQ-FT software package (Thermo Fisher). FTMS masses were analyzed using Extract software (Thermo Electron Bremen). Peptide MS^n sequence tags were assigned from MS^n data by manual *de novo* sequencing within the mass accuracy of the mass spectrometer using a mass shift list of proteinogenic amino acid monomers (**Supplementary Table 5**) and nonproteinogenic monomers (**Supplementary Tables 6**,**7**). Sequence tagging emphasized a correct assignment of 5–10-amino-acid MS^n sequence tags rather than longer, incorrect assignments for reliable genome mining. The MS^n sequence tag was further manually processed into a set of search tags depending on the degree of degeneration of the MS^n sequence tag. The MS^n sequence tag processing included differentiation of positions with identical masses (for example, isoleucine and leucine), positions with biosynthetic modifications (for example, Dha derived from serine or cysteine in RNPs; **Supplementary Tables 6**,**7**) and positions modified by MS analysis (for example, Dha derived from the cysteine of a lanthionine PTM or Dhb derived from the threonine of a macrolactone linkage). In $NaBH_4$ and $NiCl_2$-treated samples, positions were differentiated that might be chemically altered (for example, alanine derived from cysteine or alanine). The MS^n sequence tag was also differentiated in its reversed direction.

Genome mining of ribosomal peptides. A six-frame translated supercontig was searched with all possible RNP search tags from a given MS^n sequence tag in a standard text processing program. A candidate precursor peptide was defined in its N terminus by a pBLAST search of its C-terminal partial sequence to find homologs or was defined by reanalysis of the region in the supercontig in order to find missed alternative start codons that were not translated as methionine in the six-frame translation. A candidate precursor peptide was confirmed by (i) mass matching of putative core peptide sequence to the observed peptide mass by considering possible PTMs, (ii) sequence matching of the putative core peptide to the MS^n data and (iii) pBLAST analysis of the neighboring ORFs (gene cluster analysis). Based on the gene cluster components and the observed PTMs, an RNP class could usually be characterized (**Supplementary Table 1**). In cases of unusual gene cluster components during the RNP gene cluster analysis, a putative new RNP gene cluster could be defined by a search of homologous gene clusters (**Supplementary Figs. 11c**,**12c**). Finally, a structure of the RNP could be predicted based on the characterized core peptide sequence and PTMs that were characterized or predicted from the MS and bioinformatic analysis of the target peptide and its gene cluster.

Genome mining of nonribosomal peptides. A search tag that did not yield a candidate precursor peptide by six-frame translation-based genome mining was subjected to genome mining of NRP gene clusters. The mass shift sequence was reanalyzed by applying NRP monomer mass shifts (**Supplementary Table 7**) to characterize all possible NRP search tags. The supercontig of the target organism (for example, *S. hygroscopicus* ATCC 53653; **Supplementary Fig. 5**) was analyzed by NP.searcher[33] and by antiSMASH[34], and NRP search tags were compared to the predicted NRP sequences in monomers and in length. In case of a putative match, the corresponding NRP gene cluster was analyzed in its assembly-line organization in the corresponding antiSMASH output and by InterPro[50]. The accessibility of NRP families to genome mining by the NPP approach was assessed by an NP.searcher- and antiSMASH-based analysis of the GenBank files of characterized NRPS gene cluster families as described in the **Supplementary Methods**.

Additional methods. Bioinformatic prediction of PNP pathways, proteomic analysis of characterized RNPs and isolation and structure elucidation of Q027-1628 (stendomycin I) from the marine *Streptomyces* strain CNQ-027 are described in the **Supplementary Methods**.

Received 12 July 2011; accepted 27 August 2011;
published online 9 October 2011

References

1. Daffre, S. *et al.* Bioactive natural peptides. in *Studies in Natural Products Chemistry*, 1st edn., Vol. 35 (ed. Rahman, A.U.) 597–691 (Elsevier, 2008).
2. Nolan, E.M. & Walsh, C.T. How nature morphs peptide scaffolds into antibiotics. *ChemBioChem* **10**, 34–53 (2009).
3. Donadio, S., Monciardini, P. & Sosio, M. Polyketide synthases and nonribosomal peptide synthetases: the emerging view from bacterial genomics. *Nat. Prod. Rep.* **24**, 1073–1109 (2007).
4. Velásquez, J.E. & van der Donk, W.A. Genome mining for ribosomally synthesized natural products. *Curr. Opin. Chem. Biol.* **15**, 11–21 (2011).
5. Ganz, T. Defensins and host defense. *Science* **286**, 420–421 (1999).
6. Moore, B.S. Extending the biosynthetic repertoire in ribosomal peptide assembly. *Angew. Chem. Int. Edn Engl.* **47**, 9386–9388 (2008).
7. Willey, J.M. & van der Donk, W.A. Lantibiotics: peptides of diverse structure and function. *Annu. Rev. Microbiol.* **61**, 477–501 (2007).
8. Li, C. & Kelly, W.L. Recent advances in thiopeptide antibiotic biosynthesis. *Nat. Prod. Rep.* **27**, 153–164 (2010).
9. Donia, M.S., Ravel, J. & Schmidt, E.W. A global assembly line for cyanobactins. *Nat. Chem. Biol.* **4**, 341–343 (2008).
10. Duquesne, S. *et al.* Two enzymes catalyze the maturation of a lasso peptide in *Escherichia coli*. *Chem. Biol.* **14**, 793–803 (2007).
11. Duquesne, S., Petit, V., Peduzzi, J. & Rebuffat, S. Structural and functional diversity of microcins, gene-encoded antibacterial peptides from enterobacteria. *J. Mol. Microbiol. Biotechnol.* **13**, 200–209 (2007).
12. Cotter, P.D., Hill, C. & Ross, R.P. Bacteriocins: developing innate immunity for food. *Nat. Rev. Microbiol.* **3**, 777–788 (2005).
13. Oman, T.J. & van der Donk, W.A. Follow the leader: the use of leader peptides to guide natural product biosynthesis. *Nat. Chem. Biol.* **6**, 9–18 (2010).
14. Challis, G.L., Ravel, J. & Townsend, C.A. Predictive, structure-based model of amino acid recognition by nonribosomal peptide synthetase adenylation domains. *Chem. Biol.* **7**, 211–224 (2000).

15. Finn, R.D. *et al.* The Pfam protein families database. *Nucleic Acids Res.* **36**, D281–D288 (2008).
16. Winter, J.M., Behnken, S. & Hertweck, C. Genomics-inspired discovery of natural products. *Curr. Opin. Chem. Biol.* **15**, 22–31 (2011).
17. Lautru, S., Deeth, R.J., Bailey, L.M. & Challis, G.L. Discovery of a new peptide natural product by *Streptomyces coelicolor* genome mining. *Nat. Chem. Biol.* **1**, 265–269 (2005).
18. Claesen, J. & Bibb, M. Genome mining and genetic analysis of cypemycin biosynthesis reveal an unusual class of posttranslationally modified peptides. *Proc. Natl. Acad. Sci. USA* **107**, 16297–16302 (2010).
19. Li, B. *et al.* Catalytic promiscuity in the biosynthesis of cyclic peptide secondary metabolites in planktonic marine cyanobacteria. *Proc. Natl. Acad. Sci. USA* **107**, 10430–10435 (2010).
20. Kodani, S. *et al.* The SapB morphogen is a lantibiotic-like peptide derived from the product of the developmental gene *ramS* in *Streptomyces coelicolor*. *Proc. Natl. Acad. Sci. USA* **101**, 11448–11453 (2004).
21. Gressler, M., Zaehle, C., Scherlach, K., Hertweck, C. & Brock, M. Multifactorial induction of an orphan *PKS-NRPS* gene cluster in *Aspergillus terreus*. *Chem. Biol.* **18**, 198–209 (2011).
22. Ng, J. *et al.* Dereplication and *de novo* sequencing of nonribosomal peptides. *Nat. Methods* **6**, 596–599 (2009).
23. Tsur, D., Tanner, S., Zandi, E., Bafna, V. & Pevzner, P.A. Identification of post-translational modifications by blind search of mass spectra. *Nat. Biotechnol.* **23**, 1562–1567 (2005).
24. Duncan, M.W., Aebersold, R. & Caprioli, R.M. The pros and cons of peptide-centric proteomics. *Nat. Biotechnol.* **28**, 659–664 (2010).
25. Yang, Y.L., Xu, Y., Straight, P. & Dorrestein, P.C. Translating metabolic exchange with imaging mass spectrometry. *Nat. Chem. Biol.* **5**, 885–887 (2009).
26. Perkins, D.N., Pappin, D.J., Creasy, D.M. & Cottrell, J.S. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20**, 3551–3567 (1999).
27. Ueda, K. *et al.* AmfS, an extracellular peptidic morphogen in *Streptomyces griseus*. *J. Bacteriol.* **184**, 1488–1492 (2002).
28. McIntosh, J.A., Donia, M.S. & Schmidt, E.W. Ribosomal peptide natural products: bridging the ribosomal and nonribosomal worlds. *Nat. Prod. Rep.* **26**, 537–559 (2009).
29. Ohnishi, Y. *et al.* Genome sequence of the streptomycin-producing microorganism *Streptomyces griseus* IFO 13350. *J. Bacteriol.* **190**, 4050–4060 (2008).
30. Willey, J.M., Willems, A., Kodani, S. & Nodwell, J.R. Morphogenetic surfactants and their role in the formation of aerial hyphae in *Streptomyces coelicolor*. *Mol. Microbiol.* **59**, 731–742 (2006).
31. Wilkinson, B. & Micklefield, J. Biosynthesis of nonribosomal peptide precursors. *Methods Enzymol.* **458**, 353–378 (2009).
32. Romano, A., Vitullo, D., Di Pietro, A., Lima, G. & Lanzotti, V. Antifungal lipopeptides from *Bacillus amyloliquefaciens* strain BO7. *J. Nat. Prod.* **74**, 145–151 (2011).
33. Li, M.H., Ung, P.M., Zajkowski, J., Garneau-Tsodikova, S. & Sherman, D.H. Automated genome mining for natural products. *BMC Bioinformatics* **10**, 185 (2009).
34. Medema, M.H. *et al.* antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters. *Nucleic Acids Res.* **39**, W339–W346 (2011).
35. Bodanszky, M., Izdebski, J. & Muramatsu, I. Structure of the peptide antibiotic stendomycin. *J. Am. Chem. Soc.* **91**, 2351–2358 (1969).
36. Strieker, M. & Marahiel, M.A. The structural diversity of acidic lipopeptide antibiotics. *ChemBioChem* **10**, 607–616 (2009).
37. Roongsawang, N., Washio, K. & Morikawa, M. Diversity of nonribosomal peptide synthetases involved in the biosynthesis of lipopeptide biosurfactants. *Int. J. Mol. Sci.* **12**, 141–172 (2010).
38. Visca, P., Imperi, F. & Lamont, I.L. Pyoverdine siderophores: from biogenesis to biosignificance. *Trends Microbiol.* **15**, 22–30 (2007).
39. Liu, W.T., Kersten, R.D., Yang, Y.L., Moore, B.S. & Dorrestein, P.C. Imaging mass spectrometry and genome mining via short sequence tagging identified the anti-infective agent arylomycin in *Streptomyces roseosporus*. *J. Am. Chem. Soc.* (in the press).
40. Caboche, S., Leclere, V., Pupin, M., Kucherov, G. & Jacques, P. Diversity of monomers in nonribosomal peptides: towards the prediction of origin and biological activity. *J. Bacteriol.* **192**, 5143–5150 (2010).
41. Kawulka, K.E. *et al.* Structure of subtilosin A, a cyclic antimicrobial peptide from *Bacillus subtilis* with unusual sulfur to α-carbon cross-links: formation and reduction of α-thio-α-amino acid derivatives. *Biochemistry* **43**, 3385–3395 (2004).
42. Knappe, T.A., Linne, U., Xie, X. & Marahiel, M.A. The glucagon receptor antagonist BI-32169 constitutes a new class of lasso peptides. *FEBS Lett.* **584**, 785–789 (2010).
43. Bentley, S.D. *et al.* Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2). *Nature* **417**, 141–147 (2002).
44. Nett, M., Ikeda, H. & Moore, B.S. Genomic basis for natural product biosynthetic diversity in the actinomycetes. *Nat. Prod. Rep.* **26**, 1362–1384 (2009).
45. Vats, P. & Rothfield, L. Duplication and segregation of the actin (MreB) cytoskeleton during the prokaryotic cell cycle. *Proc. Natl. Acad. Sci. USA* **104**, 17795–17800 (2007).
46. Miao, V. *et al.* Daptomycin biosynthesis in *Streptomyces roseosporus*: cloning and analysis of the gene cluster and revision of peptide stereochemistry. *Microbiology* **151**, 1507–1523 (2005).
47. Koal, T. & Deigner, H.P. Challenges in mass spectrometry based targeted metabolomics. *Curr. Mol. Med.* **10**, 216–226 (2010).
48. Warren, A.S., Archuleta, J., Feng, W.C. & Setubal, J.C. Missing genes in the annotation of prokaryotic genomes. *BMC Bioinformatics* **11**, 131 (2010).
49. Röttig, M. *et al.* NRPSpredictor2—a web server for predicting NRPS adenylation domain specificity. *Nucleic Acids Res.* **39**, W362–W367 (2011).
50. Hunter, S. *et al.* InterPro: the integrative protein signature database. *Nucleic Acids Res.* **37**, D211–D215 (2009).

Acknowledgments

We thank N. Castellana and V. Bafna for providing the algorithm to enable the six-frame translations of supercontigs. We also thank M. Meehan for FTMS training. Financial support was provided by the US National Institutes of Health (GM085770 to B.S.M. and GM086283 to P.C.D.) and the Beckman Foundation.

Author contributions

R.D.K. designed and carried out experiments, analyzed data and wrote the paper. Y.-L.Y., Y.X. and S.-J.N. carried out experiments and analyzed data. P.C. and M.A.F. carried out the bioinformatic analysis and analyzed data. W.F. analyzed data. B.S.M. and P.C.D. designed experiments, analyzed data and wrote the paper.

Competing financial interests

The authors declare no competing financial interests.

Additional information

Supplementary information is available online at http://www.nature.com/naturechemicalbiology/. Reprints and permissions information is available online at http://www.nature.com/reprints/index.html. Correspondence and requests for materials should be addressed to P.C.D. or B.S.M.

SUPPLEMENTARY INFORMATION for

A mass spectrometry-guided genome mining approach for natural product peptidogenomics

Roland D. Kersten[1], Yu-Liang Yang[2], Yuquan Xu[2], Peter Cimermancic[3], Sang-Jip Nam[1], William Fenical[1,2], Michael A. Fischbach[3], Bradley S. Moore[1,2,*], Pieter C. Dorrestein[1,2,4,*]

Addresses and affiliations.

1) Center for Marine Biotechnology and Biomedicine, Scripps Institution of Oceanography, University of California at San Diego, La Jolla, California, USA.

2) Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California at San Diego, La Jolla, California, USA.

3) Department of Bioengineering and Therapeutic Sciences and California Institute of Quantitative Biosciences, University of California, San Francisco, USA.

4) Department of Pharmacology and Department of Chemistry and Biochemistry, University of California at San Diego, La Jolla, California, USA.

* to whom correspondence should addressed: pdorrestein@ucsd.edu or bsmoore@ucsd.edu.

SUPPLEMENTARY METHODS

Materials

All chemicals were purchased from Themo Fisher Scientific or Sigma-Aldrich unless otherwise noted. All *Streptomyces* strains except *Streptomyces griseus* IFO 13350 and *Streptomyces coelicolor* A3(2) were acquired from the Broad institute, MIT/Harvard, MA, USA. Amino acid standards of known configuration used to determine the absolute stereochemistry of CNQ27-1628 were purchased from Sigma-Aldrich.

***Streptomyces* strains used in this study**

Strain	Source
Streptomyces albus J1074	Broad (MIT/Harvard)
Streptomyces coelicolor A3(2)	Dorrestein lab, UCSD
Streptomyces roseosporus NRRL 15998	Broad (MIT/Harvard)
Streptomyces sviceus ATCC 29083	Broad (MIT/Harvard)
Streptomyces hygroscopicus ATCC 53653	Broad (MIT/Harvard)
Streptomyces lividans TK24	Broad (MIT/Harvard)
Streptomyces sp. E14 (WASP)	Broad (MIT/Harvard)
Streptomyces griseus IFO 13350	Moore lab (SIO-UCSD)

Bioinformatic prediction of peptide natural product pathways in microbial genomes

24 characteristic biosynthetic gene clusters of various RNP classes (SGR_6361-SGR_6366, SGR_2393-SGR_2397, SSGG_04019-SSGG_04023, SSGG_03858-SSGG_03862, SSEG_05172-SSEG_05180, SSGG_04748-SSGG_04752, SSPG_05081-SSPG_05096, SSTG_01176-SSTG_01182, SCO_2438-SCO_2451, SSHG_03588-SSHG_03592, GI: 225055344 (siomycin gene cluster), GI: 225055372 (thiostrepton gene cluster), GI: 78042196 (goadsporin gene cluster), GI: 186886571 (microviridin gene cluster), GI: 62910836 (patellamide gene cluster), GI: 215272911 (microcin B17 gene cluster), GI: 4731431 (microcin J25 gene cluster), GI: 682763 (microcin C7 gene cluster), GI: 202072960 (agr gene cluster), SPy_0738-SPy_0747 (streptolysin S gene cluster), GI: 46964 (epidermin gene cluster), GI: 299832736 (nisin gene cluster), GI: 300719225 (microbiosporicin gene cluster), GI: 20387045 (subtilosin gene cluster) were analyzed by a HMMER-based algorithm to generate Pfam domain counts in these gene clusters. The HMMER-based analysis of 24 characterized RNP biosynthetic gene clusters yielded 44 Pfam domains (**Supplementary Table 2**). Selected biosynthetic Pfam domains with multiple counts (**Supplementary Table 2,** red) were used to search all completed microbial genomes in the JGI database and the draft *Streptomyces* genomes of the Broad database (1035 total genomes) by a Hidden Markov Model (HMM)-based gene cluster identification algorithm for RNP gene clusters. A putative RNP gene cluster was assigned when ≥2 of the 9 Pfam domains most frequently found in RNP gene clusters were

present in a cluster with a threshold of >0.5. Because the training set was limited to 24 biosynthetic gene clusters representing common RNP classes, we anticipate that the predicted abundance of RNP pathways by this method is an underestimate. For the identification of NRP gene clusters in these genomes, a set of 137 known NRP gene clusters was used to train our HMM-based gene cluster ID algorithm. To estimate the number of NRP gene clusters, the predicted NRP gene clusters were then filtered with a threshold of >0.5.

Peptide extraction and enrichment

25-80 ISP2 agar plates were inoculated with spore suspension of a target *Streptomyces* strain by 4 parallel streaks. The plates were incubated for 4-10 d at 28 °C until sporulation. Then, the agar was sliced into small pieces and extracted with n-butanol for 12 h at 28 °C and 225 rpm in a 2.8 liter-Erlenmeyer flask. The n-butanol extract was separated from agar by cheesecloth filtration and, subsequently, from cell debris by centrifugation (6440 g, 10 min). N-butanol was removed with a rotovaporator (Buechi R-200). The crude extract was resuspended in 1 ml methanol and analyzed by MALDI-TOF MS to detect target putative peptides. The resuspended *Streptomyces* extract was centrifugated for 2 min at 16800 g and the supernatant was placed on a methanol-equilibrated Sephadex LH20 column (length: 30 cm). The extract was separated with a methanol mobile phase at a flow rate of 0.4 ml/min. 7.5 ml fractions were collected, lyophilized and resuspended in 200 µl methanol. Gel filtration fractions were then analyzed by MALDI-TOF MS for peptide content. Peptide-comprising gel filtration fractions were desalted and concentrated either by HPLC (Agilent) or by C18 Ziptips (Millipore) according to a modified manufacturer's protocol in which peptides were eluted with 85 % acetonitrile containing 1 % formic acid. Desalted and concentrated peptide fractions were mixed 1:1 with 50 % methanol containing 1 % formic acid and subjected to FTMS analysis and sequence tagging by tandem MS analysis.

MALDI-TOF MS analysis of peptide samples

MALDI-TOF MS was used to detect target peptides in crude extracts and fractions of gel filtration and HPLC. The sample was mixed 1:1 with a saturated solution of Universal MALDI matrix in 70 % acetonitrile containing 0.1 % TFA and spoted on a Bruker MSP 96 anchor plate . The sample was dried and inserted into the Microflex Bruker Daltonics mass spectrometer. External calibration was done as described for MALDI-imaging. Mass spectra were obtained with the FlexControl method as used for MALDI-imaging and a single spot acquisition of 200 shots. Single spot MALDI-TOF MS data was analyzed by FlexAnalysis software.

Sample preparation for sequence tagging

Desalted and concentrated peptides were resuspended in 50 µl 50 % methanol containing 1 % formic acid if they were analyzed directly and unmodified by ESI-MS. For DTT reduction, 0.05 µmol peptide was dissolved in 20 µl methanol, added with DTT to give a concentration of 0.5 µmol DTT and incubated for

10 min at 90 °C. The reduced peptide was subsequently desalted by C18 Ziptips and eluted and prepared for ESI-MS analysis as described above. For $NaBH_4/NiCl_2$–reaction, 0.05 µmol peptide was added with 0.05 µmol $NaBH_4$ and 0.05 µmol $NiCl_2$ in 50 µl methanol and incubated for 5 min at 50 °C. After 5 min, another 0.05 µmol $NaBH_4/NiCl_2$ was added and incubation at 50 °C was repeated. This was repeated a third time before MS sample preparation. The $NaBH_4/NiCl_2$- and DTT-treated peptides were desalted by C18 Ziptips and eluted and prepared for ESI-MS analysis as described above.

Genome mining accessibility of nonribosomal peptides

The accessibility of NRP families to genome mining by the NPP approach was assessed by NP.searcher[1] and NRPSpredictor2[2,3]-based analysis of corresponding gene cluster files and comparison of the predicted NRP sequences with the known NRP structure. Single representatives of NRP families from the NORINE database[4] with curated biosynthetic gene cluster files in the GenBank database were analysed (**Supplementary Fig. 1**). An NRP was marked accessible by NPP (yellow or orange bars) when at least 2 adjacent monomers of the structure were predicted correctly by NP.searcher or NRPSpredictor2 from the gene cluster (red). This sequence tag is usually sufficient to characterize a peptide in a query space of maximum 10 peptides with an average length of 8 amino acids. First, only proteinogenic amino acids were considered in the analysis (**Supplementary Table 5**, yellow bar) and, second, they were extended by the nonproteinogenic amino acid characteristic for NRPs (**Supplementary Table 7** and NORINE[4], orange bar) for NP.searcher- and NRPSpredictor2-based genome mining.

Peptide nomenclature

New peptides were named by the first letter of the producer genus and the two initial letters of the producer species and the observed mass, e.g. **SSV-2083** for class I lassopeptide with the mass **2083** Da isolated from ***S**treptomyces **sv**iceus* ATCC 20983. In case of different strains of a targeted *Streptomyces* species, the first 2 strain ID numbers were added, e.g. **SRO15-2005** for class II lassopeptide with the mass **2005** Da from ***S**treptomyces **ro**seosporus* NRRL **15**998.

Proteomic analysis of characterized RNPs

The Thermo-RAW files of the MS/MS experiments of each peptide were converted to mzXML-files by ReAdW software. For InsPecT[5] analysis, mzXML-files of the MS^n data were analyzed in blind search mode or with specified modifications for each peptide as dissected by NPP. Each search was done either in the annotated Broad protein database or with the 6-frame translations of the supercontigs of each target species. For InsPecT analysis, specified modifications (mod,[MASS in Da],[RESIDUES],[TYPE],[NAME]) were as follows: SRO15-2005: none, SRO15-2212: mod,-18,ST,opt,dehydration; mod,-34,C,opt,dethiolation (5 = number of allowed PTMs), SRO15-3108: mod,-18,ST,opt,dehydration; (9), SGR-1832: mod,-18,ST,opt,dehydration; mod,-34,C,opt,dethiolation; mod,+28,ACDEFGHILMNPQRSTVWY,nterminal,dimethylation; (5), SGR-2212: mod,-

18,ST,opt,dehydration; mod,-34,C,opt,dethiolation; (5), SAL-2242: mod,-18,ST,opt,dehydration, mod,-34,C,opt,dethiolation; (5), SCO-2138: mod,-18,ST,opt,dehydration; mod,-34,C,opt,dethiolation; (5), SLI-2138: mod,+28,M,nterminal,formylation; mod,+16,M,opt,oxidation; (2), SWA-2138: mod,+28,M,nterminal,formylation; mod,+16,M,opt,oxidation; (2), SSV-2083: mod,-32,C,opt,sulfur-loss Cys; (4). Each search was done either with the annotated Broad protein database of each target species and or with the 6-frame translations of the supercontigs of each target species. For Mascot[6] analysis, mgf-files of the MS^n data were analyzed by unrestricted Mascot search with common RNP modifications as variable modifications in the SwissProt-Actinobacteria protein database and in the NCBInr-Actinobacteria protein database. For Mascot analysis, the mzXML-files were converted to mgf-files by MzXML2Search software from the Transproteomic Pipeline package. Variable modifications in the Mascot as common RNP modifications were (Cys->Dha (C),Dehydrated (S),Dehydrated (T),Dimethyl (N-term),Formyl (N-term),Oxidation (M)).

Cultivation, extraction and isolation of Q027-1628 complex (stendomycins) from marine *Streptomyces* strain CNQ-027

Marine *Streptomyces* sp. (strain CNQ-027), isolated from a marine sediment sample collected off the coast of San Diego, CA, was cultured in forty 2.8 L Fernbach flasks each containing 1 L of a seawater-based medium (10 g starch, 4 g yeast extract, 2 g peptone, 1 g $CaCO_3$, 40 mg $Fe_2(SO_4)_3 \times 4H_2O$, 100 mg KBr) and shaken at 230 rpm at 27 °C. After seven days of cultivation, sterilized XAD-16 resin (20 g/L) was added to adsorb the organic substances, and the culture and resin were shaken at 215 rpm for 2 hours. The resin was filtered through cheesecloth and washed with deionized water, and eluted with acetone. The acetone was removed under reduced pressure, and the resulting aqueous layer was extracted with ethyl acetate (3 × 500 mL). The ethyl acetate-soluble fraction was dried *in vacuo* to yield 4.5 g of crude material from a 40 L culture. The crude extract (4.5 g) from strain CNQ-027 was fractionated by open column chromatography on silica gel (25 g) eluting with a step gradient of dichloromethane and methanol. The dichloromethane/methanol 10:1 fraction contained a mixture of stendomycins, which were purified by reversed-phase HPLC (Phenomenex Ultracarb C30, 5 µm, 100 Å, 250 × 100 mm, 2.0 ml/min, UV = 210 nm), eluting with 90% CH_3CN in H_2O to afford Q027-1628 (200.0 mg), as a pale yellow oil. The purity of the compound was verified by ^{1}H-NMR (**Supplementary Fig. 6a**).

NMR measurement of Q027-1628 (stendomycin I)

Proton and 2D NMR spectra of Q027-1628 were recorded in methanol-d_4 containing Me_4Si as internal standard on Varian Inova spectrometers at 500 or 600 MHz. ^{13}C NMR spectra were acquired on Varian Inova spectrometers at 75 MHz.

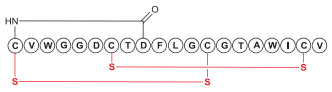
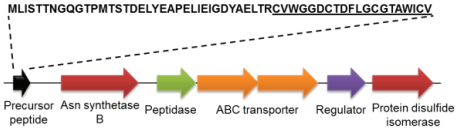
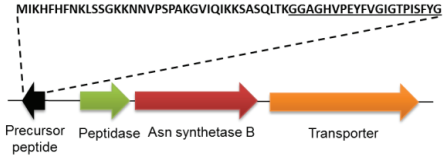
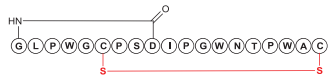
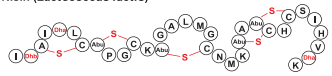
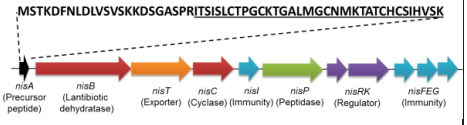
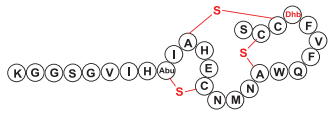
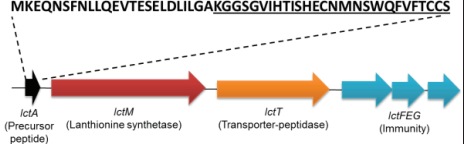
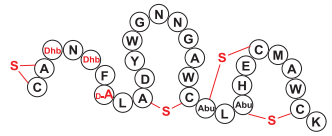
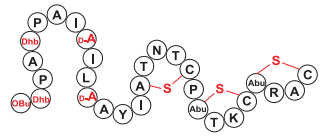
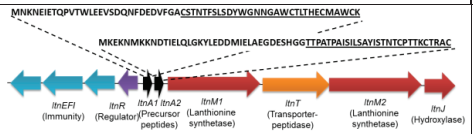
Determination of absolute configuration of Q027-1628 (stendomycin I)

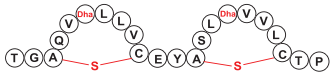
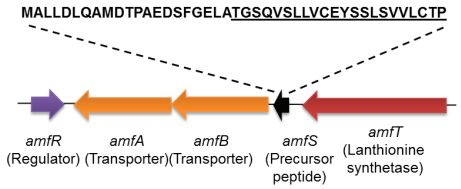
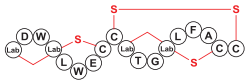
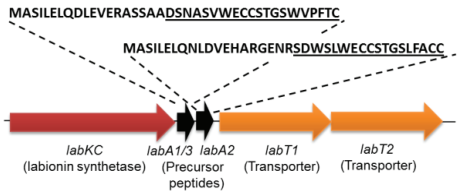
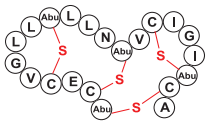
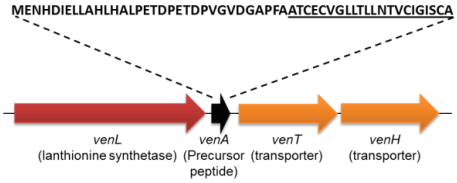
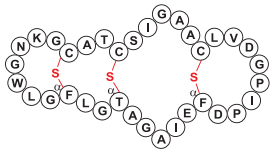
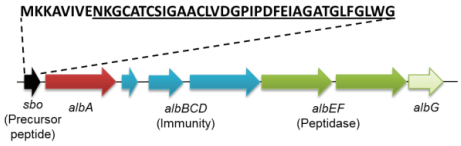
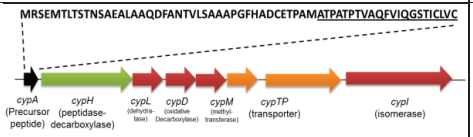
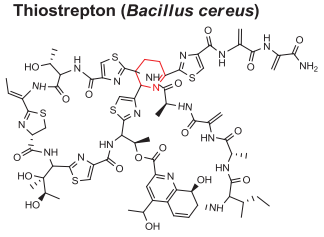
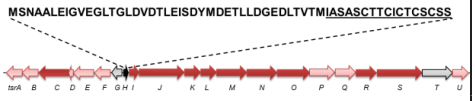
Q027-1628 (1.0 mg) was hydrolyzed in 0.4 mL of 6 N HCl at 110 ^{o}C for 12 h; the HCl was removed under vacuum; and the dry material was resuspended in 0.8 mL of water and dried three times to remove

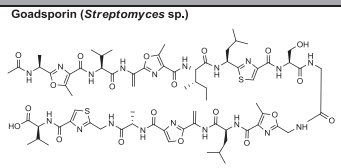
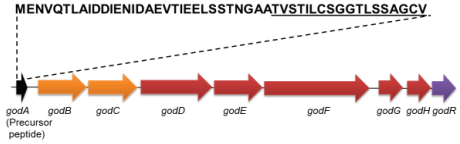
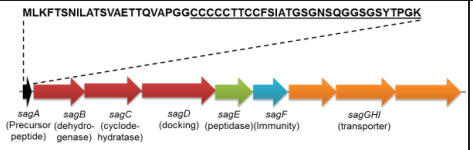
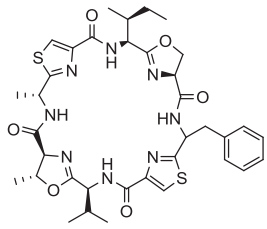
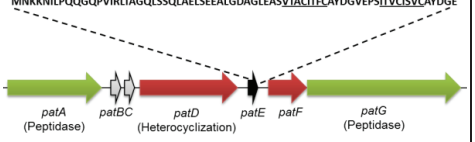
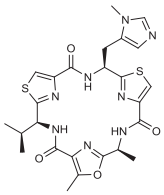
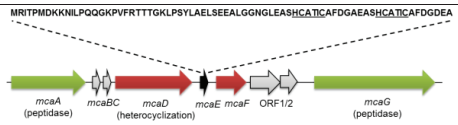
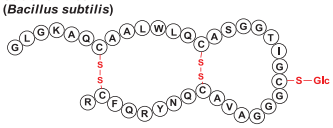
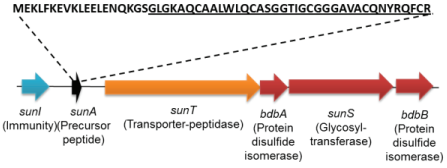
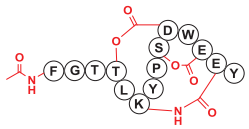
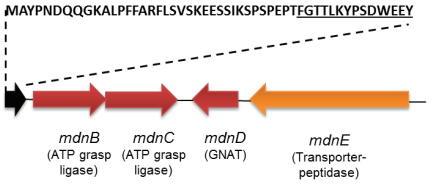
residual acid. The hydrolysate was divided into two portions and dissolved in 1 N $NaHCO_3$ (200 µL). To the first portion was added 100 µL of 1% D-FDLA (1-fluoro-2,4-dinitrophenyl-5-D-leucinamide) in acetone; to the second portion was added 25 µl of 1% D- and L-FDLA mixtures in acetone. The reaction mixtures were incubated at 60 ^{o}C for 30 min, and quenched with 200 µL of 2 N HCl. A small portion (200 µL) of CH_3CN was added to the solution to dissolve the reaction mixture. The resulting products were analyzed by reversed-phase LC-MS with a gradient solvent system (95 % aqueous CH_3CN to 50% aqueous CH_3CN (0.1% TFA) over 50 min; 50% aqueous CH_3CN to 100% aqueous CH_3CN (0.1% TFA) over 20 min; 100% CH_3CN (0.1% TFA) for 15 min). Low resolution LC-MS data were obtained using a Hewlett-Packard series 1100 LC/MS system with a reversed-phase C_{18} column (Phenomenex Luna, 4.6 mm × 100 mm, 5 µm) at a flow rate of 0.7 mL/min. D-FDLA-derivatized amino acids were detected by absorption at 340 nm. The retention times (min) of the derivatives were compared with those of authentic derivatized standards, with the exception of N-methyl-L-threonine and stendomycidine: L-Pro (42.6), L-Val (52.3), L-*allo*-Ile (55.4), L-Ala (44.3), L-Thr (39.7), L-*allo*-Thr (37.2), L-Ser (35.2); D-Pro (39.0), D-Val (44.0), D-*allo*-Ile (47.4), D-Ala (39.7), D-Thr (34.0), D-*allo*-Thr (35.2), D-Ser (34.6). The D-FDLA derivative of N-Me-Thr in hydrolyzed Q027-1628 was detected at 37.5 min, while the L- and D-FDLA derivatives of N-Me-Thr in hydrolyzed Q027-1628 were detected at 37.5 and 34.0 min, respectively. The elution order was assumed to be the same as observed for Thr. The absolute configuration of N-Me-Thr was determined as *S* (N-Me-L-Thr). Also, The D-FDLA derivative of stendomycidine (Ste) in hydrolyzed Q027-1628 was detected at 31.5 min, while the L- and D-FDLA derivatives of stendomycidine in hydrolyzed Q027-1628 were detected at 31.5 and 30.5 min, respectively. These elution orders suggested that the absolute configuration of stendomycidine is *S*. Our analysis showed that Q027-1628 possessed L-Pro, N-Me-L-Thr, D-Val, D-*allo*-Ile, D-Ala, D-*allo*-Thr, L-Val, L-Ser, and L-Ste. This aa configuration matched to stendomycin I stereochemistry. Due to this overlap, the stereochemistry of CNQ27-1628 was assigned as the one of stendomycin I[7] (**Supplementary Fig. 7a**).

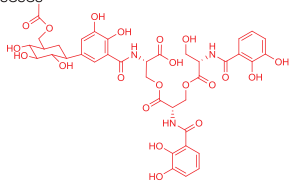
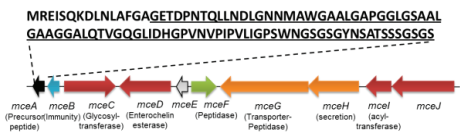
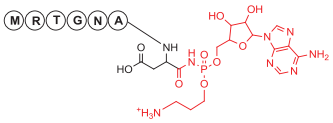
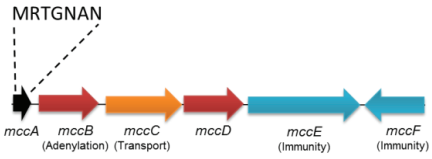
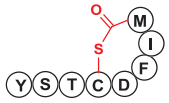
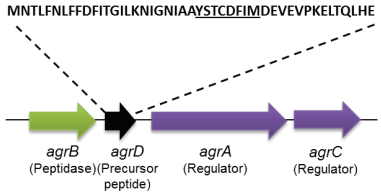
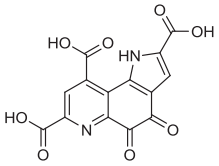
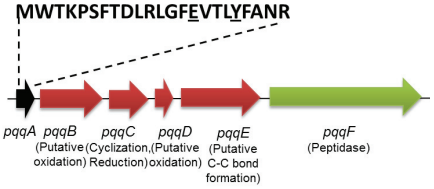
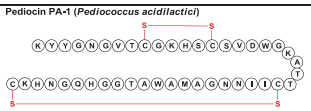
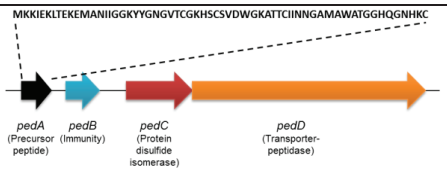
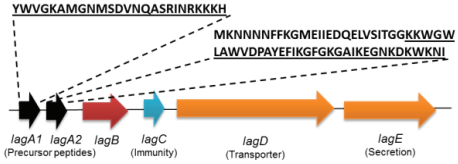
SUPPLEMENTARY RESULTS

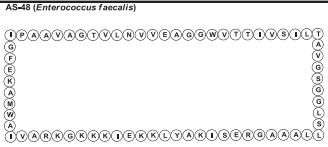
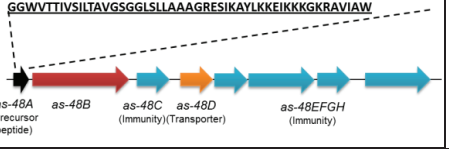
Supplementary Table 1: Current classification of prokaryotic RNPs based on biosynthetic gene clusters. PTMs in RNP chemotypes are highlighted in red. RNP genotypes show gene cluster organization, gene annotations and PPP sequences (core peptide sequence underlined) for each corresponding RNP chemotype. Genotype colors are: black – precursor peptide, red – post-translational modifying enzyme, green – peptidase, orange – transporter, purple – regulator, blue – immunity protein/transporter.

RNP class	Subclass	Chemotype	Genotype	Ref.
Lasso peptides	I	SSV-2083 (*Streptomyces sviceus*) 	MLISTTNGQGTPMTSTDELYEAPELIEIGDYAELTRCVWGGDCTDFLGCGTAWICV 	this study
	II	MccJ25 (*Escherichia coli*) 	MIKHFHFNKLSSGKKNNVPSPAKGVIQIKKSASQLTKGGAGHVPEYFVGIGTPISFYG 	[8]
	III	BI-32169 (*Streptomyces* sp.) 	Uncharacterized	[9]
Lantipeptides	I	Nisin (*Lactococcus lactis*) 	MSTKDFNLDLVSVSKKDSGASPRITSISLCTPGCKTGALMGCNMKTATCHCSIHVSK 	[10]
	II	Lacticin 481 (*Lactococcus lactis*) 	MKEQNSFNLLQEVTESELDLILGAKGGSGVIHTISHECNMNSWQFVFTCCS 	[10]
	II (two-peptide)	Lacticin 3147 A1 (*Lactococcus lactis*)  Lacticin 3147 A2 (*Lactococcus lactis*) 	MNKNEIETQPVTWLEEVSDQNFDEDVFGACSTNTFSLSDYWGNNGAWCTLTHECMAWCK MKEKNMKKNDTIELQLGKYLEDDMIELAEGDESHGGTTPATPAISILSAYISTNTCPTTKCTRAC 	[10]

RNP class	Subclass	Chemotype	Genotype	Ref.
	III	AmfS (*Streptomyces griseus*) 	MALLDLQAMDTPAEDSFGELATGSQVSLLVCEYSSLSVVLCTP 	[11]
		Labyrinthopeptin A2 (*Actinomadura namibiensis*) 	MASILELQDLEVERASSAADSNASVWECCSTGSWVPFTC MASILELQNLDVEHARGENRSDWSLWECCSTGSLFACC 	[12]
	IV	Venezuelin (*Streptomyces venezuelae*) 	MENHDIELLAHLHALPETDPETDPVGVDGAPFAATCECVGLLTLLNTVCIGISCA 	[13]
	cyclic	Subtilosin (*Bacillus subtilis*) 	MKKAVIVENKGCATCSIGAACLVDGPIPDFEIAGATGLFGLWG 	[14]
Linaridins	-	Cypemycin (*Streptomyces* sp.) 	MRSEMTLTSTNSAEALAAQDFANTVLSAAAPGFHADCETPAMATPATPTVAQFVIQGSTICLVC 	[15]
Thiopeptides	cyclic	Thiostrepton (*Bacillus cereus*) 	MSNAALEIGVEGLTGLDVDTLEISDYMDETLLDGEDLTVTMIASASCTTCICTCSCSS 	[16]-[19]

RNP class	Subclass	Chemotype	Genotype	Ref.
	linear	Goadsporin (*Streptomyces* sp.) 	MENVQTLAIDDIENIDAEVTIEELSSTNGAATVSTILCSGGTLSSAGCV 	[20]
Thiazole/ Oxazole-modified microcins	-	Streptolysin S (*Streptococcus pyogenes*) CCCCCTTCCFSIATGS GNSQGGSGSYTPGK (core peptide)	MLKFTSNILATSVAETTQVAPGGCCCCCTTCCFSIATGSGNSQGGSGSYTPGK 	[21]
Cyanobactins	Patellamides	Patellamide C (*Prochloron* sp.) 	MNKKNILPQQGQPVIRLTAGQLSSQLAELSEEALGDAGLEASVTACITFCAYDGVEPSITVCISVCAYDGE 	[22]
	Micro-cyclamides	Microcyclamide (*Microcystis aeruginosa*) 	MRITPMDKKNILPQQGKPVFRTTTGKLPSYLAELSEEALGGNGLEASHCATICAFDGAEASHCATICAFDGDEA 	[23]
S-glyco-peptides	-	Sublancin 168 (*Bacillus subtilis*) 	MEKLFKEVKLEELENQKGSGLGKAQCAALWLQCASGGTIGCGGGAVACQNYRQFCR 	[24]
Microviridins	-	Microviridin B (*Microcystis aeruginosa*) 	MAYPNDQQGKALPFFARFLSVSKEESSIKSPSPEPTFGTTLKYPSDWEEY 	[25]

RNP class	Subclass	Chemotype	Genotype	Ref.
Siderophore-RNPs (microcin)	-	**MccE492m (*Klebsiella pneumoniae*)** GETDPNTQLLNDLGNNMAWGAALGAPGGLGSAALGAAGG ALQTVGQGLIDHGPVN VPIPVLIGPSWNGSGSGYNSAT SSSGSGS 	MREISQKDLNLAFGAGETDPNTQLLNDLGNNMAWGAALGAPGGLGSAAL GAAGGALQTVGQGLIDHGPVNVPIPVLIGPSWNGSGSGYNSATSSSGSGS 	[26]
Nucleoside-RNPs (microcin)	-	**MccC7 (*Escherichia coli*)** 	MRTGNAN 	[27]
Autoinducing peptides	-	**Autoinducing peptide-I (*Staphylococcus aureus*)** 	MNTLFNLFFDFITGILKNIGNIAAYSTCDFIMDEVEVPKELTQLHE 	[28]
PQQ		**PQQ (P*seudomonas aeruginosa*)** 	MWTKPSFTDLRLGFEVTLYFANR 	[29]
Class II bacteriocins	**IIa (pediocin-like)**	**Pediocin PA-1 (*Pediococcus acidilactici*)** 	MKKIEKLTEKEMANIIGGKYYGNGVTCGKHSCSVDWGKATTCIINNGAMAWATGGHQGNHKC 	[30]
	IIb (two-peptide)	**Lactococcin G (*Lactococcus lactis*)** **a) GTWDDIGQGIGRVAYWVGKAM GNMSDVNQ ASRINRKKKH** **b) KKWGWLAWVDPAYEFIKGF GKGAIKEGNKDKWKNI**	MKELSEKELRECVGGGTWDDIGQGIGRVA YWVGKAMGNMSDVNQASRINRKKKH MKNNNNFFKGMEIIEDQELVSITGGKKWGW LAWVDPAYEFIKGFGKGAIKEGNKDKWKNI 	[31]

RNP class	Subclass	Chemotype	Genotype	Ref.
	IIc (cyclic)	AS-48 (*Enterococcus faecalis*) 	MVKENKFSKIFILMALSFLGLALFSASLQFLPIAHMAKEFGIPAAVAGTVLNVVEA GGWVTTIVSILTAVGSGGLSLLAAAGRESIKAYLKKEIKKKGKRAVIAW 	[32]

Supplementary Table 2: HMMER-extracted Pfam domains from known RNP biosynthetic pathways used in RNP pathway prediction. 9 most common Pfam domains (red) were used for RNP gene cluster search in JGI database genomes.

Pfam domain ID	Pfam domain name	HMMER-Extraction count
PF04738	Lant_dehyd_C	10
PF05147	LANC_like	7
PF00082	Peptidase_S8	7
PF04737	Lant_dehyd_N	5
PF02624	YcaO	5
PF00881	Nitroreductase	5
PF00733	Asn_synthase	5
PF00501	AMP-binding	5
PF04055	Radical_SAM	3
PF02441	Flavoprotein	3
PF02052	Gallidermin	3
PF01039	Carboxyl_trans	3
PF00583	Acetyltransf_1	3
PF00067	p450	3
PF12146	Hydrolase_4	2
PF08443	RimK	2
PF08241	Methyltransf_11	2
PF08028	Acyl-CoA_dh_2	2
PF07366	SnoaL	2
PF02771	Acyl-CoA_dh_N	2
PF02310	B12-binding	2
PF01494	FAD_binding_3	2
PF00155	Aminotran_1_2	2
PF00106	adh_short	2
PF11420	Subtilosin_A	1
PF10503	Esterase_phd	1
PF05931	AgrD	1
PF05402	PqqD	1
PF05193	Peptidase_M16_C	1
PF04820	Trp_halogenase	1
PF04647	AgrB	1
PF03441	FAD_binding_7	1
PF03412	Peptidase_C39	1
PF02784	Orn_Arg_deC_N	1
PF02016	Peptidase_S66	1
PF01613	Flavin_Reduct	1
PF00903	Glyoxalase	1
PF00899	ThiF	1
PF00756	Esterase	1
PF00675	Peptidase_M16	1
PF00657	Lipase_GDSL	1
PF00156	Pribosyltran	1
PF00383	dCMP_cyt_deam_1	1
PF00278	Orn_DAP_Arg_deC	1

Supplementary Table 3: Bioinformatic prediction of PNP pathways in microbial genomes.

Genome with predicted PNP gene cluster	Absolute number of genomes (total: 1035)	[%] of total genomes
Genomes with candidate RNP gene clusters	735	71
Genomes with confident RNP gene clusters	322	31
Genomes with candidate NRP gene clusters	714	69
Genomes with candidate hybrid NRPS-PKS gene clusters	549	53

Supplementary Table 4: Comparison of RNP genome mining results of the manual *de novo* NPP workflow and current automated proteomic workflows.

Method	Mascot	InsPecT				NPP	
Peptide	NCBI & SwissProt Actinobacteria Protein Database	Annotated protein database		6-frame translation		Annotated protein database	6-frame translation
		blind search	specific search	blind search	specific search		
SRO15-2005	No	No	No	No	No	Yes	Yes
SRO15-2212	No	No	No	No	No	Yes	Yes
SRO15-3108	No	No	No	No	No	Yes	Yes
SGR-1832	No	No	No	No	No	No	Yes
SGR-2212 (AmfS)	No	No	No	No	No	Yes	Yes
SAL-2242	No	No	No	No	No	Yes	Yes
SCO-2138	No	No	No	No	Yes	No	Yes
SLI-2138	No	No	Yes	No	Yes	Yes	Yes
SWA-2138	No	No	No	No	No	Yes	Yes
SSV-2083	No	No	No	No	No	Yes	Yes

Supplementary Table 5: Mass shifts of proteinogenic amino acids.

Mass	Residue	Abbreviation
57.022	Glycine	Gly, G
71.037	Alanine	Ala, A
87.032	Serine	Ser, S
97.053	Proline	Pro, P
99.068	Valine	Val, V
101.048	Threonine	Thr, T
103.009	Cysteine	Cys, C
113.084	Isoleucine	Ile, I
	Leucine	Leu, L
114.043	Asparagine	Asn, N
115.027	Aspartic acid	Asp, D
128.059	Glutamine	Gln, Q
128.095	Lysine	Lys, K
129.042	Glutamic acid	Glu, E
131.041	Methionine	Met, M
137.059	Histidine	His, H
147.068	Phenylalanine	Phe, F
156.101	Arginine	Arg, R
163.063	Tyrosine	Tyr, Y
186.079	Tryptophan	Trp, W

* - after $NaBH_4/NiCl_2$-treatment

Supplementary table 6: Mass shifts of RNP monomers and modifications and their corresponding precursor amino acids in precursor peptides.

See as stand-alone Supplementary Table file.

Supplementary Table 7: Mass shifts of NRP monomers (NORINE monomer list excluding lipids) and corresponding NRPS monomers and genome mining accessibility by NRPSpredictor2 (AntiSMASH, * - biosynthetic monomers can be putative due to lack of biosynthetic knowledge).
See as stand-alone Supplementary Table file.

Supplementary table 8: ^{1}H and ^{13}C NMR Data for Q027-1628 (stendomycin I) in CD_3OD (δ in ppm)[a]

Residue	Position	δ_C	δ_H	Residue	Position	δ_C	δ_H
Pro_1	α	57.9	4.92	Val_{10}	NH		7.46
	β1	47.9	3.61		α	63.3	3.66
	β2	47.9	3.71		β	29.5	1.9
	γ	28.4	2.37		γ	17.7	0.98
	δ1	25.1	2.17			19.5	1.03
	δ2	24.7	2.05		C=O	174.5	
	C=O	176.6		Thr_{11}	NH		7.79
Thr_2	N-CH_3	40.0	3.45		α	61.8	4.41
	α	71.8	4.08		β	67.6	4.34
	β	66.9	4.50		γ	19.2	1.34
	γ	20.7	1.33		C=O	170.8	
	C=O	171.3		Ser_{12}	NH		7.69
Gly_3	NH		8.85		α	55.5	4.7
	α	44.5	3.85, 3.95		β	61.8	3.85, 4.25
	C=O	174.5			C=O	172.0	
Val_4	NH		7.94	Ile_{13}	NH		7.70
	α	64.5	3.57		α	57.7	4.46
	β	29.7	1.92		β	36.4	2.14
	γ	18.3	0.98		γ_1(Me)	13.7	1.05
		19.9	1.08		γ_2	25.8	1.34
	C=O	174.7			δ	10.6	0.88
Ile_5	NH		7.81		C=O	173.4	
	α	62.4	3.77	Ste*$_{14}$	NH		7.39
	β	34.8	2.02		α	51.9	5.26
	γ_1(Me)	15.8	1.06		β	58.1	4.00
	γ_2	25.4	1.52		γ	24.4	1.86, 2.09
	δ	15.8	1.28		δ	36.4	3.12, 3.35
	C=O	176.0			N_δH		7.33
Ala_6	NH		8.18		Cε	156.3	
	α	53.8	4.06		N_β-CH3	27.2	2.89
	β	15.5	1.6		N_ε-CH3	35.4	3.08
	C=O	176.9			C=O	167.1	
Dhb_7	NH		9.3	acyl[b]	C=O	173.2	
	α	131.6			α	24.7	1.6
	β	131.5	6.47				
	γ	13.1	1.82				
	C=O	168.4					
Thr_8	NH		8.15				
	α	61.7	4.14				
	β	69.6	5.77				
	γ	17.7	1.31				
	C=O	174.0					
Val_9	NH		8.51				
	α	64.8	3.34				
	β	29.0	2.24				
	γ	18.8	0.99				
		19.8	0.99				
	C=O	172.7					

[a] Assignment based on ^{1}H NMR, ^{13}C NMR, ^{1}H-^{1}H COSY, ^{1}H-^{1}H TOCSY, ^{1}H-^{13}C HSQC, ^{1}H-^{13}C HMBC data. [b] Stendomycin I acyl chain is assumed to be 12-methyltridecanoic acid based on literature[7]. * Ste – stendomycidine.

Supplementary table 9: MS^n analysis of SHY-1586a

Detector	Precursor ion [m/z]	Observed mass [Da]	Calculated mass [Da]	Difference [Da]	Error	Species
FT	1587 (1+)	1403.812975	1403.813675	-0.0007	-5E-07	y14 (hydrolyzed lactone)
FT	1587 (1+)	1306.760175	1306.760875	-0.0007	-5.4E-07	y13 (hydrolyzed lactone)
FT	1587 (1+)	1191.698175	1191.697575	0.0006	5.03E-07	y12 (hydrolyzed lactone)
FT	1587 (1+)	1134.675775	1134.676075	-0.0003	-2.6E-07	y11 (hydrolyzed lactone)
FT	1587 (1+)	1035.607475	1035.607675	-0.0002	-1.9E-07	y10 (hydrolyzed lactone)
FT	1587 (1+)	922.523175	922.523575	-0.0004	-4.3E-07	y9 (hydrolyzed lactone)
FT	1587 (1+)	851.485475	851.486475	-0.001	-1.2E-06	y8 (hydrolyzed lactone)
FT	1587 (1+)	768.448775	768.449375	-0.0006	-7.8E-07	y7 (hydrolyzed lactone)
FT	1587 (1+)	685.411475	685.412275	-0.0008	-1.2E-06	y6 (hydrolyzed lactone)
FT	1587 (1+)	586.343475	586.343875	-0.0004	-6.8E-07	y5 (hydrolyzed lactone)
FT	1587 (1+)	487.275075	487.275475	-0.0004	-8.2E-07	y4 (hydrolyzed lactone)
FT	1587 (1+)	386.226875	386.227775	-0.0009	-2.3E-06	y3 (hydrolyzed lactone)
IT	1587 (1+), 488 (1+)	299.352175	299.195775	0.1564	0.000523	y2 (hydrolyzed lactone)
IT	1587 (1+), 488 (1+)	200.172175	200.127375	0.0448	0.000224	y1 (hydrolyzed lactone)
IT	1587 (1+), 395 (1+)	279.172175	279.219875	-0.0477	-0.00017	b1 (hydrolyzed lactone)
FT	1587 (1+)	394.282675	394.283175	-0.0005	-1.3E-06	b2 (hydrolyzed lactone)
FT	1587 (1+)	433.293075	433.294075	-0.001	-2.3E-06	b3 (hydrolyzed lactone)
FT	1587 (1+)	550.373075	550.373075	0	0	b4 (hydrolyzed lactone)
FT	1587 (1+)	663.456275	663.457075	-0.0008	-1.2E-06	b5 (hydrolyzed lactone)
FT	1587 (1+)	734.493875	734.494175	-0.0003	-4.1E-07	b6 (hydrolyzed lactone)
FT	1587 (1+)	817.532375	817.531375	0.001	1.22E-06	b7 (hydrolyzed lactone)
FT	1587 (1+)	900.567175	900.568475	-0.0013	-1.4E-06	b8 (hydrolyzed lactone)
FT	1587 (1+)	999.635875	999.636875	-0.001	-1E-06	b9 (hydrolyzed lactone)
FT	1587 (1+)	1098.703775	1098.705275	-0.0015	-1.4E-06	b10 (hydrolyzed lactone)
FT	1587 (1+)	1199.752575	1199.752975	-0.0004	-3.3E-07	b11 (hydrolyzed lactone)
FT	1587 (1+)	1268.777675	1268.774375	0.0033	2.6E-06	b12 (hydrolyzed lactone)
FT	1587 (1+)	1367.839575	1367.842775	-0.0032	-2.3E-06	b13 (hydrolyzed lactone)

Supplementary table 10: MS^n analysis of SHY-1586b

Detector	Precursor ion [m/z]	Observed mass [Da]	Calculated mass [Da]	Difference [Da]	Error	Species
IT	1587 (1+), 502 (1+)	200.172175	200.127375	0.0448	0.00022	y1 (hydrolyzed lactone)
IT	1587 (1+), 395 (1+)	279.172175	279.219875	-0.0477	-0.0002	b1 (hydrolyzed lactone)
IT	1587 (1+), 502 (1+)	313.172175	313.211375	-0.0392	-0.0001	y2 (hydrolyzed lactone)
FT	1587 (1+)	394.282675	394.283175	-0.0005	-1E-06	b2 (hydrolyzed lactone)
FT	1587 (1+)	400.242675	400.243375	-0.0007	-2E-06	y3 (hydrolyzed lactone)
FT	1587 (1+)	433.293075	433.294075	-0.001	-2E-06	b3 (hydrolyzed lactone)
FT	1587 (1+)	501.290575	501.291075	-0.0005	-1E-06	y4 (hydrolyzed lactone)
FT	1587 (1+)	550.373075	550.373075	0	0	b4 (hydrolyzed lactone)
FT	1587 (1+)	600.359075	600.359475	-0.0004	-7E-07	y5 (hydrolyzed lactone)
FT	1587 (1+)	649.441275	649.441475	-0.0002	-3E-07	b5 (hydrolyzed lactone)
FT	1587 (1+)	699.427675	699.427875	-0.0002	-3E-07	y6 (hydrolyzed lactone)
FT	1587 (1+)	720.477875	720.478575	-0.0007	-1E-06	b6 (hydrolyzed lactone)
FT	1587 (1+)	782.464175	782.465075	-0.0009	-1E-06	y7 (hydrolyzed lactone)
FT	1587 (1+)	803.515275	803.515675	-0.0004	-5E-07	b7 (hydrolyzed lactone)
FT	1587 (1+)	865.501475	865.502175	-0.0007	-8E-07	y8 (hydrolyzed lactone)
FT	1587 (1+)	886.552575	886.552775	-0.0002	-2E-07	b8 (hydrolyzed lactone)
FT	1587 (1+)	936.538175	936.539275	-0.0011	-1E-06	y9 (hydrolyzed lactone)
FT	1587 (1+)	985.620775	985.621175	-0.0004	-4E-07	b9 (hydrolyzed lactone)
FT	1587 (1+)	1035.607475	1035.607675	-0.0002	-2E-07	y10 (hydrolyzed lactone)
FT	1587 (1+)	1084.688875	1084.689575	-0.0007	-6E-07	b10 (hydrolyzed lactone)
FT	1587 (1+)	1134.675775	1134.676075	-0.0003	-3E-07	y11 (hydrolyzed lactone)
FT	1587 (1+)	1185.736975	1185.737275	-0.0003	-3E-07	b11 (hydrolyzed lactone)
FT	1587 (1+)	1191.698175	1191.697575	0.0006	5E-07	y12 (hydrolyzed lactone)
FT	1587 (1+)	1254.761675	1254.758775	0.0029	2.3E-06	b12 (hydrolyzed lactone)
FT	1587 (1+)	1306.760175	1306.760875	-0.0007	-5E-07	y13 (hydrolyzed lactone)
FT	1587 (1+)	1367.839575	1367.842775	-0.0032	-2E-06	b13 (hydrolyzed lactone)
FT	1587 (1+)	1403.812975	1403.813675	-0.0007	-5E-07	y14 (hydrolyzed lactone)

Supplementary table 11: MS^n analysis of SHY-1600

Detector	Precursor ion [m/z]	Observed mass [Da]	Calculated mass [Da]	Difference [Da]	Error	Species
IT	1601 (1+), 502 (1+)	200.172175	200.127375	0.0448	0.00022	y1 (hydrolyzed lactone)
FT	801 (2+	207.219175	207.219875	-0.0007	-3E-06	b1 (hydrolyzed lactone)
IT	1601 (1+), 502 (1+)	313.262175	313.211375	0.0508	0.00016	y2 (hydrolyzed lactone)
FT	1601 (1+)	394.283175	394.283175	0	0	b2 (hydrolyzed lactone)
FT	1601 (1+)	400.243475	400.243375	0.0001	2.5E-07	y3 (hydrolyzed lactone)
FT	1601 (1+)	433.293975	433.294075	-1E-04	-2E-07	b3 (hydrolyzed lactone)
FT	1601 (1+)	501.291475	501.291075	0.0004	8E-07	y4 (hydrolyzed lactone)
FT	1601 (1+)	550.373375	550.373075	0.0003	5.5E-07	b4 (hydrolyzed lactone)
FT	1601 (1+)	600.360475	600.359475	0.001	1.7E-06	y5 (hydrolyzed lactone)
FT	1601 (1+)	663.458375	663.457075	0.0013	2E-06	b5 (hydrolyzed lactone)
FT	1601 (1+)	699.429275	699.427875	0.0014	2E-06	y6 (hydrolyzed lactone)
FT	1601 (1+)	734.495575	734.494175	0.0014	1.9E-06	b6 (hydrolyzed lactone)
FT	1601 (1+)	782.466475	782.465075	0.0014	1.8E-06	y7 (hydrolyzed lactone)
FT	1601 (1+)	817.533675	817.531375	0.0023	2.8E-06	b7 (hydrolyzed lactone)
FT	1601 (1+)	865.503875	865.502175	0.0017	2E-06	y8 (hydrolyzed lactone)
FT	1601 (1+)	900.570775	900.568475	0.0023	2.6E-06	b8 (hydrolyzed lactone)
FT	1601 (1+)	936.541575	936.539275	0.0023	2.5E-06	y9 (hydrolyzed lactone)
FT	1601 (1+)	999.639275	999.636875	0.0024	2.4E-06	b9 (hydrolyzed lactone)
FT	1601 (1+)	1049.626375	1049.623375	0.003	2.9E-06	y10 (hydrolyzed lactone)
FT	1601 (1+)	1098.709875	1098.705275	0.0046	4.2E-06	b10 (hydrolyzed lactone)
FT	1601 (1+)	1148.697575	1148.691775	0.0058	5E-06	y11 (hydrolyzed lactone)
FT	1601 (1+)	1199.756675	1199.752975	0.0037	3.1E-06	b11 (hydrolyzed lactone)
FT	1601 (1+)	1205.714975	1205.713175	0.0018	1.5E-06	y12 (hydrolyzed lactone)
FT	1601 (1+)	1268.779875	1268.774375	0.0055	4.3E-06	b12 (hydrolyzed lactone)
FT	1601 (1+)	1320.782275	1320.776575	0.0057	4.3E-06	y13 (hydrolyzed lactone)
FT	1601 (1+)	1381.859075	1381.858475	0.0006	4.3E-07	b13 (hydrolyzed lactone)
FT	1601 (1+)	1417.836475	1417.829275	0.0072	5.1E-06	y14 (hydrolyzed lactone)

Supplementary table 12: MS^n analysis of SHY-1614a

Detector	Precursor ion [m/z]	Observed mass [Da]	Calculated mass [Da]	Difference [Da]	Error	Species
IT	1615 (1+), 502 (1+)	200.172175	200.107375	0.0448	0.000224	y1 (hydrolyzed lactone)
FT	808 (2+)	293.235075	293.235475	-0.0004	-0.000001	b1 (hydrolyzed lactone)
IT	1615 (1+), 502 (1+)	313.262175	313.211375	0.0508	0.000162	y2 (hydrolyzed lactone)
FT	1615 (1+)	400.241475	400.243375	-0.0019	-0.000005	y3 (hydrolyzed lactone)
FT	808 (2+)	408.296775	408.298775	-0.002	-0.000005	b2 (hydrolyzed lactone)
FT	808 (2+)	447.307575	447.309675	-0.0021	-0.000005	b3 (hydrolyzed lactone)
FT	1615 (1+)	501.288675	501.291075	-0.0024	-0.000005	y4 (hydrolyzed lactone)
FT	1615 (1+)	564.385775	564.388675	-0.0029	-0.000005	b4 (hydrolyzed lactone)
FT	1615 (1+)	600.356575	600.359475	-0.0029	-0.000005	y5 (hydrolyzed lactone)
FT	1615 (1+)	677.468875	677.472775	-0.0039	-0.000006	b5 (hydrolyzed lactone)
FT	1615 (1+)	699.424175	699.427875	-0.0037	-0.000005	y6 (hydrolyzed lactone)
FT	1615 (1+)	748.505575	748.509875	-0.0043	-0.000006	b6 (hydrolyzed lactone)
FT	1615 (1+)	782.460175	782.465075	-0.0049	-0.000006	y7 (hydrolyzed lactone)
FT	1615 (1+)	831.542175	831.546975	-0.0048	-0.000006	b7 (hydrolyzed lactone)
FT	1615 (1+)	865.496375	865.502175	-0.0058	-0.000007	y8 (hydrolyzed lactone)
FT	1615 (1+)	914.578775	914.584075	-0.0053	-0.000006	b8 (hydrolyzed lactone)
FT	1615 (1+)	936.532575	936.539275	-0.0067	-0.000007	y9 (hydrolyzed lactone)
FT	1615 (1+)	1013.645575	1013.652475	-0.0069	-0.000007	b9 (hydrolyzed lactone)
FT	1615 (1+)	1049.615675	1049.623375	-0.0077	-0.000007	y10 (hydrolyzed lactone)
FT	1615 (1+)	1112.713375	1112.720875	-0.0075	-0.000007	b10 (hydrolyzed lactone)
FT	1615 (1+)	1148.684875	1148.691775	-0.0069	-0.000006	y11 (hydrolyzed lactone)
FT	1615 (1+)	1205.701275	1205.713175	-0.0119	-0.000010	y12 (hydrolyzed lactone)
FT	1615 (1+)	1213.756875	1213.768575	-0.0117	-0.000010	b11 (hydrolyzed lactone)
FT	1615 (1+)	1282.778475	1282.790075	-0.0116	-0.000009	b12 (hydrolyzed lactone)
FT	1615 (1+)	1320.765575	1320.776575	-0.011	-0.000008	y13 (hydrolyzed lactone)
FT	1615 (1+)	1395.858975	1395.874075	-0.0151	-0.000011	b13 (hydrolyzed lactone)
FT	1615 (1+)	1417.818875	1417.829275	-0.0104	-0.000007	y14 (hydrolyzed lactone)

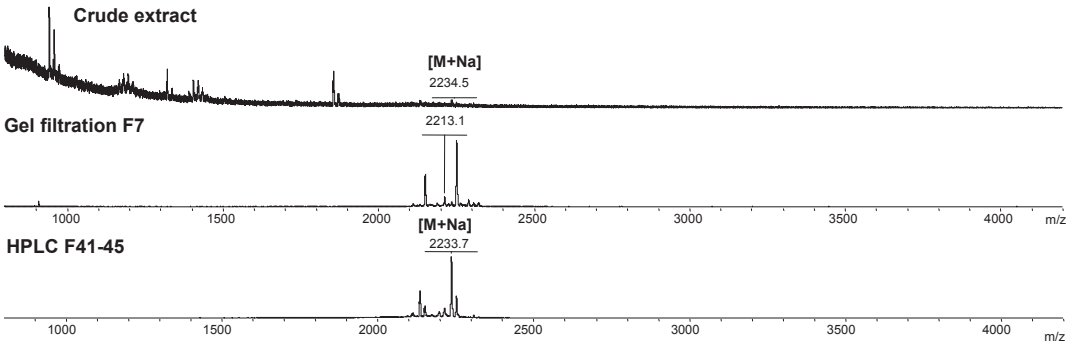
Supplementary table 13: MS^n analysis of SHY-1614b

Detector	Precursor ion [m/z]	Observed mass [Da]	Calculated mass [Da]	Difference [Da]	Error	Species
IT	1615 (1+), 502 (1+)	200.172175	200.127375	0.0448	0.00022	y1 (hydrolyzed lactone)
FT	808 (2+)	307.250675	307.250275	0.0004	1.3E-06	b1 (hydrolyzed lactone)
IT	1615 (1+), 502 (1+)	313.172175	313.211375	-0.0392	-0.00013	y2 (hydrolyzed lactone)
FT	1615 (1+)	400.241475	400.243375	-0.0019	-4.7E-06	y3 (hydrolyzed lactone)
FT	1615 (1+)	422.313975	422.314375	-0.0004	-9.5E-07	b2 (hydrolyzed lactone)
FT	1615 (1+)	461.325175	461.324975	0.0002	4.3E-07	b3 (hydrolyzed lactone)
FT	1615 (1+)	501.288675	501.291075	-0.0024	-4.8E-06	y4 (hydrolyzed lactone)
FT	1615 (1+)	578.401975	578.404875	-0.0029	-5E-06	b4 (hydrolyzed lactone)
FT	1615 (1+)	600.356575	600.359475	-0.0029	-4.8E-06	y5 (hydrolyzed lactone)
FT	1615 (1+)	677.468875	677.472775	-0.0039	-5.8E-06	b5 (hydrolyzed lactone)
FT	1615 (1+)	699.424175	699.427875	-0.0037	-5.3E-06	y6 (hydrolyzed lactone)
FT	1615 (1+)	748.505575	748.509875	-0.0043	-5.7E-06	b6 (hydrolyzed lactone)
FT	1615 (1+)	782.460175	782.465075	-0.0049	-6.3E-06	y7 (hydrolyzed lactone)
FT	1615 (1+)	831.542175	831.546975	-0.0048	-5.8E-06	b7 (hydrolyzed lactone)
FT	1615 (1+)	865.496375	865.502175	-0.0058	-6.7E-06	y8 (hydrolyzed lactone)
FT	1615 (1+)	914.578775	914.584075	-0.0053	-5.8E-06	b8 (hydrolyzed lactone)
FT	1615 (1+)	936.532575	936.539275	-0.0067	-7.2E-06	y9 (hydrolyzed lactone)
FT	1615 (1+)	1013.645575	1013.652475	-0.0069	-6.8E-06	b9 (hydrolyzed lactone)
FT	1615 (1+)	1035.600675	1035.607675	-0.007	-6.8E-06	y10 (hydrolyzed lactone)
FT	1615 (1+)	1112.713375	1112.720875	-0.0075	-6.7E-06	b10 (hydrolyzed lactone)
FT	1615 (1+)	1134.667675	1134.676075	-0.0084	-7.4E-06	y11 (hydrolyzed lactone)
FT	1615 (1+)	1191.688275	1191.697575	-0.0093	-7.8E-06	y12 (hydrolyzed lactone)
FT	1615 (1+)	1213.756875	1213.768575	-0.0117	-9.6E-06	b11 (hydrolyzed lactone)
FT	1615 (1+)	1282.778475	1282.790075	-0.0116	-9E-06	b12 (hydrolyzed lactone)
FT	1615 (1+)	1306.768575	1306.760875	0.0077	5.9E-06	y13 (hydrolyzed lactone)
FT	1615 (1+)	1395.858975	1395.874075	-0.0151	-1.1E-05	b13 (hydrolyzed lactone)

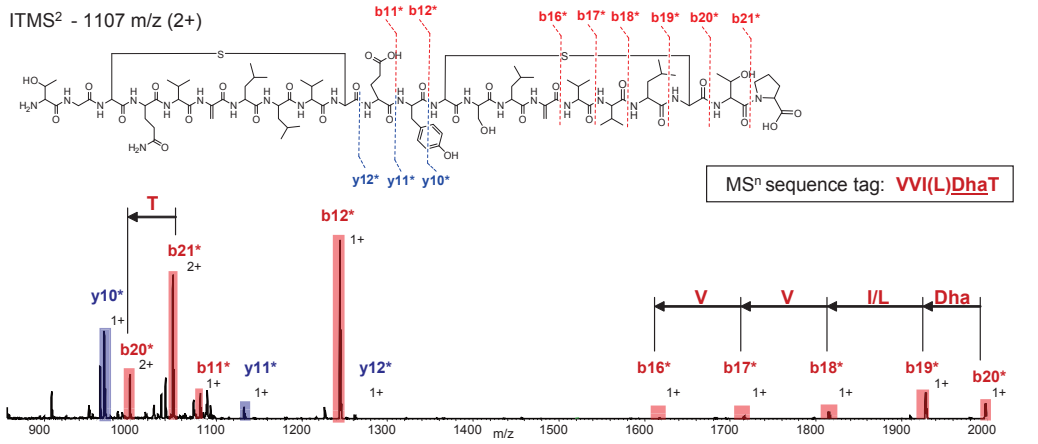
NRP representative	Mass [Da]	Monomers	Family	Class	NRP sequence (linear representation as in NORINE)	NRP prediction (top NP.searcher, bottom - NRPSpredictor2/antiSMASH)	Reference
A54145 A	1643.8	14	A54145	partial cyclic lipopeptide	fC10:0_Trp_D-Glu_OH-Asn_[_Thr_NMe-Gly_Ala_Asp_D-Lys_OMe-Asp_Gly_D-Asn_Glu_Ile_]	nrp_trp_tyr_asn_thr_gly_gly_asp_trp_asp_gly_asn_dhb_ile pk_trp_asp_asn_thr_gly_gly_asp_glu_asp_gly_asn_asp_tyr	DQ118863.1
Actinomycin D	1254.6	11	Actinomycin	double-cyclic chromopeptide	[_D-Val_Pro_NMe-Gly_NMe-Val_Thr_]_ChrAct_[_Thr_D-Val_Pro_NMe-Gly_NMe-Val_]	ile_pro_gly_val_val pip_thr_val_pro_gly_val	HM038106.1
ACV	363.2	3	ACV	linear peptide	Aad_Cys_D-Val	aad_cys_val aad_cys_val	AM920436.1
Arthrofactin	1353.8	12	Amphisin	partial cyclic lipopeptide	C10:0-OH(3)_D-Leu_D-Asp_[_D-Thr_D-Leu_D-Leu_D-Ser_Leu_D-Ser_Ile_Ile_Asp_]	leu_asp_thr_leu_leu_ser_leu_ser_ile_ile_asp leu_asp_thr_leu_leu_ser_leu_ser_ile_ile_asp	AB107223.1
AM-toxin I	445.2	4	AM-toxin	cyclic peptide	[_Amv_Hiv_Ala_dh-Ala_]	val_phe_nrp val_hyv-d_ala	AF184074.1
Anabaenopeptilide 202A	1049.5	8	Anabaenopeptilide	partial cyclic peptide	NFo-Pro_Gln_[_Thr_Hty_Ahp_Thr_NMe-OMe-Tyr_Ile_]	gln_thr_arg_orn_thr_tyr_ile gln_thr_hty_ahp_thr_tyr	AJ269505.1
Anabaenopeptin A	843.4	7	Anabaenopeptin	partial cyclic peptide	[_Phe_NMe-Ala_Hty_Val_D-Lys_]_CO_Tyr	lys_lys_tyr_lys_val_val_gly_phe lys_lys_tyr_lys_val_leu_ala_phe	GU174493.1
Andrimid	479.2	4	Andrimid	linear peptide	C8:3(t2,t4,t6)_D-bPhe_Val_Me-Suc	nrp_val_gly leu_val_gly	AY192157.1
Aureobasidin A	1100.7	9	Aureobasidin	cyclic peptide	[_D-Hmp_NMe-Val_Phe_NMe-Phe_Pro_alle_NMe-Val_Leu_bOH-NMe-Val_]	tyr_val_phe_phe_val_leu_val_leu_val tyr_val_phe_phe_ser_phe_val_leu_val	EU886741.1
Bacillibactin	882.3	9	Bacillibactin	partial cyclic peptide	diOH-Bz_Gly_[_Thr_Thr_(_Gly_diOH-Bz_)_Thr_]_Gly_diOH-Bz	gly_thr no prediction	AF184977.1
Bacilysin	270.1	2	Bacilysin	linear peptide	Ala_Aca	ala no prediction	AF396778.1
Bacitracin A1	1422.7	12	Bacitracin	peptide	[_Ile_Cys_]_Leu_D-Glu_Ile_[_Lys_D-Orn_Ile_D-Phe_His_D-Asp_Asn_]	cys_leu_glu_ile_lys_orn_ile_ile_phe_his_asp_asn ile_cys_leu_glu_ile_lys_orn_ile_phe_his_asp_asn	AF007865.2
Balhimycin	1445.4	9	Balhimycin	glycopeptide group I	Asn_bOH-Cl-Tyr_NMe-Leu_Hpg_D-Glc_bOH-Cl-Tyr_4oxo-Van_Dhpg_Hpg	leu_bht_asn_hpg_phg_bht_dhpg_tyr_nrp leu_bht_asn_hpg_hpg_bht_dhpg_tyr	Y16952.3
Beauvericin	783.4	6	Beauvericin	cyclic peptide	[_NMe-Phe_D-Hiv_NMe-Phe_D-Hiv_NMe-Phe_D-Hiv_]	hyv-d_nrp thyv_phe	EU886196.1
CDA2a	1575.5	12	Ca-dependent antibiotic	partial cyclic lipopeptide	C6:0-Ep(2)_Ser_[_Thr_D-Trp_Asp_Asp_D-Hpg_Asp_Gly_D-PO-Asn_3Me-Glu_dh-Trp_]	ser_thr_trp_asp_asp_ser_asp_gly_asn_3me-Glu_trp_val_nrp no prediction	AL645882.2
Capreomycin IA	668.4	6	Capreomycin	partial cyclic peptide	bUys_[_Dpr_bU-dAla_Cap_Dpr_Ser_]	ser_ser_glu_gln_lys-b ser_ser_gln_asp_lys (Minowa: ser_ser_dpr_cap_bUys)	EF472579.1
Chrysobactin	369.2	3	Chrysobactin	linear peptide	diOH-Bz_D-Lys_Ser	orn_ser pro_gly	NC_014500.1
Coronatine	319.2	2	Coronatine	linear peptide	CMA_CFA	val no prediction	AY391839.1
Cyclosporin A	1201.8	11	Cyclosporin	cyclic peptide	[_D-Ala_NMe-Leu_NMe-Leu_NMe-Val_NMe-Bmt_Abu_NMe-Gly_NMe-Leu_Val_NMe-Leu_Ala_]	ala-d_val_leu_val_phe_ile_gly_phe_val_val_ala ala-d_leu_leu_val_bmt_abu_gly_leu_val_leu_ala	Z28383.1
Daptomycin	1619.7	14	Daptomycin	partial cyclic peptide	C10:0_Trp_D-Asn_Asp_[_Thr_Gly_Orn_Asp_D-Ala_Asp_Gly_D-Ser_3Me-Glu_Kyn_]	nrp_arg_asn_asp_thr_gly_glu_asp_nrp_asp_gly_ser_dhb_phe pk_arg_asn_asp_thr_gly_glu_asp_ala_asp_gly_ser_glu_phe	AY787762.1
Enterobactin	669.1	6	Enterobactin	partial cyclic peptide	diOH-Bz_[_Ser_Ser_(_diOH-Bz_)_Ser_]_diOH-Bz	no prediction no prediction	M24148.1
Fengycin A	1462.8	11	Fengycin	partial cyclic lipopeptide	C16:0-OH(3)_Glu_D-Orn_[_Tyr_D-aThr_Glu_D-Ala_Pro_Glu_D-Tyr_Ile_]	pro_glu_tyr_glu_val_ile pro_glu_tyr_glu_val_ile	AF023464.2
Gramicidin S	1140.7	10	Gramicidin	cyclic peptide	[_Val_Orn_Leu_D-Phe_Pro_Val_Orn_Leu_D-Phe_Pro_]	val_orn_leu_pro_phe no prediction	NC_012491.1
HC-toxin	436.2	4	HC-toxin	cyclic peptide	[_D-Pro_Ala_D-Ala_C10:0-NH2(2)-Ep(9)-oxo(8)_]	pro_ala_ala_aeo pro_ala_ala_aeo	M98024.2
Iturin A-1	1028.5	8	Iturin	cyclic lipopeptide	[_C13:0-NH2(3)_Asn_D-Tyr_D-Asn_Gln_Pro_D-Asn_Ser_]	mal_nrp_asn_tyr_asn_gln_pro_asn_ser pk_asn_tyr_asn_gln_pro_asn_ser	AB050629.1
Microcystin LR	994.6	7	Microcystin	cyclic peptide	[_D-Ala_Leu_D-bMe-Asp_Arg_Adda_D-Glu_NMe-Dha_]	mal_mal_glu_nrp_mal mal_glu_tcl_mal_pk	AB032549.2
Nocardicin A	500.2	4	Nocardicin	partial cyclic peptide	Hse_OH-dHpg_[_Ser_Hpg_]	glu_orn_ser_ser_glu leu_gly_glu_ser_gly (minowa: hpg_orn_hpg_ser_hpg)	AY541063.1
Phosphinothricin	323.1	3	Phosphinothricin tripeptide	linear peptide	PT_Ala_Ala	ser_pro_nrp ser_pro_bht	X65195.2
Polymyxin B1	1202.8	11	Polymyxin	partial cyclic lipopeptide	aC9:0_Dab_Thr_Dab_[_Dab_Dab_D-Phe_Leu_Dab_Dab_Thr_]	leu_thr_thr_orn_orn_orn_thr_orn_orn_orn lys_thr_val_val_thr_asn_ala	CP000154.1
Pristinamycin IA	866.4	7	Pristinamycin I	partial cyclic peptide	Hpa_[_Thr_D-Abu_Pro_NMe-Me2A-Phe_4oxo-Hpr_Ph-Gly_]	pro_phe_thr_val_pro_phe_pip_phg no prediction	Y11547/8.1,
Echinomycin	1100.4	10	Quinomycin	partial cyclic peptide	COOH-Qui_D-Ser_Ala_diMe-Cys_NMe-Val_D-Ser_COOH-Qui_Ala_NMe-Cys_NMe-Val	phe_ser_gly_cys_val no prediction	AB211309.1
Serrawettin W1	514.3	4	Serrawettin W	cyclic lipopeptide	[_C10:0-OH(3)_Ser_C10:0-OH(3)_Ser_]	ser ser	AB193098.2
Surfactin aC15	1035.7	8	Surfactin	cyclic lipopeptide	[_aC15:0-OH(3)_Glu_Leu_D-Leu_Val_Asp_D-Leu_Leu_]	glu_leu_leu_val_asp_leu_leu no prediction	AJ575642.1
Syringomycin E	1224.6	10	Syringomycin	partial cyclic lipopeptide	C12:0-OH(3)_[_Ser_D-Ser_Dab_D-Dab_Arg_Phe_dhAbu_OH-Asp_4Cl-Thr_]	ser_ser_dab_dab_arg_val_thr_asp no prediction	AF047828.1
Triostin A	1086.4	10	Triostin	partial cyclic peptide	COOH-Qui_D-Ser_Ala_NMe-Cys_NMe-Val_D-Ser_COOH-Qui_Ala_NMe-Cys_NMe-Val	ser_pk_gly_cys_val_phe dhb_phe_ser_gly_cys_val	AB366635.1
viomycin	685.3	6	Tuberactinomycin	partial cyclic peptide	bUys_[_Dpr_Ser_Ser_bU-dAla_5OH-Cap_]	ser_ser_gln_gln_lys-b ser_ser_gln_gln_bUys	AY263398.1
Tyrocidine A	1269.7	10	Tyrocidine	cyclic peptide	[_D-Phe_Pro_Phe_D-Phe_Asn_Gln_Tyr_Val_Orn_Leu_]	phe_pro_phe_phe_asn_gln_tyr_val_orn_leu phe_pro_phe_phe_asn_gln_tyr_val_orn_leu	AF004835.1
Vancomycin	1447.4	9	Vancomycin	glycopeptide group I	Asn_bOH-Cl-Tyr_NMe-Leu_Hpg_D-Glc_Van_bOH-Cl-Tyr_Dhpg_Hpg	tyr_leu_bht_asn_hpg_phg_bht_dhpg no prediction	AJ223998.1
Vibriobactin	705.3	4	Vibriobactin	branched peptide	diOH-Bz_NSPD_(_DMOG_)_DMOG	no prediction no prediction	U52150.2
Massetolide A	1139.7	10	Viscosin	partial cyclic lipopeptide	C10:0-OH(3)_Leu_D-Glu_[_D-aThr_D-alle_Leu_D-Ser_Leu_D-Ser_Ile_]	leu_val_thr_ile_leu_ser_leu_ser_ile leu_asp_thr_ile_leu_ser_leu_ser_ile	EU199080/1.2
Average mass	948.1						
Average monomer #		8					
NPP-Accessibility of NRP families with mass shifts of proteinogenic amino acids (Suppl. Table 5)					50%		
NPP-Accessibility of NRP families with mass shifts of NRP monomers (Suppl. Table 7)					68%		

Supplementary Figure 1. Accessibility of NRP families in NORINE database to genome mining by NPP. Representatives of NRP families with curated biosynthetic gene cluster files in the GenBank database were analyzed. An NRP was marked accessible by NPP (yellow or orange) when at least 2 adjacent monomers were predicted correctly by NP.searcher and/or antiSMASH from the gene cluster. Either only proteinogenic amino acids (**Supplementary Table 5**, yellow) or additionally nonproteinogenic NRP monomers from the NORINE database were considered (**Supplementary Table 7**, orange). See corresponding supplementary tables for monomer abbreviations. Failed NRP prediction could be due to the minimal size limits of GenBank files for antiSMASH analysis or a failed single substrate prediction by either NP.searcher or NRPSpredictor2/antiSMASH based on an unusual A domain specificity code or unrecognized A domain sequences.

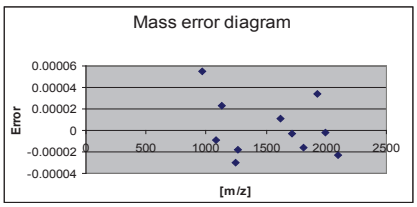
a

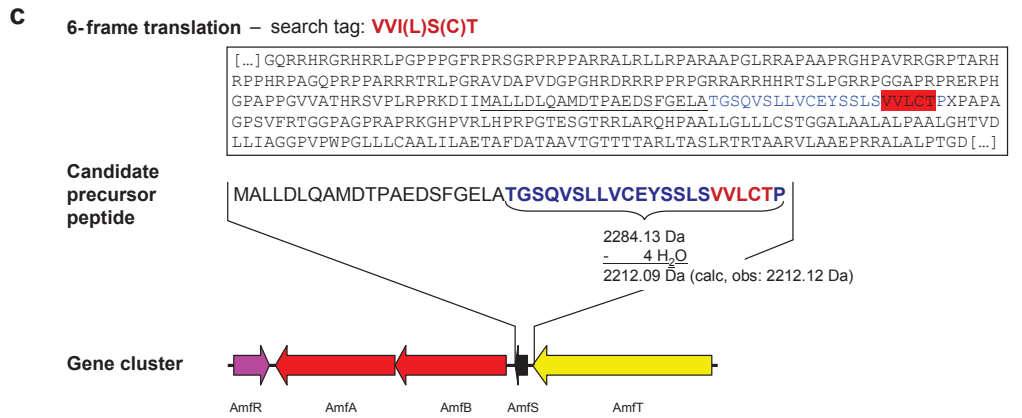


b



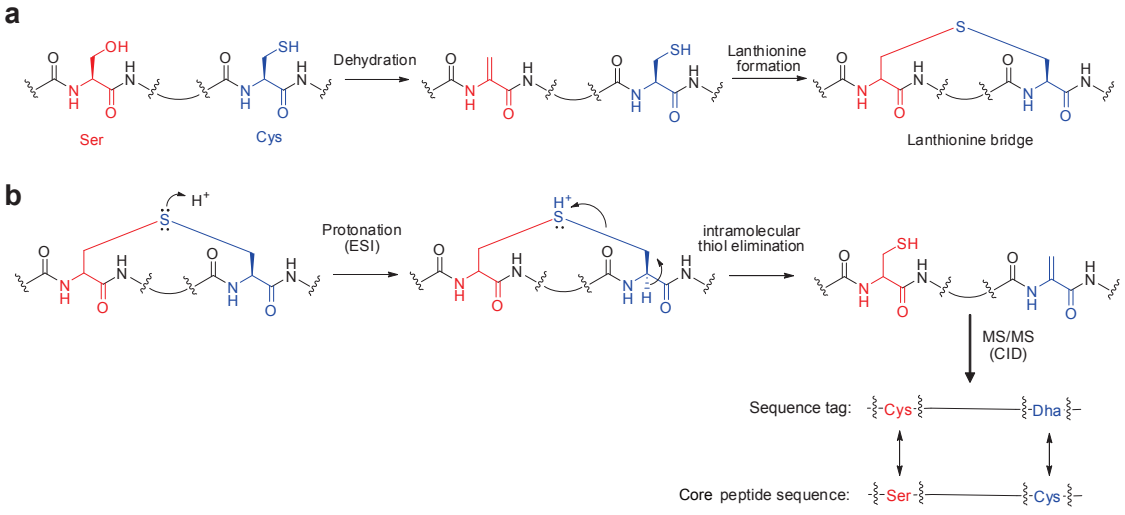
Detector	Observed mass [Da]	Calculated mass [Da]	Difference [Da]	Error	Species
IT	969.4546	969.50798	0.05338	5.50589E-05	y10-2H2O
IT	1081.5455	1081.53526	-0.01024	-9.46802E-06	b11-2H2O
IT	1132.5455	1132.57131	0.02581	2.27889E-05	y11-2H2O
IT	1244.6364	1244.59859	-0.03781	-3.03793E-05	b12-2H2O
IT	1261.6364	1261.6139	-0.0225	-1.78343E-05	y12-2H2O
IT	1616.7273	1616.74534	0.01804	1.11582E-05	b16-4H2O+S+2H
IT	1715.8182	1715.81376	-0.00444	-2.58769E-06	b17-4H2O+S+2H
IT	1814.9091	1814.88084	-0.02826	-1.55713E-05	b18-4H2O+S+2H
IT	1927.9014	1927.96624	0.06484	3.36313E-05	b19-4H2O+S+2H
IT	1996.9922	1996.98769	-0.00451	-2.2584E-06	b20-4H2O
IT	2098.0832	2098.03537	-0.04783	-2.27975E-05	b21-4H2O





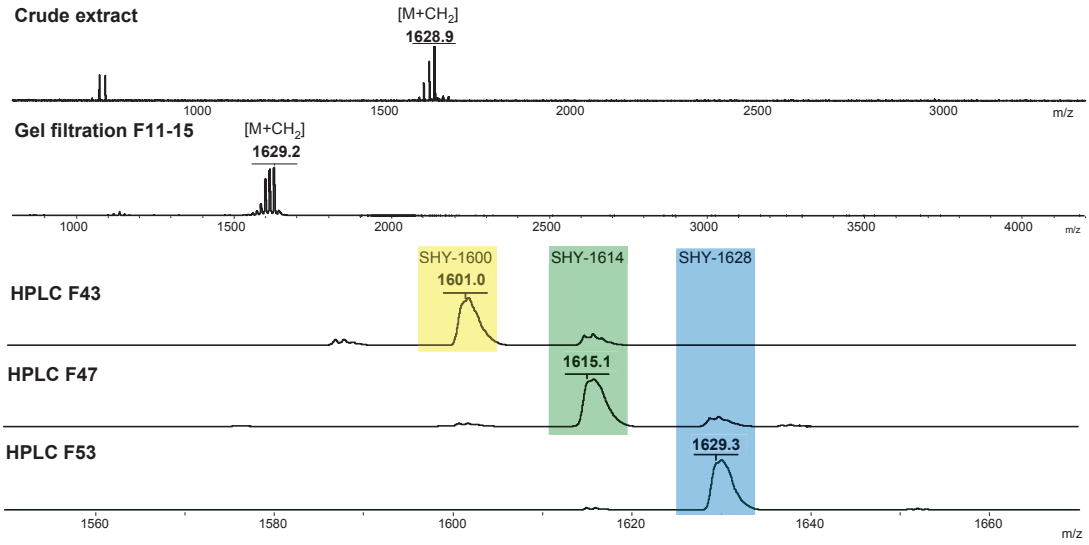
Gene	Size [aa]	Predicted function	Closest homolog [sim%/id%]
SGR_2393	201	AmfR - transcriptional regulator	ref|ZP_06281789.1| two component transcriptional regulator, LuxR family [*Streptomyces* sp. ACT-1] (100/100)
SGR_2394	634	AmfA - ABC transporter ATP-binding protein	ref|ZP_06281788.1| ABC transporter related protein [*Streptomyces* sp. ACT-1] (99/99)
SGR_2395	638	AmfB - ABC transporter ATP-binding membrane translocator	ref|ZP_06281787.1| ABC transporter related protein [*Streptomyces* sp. ACT-1] (97/97)
SGR_2396	43	AmfS - precursor peptide	ref|ZP_04696001.1| AmfS protein [*Streptomyces roseosporus* NRRL 15998] ref|ZP_04711182.1| AmfS protein [*Streptomyces roseosporus* NRRL 11379] (100/100)
SGR_2397	875	AmfT	ref|ZP_06281785.1| serine/threonine protein kinase [*Streptomyces* sp. ACT-1] (99/98)

Supplementary Figure 2. Characterization of AmfS and its biosynthetic gene cluster by the NPP approach. (a) AmfS purification. **(b)** AmfS sequence tagging. **(c)** Genome mining of AmfS gene cluster.

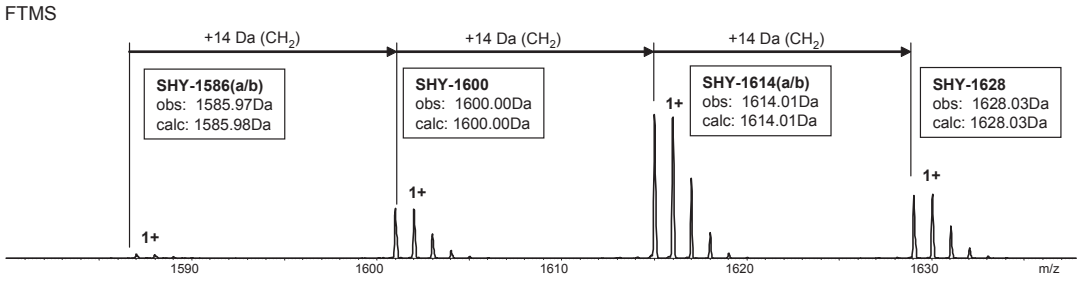


Supplementary Figure 3. Putative thiol elimination mechanism of lanthionine PTMs observed in ESI-MS/MS analysis of class III lantipeptides. (a) Biosynthesis of a lanthionine bridge from a Ser and Cys side chain by dehydration and nucleophilic attack. **(b)** Lanthionine bridge cleavage upon putative protonation (ESI) and subsequent intramolecular thiol elimination to yield a Cys and Dha at the position of the Ser and Cys, respectively, in the core peptide.

a

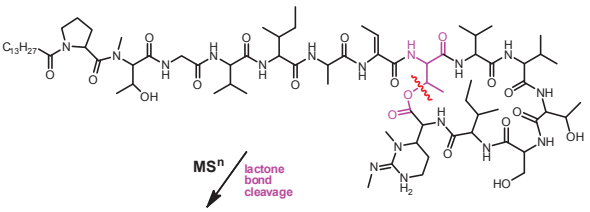

Crude extract
[M+CH2]
1628.9
1000
1500
2000
2500
3000
m/z
Gel filtration F11-15
[M+CH2]
1629.2
1000
1500
2000
2500
3000
3500
4000
m/z
SHY-1600
1601.0
SHY-1614
SHY-1628
HPLC F43
1615.1
HPLC F47
1629.3
HPLC F53
1560
1580
1600
1620
1640
1660
m/z

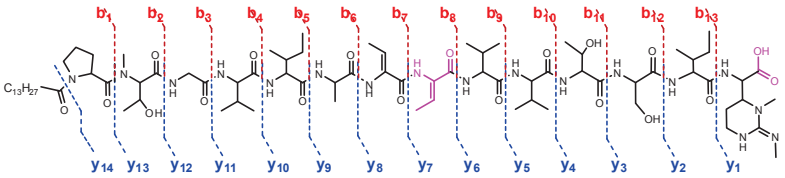
b


FTMS
+14 Da (CH2)
+14 Da (CH2)
+14 Da (CH2)
SHY-1586(a/b)
obs: 1585.97Da
calc: 1585.98Da
SHY-1600
obs: 1600.00Da
calc: 1600.00Da
SHY-1614(a/b)
obs: 1614.01Da
calc: 1614.01Da
SHY-1628
obs: 1628.03Da
calc: 1628.03Da
1+
1+
1+
1+
1590
1600
1610
1620
1630
m/z

b, continued

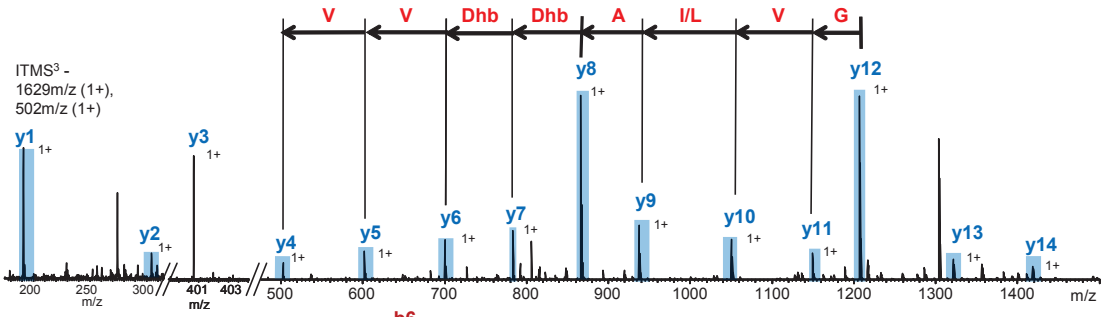
MSⁿ – SHY-1628 – Stendomycin I

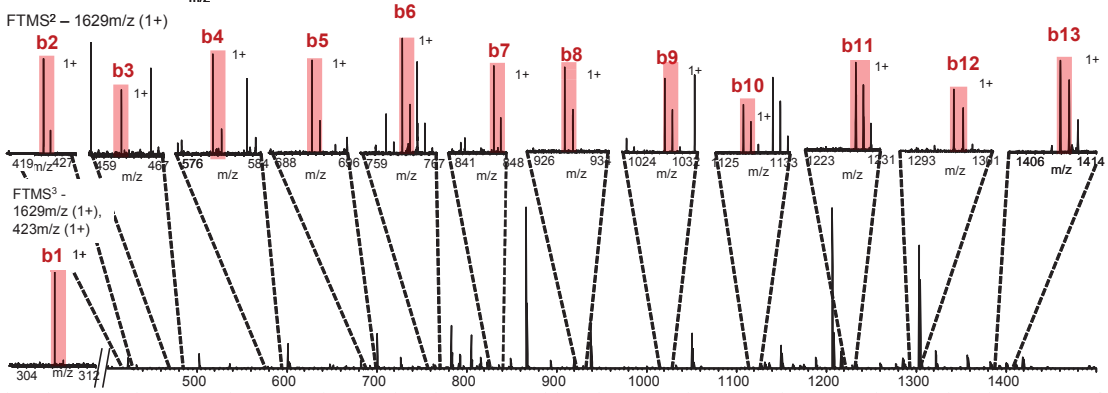




FTMS² – 1629m/z (1+)

Initial MSⁿ sequence tag: G-V-I/L-A-Dhb-Dhb-V-V-T

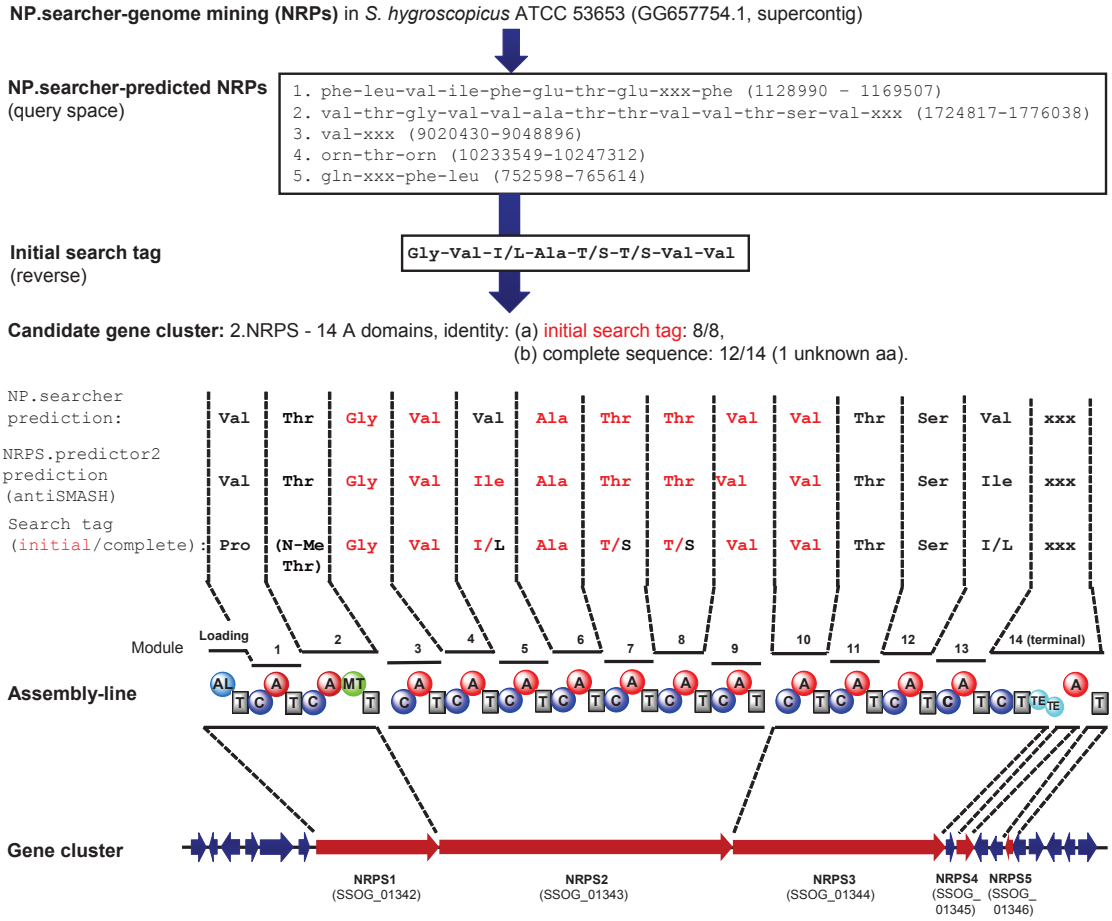




Detector	Precursor ion [m/z]	Observed mass [Da]	Calculated mass [Da]	Difference [Da]	Error	Species
IT	1629 (1+), 502 (1+)	200.172175	200.127375	0.0448	0.00022	y1 (hydrolyzed lactone)
FT	1629 (1+), 423 (1+)	307.250375	307.250275	1E-04	3.3E-07	b1 (hydrolyzed lactone)
IT	1629 (1+), 502 (1+)	313.262175	313.211375	0.0508	0.00016	y2 (hydrolyzed lactone)
FT	1629 (1+)	400.243275	400.243375	-1E-04	-2E-07	y3 (hydrolyzed lactone)
FT	1629 (1+)	422.314075	422.314375	-0.0003	-7E-07	b2 (hydrolyzed lactone)
FT	1629 (1+)	461.324975	461.324975	0	0	b3 (hydrolyzed lactone)
FT	1629 (1+)	501.291275	501.291075	0.0002	4E-07	y4 (hydrolyzed lactone)
FT	1629 (1+)	578.404175	578.404875	-0.0007	-1E-06	b4 (hydrolyzed lactone)
FT	1629 (1+)	600.360175	600.359475	0.0007	1.2E-06	y5 (hydrolyzed lactone)
FT	1629 (1+)	691.488675	691.489475	-0.0008	-1E-06	b5 (hydrolyzed lactone)
FT	1629 (1+)	699.428875	699.427875	0.001	1.4E-06	y6 (hydrolyzed lactone)
FT	1629 (1+)	762.525675	762.526775	-0.0011	-1E-06	b6 (hydrolyzed lactone)
FT	1629 (1+)	782.465975	782.465075	0.0009	1.2E-06	y7 (hydrolyzed lactone)

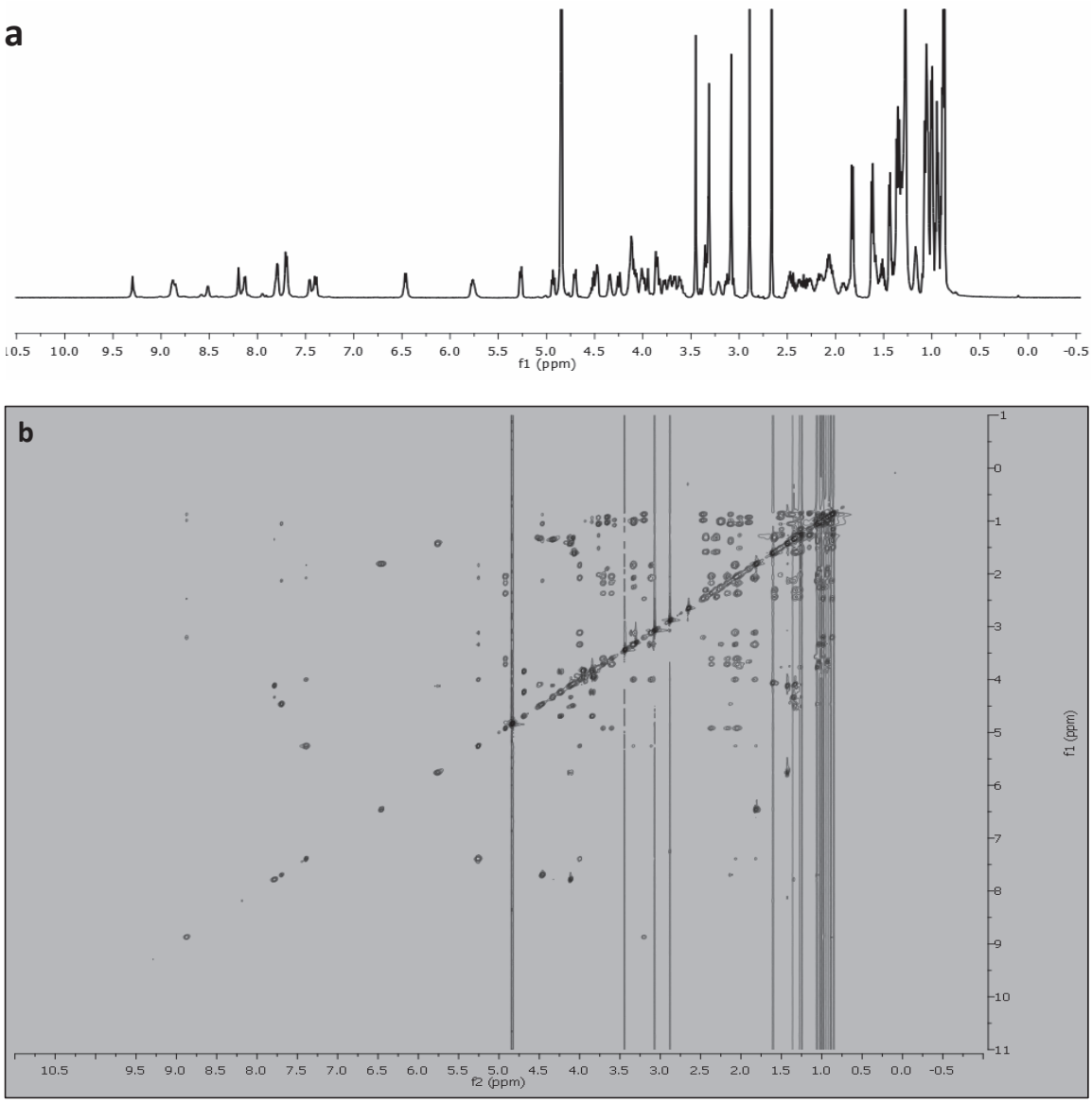
Detector	Precursor ion [m/z]	Observed mass [Da]	Calculated mass [Da]	Difference [Da]	Error	Species
FT	1629 (1+)	845.562675	845.563375	-0.0007	-8E-07	b7 (hydrolyzed lactone)
FT	1629 (1+)	865.503075	865.502175	0.0009	1E-06	y8 (hydrolyzed lactone)
FT	1629 (1+)	928.601275	928.599975	0.0013	1.4E-06	b8 (hydrolyzed lactone)
FT	1629 (1+)	936.540475	936.539275	0.0012	1.3E-06	y9 (hydrolyzed lactone)
FT	1629 (1+)	1027.668675	1027.670175	-0.0015	-1E-06	b9 (hydrolyzed lactone)
FT	1629 (1+)	1049.624675	1049.623375	0.0013	1.2E-06	y10 (hydrolyzed lactone)
FT	1629 (1+)	1126.739275	1126.738975	0.0003	2.7E-07	b10 (hydrolyzed lactone)
FT	1629 (1+)	1148.695575	1148.691775	0.0038	3.3E-06	y11 (hydrolyzed lactone)
FT	1629 (1+)	1205.714975	1205.713175	0.0018	1.5E-06	y12 (hydrolyzed lactone)
FT	1629 (1+)	1227.786675	1227.787675	-0.001	-8E-07	b11 (hydrolyzed lactone)
FT	1629 (1+)	1296.809775	1296.807975	0.0018	1.4E-06	b12 (hydrolyzed lactone)
FT	1629 (1+)	1320.780475	1320.776575	0.0039	3E-06	y13 (hydrolyzed lactone)
FT	1629 (1+)	1409.890975	1409.892875	-0.0019	-1E-06	b13 (hydrolyzed lactone)
FT	1629 (1+)	1417.834775	1417.829275	0.0055	3.9E-06	y14 (hydrolyzed lactone)

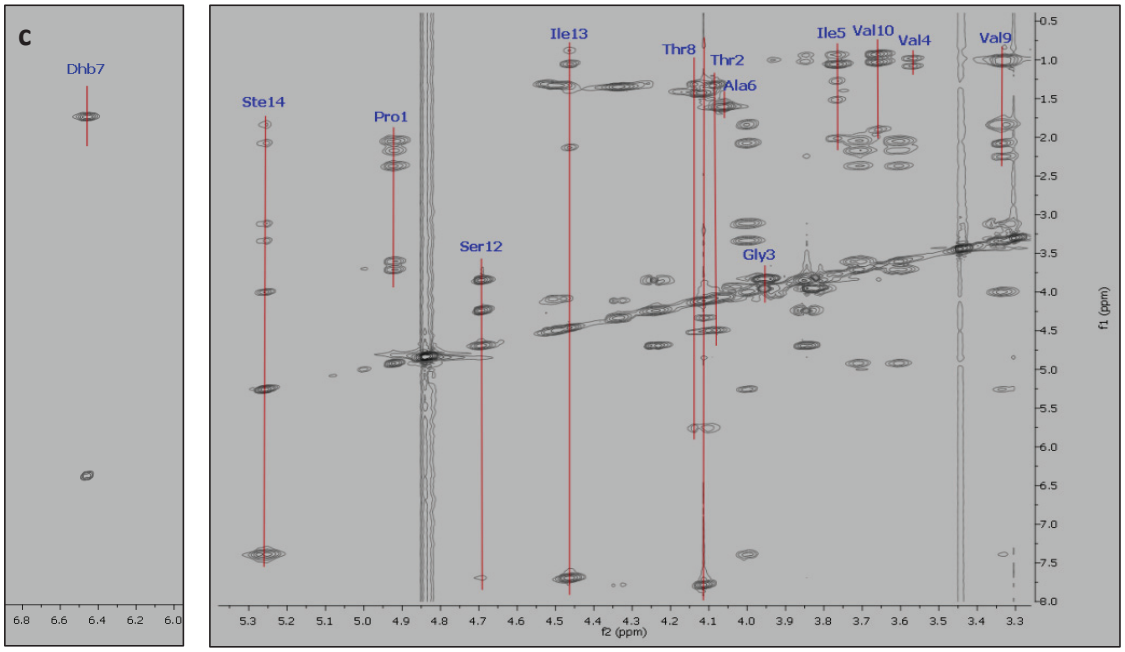
Supplementary Figure 4. Purification and sequence tagging of novel nonribosomal stendomycin lipopeptides from *Streptomyces hygroscopicus* ATCC 53653. (a) SHY-1614 purification of SHY-1600, SHY-1614 and SHY-1628. A compound complex with the predominant mass of 1614 Da was detected by MALDI-imaging around a sporulating *Streptomyces hygroscopicus* ATCC 53653 colony on ISP2 agar. The compound complex SHY-1614 was extracted with n-butanol and partially purified by gel filtration and desalted and concentrated by HPLC. Three mass derivatives (SHY-1600, SHY-1614, SHY-1628) could be further purified by HPLC. **(b)** SHY-1628 sequence tagging. FTMS analysis revealed a (+CH_2/14 Da)-pattern within the SHY-1614 complex indicating the presence of a varying acyl substituent as in lipopeptides. MS^n analysis of SHY-1614 yielded an initial MS^n 8-aa-long sequence tag. Part of the initial MS^n sequence tag and subsequent NMR comparison enabled to dereplicate the SHY-1614 complex in parallel with a lipopeptide complex from marine *Streptomyces* sp. CNQ027 as stendomycin lipopeptides[7]. MS^n analysis verified the full stendomycin I sequence which matched the predicted NRP sequence in 12 of 13 aa.

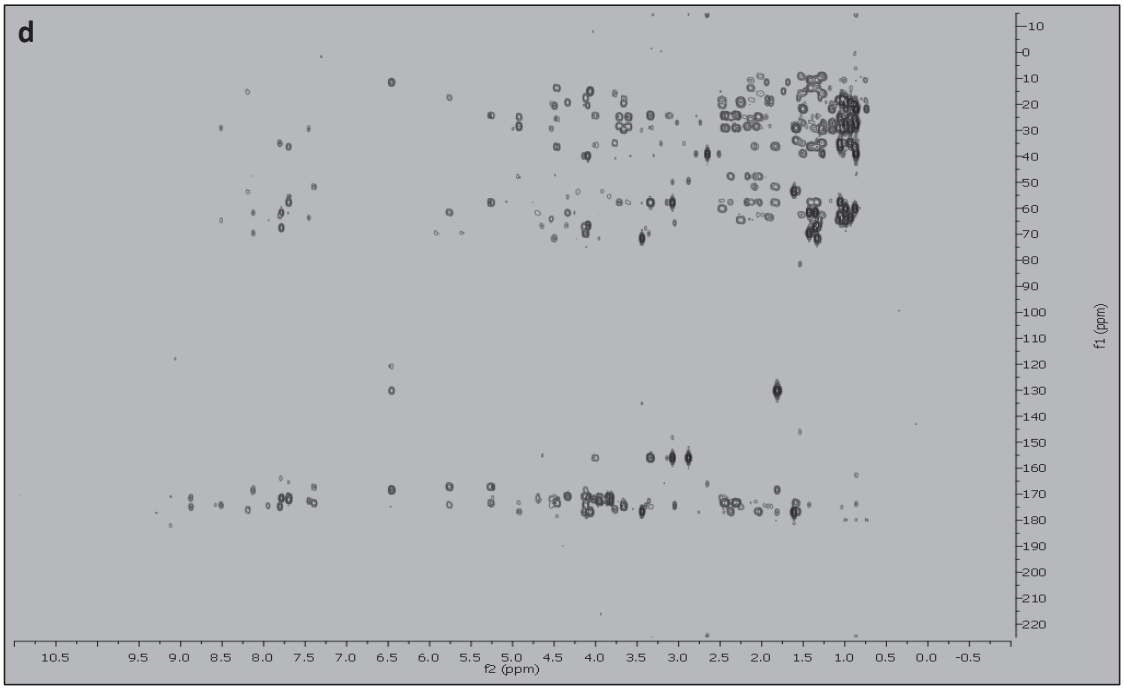


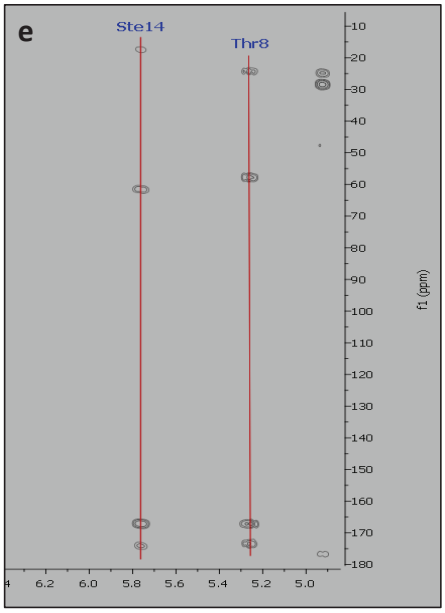
Gene	Size [aa]	Predicted function	Closest homolog [sim%/id%]
SSOG_01336	334	Putative methyltransferase	ref|ZP_07603413.1| hypothetical protein StrviDRAFT_1097 [Streptomyces violaceusniger Tu 4113] (80/68)
SSOG_01337	53	two-component system response regulator	ref|YP_003680358.1| LuxR family transcriptional regulator [Nocardiopsis dassonvillei subsp. dassonvillei DSM 43111] (91/85)
SSOG_01338	429	putative two-component system sensor kinase	ref|NP_824808.1| two-component system sensor kinase [Streptomyces avermitilis MA-4680] (84/71)
SSOG_01339	254	ABC transporter, AMP-binding	ref|NP_824811.1| ABC transporter ATP-binding protein [Streptomyces avermitilis MA-4680] (92/86)
SSOG_01340	832	ABC transporter, integral membrane protein	ref|ZP_06918409.1| ABC transporter integral membrane protein [Streptomyces sviceus ATCC 29083](82/75)
SSOG_01341	212	AMP-dependent ligase	ref|YP_003899809.1| AMP-dependent synthetase and ligase [Cyanothece sp. PCC 7822] (58/43)
SSOG_01342	3115	NRPS1 (AL-T-C-A-T-C-A-MT-T)	dbj|BAH22764.1| nonribosomal peptide synthetase [Microcystis aeruginosa K-139] (53/35)
SSOG_01343	7593	NRPS2 (C-A-T-C-A-T-C-A-T-C-A-T-C-A-T-C-A-T-C-A-T)	gb|AAO72425.1| syringopeptin synthetase C [Pseudomonas syringae pv. syringae] (64/49)
SSOG_01344	5530	NRPS3 (C-A-T-C-A-T-C-A-T-C-A-T-C-T-TE)	ref|ZP_07293262.1| non-ribosomal peptide synthetase, component [Streptomyces hygroscopicus ATCC 53653] (80/71)
SSOG_01345	70	MbtH	ref|ZP_04702846.1| MbtH domain-containing protein [Streptomyces albus J1074] (77/68)
SSOG_01346	508	NRPS4 (A domain)	ref|ZP_07654918.1| amino acid adenylation domain protein [Methylobacter tundripaludum SV96] (58/42)
SSOG_01347	222	acyl-CoA dehydrogenase/oxidase	ref|YP_604645.1| acyl-CoA dehydrogenase-like protein [Deinococcus geothermalis DSM 11300](49/35)
SSOG_01348	360	acyl-CoA dehydrogenase/oxidase	ref|ZP_06918375.1| pimeloyl-CoA dehydrogenase, large subunit [Streptomyces sviceus ATCC 29083](50/35)
SSOG_01349	84	NRPS5 (T domain)	ref|ZP_04689798.1| amino acid adenylation domain-containing protein [Streptomyces ghanaensis ATCC 14672] (66/48)
SSOG_01350	232	Thioesterase, type II	ref|NP_107144.1| short-chain dehydrogenase [Mesorhizobium loti MAFF303099] (42/28)
SSOG_01351	373	SAM-dependent methyltransferase	ref|ZP_07308876.1| conserved hypothetical protein [Streptomyces griseoflavus Tu4000] (59/42)
SSOG_01352	329	transcriptional regulator	ref|ZP_07606462.1| transcriptional regulator, SARP family [Streptomyces violaceusniger Tu 4113](52/34)
SSOG_01353	234	transcriptional regulator, TetR	ref|YP_003494470.1| TetR family regulator [Streptomyces scabiei 87.22] (84/75)
SSOG_01354	502	resistance protein	ref|YP_003494468.1| ABC transporter ransmembrane protein [Streptomyces scabiei 87.22] (88/81)

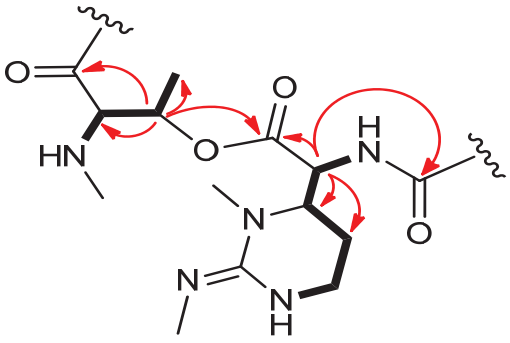
Supplementary Figure 5. Genome mining of novel nonribosomal stendomycin lipopeptides from *Streptomyces hygroscopicus* ATCC 53653. No candidate precursor peptide with the corresponding tag could be identified from the 6-frame translation of the *S. hygroscopicu s* genome. NP.searcher and antiSMASH analysis of the *S. hygroscopicus* supercontig yielded 5 predicted NRPs. Comparison of the search tag with the predicted NRPs identified a predicted 14-aa long NRP with similarity of the predicted NRP substrates and the observed sequence tag. The predicted NRPS gene cluster contained a 14 module assembly line on 5 ORFs with an acyl ligase domain in the loading module. NRPS.predictor2 analysis of the A domains yielded a predicted NRP sequence which matched in 8-aa with the observed initial 8-aa sequence tag.


a
10.5 10.0 9.5 9.0 8.5 8.0 7.5 7.0 6.5 6.0 5.5 5.0 4.5 4.0 3.5 3.0 2.5 2.0 1.5 1.0 0.5 0.0 -0.5
f1 (ppm)
b
1 0 1 2 3 4 5 6 7 8 9 10 11
f1 (ppm)
10.5 9.5 9.0 8.5 8.0 7.5 7.0 6.5 6.0 5.5 5.0 4.5 4.0 3.5 3.0 2.5 2.0 1.5 1.0 0.5 0.0 -0.5
f2 (ppm)


c
Dhb7
Ste14
Pro1
Ser12
Ile13
Thr8
Thr2
Ala6
Gly3
Ile5
Val10
Val4
Val9
f2 (ppm)
f1 (ppm)

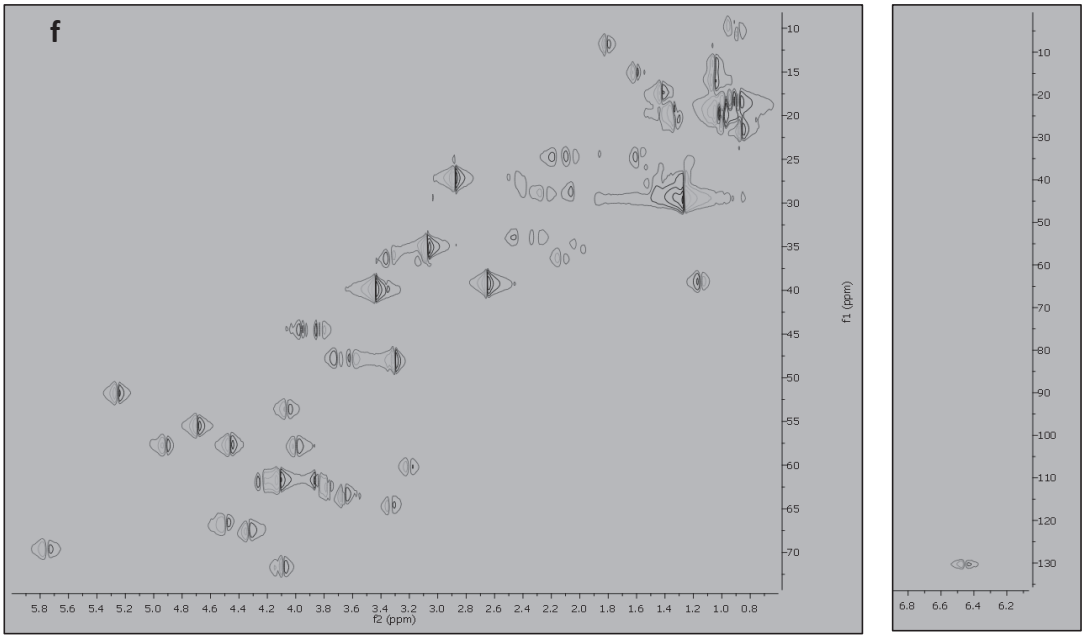

d
f2 (ppm)
f1 (ppm)

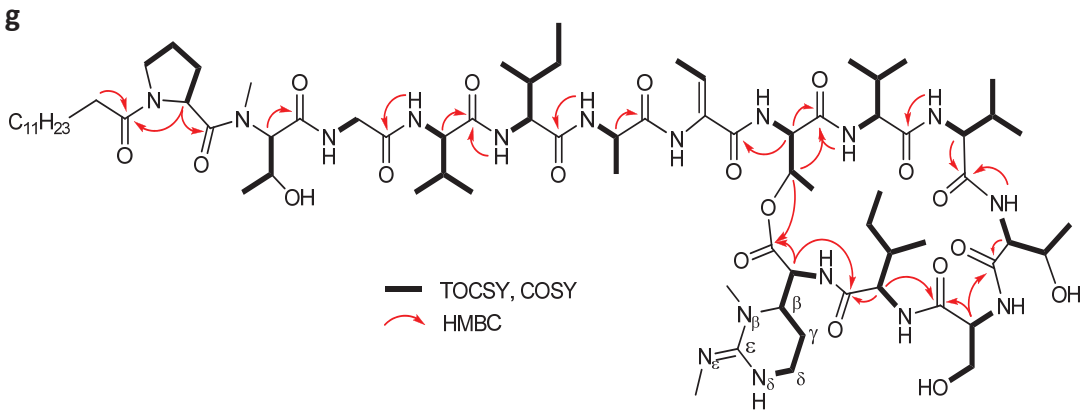




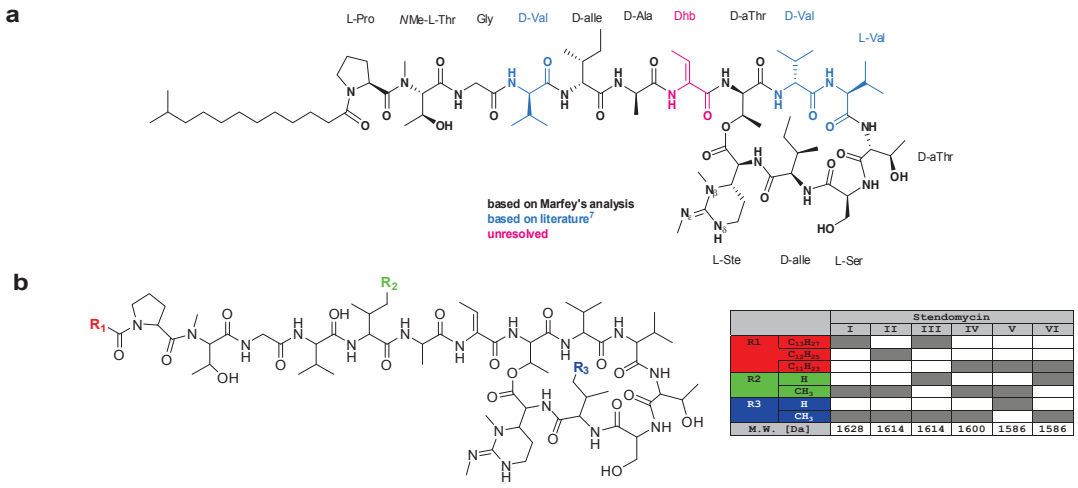
TOCSY, COSY

HMBC

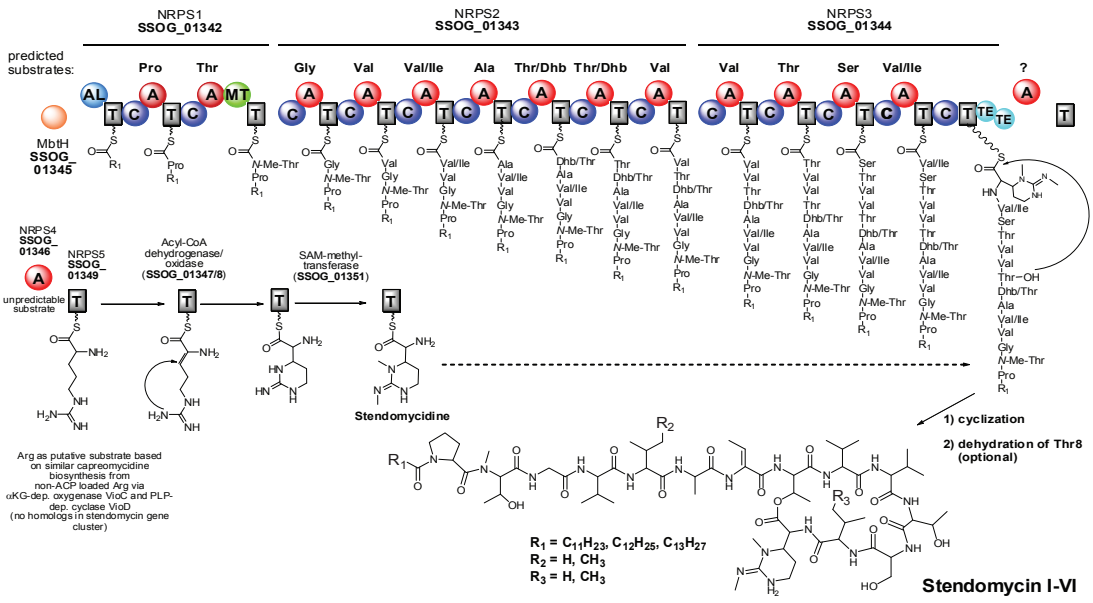




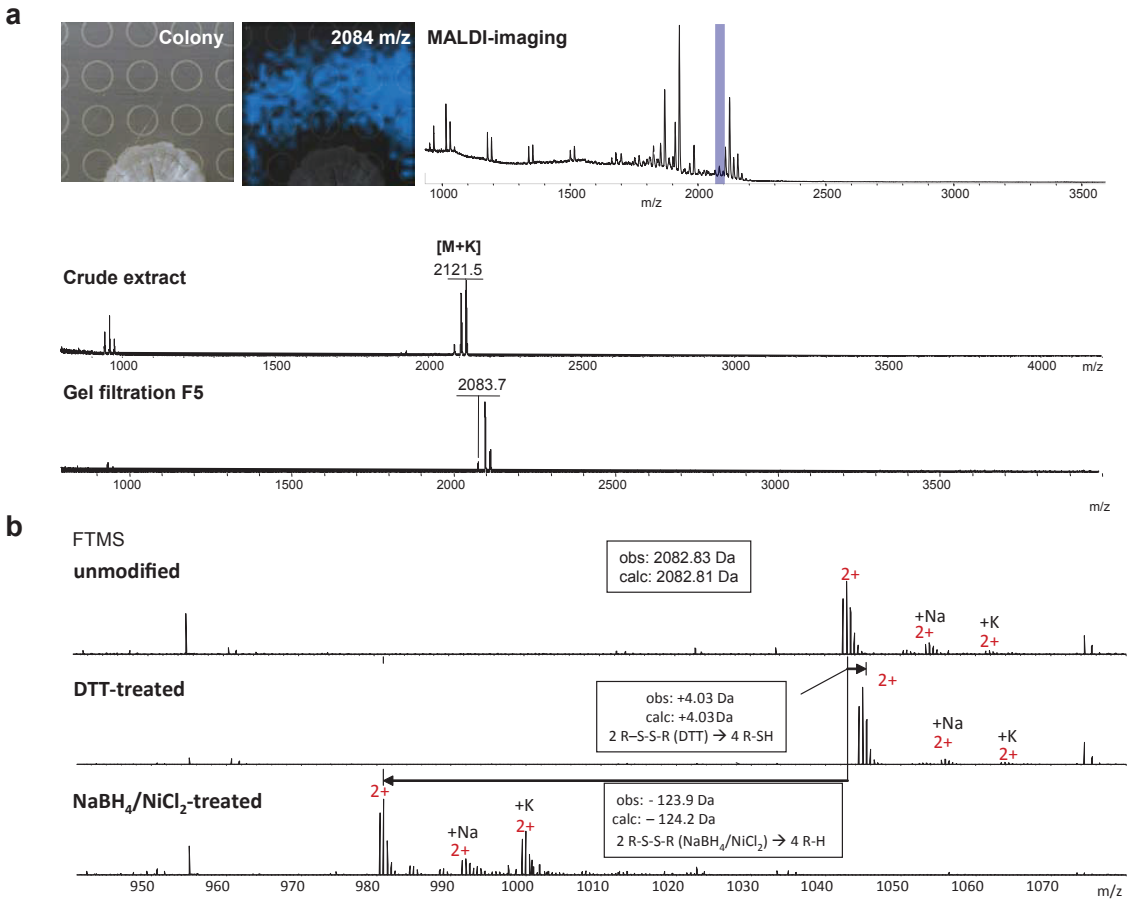
Supplementary Fig. 6. NMR spectra of CNQ27-1628. NMR analysis of the dominant component of the lipopeptide complexes in *S. hygroscopicus* ATCC 53653 and *S.* sp. CNQ27 was done with the 1628 Da-derivative from *S.* sp. CNQ27 (Q027-1628) which matched in 1D and 2D NMR spectra with SHY-1628. 1D and 2D NMR analysis of purified Q027-1628 (stendomycin I) verified the MS^n structural assignment and furthermore the Thr8→Ste14 macrolactone bond, the N-methylation of Thr3, the stendomycidine residue at position 14 and the Dhb residue at position 7. (**a**) ^1H NMR spectrum of CNQ27-1628. The spectrum was observed in CD_3OD, 600 MHz. The detailed annotations were listed in **Supplementary Table 13**. **(b), (c)** ^1H-^1H TOCSY spectra and annotations of CNQ27-1628. The spectrum was observed in CD_3OD, 600 MHz, with mixing time = 90 ms. Subfigure b is a full spectrum, subfigure c is a zoom in the spectrum with annotations. **(d), (e)** ^1H-^{13}C HMBC spectrum of CNQ27-1628. The spectrum was observed in CD_3OD, 600 MHz, with $^{2,3}J_{1H/13C}$ = 7 Hz. The full ^1H-^{13}C HMBC spectrum is shown. The annotations for critical signals supporting the Thr8→Ste14-macrolactone bond are displayed in subfigure e. **(f)** ^1H-^{13}C HSQC spectrum of CNQ27-1628. The spectrum was observed in CD_3OD, 600 MHz, with $^1J_{1H/13C}$ = 145 Hz. ^1H-^{13}C HSQC spectrum was collected to assist in HMBC spectrum annotation. **(g)** Representation of NMR assignment of CNQ27-1628.



Supplementary Figure 7. Characterization of novel nonribosomal stendomycin lipopeptides from *Streptomyces hygroscopicus* ATCC 53653 and *Streptomyces* sp. CNQ27. (a) Stereochemical assignment of CNQ27-1628 based on Marfey's analysis and literature[7] (see **Supplementary Methods**). **(b)** Characterization of 5 novel stendomycin derivatives by MS^n. MS^n analysis of other components of the *S. hygroscopicus* stendomycin complex enabled the structure elucidation of 5 new stendomycin derivatives (II-VI) which differ either in the acyl chain or Val/Ile at position 5 and 13 (**Supplementary Table 8-12**). The acyl chains of stendomycin I-VI were assumed to be the reported stendomycin-like branched fatty acids[7] based on the MS-determined molecular formulas ($C_{11}H_{23}$ – 10-methylundecanoic acid, $C_{12}H_{25}$ – 11-methyldodecanoic acid, $C_{13}H_{27}$ – 12-methyltridecanoic acid).



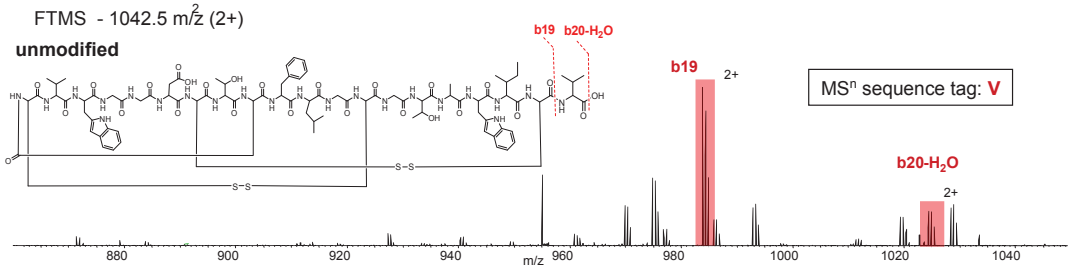
Supplementary Figure 8. Biosynthetic schem of stendomycin lipopeptides based on the corresponding *Streptomyces hygroscopicus* ATCC 53653 gene cluster. Based on the elucidated stendomycin structures and the characterized biosynthetic gene cluster in *S. hygroscopicus* ATCC 53653, a biosynthetic schem could be predicted. Herein, the acyl ligase is responsible for the observed variability of the N-acyl chains as acyl ligases are known to have promiscuity in integrating various acyl substrates into NRPs[33]. Given the high degree of D-aa and allo-aa in stendomycin lipopeptides[7], CLUSTALW2 analysis of all C domains revealed that C domains in NRPS2 (SSOG_01343) and NRPS3 (SSOG_01344) which incorporate for all D-aa and *allo*-aa cluster closest to C domains with dual epimerization/condensation activity and, thus, might catalyze epimerization of corresponding L-aa substrates. The biosynthesis of Dhb and of stendomycidine in the absence of similar capreomycidine biosynthetic genes[34] are open questions in the stendomycin biosynthesis which might be answered with the biosynthetic gene cluster in hand.


a
Colony
2084 m/z
MALDI-imaging
1000
1500
2000
2500
3000
3500
m/z
[M+K]
2121.5
Crude extract
1000
1500
2000
2500
3000
3500
4000
m/z
2083.7
Gel filtration F5
1000
1500
2000
2500
3000
3500
m/z
b
FTMS
unmodified
obs: 2082.83 Da
calc: 2082.81 Da
2+
+Na
2+
+K
2+
DTT-treated
obs: +4.03 Da
calc: +4.03Da
2 R–S-S-R (DTT) → 4 R-SH
2+
+Na
2+
+K
2+
NaBH4/NiCl2-treated
2+
+Na
2+
+K
2+
obs: - 123.9 Da
calc: – 124.2 Da
2 R-S-S-R (NaBH4/NiCl2) → 4 R-H
950
960
970
980
990
1000
1010
1020
1030
1040
1050
1060
1070
m/z

b, continued

FTMS - 1042.5 m/z (2+)

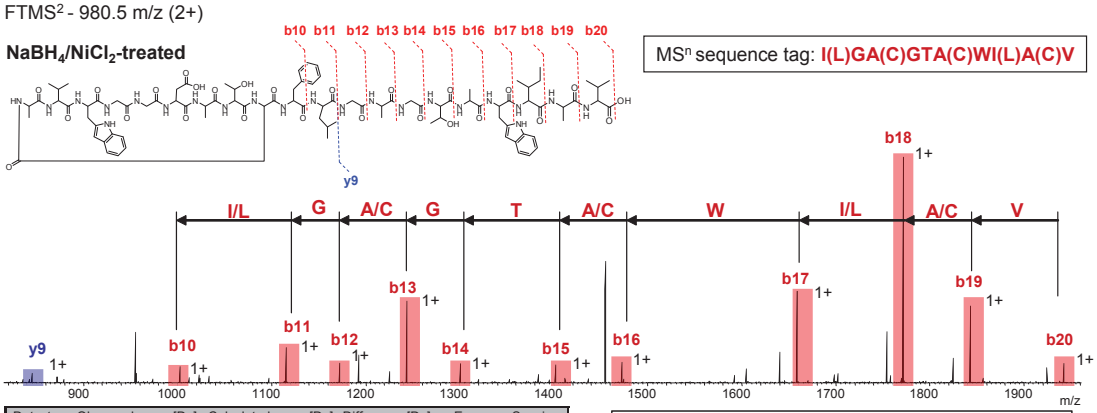
unmodified



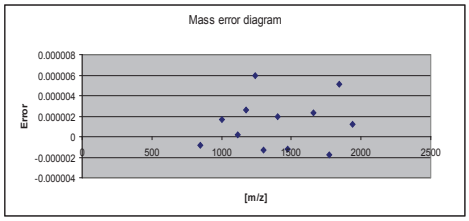
Detector	Observed mass [Da]	Calculated mass [Da]	Difference [Da]	Error	Species
FT	2047.80346	2047.79637	0.00709	3.46226E-06	b20-H2O
FT	1966.740255	1966.73852	0.001735	8.82171E-07	b19

FTMS^2 - 980.5 m/z (2+)

$NaBH_4$/$NiCl_2$-treated



Detector	Observed mass [Da]	Calculated mass [Da]	Difference [Da]	Error	Species
FT	1941.95227	1941.94997	0.0023	1.18E-06	b20
FT	1842.8909	1842.88155	0.00935	5.07E-06	b19
FT	1771.84123	1771.84444	-0.00321	-1.81E-06	b18
FT	1658.76429	1658.76038	0.00391	2.36E-06	b17
FT	1472.67931	1472.68106	-0.00175	-1.19E-06	b16
FT	1401.6467	1401.64395	0.00275	1.96E-06	b15
FT	1300.59462	1300.59627	-0.00165	-1.27E-06	b14
FT	1243.58225	1243.57481	0.00744	5.98E-06	b13
FT	1172.54079	1172.5377	0.00309	2.64E-06	b12
FT	1115.51643	1115.51623	0.0002	1.79E-07	b11
FT	1002.43382	1002.43217	0.00165	1.65E-06	b10
FT	845.45149	845.45217	-0.00068	-8.04E-07	y9



C

6-frame translation – search tag: **I(L)GA(C)GTA(C)WI(L)A(C)V**

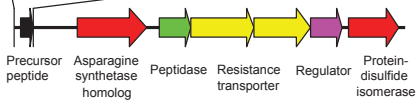
```
[…]PSAYRWTARGCRAPWWARSSSXPSSGPSAPGWRSSCGARRCPSACXSRSTSSCRPSCSTSPASRRSPT
TSPPRRARRSWRWPSRATRPCPRAPGWPCCSGGRPPRSSAATSPCAPGTCDPGPLRDPXEDNRPGRPGHRA
TYVAGGHRPVLISTTNGQGTPMTSTDELYEAPELIEIGDYAELTRCVWGGDCTDFLGCGTAWICVXPTDRL
TAPXGRSPTTXWWRQAGQQPASTVRLLNGRDEVNTRMDFLILPDREEAARLLPDHLARSDDEMIRHPSGRP
WLLGRWEPHELTVVTAGARRLVMLGPTRIDHPTVERVLGRARTLHGLDAVARSLPGNPYLIASMDGQV[…]
```

Candidate precursor peptide

MLISTTNGQGTPMTSTDELYEAPELIEIGDYAELTR**CVWGGDCTDFLGCGTAWICV**

2104.85 Da
- H_2O (cyclization)
- 4 H (2 DSB formations)
2082.81 Da (calc, obs: 2082.83 Da)

Gene cluster



Gene	Size [aa]	Predicted function	Closest homolog [sim%/id%]
SSEG_05172	56	precursor peptide	ref|YP_003101077.1| transcriptional regulator, AraC family [*Actinosynnema mirum* DSM 43827] (83/62)
SSEG_09672	374	lactamase	ref|NP_821433.1| asparagine synthetase [*Streptomyces avermitilis* MA-4680] (52/41)
SSEG_09764	165	peptidase	ref|YP_003101080.1| hypothetical protein Amir_3329 [*Actinosynnema mirum* DSM 43827] (66/54)
SSEG_09675	325	antibiotic resistance ABC transporter	ref|YP_003298018.1| daunorubicin resistance ABC transporter ATPase subunit [*Thermomonospora curvata* DSM 43183] (72/61)
SSEG_05178	294	antibiotic resistance ABC transporter	ref|YP_003298019.1| ABC-2 type transporter [*Thermomonospora curvata* DSM 43183] (69/48)
SSEG_05179	162	Doxx family protein	gb|ACU95843.1| DoxX protein [*Saccharomonospora viridis* DSM 43017] (70/57)
SSEG_05180	272	protein disulfide isomerase	ref|YP_003098760.1| protein-disulfide isomerase-like protein [*Actinosynnema mirum* DSM 43827] (54/41)

Asparagine synthetase homolog

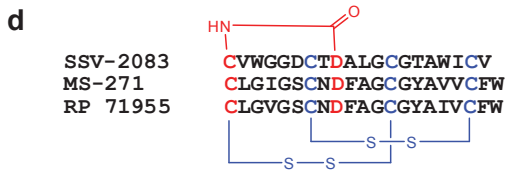
```
S.sviceus       ---------------------------------------------MDSTTLCFVAAAEGA 15
AS              MCSIFGVFDIKTDAVELRKKALELSRLMRHRGPDWSGIYASDNAILAHERLSIVDVNAGA 60
                                                             :    *.:* .  **

S.sviceus       DLLTYHVEPLDRANEDARWARLAAERLTDARHLTVPSQGASSWFDLPVSDG--------- 66
AS              QPLYNQQKTHVLAVNGEIYNHQALRAEYGDRYQFQTGSDCEVILALYQEKGPEFLDDLQG 120
                : *  : ::.   * :.  : : * .   . *:   ......  : *  ..*
```

A cysteine at position 2 indicates a Gln-hydrolyzing Asn synthetase, any other residue at position 2 indicates a β-lactamase homolog as in lassopeptide biosynthetic pathways.

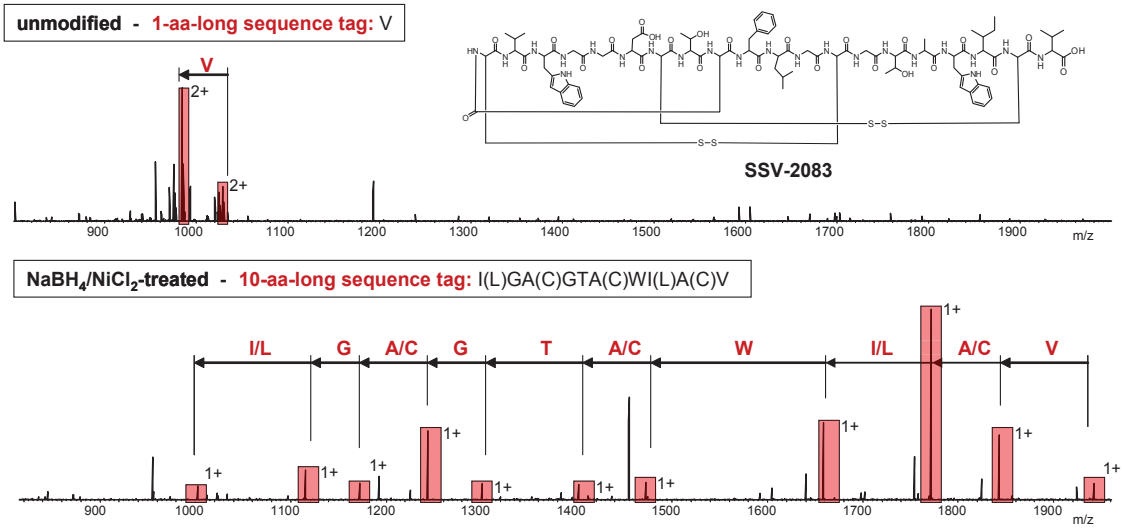
Peptidase

```
S.sviceus peptidase  ----------------------MSIPVAVEGSVPLPWRRRVTARSAAGA 27
CapB                 MQRSRYFLYDEACAKAFADHYLDFKPIDAPHALKPLISDRIVVAASPASV 50
McjB                 --MIRYCLTSYREDLVILDIINDSFSIVPDAGSLLKERDKLLKEFPQLSY 48
                                                   .  .       : :   .  .

S.sviceus peptidase  ARLLVR-----------------------------LPPRRLCQ-VLRFVS 47
CapB                 PKRIADYRGWAFDTFDSGIWASRTLGERSAAGFEWLPFWRIVRGAVSLKM 100
McjB                 FFDSEYHIGSVSRNSDTSFLEERWFLPEPDKTLYKCSLFKRFILLLKVFY 98
                                                     .  :     : .

S.sviceus peptidase  RGS---RPADMERALAARQAVVTVSLRCAGIAG----------------- 77
CapB                 RGF---RALSALDRLARLDAGAEQRVRADGGPSRTAERYLRASIWSPFRI 147
McjB                 YSWNIEKKGMAWIFISNKKENRLYSLNEEHLIRKEISNLSIIFHLNIFKS 148
                      .    :   *   ::  .      :.

S.sviceus peptidase  -CLQRSVATALLCRLAGRWPDWCSGFRTRPPGAHAWVEVDGTAIGEPGDM 126
CapB                 TCLQMSFALATHLRRENVPAQLVIGVRPMPFVAHAWVEIDGRVCGD---- 193
McjB                 DCLTYSYALKRILNSRNIDAHLVIGVKTQPIYSHSWVEVGGQVIND---- 194
                      **  * *     .  .  ..   *.*. ** :*:***:.* . .:

S.sviceus peptidase  TLFHTVLSVRHQDRDQHLHQGRRQARRQARAGRHEGRQP 165
CapB                 ----------EPELKKSYGEIYRTPRHDERAGPFGLAA- 221
McjB                 ----------APNMRDKLSVIAEI--------------- 208
```

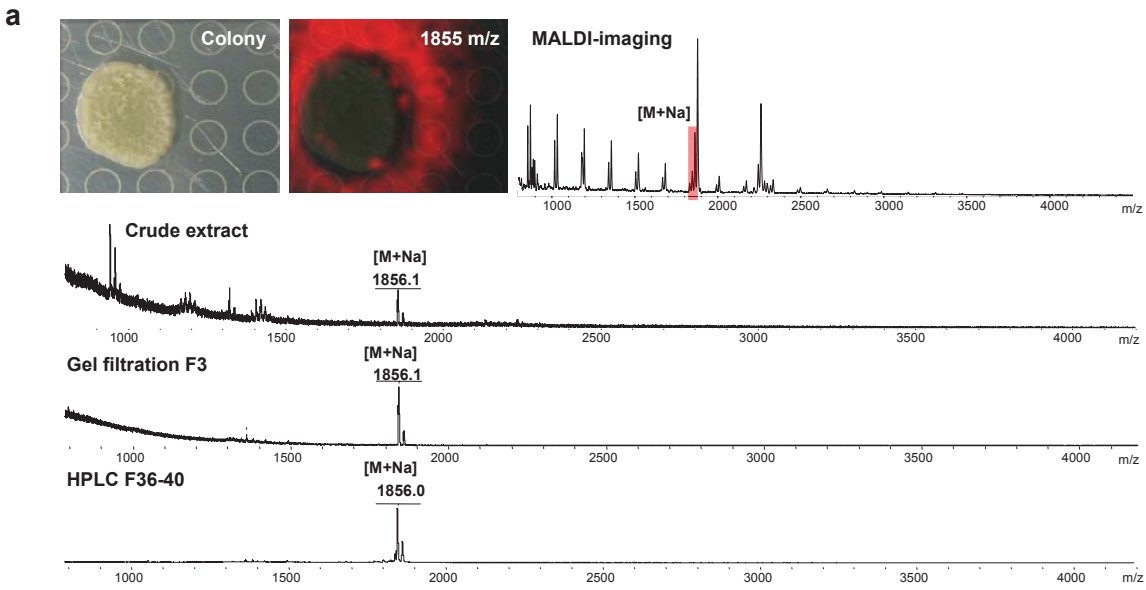
Conserved residues with CapB and McjB (lassopeptide peptidases)
Putative catalytic triad residues including cysteine or serine

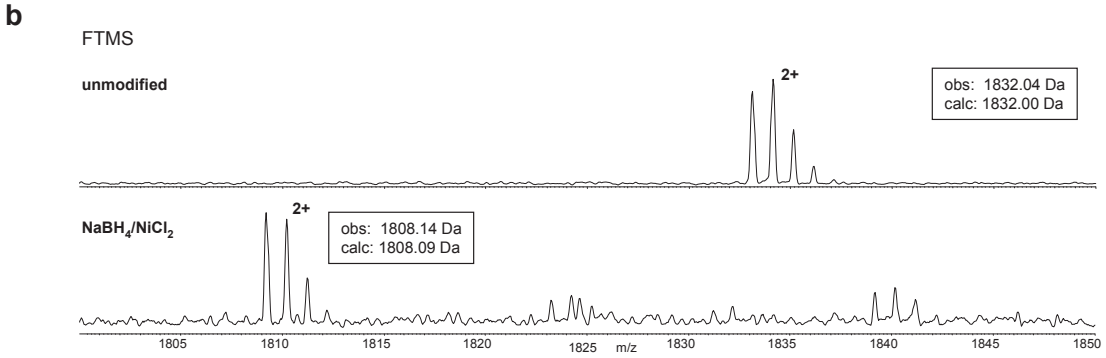


Supplementary Figure 9. Characterization of class I lassopeptide SSV-2083 from *Streptomyces sviceus* ATCC 20983 and its biosynthetic gene cluster by NPP. (a) SSV-2083 imaging and purification. MALDI-imaging of sporulating *Streptomyces sviceus* ATCC 20983 revealed a secreted mass of 2083 Da. **(b)** SSV-2083 sequence tagging. SSV-2083 showed a poor MS/MS fragmentation behaviour indicating a cyclic or constrained structure. Treatment of SSV-2083 with DTT and $NaBH_4/NiCl_2$ resulted in a +4 Da and -124 Da mass shift, respectively, suggesting the presence of 2 disulfide bonds (DSB). MS^2 analysis of $NaBH_4/NiCl_2$-treated, i.e. DSB-deconstrained SSV-2083 yielded a 10-aa-long MS^n sequence tag. **(c)** Genome mining of SSV-2083 gene cluster. A candidate precursor peptide of 56-aa comprised the search tag and 4 cysteines in the C-terminal region. The putative core peptide ^{37}C-^{56}V had a calculated mass of 2104.85 Da which gave the observed mass of 2082.81 Da after subtraction of 4 Da (loss of 4 protons via 2 DSB formations) and 18 Da (water loss via cyclization). The gene cluster contained a homolog of an Asn synthetase, a protease, a protein disulfide isomerase (PDI) and resistance transporters. The homolog of an Asn synthetase and the peptidase indicated a lassopeptide[8] and the PDI, the 2 observed DSBs and the Cys at position 1 in the core peptide sequence suggested a class I lassopeptide[9]. **(d)** SSV-2083 sequence alignment with 2 other class I lassopeptides showed that the cysteines are at class I lassopeptide-conserved sites in SSV-2083[35,36]. The SSV-2083 gene cluster is the first reported gene cluster of a class I lassopeptide.



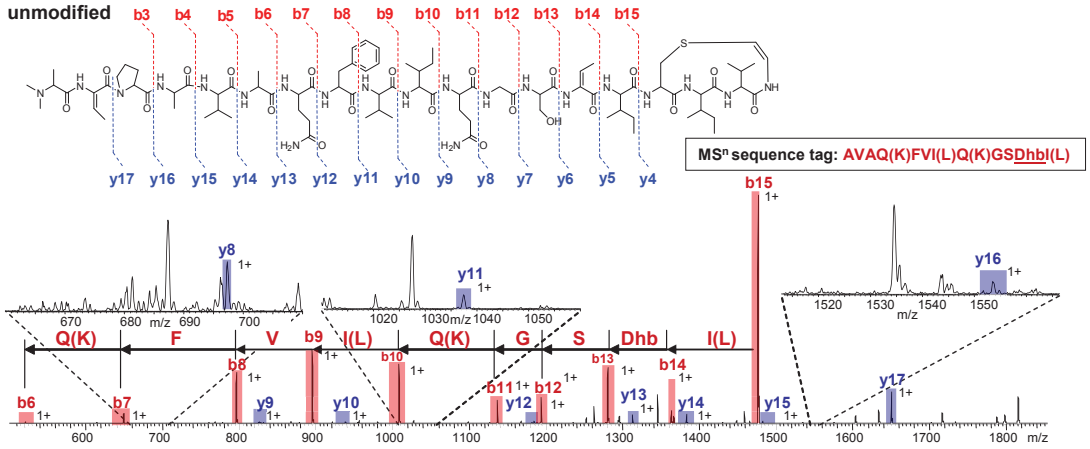
Supplementary Figure 10. Exemplified effect of structural deconstraining by $NaBH_4/NiCl_2$-treatment on sequence tagging of class I lassopeptide SSV-2083 from *Streptomyces sviceus* ATCC 20983.


a
Colony
1855 m/z
MALDI-imaging
[M+Na]
1000
1500
2000
2500
3000
3500
4000
m/z
Crude extract
[M+Na]
1856.1
Gel filtration F3
[M+Na]
1856.1
HPLC F36-40
[M+Na]
1856.0

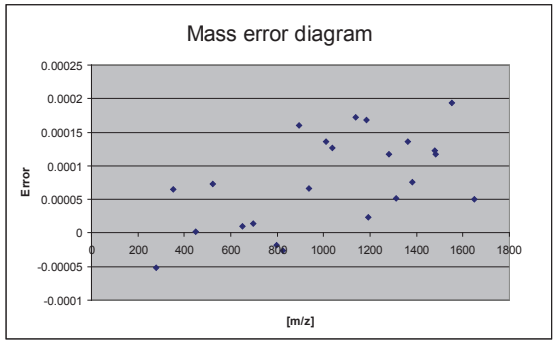

b
FTMS
unmodified
2+
obs: 1832.04 Da
calc: 1832.00 Da
NaBH4/NiCl2
2+
obs: 1808.14 Da
calc: 1808.09 Da
1805
1810
1815
1820
1825
m/z
1830
1835
1840
1845
1850

b, continued

ITMS^2 - 917 m/z (2+)



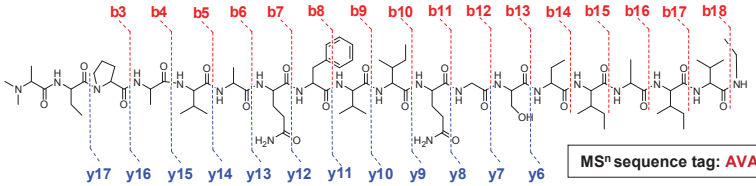
Detector	Observed mass [Da]	Calculated mass [Da]	Difference	Error	Species
IT	280.18	280.16557	-0.01443	-5.15053E-05	b3
IT	351.18	351.20268	0.02268	6.45781E-05	b4
IT	450.27	450.2711	0.0011	2.44297E-06	b5
IT	521.27	521.30821	0.03821	7.32964E-05	b6
IT	649.36	649.36679	0.00679	1.04563E-05	b7
IT	697.36	697.37016	0.01016	1.4569E-05	y8
IT	796.45	796.4352	-0.0148	-1.85828E-05	b8
IT	825.45	825.42874	-0.02126	-2.57563E-05	y9
IT	895.36	895.50361	0.14361	0.000160368	b9
IT	938.45	938.5128	0.0628	6.69144E-05	y10
IT	1008.45	1008.58768	0.13768	0.000136508	b10
IT	1037.45	1037.58121	0.13121	0.000126458	y11
IT	1136.45	1136.64626	0.19626	0.000172666	b11
IT	1184.45	1184.64963	0.19963	0.000168514	y12
IT	1193.64	1193.66772	0.02772	2.32225E-05	b12
IT	1280.55	1280.69975	0.14975	0.000116928	y13
IT	1312.64	1312.7082	0.0682	5.19537E-05	b13
IT	1363.55	1363.73631	0.18631	0.000136617	y14
IT	1383.64	1383.74532	0.10532	7.61123E-05	b14
IT	1476.64	1476.82093	0.18093	0.000122513	y15
IT	1482.64	1482.81373	0.17373	0.000117162	b15
IT	1553.55	1553.85085	0.30085	0.000193616	y16
IT	1650.82	1650.90361	0.08361	5.0645E-05	y17



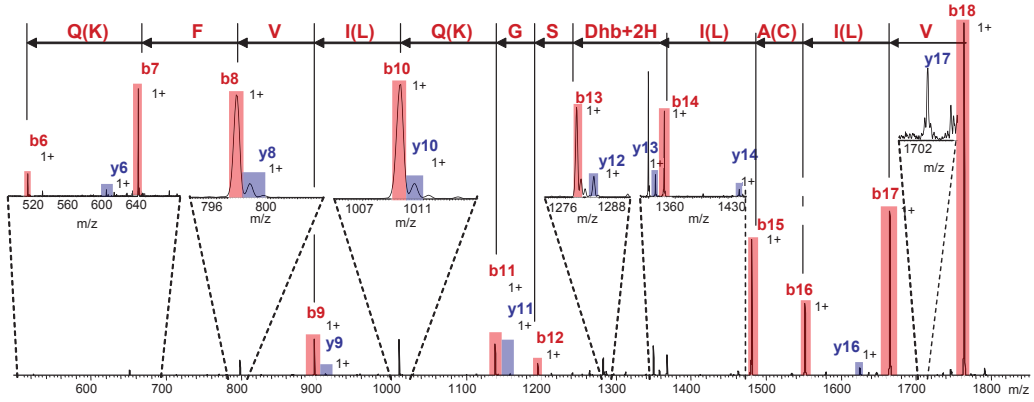
b, continued

ITMS^2 - 905 m/z (2+)

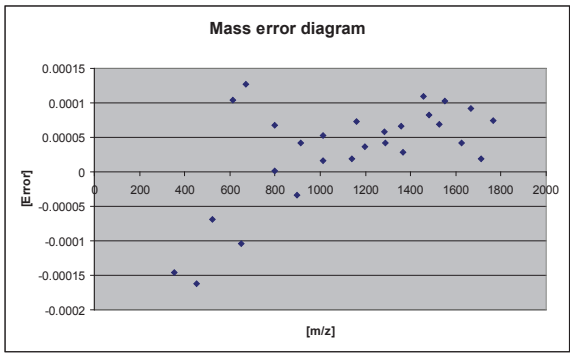
$NaBH_4/NiCl_2$



MS^n sequence tag: AVAQ(K)FVI(L)Q(K)GSDhb+2HI(L)A(C)I(L)V



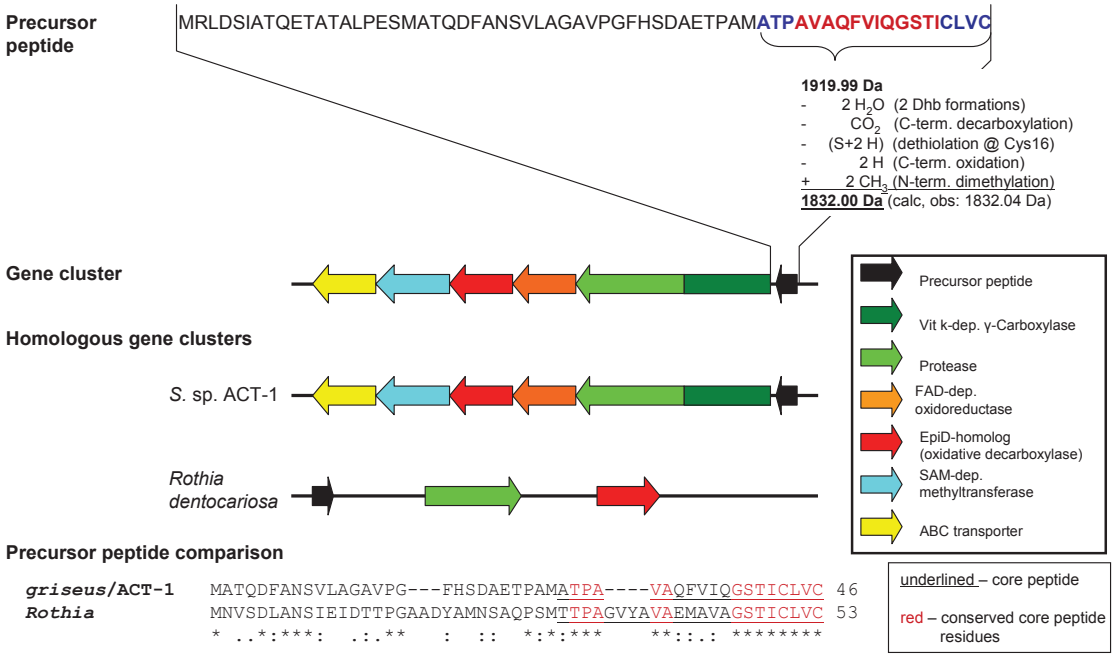
Detector	Observed mass [Da]	Calculated mass [Da]	Difference	Error	Species
IT	353.27	353.21833	-0.05167	-0.000146283	b4
IT	452.36	452.28675	-0.07325	-0.000161955	b5
IT	523.36	523.32386	-0.03614	-6.90586E-05	b6
IT	614.36	614.42357	0.06357	0.000103463	y6
IT	651.45	651.38238	-0.06761	-0.000103801	b7
IT	671.36	671.44504	0.08504	0.000126652	y7
IT	798.45	798.45085	0.00085	1.06456E-06	b8
IT	799.45	799.50361	0.05361	6.70541E-05	y8
IT	897.55	897.51926	-0.03074	-3.425E-05	b9
IT	912.55	912.58768	0.03768	4.12892E-05	y9
IT	1010.55	1010.60333	0.05333	5.27705E-05	b10
IT	1011.64	1011.65609	0.01609	1.59046E-05	y10
IT	1138.64	1138.66191	0.02191	1.92419E-05	b11
IT	1158.64	1158.72451	0.08451	7.29336E-05	y11
IT	1195.64	1195.68337	0.04337	3.62721E-05	b12
IT	1282.64	1282.7154	0.0754	5.87816E-05	b13
IT	1286.73	1286.78308	0.05308	4.12502E-05	y12
IT	1357.73	1357.8202	0.0902	6.643E-05	y13
IT	1367.73	1367.76816	0.03816	2.78995E-05	b14
IT	1456.73	1456.88861	0.15861	0.000108869	y14
IT	1480.73	1480.85223	0.12223	8.25403E-05	b15
IT	1527.82	1527.92573	0.10573	6.91984E-05	y15
IT	1551.73	1551.88934	0.15934	0.000102675	b16
IT	1624.91	1624.97849	0.06849	4.21483E-05	y16
IT	1664.82	1664.9734	0.1534	9.21336E-05	b17
IT	1710	1710.03125	0.03125	1.82745E-05	y17
IT	1763.91	1764.04182	0.13182	7.47261E-05	b18



c

6-frame translation – search tag: **AVAQ(K)FVI(L)Q(K)GSTI(L)**

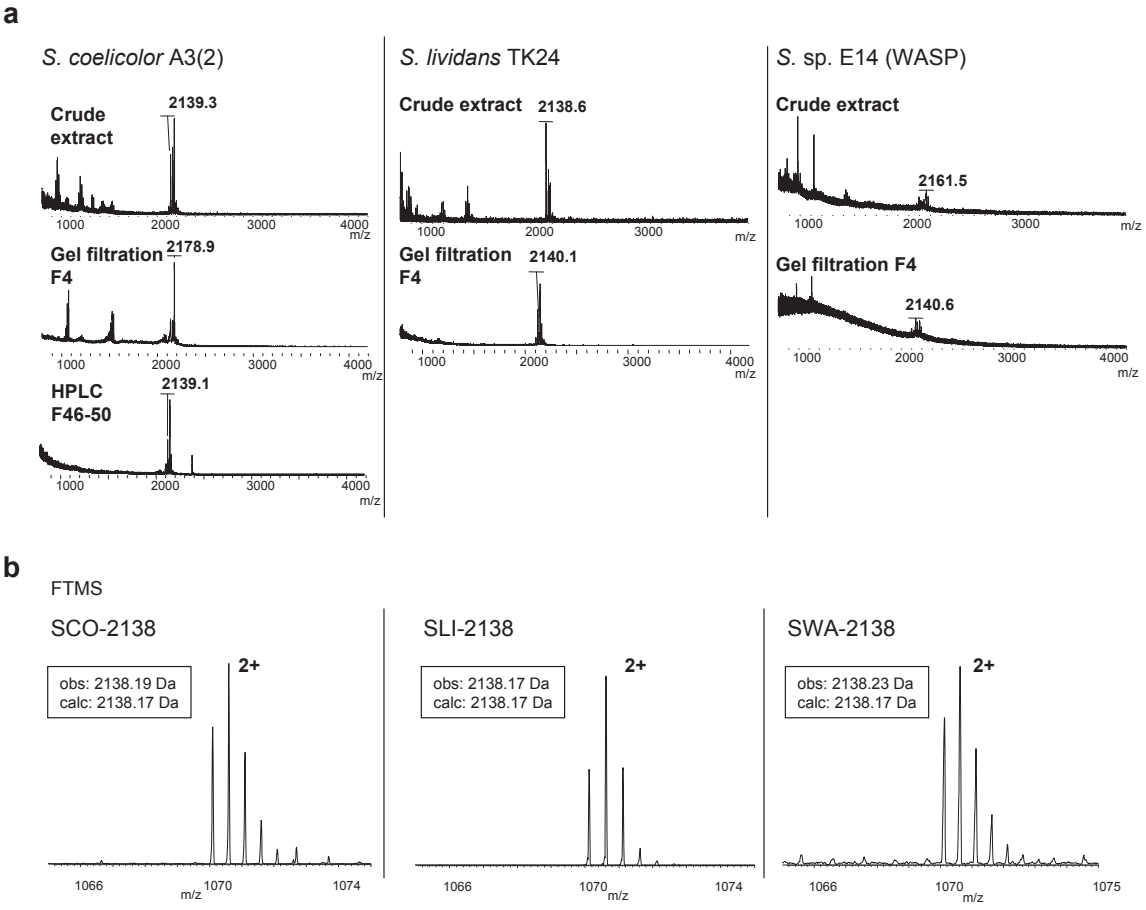
```
[…]VQGLGPSPHKTRTRSGPCSGASAHARRPAPGHRPTREGPGERPGVRRNRHRHEIARQARGNLSKFSGT
XGLXSXYSCDGWGVDGTRTSVYAGSAMLITSFAGFPQIGWPTXCIARSIQSWRRGMGKMGPPSPFRSAXVP
RKDSEMRLDSIATQETATALPESMATQDFANSVLAGAVPGFHSDAETPAMATPAVAQFVIQGSTICLVCXW
WADSRVGGVRPGAAPPARASGWLTGTFEFATERSXRASXDFDGGRGALRARPLPKGLPIASSCRSRSDRFS
LLPRPXASSTFSTTTDISEAYRPRNTLREVCDTGSAARSTIRVPRKCSPGXHSRRRPGWPSGAA[…]
```



S. griseus **gene cluster**

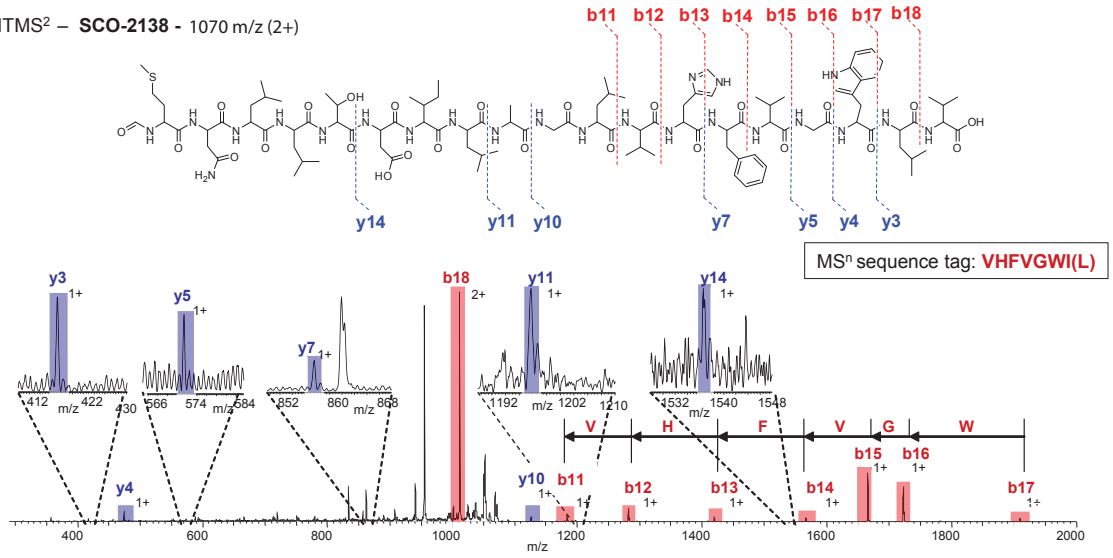
Gene	Size	Predicted function	Closest homolog [sim/id]
SGR_6361	211	putative ABC transporter ATP-binding protein	ref|ZP_06282342.1| ABC transporter related protein [*Streptomyces* sp. ACT-1](99/99)
SGR_6362	246	putative methyltransferase	ref|ZP_06282341.1| Methyltransferase type 11 [*Streptomyces* sp. ACT-1](100/99)
SGR_6363	191	putative phosphopantothenoyl-cysteine synthetase/decarboxylase (EpiD homolog)	ref|ZP_06282340.1| flavoprotein [*Streptomyces* sp. ACT-1] (100/100)
SGR_6364	189	FAD dependent oxidoreductase	ref|ZP_06282339.1| hypothetical protein SACT1DRAFT_6364 [*Streptomyces* sp. ACT-1] (100/100) ref|YP_003061191.1| FAD dependent oxidoreductase [*Hirschia baltica* ATCC 49814] (48/25)
SGR_6365	586	Protease	ref|ZP_06282338.1| hypothetical protein SACT1DRAFT_6363 [*Streptomyces* sp. ACT-1] (100/99) ref|ZP_06281095.1| alpha/beta hydrolase fold protein [Streptomyces sp. ACT-1](43/29)
N/A	46	precursor peptide	ref|ZP_06282337.1| hypothetical protein SACT1DRAFT_6362 [*Streptomyces* sp. ACT-1] (100/100) ref|ZP_06905928.1| conserved hypothetical protein [*Rothia dentocariosa* ATCC 17931](54/41)

Supplementary Figure 11. Characterization of novel ribosomal peptide SGR-1832 from *Streptomyces griseus* IFO 13350 and its biosynthetic gene cluster. (a) SGR-1832 imaging and purification. MALDI-imaging of sporulating *Streptomyces griseus* IFO 13350 detected a secreted mass of 1854 Da. **(b)** SGR-1832 sequence tagging and structure elucidation. FTMS analysis revealed the mass of the purified peptide as 1832.0 Da. MS^2 analysis of unmodified SGR-1832 yielded a 12-aa-long MS^n sequence tag. In addition, MS^2 analysis of SGR-1832 revealed 2 Dhb residues. The N-terminus has a +28.0 Da mass shift which corresponds to a *N,N*-dimethylation. No b_{16}-b_{18}-ions were detected in the MS^2 spectrum of unmodified SGR-1832 which could be due to a cyclic structure. MS^2 analysis of unmodified and $NaBH_4/NiCl_2$-treated SGR-1832 showed the presence of a C-terminal aminovinylcysteine (AviCys[37]) modification derived from 2 Cys. The (Cys16→Cys19)-derived AviCys-modification and the *N,N*-dimethylation at Ala1 was confirmed by a recent study on the novel linaridin RP class of cypemycin which also comprises the observed *N,N*-dimethylation, 2 Dhb and a C-terminal AviCys PTM derived from 2 Cys. **(c)** Genome mining of SGR-1832 gene cluster and its homologs. The corresponding search tag could not be found in the NCBI protein database of *Streptomyces griseus* IFO 13350 but a candidate precursor peptide could be identified in the 6-frame translated supercontig of *Streptomyces griseus*. In addition to the recently predicted SGR-1832 homologous gene clusters[15], a homologous gene cluster was identified in a microbiome-derived organism, *Rothia dentocariosa*, by BLAST analysis.

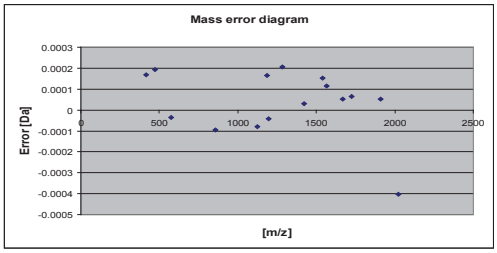

a
S. coelicolor A3(2)
Crude extract
2139.3
Gel filtration F4
2178.9
HPLC F46-50
2139.1
S. lividans TK24
Crude extract
2138.6
Gel filtration F4
2140.1
S. sp. E14 (WASP)
Crude extract
2161.5
Gel filtration F4
2140.6
1000
2000
3000
4000
m/z
b
FTMS
SCO-2138
obs: 2138.19 Da
calc: 2138.17 Da
2+
SLI-2138
obs: 2138.17 Da
calc: 2138.17 Da
SWA-2138
obs: 2138.23 Da
calc: 2138.17 Da
1066
1070
1074
1075

b, continued

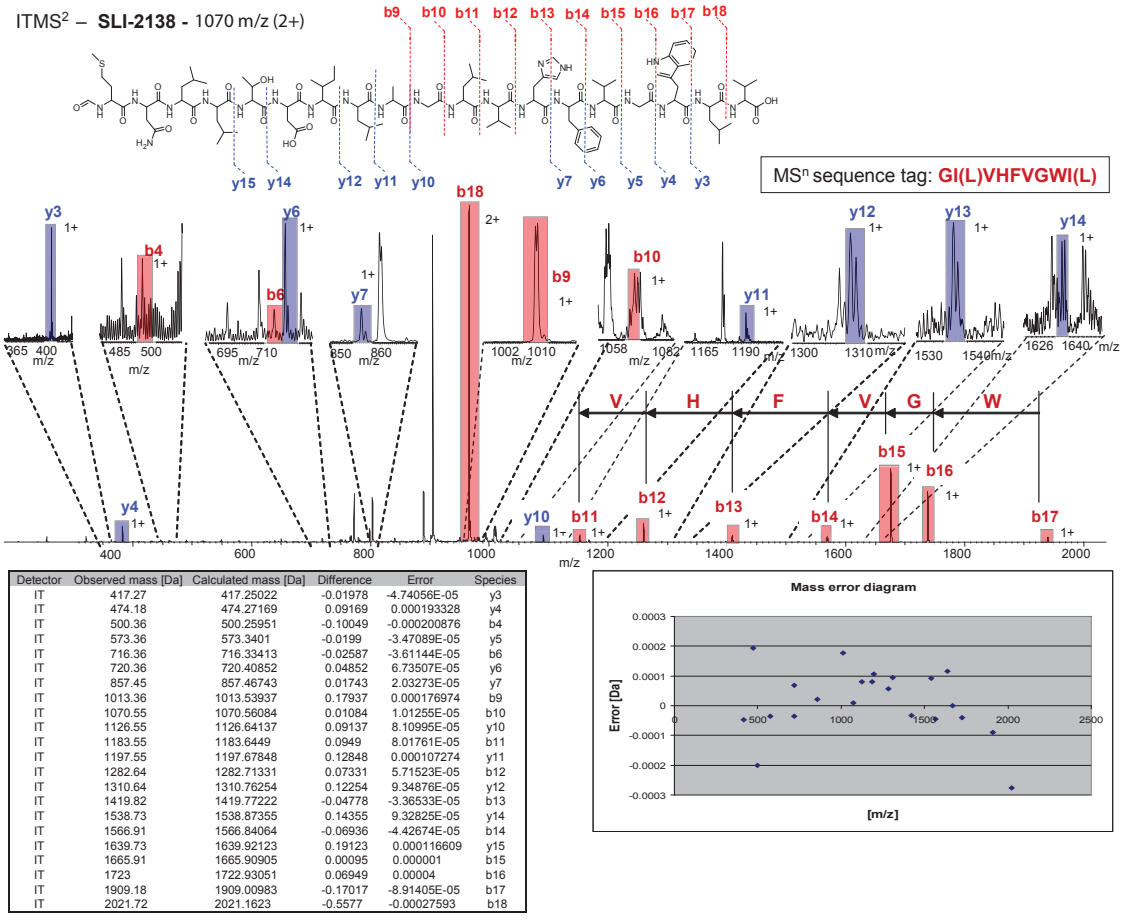
ITMS^2 – **SCO-2138 -** 1070 m/z (2+)



Detector	Observed mass [Da]	Calculated mass [Da]	Difference	Error	Species
IT	417.18	417.25022	0.07022	0.000168	y3
IT	474.18	474.27169	0.09169	0.000193	y4
IT	573.36	573.3401	-0.0199	-3.47E-05	y5
IT	857.55	857.46743	-0.08257	-9.63E-05	y7
IT	1126.73	1126.64137	-0.08863	-7.87E-05	y10
IT	1183.45	1183.645	0.195	0.000165	b11
IT	1197.73	1197.67848	-0.05152	-4.3E-05	y11
IT	1282.45	1282.713	0.263	0.000205	b12
IT	1419.73	1419.772	0.042	2.96E-05	b13
IT	1538.64	1538.87355	0.23355	0.000152	y14
IT	1566.73	1566.909	0.179	0.000114	b14
IT	1665.82	1665.909	0.089	5.34E-05	b15
IT	1722.82	1722.931	0.111	6.44E-05	b16
IT	1908.91	1909.01	0.1	5.24E-05	b17
IT	2021.9	2021.08608	-0.81391	-0.000403	b18

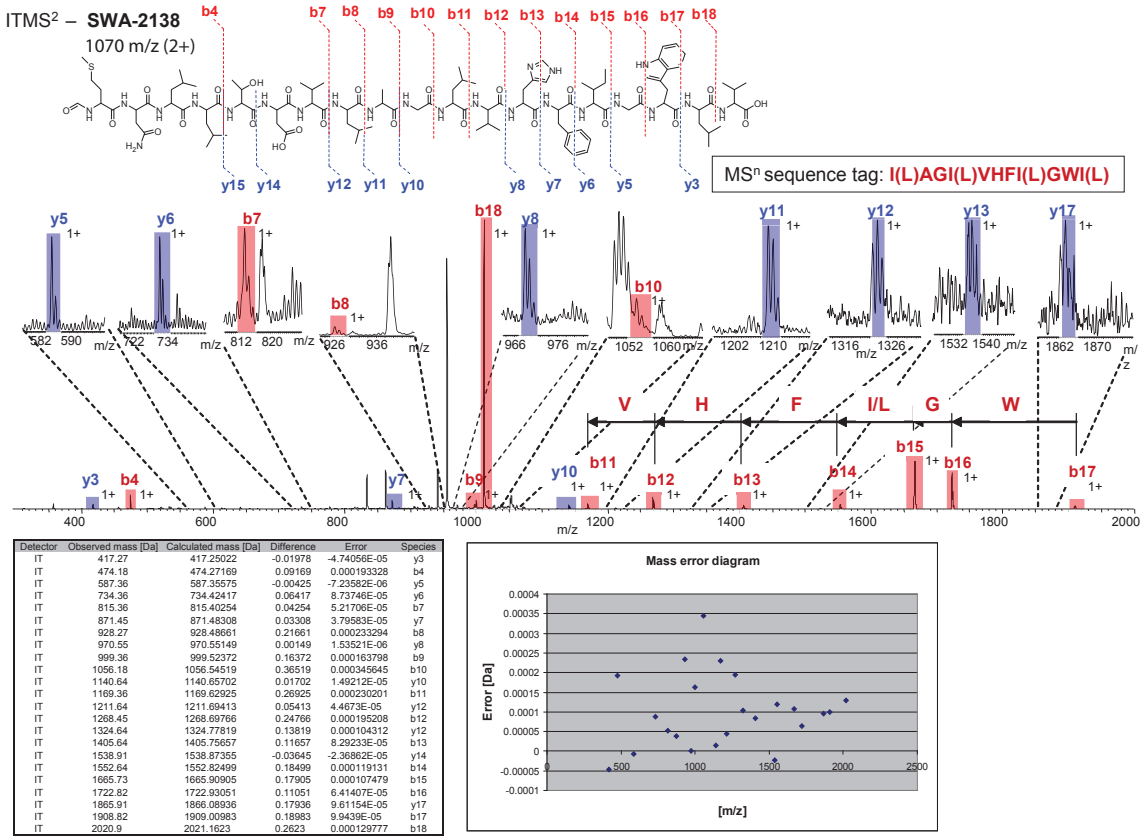


b, continued



Detector	Observed mass [Da]	Calculated mass [Da]	Difference	Error	Species
IT	417.27	417.25022	-0.01978	-4.74056E-05	y3
IT	474.18	474.27169	0.09169	0.000193328	y4
IT	500.36	500.25951	-0.10049	-0.000200876	b4
IT	573.36	573.3401	-0.0199	-3.47089E-05	y5
IT	716.36	716.33413	-0.02587	-3.61144E-05	b6
IT	720.36	720.40852	0.04852	6.73507E-05	y6
IT	857.45	857.46743	0.01743	2.03273E-05	y7
IT	1013.36	1013.53937	0.17937	0.000176974	b9
IT	1070.55	1070.56084	0.01084	1.01255E-05	b10
IT	1126.55	1126.64137	0.09137	8.10995E-05	y10
IT	1183.55	1183.6449	0.0949	8.01761E-05	b11
IT	1197.55	1197.67848	0.12848	0.000107274	y11
IT	1282.64	1282.71331	0.07331	5.71523E-05	b12
IT	1310.64	1310.76254	0.12254	9.34876E-05	y12
IT	1419.82	1419.77222	-0.04778	-3.36533E-05	b13
IT	1538.73	1538.87355	0.14355	9.32825E-05	y14
IT	1566.91	1566.84064	-0.06936	-4.42674E-05	b14
IT	1639.73	1639.92123	0.19123	0.000116609	y15
IT	1665.91	1665.90905	0.00095	0.000001	b15
IT	1723	1722.93051	0.06949	0.00004	b16
IT	1909.18	1909.00983	-0.17017	-8.91405E-05	b17
IT	2021.72	2021.1623	-0.5577	-0.00027593	b18

b, continued



Detector	Observed mass [Da]	Calculated mass [Da]	Difference	Error	Species
IT	417.27	417.25022	-0.01978	-4.74056E-05	y3
IT	474.18	474.27169	0.09169	0.000193328	b4
IT	587.36	587.35575	-0.00425	-7.23582E-06	y5
IT	734.36	734.42417	0.06417	8.73746E-05	y6
IT	815.36	815.40254	0.04254	5.21706E-05	b7
IT	871.45	871.48308	0.03308	3.79583E-05	y7
IT	928.27	928.48661	0.21661	0.000233294	b8
IT	970.55	970.55149	0.00149	1.53521E-06	y8
IT	999.36	999.52372	0.16372	0.000163798	b9
IT	1056.18	1056.54519	0.36519	0.000345645	b10
IT	1140.64	1140.65702	0.01702	1.49212E-05	y10
IT	1169.36	1169.62925	0.26925	0.000230201	b11
IT	1211.64	1211.69413	0.05413	4.4673E-05	y12
IT	1268.45	1268.69766	0.24766	0.000195208	b12
IT	1324.64	1324.77819	0.13819	0.000104312	y12
IT	1405.64	1405.75657	0.11657	8.29233E-05	b13
IT	1538.91	1538.87355	-0.03645	-2.36862E-05	y14
IT	1552.64	1552.82499	0.18499	0.000119131	b14
IT	1665.73	1665.90905	0.17905	0.000107479	b15
IT	1722.82	1722.93051	0.11051	6.41407E-05	b16
IT	1865.91	1866.08936	0.17936	9.61154E-05	y17
IT	1908.82	1909.00983	0.18983	9.9439E-05	b17
IT	2020.9	2021.1623	0.2623	0.000129777	b18

C

6-frame translation

SCO-2138 – search tag: **VHFVGWI(L)**

```
[…]RARRPGPPGTAPPPGXSTSRTGSASRCXPWWTHRARPTTRRRSGRGRVRRSRTCSAPWPRRVRRSPRW
XSARAARAGRWPWPRPAARGPPRTATSPSSRRSTRRRSXSARRRRWRRRPVNCGCGRRTWQHLAXSEQASS
CSLEQVIVALKNACDCRDQRYLRCASNGMQTVVDAHVPSSPGGARWVPHLNSARSCTIMNLLTDILAGLVH
FVGWLVXRSEKPTAPPPRPREAAPSACPRPDRGLPHGHRPHDLLRDRAAPVVARRGPQRDVLGIARHAQRP
VPGGLPVAAAQPDLVPGRPAVVGRLQLSAAELPGVRGARHVRGAETDPQLLAGAHLPREVGVDVEGEV[…]
```

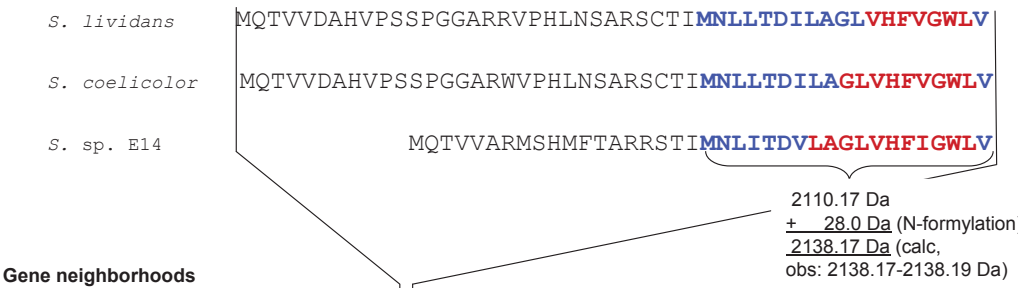
SLI-2138 – search tag: **GI(L)VHFVGWI(L)**

```
[…]GTRRACCAASARTRGARWRTPRRRARRPGPPGTAPPPGXSTSRTGSASRCXPWWTHRARPTTRRRSGR
GRVRRSRTCSAPWPRRVRRSPRWXSARAARAGRWPWPRPAARGPPRTATSPSSRRNTRRRSXSARRRRWRR
RPVNCGCDRRTWQHLAXSEQASSCSLEQVIVALKNACDCRDQRYLRCASNGMQTVVDAHVPSSPGGARRVP
HLNSARSCTIMNLLTDILAGLVHFVGWLVXRSENPTAPPPRPREAAPSACPRPDRGLPHGHRPHDLLRDRA
APVVARRGPQRDVLGIARHAQRPVPGGLRVAAAQPDLVPGRPAVVGRLQLSAAELPGVRGARHVRGAETDP
QLLAGAHLPREVGVDVEGEVDQRQEGPPVVGRGRRPEGPLRVALRRQRPACRVQRRVGDEPVGVLVGR[…]
```

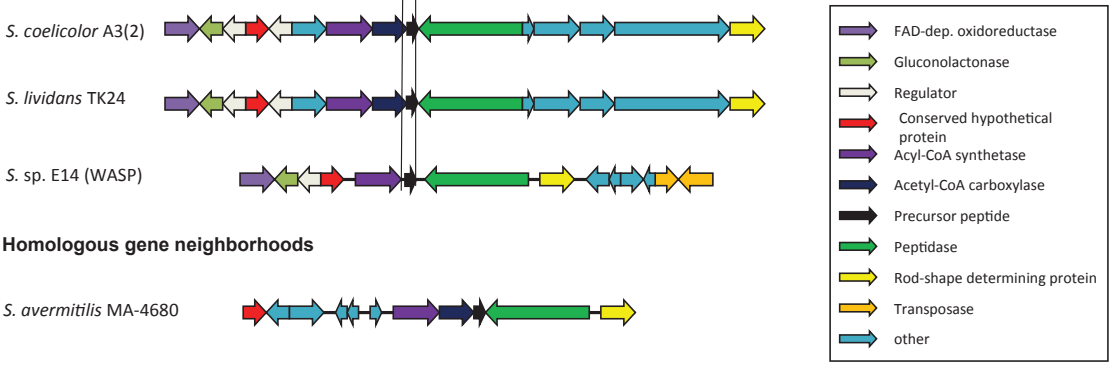
SWA-2138 – search tag: **I(L)AGI(L)VHFI(L)GWI(L)**

```
[…]SAPGTWRCARPTGMCASSAARPPIXSRAAVTRSAPARSRTHCSNIRGCGRPPSPGNRTPTSGSGSSPG
SSRPTRPRRPGKGSWPTMWPGGWPRTSAPASSGSSTRCPGTTWGRSXSGRCRRGEGAGLPSRXSEHGIRCS
LEQVIVAHTTTXTVGADTYLRCARDRMQTVVARMSHMFTARRSTIMNLITDVLAGLVHFIGWLVXRPPTPT
AAPPPRPREAAPPACSRSPGLLHARGPDDLLPYRAALVVARARPQRDGGGAAGHGERAAPLGVPAGEPYRP
PGRPGVVGHLQAPVPDRPGARRARRIRGAQPDRHPASGPEGAREVGVEVEGEVDQGQRGPVVVGRGRR[…]
```

Putative precursor peptides



Gene neighborhoods



Homologous gene neighborhoods

d

S. coelicolor A3(2)

Gene	Size [aa]	Predicted function
SCO_2438	468	FAD-dependent oxidoreductase
SCO_2441	200	conserved hypothetical protein
SCO_2444	485	acyl-CoA synthetase
SCO_2445	458	acetyl-CoA carboxylase
SCO_2446	1220	peptidase
N/A	49	precursor peptide
SCO_2451	360	rod-shape determining protein, MreB

S. lividans TK24

Gene	Size [aa]	Predicted function
SSPG_05081	360	rod-shape determining protein, MreB
SSPG_05087	1220	peptidase
N/A	49	precursor peptide
SSPG_05088	485	acetyl-CoA carboxylase
SSPG_05089		
SSPG_05090	159	acyl-CoA synthetase
SSPG_05093	196	conserved hypothetical protein
SSPG_05096	454	FAD-dependent oxidoreductase

S. sp. E14 (WASP)

Gene	Size [aa]	Predicted function
SSTG_01176	447	FAD-dependent oxidoreductase
SSTG_01179	200	conserved hypothetical protein
SSTG_01180	443	acyl-CoA synthetase
SSTG_01181	1219	peptidase
N/A	38	precursor peptide
SSTG_01182	346	rod-shape determining protein, MreB

S. avermitilis MA-4680

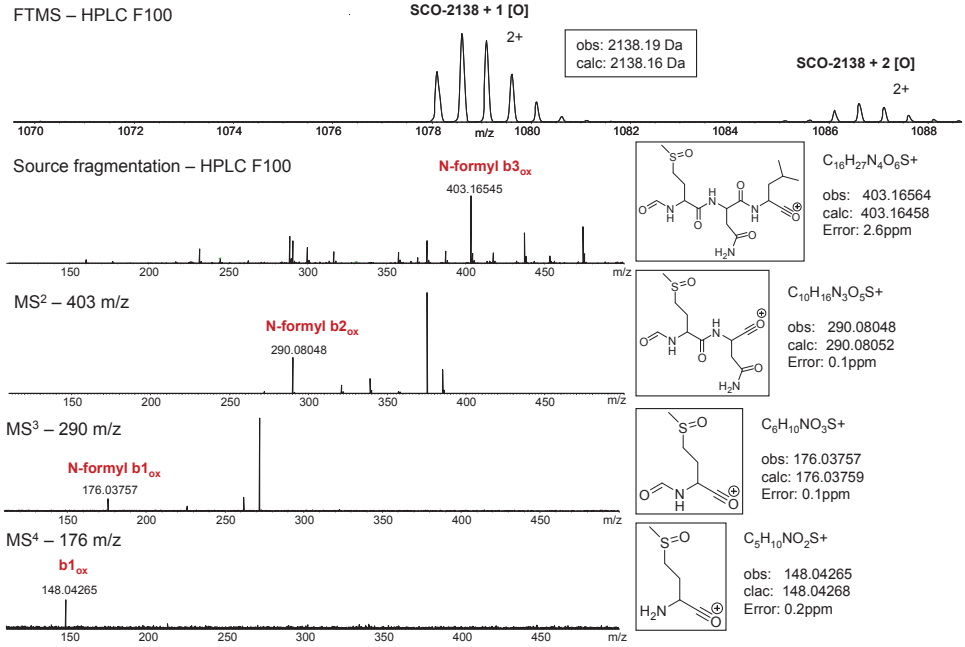
Gene	Size [aa]	Predicted function
SAV_5729	220	conserved hypothetical protein
SAV_5735	489	acyl-CoA synthetase
SAV_5736	465	acetyl-CoA carboxylase
N/A	21	precursor peptide (truncated)
SAV_5737	1208	peptidase
SAV_5738	345	rod-shape determining protein, MreB

S. sp. SPB78

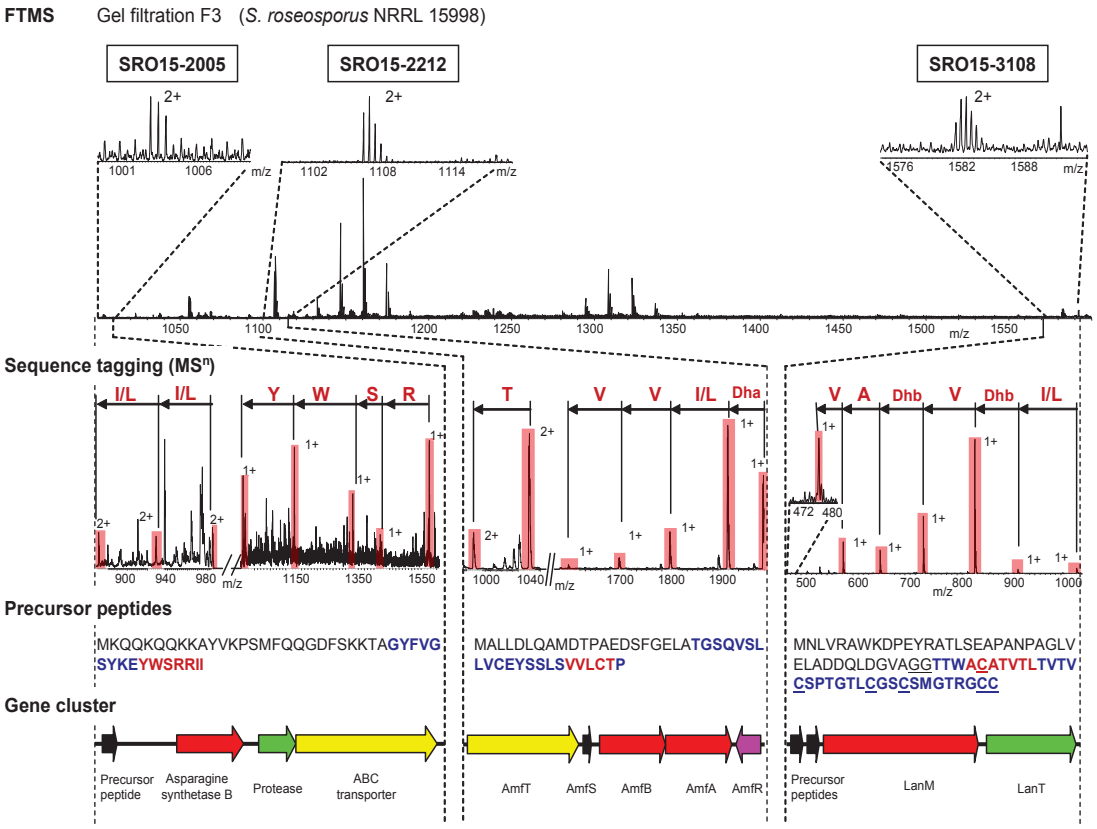
Gene	Size [aa]	Predicted function
SSLG_01586	202	conserved hypothetical protein
N/A	37	precursor peptide (truncated)
SSLG_01587	1248	peptidase
SSLG_01609	362	rod-shape determining protein, MreB

Organism	Predicted precursor peptide
S. coelicolor A3(2)	**MQTVVDAHVPSSPGGARWVPHLNSARSCTI**MNLLTDILAGLVHFVGWLV
S. lividans TK24	**MQTVVDAHVPSSPGGARRVPHLNSARSCTI**MNLLTDILAGLVHFVGWLV
S. sp. E14 (WASP)	**MQTVVARMSHMFTARRSTI**MNLITDVLAGLVHFIGWLV

e

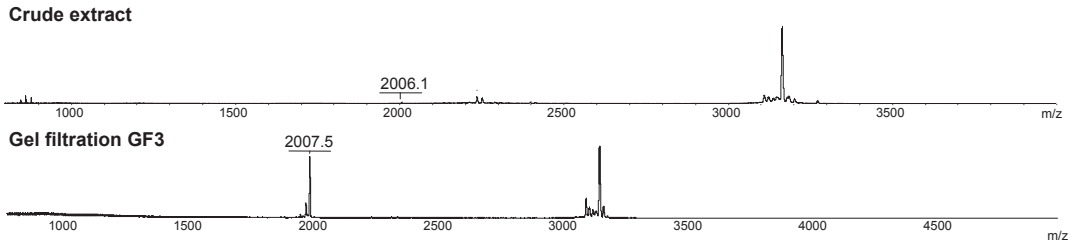


Supplementary Figure 12. Characterization of novel ribosomal peptides SCO-2138, SLI-2138 and SWA-2138 and their gene neighborhoods by NPP from *Streptomyces coelicolor* A3(2), *Streptomyces lividans* TK24, and *Streptomyces* sp. E14 (WASP), respectively. (a) Purification of SCO-2138, SLI-2138 and SWA-2138. N-butanol extraction of 80 plates of each sporulating *Streptomyces* culture resulted in the isolation of a compound with the mass 2138 Da in each respective crude extract. **(b)** Sequence tagging of SCO-2138, SLI-2138 and SWA-2138. MS^n analysis of SCO-2138, SLI-2138 and SWA-2138 yielded the MS^n sequence tags VHFVGWI(L), GI(L)VHFVGWI(L), I(L)AGI(L)VHFI(L)GWI(L), respectively. Furthermore, MS^2 analysis of SCO-2138, SLI-2138, and SWA-2138 characterized them as linear peptides with a +28.0 Da modification at the N-terminus. **(c)** Genome mining of SCO-2138, SLI-2138 and SWA-2138 gene neighborhoods. In the NCBI databases of *S. coelicolor* A3(2), *S. lividans* TK24 and *S.* sp. E14 (WASP), no precursor peptides with the characterized sequence tag could be identified. However, in the corresponding 6frame-translated supercontigs, a ribosomal peptide-encoding gene could be characterized for each strain. SCO-2138 and SLI-2138 have identical core peptide sequences. The SWA-2138 core peptide sequence is deviating in 2 positions from SCO-2138/SLI-2138 (I7V, V15I). **(d)** Bioinformatic analysis of homologous linear peptide gene neighborhoods which all contained a peptidase, a conserved hypothetical protein and a rod-shape determining protein (MreB) indicating a morphogenetic function[38]. The predicted precursor peptides of SCO-2138, SLI-2138 and SWA-2138 have similar putative leader peptide sequences. **(e)** Characterization of N-terminal formyl group of SCO-2138 (1x oxidized at Met1[39,40]) by $FTMS^n$ characterization of a CO-loss at Met-N were detected with sub-ppm accuracy.



Supplementary Figure 13. Characterization of 3 ribosomal peptides in one NPP experiment. FTMS analysis of a gel filtration fraction (F3) from *Streptomyces roseosporus* NRRL 15998 revealed multiple mass signals. 3 mass signals (2005 Da, 2+ ion; 2212 Da, 2+ ion; 3162 Da, 2+ ion) yielded 5-6-aa-long MS^n peptide sequence tags by $ITMS^n$ analysis of the same sample. The corresponding sequence tags enabled the identification of the precursor peptides and the gene clusters of the 3 observed peptide mass signals. The 2005 Da mass signal was characterized as class I lassopeptide SRO15-2005 (**Supplementary Fig. 14**), the 2212 Da mass signal was characterized as class III lantipeptide SRO15-2212 (AmfS, **Supplementary Fig. 15**), and the 3162 Da mass signal was characterized as a 3x hydrated derivative of class II lantipeptide SRO15-3108 (**Supplementary Fig. 16**).

a



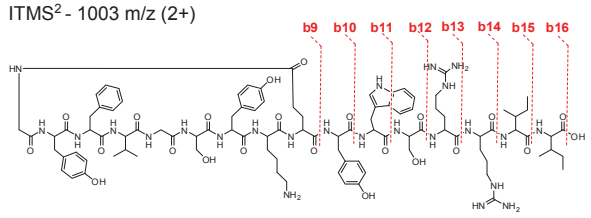
b

FTMS

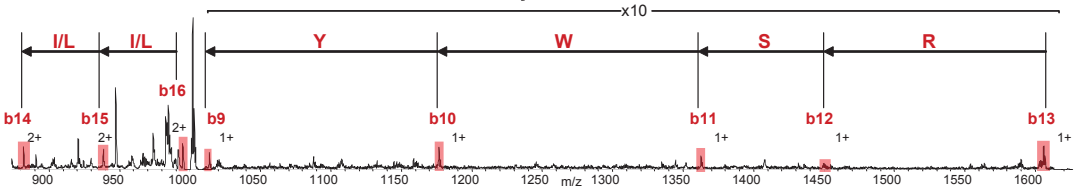
unmodified

obs: 2005.09 Da
calc: 2005.02 Da

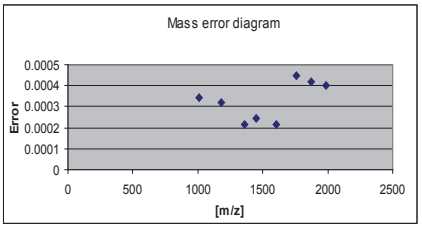
$ITMS^2$ - 1003 m/z (2+)



MS^n sequence tag: **YWSRRI(L)I(L)**



Detector	Observed mass [Da]	Calculated mass [Da]	Difference [Da]	Error	Species
IT	1013.82	1013.4733	0.3467	0.000342	b9
IT	1176.91	1176.53663	0.37337	0.000317	b10
IT	1362.91	1362.61595	0.29405	0.000216	b11
IT	1450	1449.64797	0.35203	0.000243	b12
IT	1606.09	1605.74908	0.34092	0.000212	b13
IT	1762.64	1761.85019	0.78981	0.000448	b14
IT	1875.72	1874.93426	0.78574	0.000419	b15
IT	1988.82	1988.0168	0.8032	0.000404	b16



c

6-frame translation – search tag: **YWSRRI(L)I(L)**

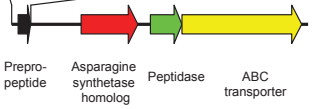
```
[…]PRPYGSTFRRTSTEGGGAHAVSGVGRSWSRSGDRSRLVPRFPGLPFPLLPAHCPVRRPQATIGPVHRV
ISAGAVCDCTGWPAGRSAPGGGVPDRPAAHLAARAIQPTSCGKSESCVPVRRPDMMALVFAXPTPVATPVA
RITPVMQKSVGHNGRQPRRREGVMKQQKQQKKAYVKPSMFQQGDFSKKTAGYFVGSYKEYWSRRIIXSADP
FPVHRVRTSERXRCPASCVTTSSSCRTARRERQRPGSFPYRERLPGGPPRGTPPGPPGRWAVPXLSPTPRA
GPGSSPARWSARSAISPGARTPSSSSARTASPTPCSPGSWRAPATGPAWNAASPGCPASTTWSPGCPG[…]
```

Candidate precursor peptide

MKQQKQQKKAYVKPSMFQQGDFSKKTA**GYFVGSYKEYWSRRII**

2023.03 Da
- H_2O (cyclization)
2005.02 Da (calc, obs: 2005.09 Da)

Gene cluster



Gene	Size [aa]	Predicted function	Closest homolog [sim%/id%]
SSGG_03858	43	precursor peptide	-
SSGG_03859	312	lactamase	ref|ZP_06920897.1| conserved hypothetical protein [*Streptomyces sviceus* ATCC 29083] (58/44)
SSGG_03861	147	peptidase	ref|YP_003314220.1| hypothetical protein Sked_14500 [*Sanguibacter keddieii* DSM 10542] (60/50)
SSGG_03862	624	ABC transporter	ref|YP_001824736.1| putative ABC transporter ATPase and permease component [*Streptomyces griseus* subsp. griseus NBRC 13350] (93/89)

Asparagine synthetase homolog

```
>SSGG_03859 | asparagine synthetase (translation) (312 aa)
MNDEPLPVAPGRARAELLLGRAHATGSRYHLTGYGGDEIFLGLPHVYQDLFHGNPFTAWS
```

A cysteine at position 2 indicates a Gln-hydrolyzing Asn synthetase, any other residue at position 2 indicates a β-lactamase homolog as in lassopeptide biosynthetic pathways.

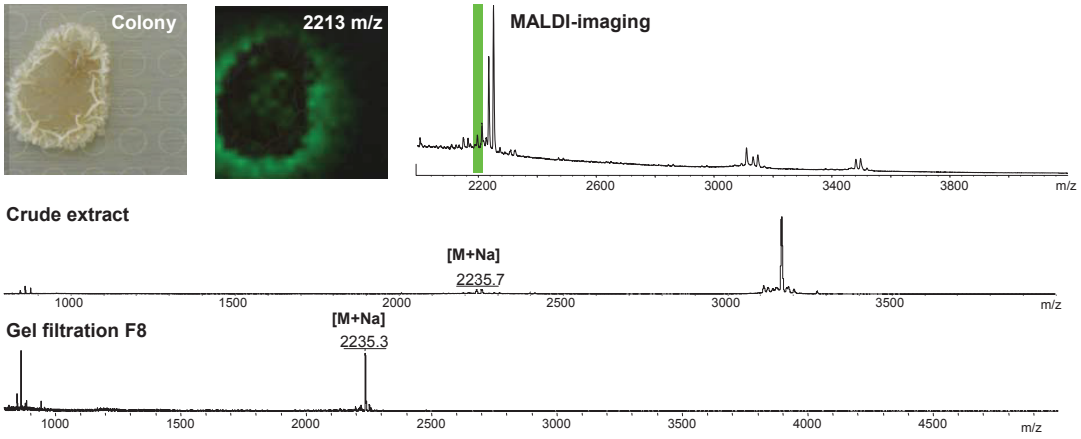
Peptidase

```
S.roseosporus peptidase  […]-------------RCASRYGCLPRSVAVALACRMSGVWPDWCAGVWATPPFSPHAWVLAE
McjB                     […]IRKEISNLSIIFHLNIFKSDCLTYSYALKRILNSRNIDAHLVIGVRTQP-FYSHSWVEVG
CapB                     […]RTAERYLRASIWSP--FRITCLQMSFALATHLRRENVPAQLVIGVRPMP-FVAHAWVEID

S.roseosporus peptidase  GRTVGEQAEAADLRPLMVVTVREGVQGER-----
McjB                     GQVINDAPNMRDKLSVIAEI--------------
CapB                     GRVCGDEPELKKSYGEIYRTPRHDERAGPFGLAA
```

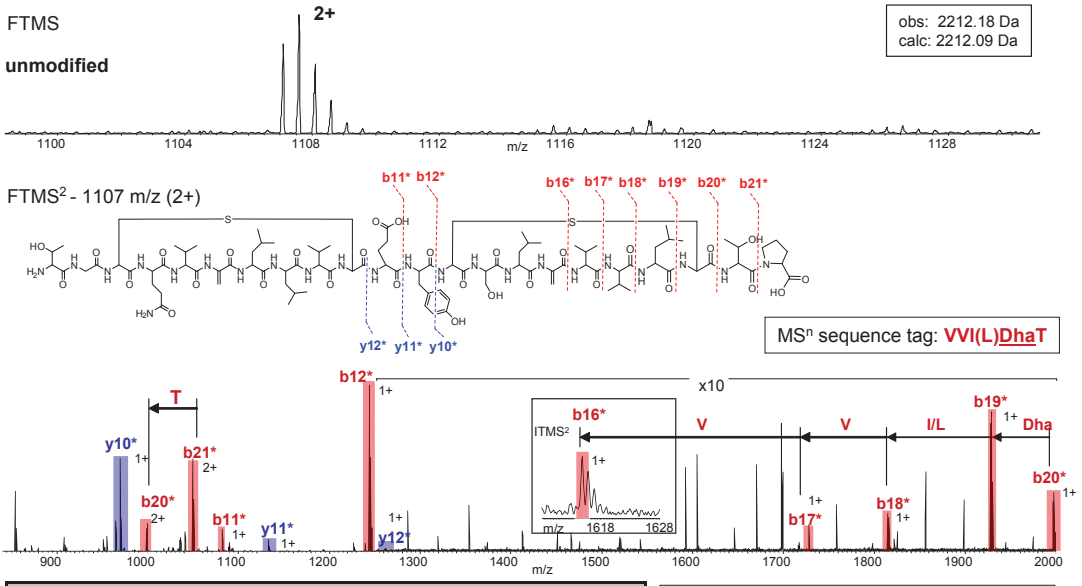
Conserved residues with CapB and McjB (lassopeptide peptidases)
Putative catalytic triad residues including cysteine or serine

Supplementary Figure 14. Characterization of class II lassopeptide SRO15-2005 from *Streptomyces roseosporus* NRRL 15998. (a) SRO15-2005 purification. A compound of the mass 2005 Da was isolated by n-butanol extraction of 80 plates of sporulating *Streptomyces roseosporus* NRRL 15998. The putative peptide was partially purified from the crude extract by gel filtration. **(b)** SRO15-2005 sequence tagging. MS^n analysis of SRO15-2005 in gel filtration fraction 3 indicated a cyclic or constrained peptide, as only few weak peptide fragment signals were detected, and yielded the 6-aa-long MS^n sequence tag YWSRRI(L)I(L). **(c)** Genome mining of SRO15-2005 gene cluster. The search tag YWSRRI(L)I(L) enabled the identification of a 43-aa-long candidate precursor peptide in the *S. roseosporus* NRRL 15998 genome. The respective precursor peptide had a putative C-terminal core peptide ^{28}G-^{43}I of the mass 2023.01 Da which matched the observed mass of 2005.09 Da after loss of water. The gene cluster contained a homolog of an Asn synthetase, a peptidase and an ABC transporter. The gene cluster prediction, the MS^n data of the peptide and the glycine at position 1 of the core peptide were indicative of a class II lassopeptide[9]. The core peptide is only 16-aa-long. It is the smallest characterized lassopeptide to date[9].

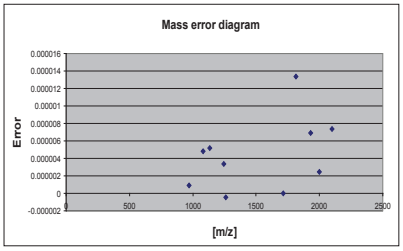
a



b

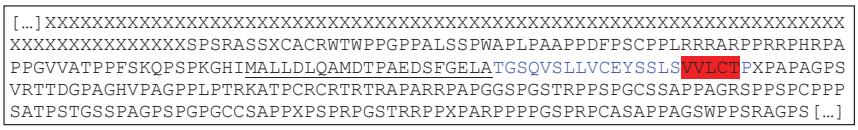


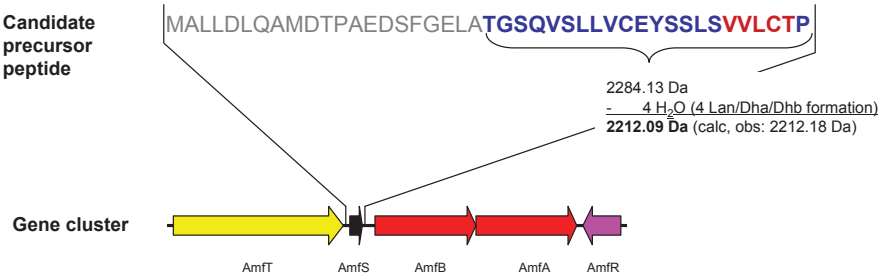
Detector	Observed mass [Da]	Calculated mass [Da]	Difference [Da]	Error	Species
FT	969.50886	969.50798	0.00088	9.1E-07	y10-2H₂O
FT	1081.54044	1081.53526	0.00518	4.8E-06	b11-2H₂O
FT	1132.57715	1132.57131	0.00584	5.2E-06	y11-2H₂O
FT	1244.60278	1244.59859	0.00419	3.4E-06	b12-2H₂O
FT	1261.61333	1261.6139	-0.00057	-4.5E-07	y12-2H₂O
IT	1616.19	1616.74534	-0.55534		b16-4H₂O+S
FT	1715.81383	1715.81376	7E-05	4.1E-08	b17-4H₂O+S
FT	1814.90503	1814.88084	0.02419	1.3E-05	b18-4H₂O+S
FT	1927.97954	1927.96624	0.0133	6.9E-06	b19-4H₂O+S
FT	1996.99263	1996.98769	0.00494	2.5E-06	b20-4H₂O
FT	2098.05091	2098.03537	0.01554	7.4E-06	b21-4H₂O



c

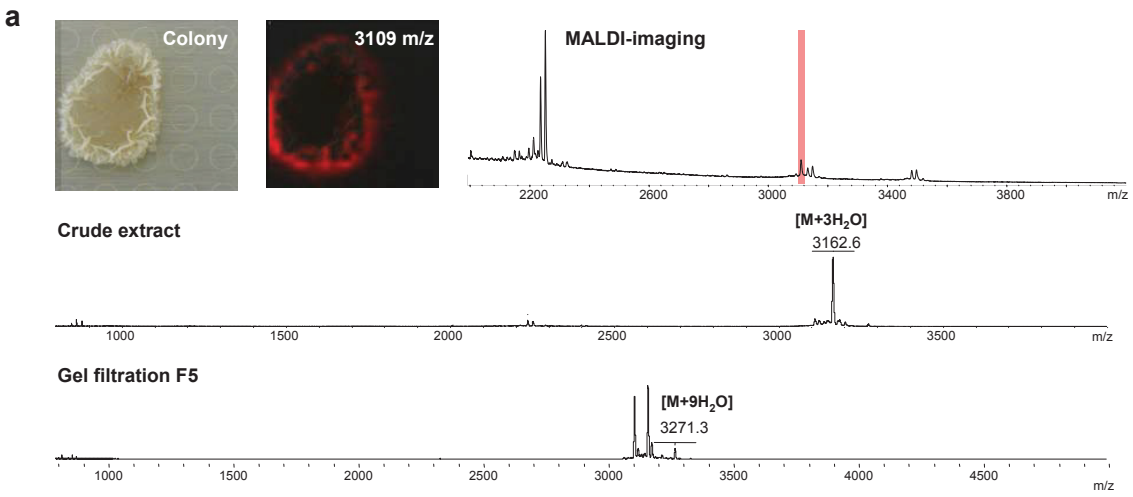
6-frame translation – search tag: **VVI(L)S(C)T**

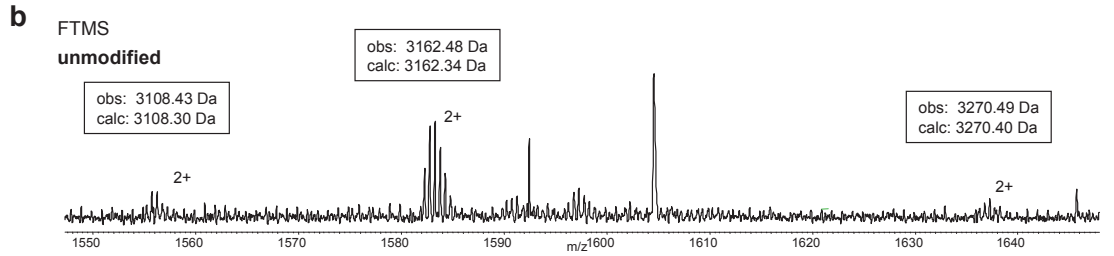




Gene	Size [aa]	Predicted function	Closest homolog [sim%/id%]
SSGG_04748	582	AmfT	dbj|BAB62264.1| membrane translocator [*Streptomyces griseus*](92/89)
SSGG_04750	43	AmfS	sp|Q07642.1|LANSB_STRGR Lanthionine-containing peptide sapB precursor (Spore-associated protein B) (Morphogen sapB) (Rapid aerial mycelium protein S) (97/97)
SrosN15_010100023901	355	AmfB	YP_001823907.1 membrane translocator, AmfB [*Streptomyces griseus* subsp. griseus NRBC 13350] (92/88)
SSGG_04751	561	AmfA	ref|YP_001823906.1| membrane translocator [*Streptomyces griseus* subsp. griseus NBRC 13350] (86/83)
SSGG_04752	201	AmfR	ref|YP_001823905.1| transcriptional regulator [*Streptomyces griseus* subsp. griseus NBRC 13350] (95/93)

Supplementary Figure 15. Characterization of class III lantipeptide SRO15-2212 from *Streptomyces roseosporus* NRRL 15998. (a) SRO15-2212 imaging and purification. A compound with the mass 2212 Da was detected by MALDI-imaging around a pre-sporulating colony of *Streptomyces roseosporus* NRRL 15998 The compound was extracted with n-butanol and partially purified by gel filtration. **(b)** SRO15-2212 sequence tagging. MS^2 analysis of SRO15-2212 yielded the 5-aa-long MS^n sequence tag VVI(L)DhaT. **(c)** Genome mining of SRO15-2212 gene cluster. The search tag VVI(L)S(C)T enabled the identification of a 43-aa-long candidate precursor peptide in the *S. roseosporus* NRRL 15998 genome. The putative core peptide ^{22}T-^{43}P had the calculated mass of 2284.13 Da which matched the observed mass 2212.18 Da after loss of 4 H_2O. The SRO15-2212 precursor peptide sequence and the observed MS^2 fragments were identical with AmfS from *Streptomyces griseus* IFO 13350[41] (**Supplementary Fig. 2**). The SRO15-2212 gene cluster contained an AmfA, AmfB, AmfR and AmfT gene and is ~99% homologous to the AmfS gene cluster[42]. SRO15-2212 was predicted to have a AmfS structure (**Table 1**) instead of an alternative labyrinthopeptin structure[12] because of the lacking labyrinthopeptin and DSB (**Supplementary Table 1**) and the AmfS sequence identity.

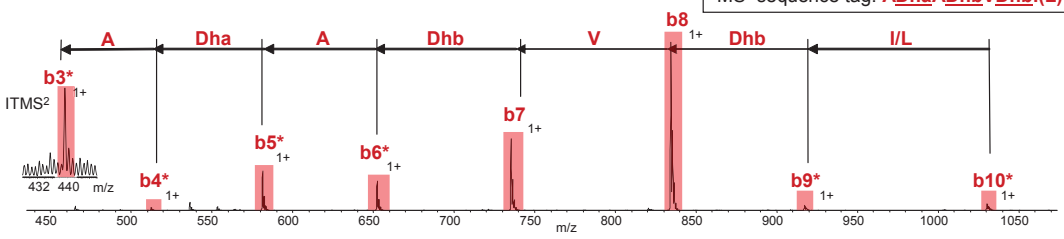




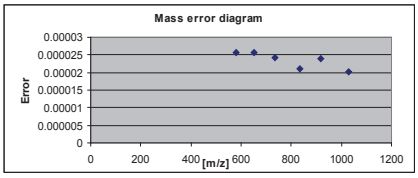
FTMS^2 - 1055 m/z (3+), 1049 m/z (3+)

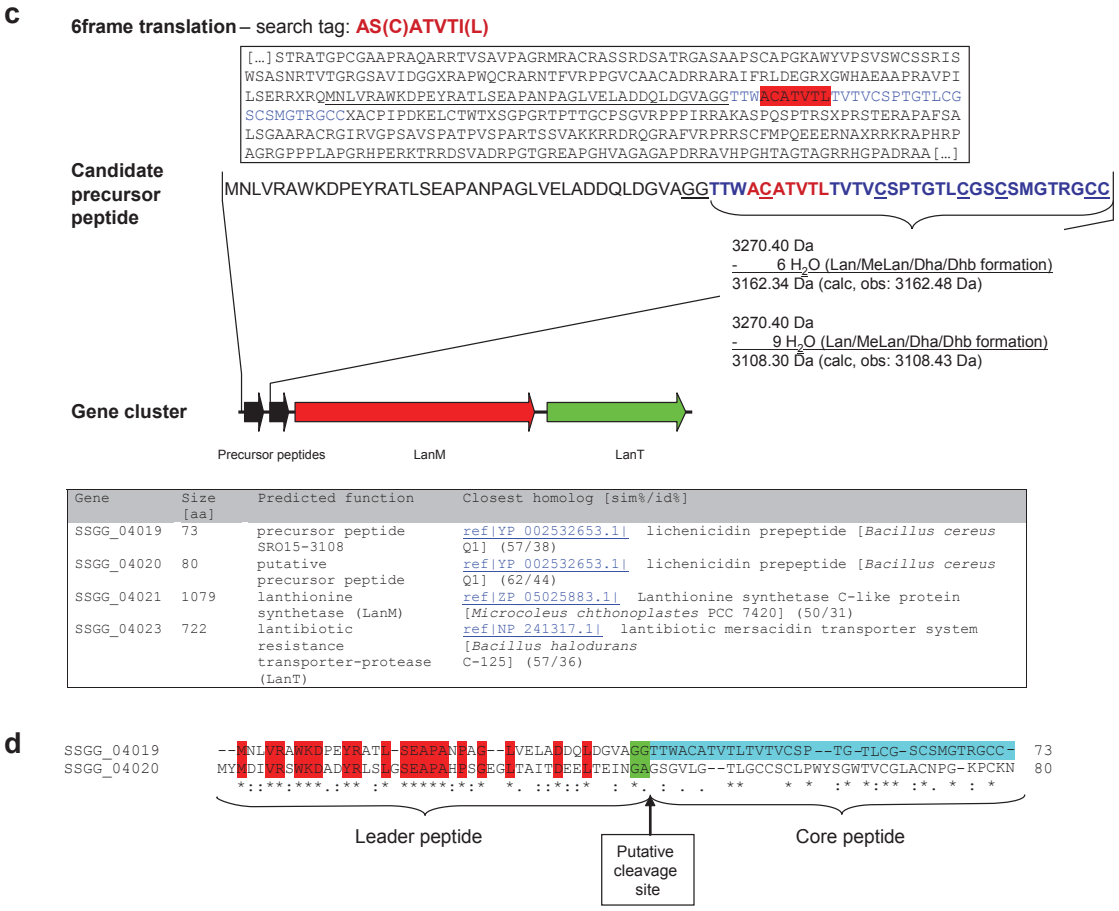
SRO15-3108 core peptide

MS^n sequence tag: ADhaADhbVDhbI(L)



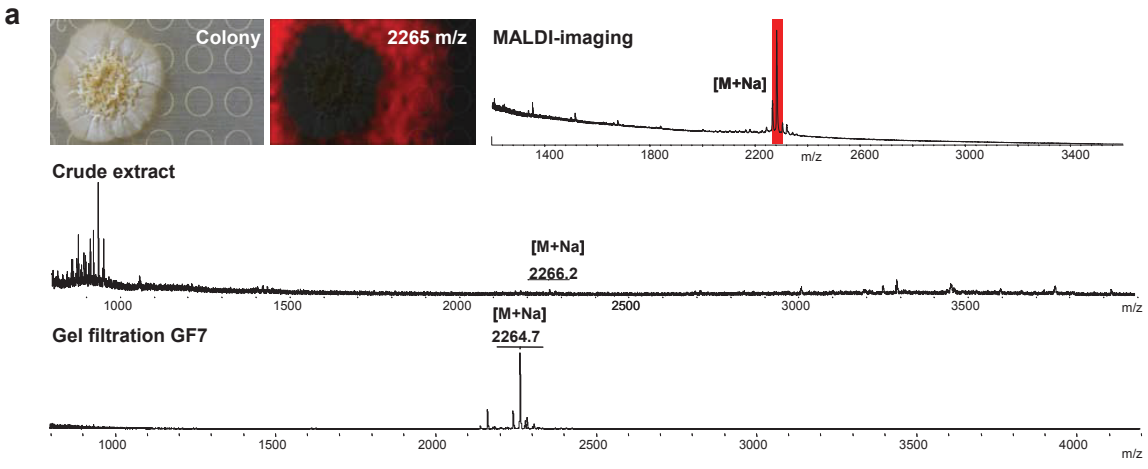
Detector	Observed mass [Da]	Calculated mass [Da]	Difference [Da]	Error	Species
IT	441.27	441.18083	0.08917	0.000202	b3+H2O+S+2H
IT	512.46	512.21794	0.24206	0.000473	b4+H2O+S+2H
FT	581.25439	581.2394	0.01499	2.58E-05	b5+H2O
FT	652.29321	652.27651	0.0167	2.56E-05	b6+H2O
FT	735.33148	735.31363	0.01785	2.43E-05	b7
FT	834.39969	834.38204	0.01765	2.12E-05	b8
FT	917.44111	917.41916	0.02195	2.39E-05	b9-H2O
FT	1030.52411	1030.50323	0.02088	2.03E-05	b10-H2O

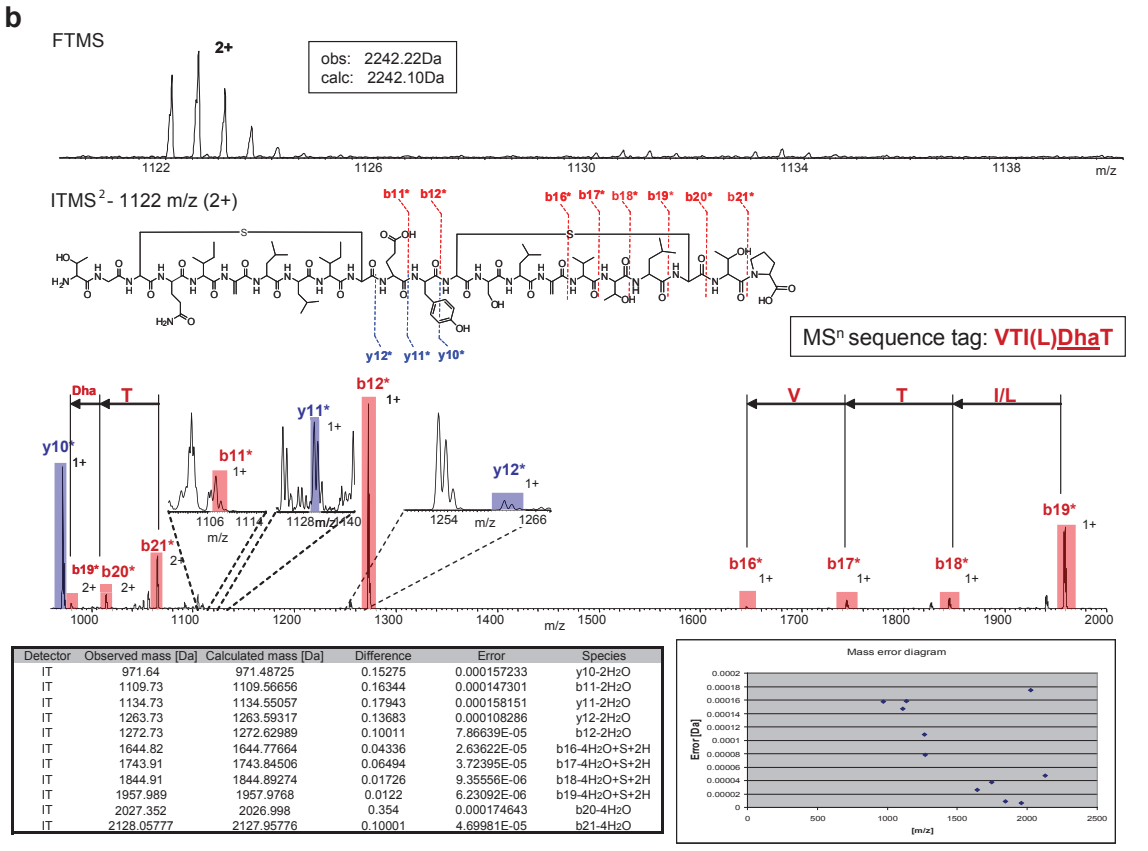




Gene	Size [aa]	Predicted function	Closest homolog [sim%/id%]
SSGG_04019	73	precursor peptide SRO15-3108	ref|YP_002532653.1| lichenicidin prepeptide [*Bacillus cereus* Q1] (57/38)
SSGG_04020	80	putative precursor peptide	ref|YP_002532653.1| lichenicidin prepeptide [*Bacillus cereus* Q1] (62/44)
SSGG_04021	1079	lanthionine synthetase (LanM)	ref|ZP_05025883.1| Lanthionine synthetase C-like protein [*Microcoleus chthonoplastes* PCC 7420] (50/31)
SSGG_04023	722	lantibiotic resistance transporter-protease (LanT)	ref|NP_241317.1| lantibiotic mersacidin transporter system [*Bacillus halodurans* C-125] (57/36)

Supplementary Figure 16. Characterization of class II lantipeptide SRO15-3108 from *Streptomyces roseosporus* NRRL 15998. (a) SRO15-3108 imaging and purification. MALDI-imaging detected a compound with a mass of 3108 Da around the aerial hyphae of a pre-sporulating *Streptomyces roseosporus* NRRL 15998. **(b)** SRO15-3108 sequence tagging. Additional compounds of 3162 Da and 3270 Da were detected in gel filtration fractions which were assumed to be biosynthetic precursors of SRO15-3108 since they differed in mass by 6 H_2O and 9 H_2O, respectively, from the mature natural peptide SRO15-3108. MS^2 analysis of unmodified SRO15-3162 yielded a 7-aa-long MS^2 sequence tag, ADhaADhbVDhbI(L). **(c)** Genome mining of SRO15-3108 gene cluster. The search tag identified a 73-aa-long candidate precursor peptide in the *S. roseosporus* genome that contained the sequence tag in the C-terminal region. The putative core peptide ^{41}T-^{73}C had the calculated mass of 3270.40 Da. The putative core peptide-(6 H_2O) had the calculated mass of 3162.34 Da and the core peptide-(9 H_2O) had the calculated mass of 3108.30 Da. The SRO15-3108 gene cluster contained a Lan-synthetase (LanM) and a lantibiotic resistance transporter-peptidase (LanT). **(d)** Alignment of putative precursor peptides from SRO15-3108 gene cluster. Another small ORF (80 aa) was located next to the precursor peptide on the genome comprising a C-terminal region with multiple Ser/Thr/Cys-residues and a N-terminal region with high homology to the N-terminal leader peptide of SRO15-3108 precursor peptide. This ORF could be the second precursor peptide of a two-component lantipeptide system with a single core peptide-substrate promiscuous LanM enzyme[43].

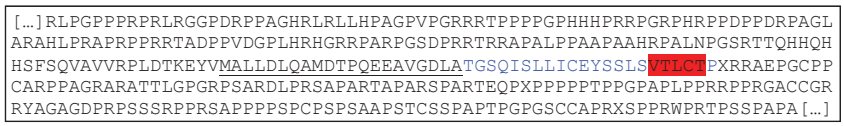


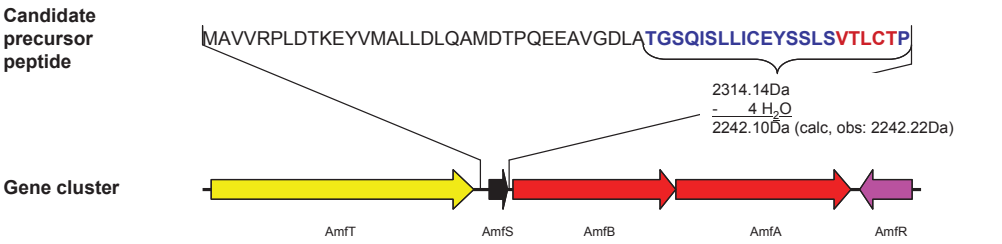


Detector	Observed mass [Da]	Calculated mass [Da]	Difference	Error	Species
IT	971.64	971.48725	0.15275	0.000157233	y10-2H2O
IT	1109.73	1109.56656	0.16344	0.000147301	b11-2H2O
IT	1134.73	1134.55057	0.17943	0.000158151	y11-2H2O
IT	1263.73	1263.59317	0.13683	0.000108286	y12-2H2O
IT	1272.73	1272.62989	0.10011	7.86639E-05	b12-2H2O
IT	1644.82	1644.77664	0.04336	2.63622E-05	b16-4H2O+S+2H
IT	1743.91	1743.84506	0.06494	3.72395E-05	b17-4H2O+S+2H
IT	1844.91	1844.89274	0.01726	9.35556E-06	b18-4H2O+S+2H
IT	1957.989	1957.9768	0.0122	6.23092E-06	b19-4H2O+S+2H
IT	2027.352	2026.998	0.354	0.000174643	b20-4H2O
IT	2128.05777	2127.95776	0.10001	4.69981E-05	b21-4H2O

C

6-frame translation – search tag: **VTI(L)S(C)T**





Gene	Size [aa]	Predicted function	
SSHG_03588	903	AmfT	ZP_06774427.1 - AmfT protein [*Streptomyces clavuligerus* ATCC27064] (64/55)
SSHG_03589	56 (43)	AmfS – precursor peptide (short ORF)	YP_001823908.1 - AmfS protein [*Streptomyces griseus* subsp. griseus IFO 13350] (93/81)
SSHG_03590	607	AmfB - ABC transporter ATP-binding membrane translocator	ADI10716.1 – ABC transporter ATP-binding membrane translocator, AmfB [*Streptomyces bingchenggensis* BCW-1] (60/51)
SSHG_03591	643	AmfA - ABC transporter ATP-binding protein	ADI10717.1 - ABC transporter ATP-binding protein, AmfA [*Streptomyces bingchenggensis* BCW-1] (54/44)
SSHG_03592	201	AmfR - transcriptional regulator	ZP_06527126.1 – two-component system response regulator [*Streptomyces lividans* TK24] (73/57)

Supplementary Figure 17. Characterization of class III lantipeptide SAL-2242 from *Streptomyces albus* J1074. (a) SAL-2242 imaging and purification. A compound with a mass of 2242 Da was detected by MALDI-imaging around a sporulating *Streptomyces albus* J1074 colony. The compound was extracted from 80 plates of sporulating *S. albus* and partially purified by gel filtration and desalted and concentrated by HPLC. **(b)** SAL-2242 sequence tagging. The MS^n analysis yielded the 5-aa-long MS^n sequence tag VTI(L)DhaT. **(c)** Genome mining of SAL-2242 gene cluster. The search tag VTI(L)S(C)T enabled to identify a 43-aa-long candidate precursor peptide with the sequence tag in the C-terminal region. The putative core peptide ^{22}T-^{43}P had the mass 2314.14 Da which matched the observed mass of 2242.22 Da after the loss of 4 H_2O (formation of 2 Lan and 2 Dha). The SAL-2242 gene cluster was homologous to the AmfS gene cluster. It contained a AmfT, AmfA, AmfB and AmfR gene. Based on the results and the sequence similarity to other known AmfS-like class III lantipeptides (not shown, see [41] and **Supplementary Table 1**), a SAL-2242 structure with 2 Lan-bridges and 2 Dha-PTMs could be predicted (**Table 1**).

References (Supplementary information)

1. Li, M.H., Ung, P.M., Zajkowski, J., Garneau-Tsodikova, S., Sherman, D.H. Automated genome mining for natural products. *BMC Bioinformatics*. **10**, 185 (2009).
2. Röttig, M. et al. NRPSpredictor2—a web server for predicting NRPS adenylation domain specificity. Nucl. Acids Res. (2011).
3. Medema, M. H. et al. antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. Nucl. Acids Res. 39, 339-346 (2011).
4. Caboche, .S, Pupin, M., Leclère, V., Fontaine, A., Jacques, P., Kucherov, G. NORINE: a database of nonribosomal peptides. *Nucl. Acids Res.* **36**, 326-331 (2008).
5. Tsur, D., Tanner, S., Zandi, E., Bafna, V., Pevzner, P.A. Identification of post-translational modifications by blind search of mass spectra. *Nat. Biotechnol.* **23**, 1562-1567 (2005).
6. Perkins, D.N., Pappin, D.J., Creasy, D.M. & Cottrell, J.S. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20**, 3551–3567 (1999).
7. Bodanszky, M., Izdebski, J., Muramatsu, I. Structure of the peptide antibiotic stendomycin. *J. Am. Chem. Soc.***91**, 2351–2358 (1969).
8. Duquesne, S. *et al.* Two enzymes catalyze the maturation of a lasso peptide in *Escherichia coli*. *Chem. Biol.* **14**, 793-803 (2007).
9. Knappe, T.A., Linne, U., Xie, X., Marahiel, M.A. The glucagon receptor antagonist BI-32169 constitutes a new class of lasso peptides. *FEBS Lett.* **584,** 785-789 (2010).
10. Willey, J.M., van der Donk, W.A. Lantibiotics: Peptides of Diverse Structure and Function. *Annu. Rev. Microbiol.*, **61**, 477-501 (2007).
11. Kodani, S. *et al.* The SapB morphogen is a lantibiotic-like peptide derived from the product of the developmental gene ramS in Streptomyces coelicolor. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 11448-11453 (2004).
12. Meindl, K. et al. Labyrinthopeptins: A New Class of Carbacyclic Lantibiotics. *Angew. Chem. Int. Ed. Engl.* **49**, 1151-1154 (2010).
13. Goto, Y. *et al.* Discovery of Unique Lanthionine Synthetases Reveals New Mechanistic and Evolutionary Insights. *PLoS Biology* **8**, e1000339 (2010).
14. Zheng, G., Yan, L.Z., Vederas, J.C., Zuber, P. Genes of the sbo-alb Locus of Bacillus subtilis Are Required for Production of the Antilisterial Bacteriocin Subtilosin. *J. Bacteriol.* **181**, 7346–7355 (1999).
15. Claesen, J., Bibb, M. Genome mining and genetic analysis of cypemycin biosynthesis reveal an unusual class of posttranslationally modified peptides. *Proc. Natl. Acad. Sci. U. S. A.* 107, 16297-16302 (2010).

16. Wieland-Brown, L.C., Acker, M.G., Clardy, J., Walsh, C. Fischbach, M.A. Thirteen posttranslational modifications convert a 14-residue peptide into the antibiotic thiocillin. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 2549–2553 (2009).
17. Liao, R. *et al.* Thiopeptide Biosynthesis Featuring Ribosomally Synthesized Precursor Peptides and Conserved Posttranslational Modifications. *Chem. Biol.* **16**, 141–147 (2009).
18. Kelly, W.L., Pan, L., Li, C. Thiostrepton Biosynthesis: Prototype for a New Family of Bacteriocins *J. Am. Chem. Soc.* **131**, 4327–4334 (2009).
19. Morris, R.P., *et al.* Ribosomally Synthesized Thiopeptide Antibiotics Targeting Elongation Factor Tu. *J. Am. Chem. Soc.* **131**, 5946–5955 (2009).
20. Onaka, H. Biosynthesis of Indolocarbazole and Goadsporin, Two Different Heterocyclic Antibiotics Produced by Actinomycetes. *Biosci. Biotechnol. Biochem.* **73**, 2149-2155 (2009).
21. Lee, S.W. *et al.* Discovery of a widely distributed toxin biosynthetic gene cluster. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 5879-5884 (2008).
22. Donia, M.S., Ravel, J., Schmidt, E.W. A global assembly line for cyanobactins. *Nat. Chem. Biol.* **4**, 341-343 (2008).
23. Ziemert, N. *et al.* Microcyclamide Biosynthesis in Two Strains of Microcystis aeruginosa: from Structure to Genes and Vice Versa. *Appl. Environ. Microbiol.* **74**, 1791–1797 (2008).
24. Oman, T.J., Boettcher, J.M., Wang, H., Okalibe, X.N., van der Donk, W.A. Sublancin is not a lantibiotic but an S-linked glycopeptide. *Nat. Chem. Biol.* **7**, 78-80 (2011).
25. Ziemert, N., Ishida, K., Liaimer, A., Hertweck, C., Dittmann, E. Ribosomal synthesis of tricyclic depsipeptides in bloom-forming cyanobacteria. *Angew. Chem. Int. Ed. Engl.* **47**, 7756-7759 (2008).
26. Lagos, R. *et al.* Structure, organization and characterization of the gene cluster involved in the production of microcin E492, a channel-forming bacteriocin. *Mol. Microbiol.* **42**, 229-243 (2001).
27. Gonzalez-Pastor, J.E., San Millan, J.L., Castilla. M.A., Moreno, F. Structure and Organization of Plasmid Genes Required To Produce the Translation Inhibitor Microcin C7. *J. Bacteriol.* **177**, 7131-7140 (1995).
28. Dufour, P. *et al.* High Genetic Variability of the agr Locus in Staphylococcus Species High Genetic Variability of the agr Locus in Staphylococcus Species. *J. Bacteriol.* **184**, 1180-1184 (2002).
29. Puehringer, S. Metlitzky, M. Schwarzenbacher, R. The pyrroloquinoline quinone biosynthesis pathway revisited: A structural approach. *BMC Biochemistry* **9** (2008).
30. Eijsink, V.G.H., Skeie, M., Middelhoven, H., Brurberg, M.B. Nes, I.F. Comparative Studies of Class IIa Bacteriocins of Lactic Acid Bacteria. *Appl. Environ. Microbiol.* **64**, 3275-3281 (1998).
31. Oppengard, C. *et al.* The Two-Peptide Class II Bacteriocins: Structure, Production, and Mode of Action. *J. Mol. Microbiol. Biotechnol.* **13**, 210-219 (2007).
32. Diaz, M. et al. Characterization of a New Operon, as-48EFGH, from the as-48 Gene Cluster Involved in Immunity to Enterocin AS-48. *Appl. Environ. Microbiol.* **69**, 1229-1236 (2003).

33. Hicks, L.M., Moffitt, M.C., Beer, L.L., Moore, B.S., Kelleher, N.L. Structural characterization of in vitro and in vivo intermediates on the loading module of microcystin synthetase. *ACS Chem. Biol.* **1**, 93-102 (2006).
34. Ju, J., Ozanick, S.G., Shen, B., Thomas, M.G. Conversion of (2S)-arginine to (2S,3R)-capreomycidine by VioC and VioD from the viomycin biosynthetic pathway of *Streptomyces* sp. strain ATCC11861. *Chembiochem*, **5**, 1281-1285 (2004).
35. Katahira, R., Yamasaki, M., Matsuda, Y., Yoshida, M. MS-271, a novel inhibitor of calmodulin-activated myosin light chain kinase from *Streptomyces* sp.–II. Solution structure of MS-271: characteristic features of the "lasso' structure. *Bioorg. Med. Chem.* **4,** 121–129 (1996).
36. D. Frechet *et al.* Solution structure of RP 71955, a new 21 amino acid tricyclic peptide active against HIV-1 virus. *Biochemistry* **33,** 42–50 (1994).
37. McIntosh, J.A., Donia, M.S., Schmidt, E.W. Ribosomal peptide natural products: bridging the ribosomal and nonribosomal worlds. *Nat. Prod. Rep.* **26**, 537-559 (2009).
38. Vats, P., Rothfield, L. Duplication and segregation of the actin (MreB) cytoskeleton during the prokaryotic cell cycle. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 17795-17800 (2007).
39. Schey, K.L., Finley, E.L. Identification of Peptide Oxidation by Tandem Mass Spectrometry. *Acc. Chem. Res.* **33**, 299-306 (2000).
40. Bridgewater, J.D., Srikanth, R., Lim, J., Vachet, R.W. The Effect of Histidine Oxidation on the Dissociation Patterns of Peptide Ions. *J. Am. Soc. Mass. Spectrom.* **18**, 553–562 (2007).
41. Willey, J.M., Willems, A., Kodani, S., Nodwell, J.R. Morphogenetic surfactants and their role in the formation of aerial hyphae in *Streptomyces coelicolor*. *Mol. Microbiol.* **59**, 731-742 (2006).
42. Ueda, K. AmfS, an extracellular peptidic morphogen in *Streptomyces griseus*. J Bacteriol. **184**, 1488-1492 (2002).
43. Li, B. *et al.* Catalytic promiscuity in the biosynthesis of cyclic peptide secondary metabolites in planktonic marine cyanobacteria. *Proc. Natl. Acad. Sci. U. S. A.* **107,** 10430-10435 (2010).

Supplementary Table 6 | Mass shifts of RNP monomers and modifications and their corresponding precursor amino acids in precursor peptides

Mass shift [Da]	Monomer (Abbreviation)	Biosynthetic precursor in prepropeptide	Mass shift [Da]	Monomer (Abbreviation)	Biosynthetic precursor in prepropeptide
69.021	Dehydroalanine (Dha)	Ser or Cys	195.113	Leucine-methyloxazole	Leu-Thr
71.037	D-Ala	L-Ser	195.113	Isoleucine-methyloxazole	Ile-Thr
72.021	oxopropionic acid	Ser (N-terminal in propeptide)	196.072	Gln-oxazole	Gln-Ser
74.036	2-hydroxypropionic acid	Ser (N-terminal in propeptide)	196.072	Asn-methyloxazole	Asn-Thr
83.037	Dehydrobutyrine (Dhb)	Thr	196.108	Lys-oxazole	Lys-Ser
86.036	2-oxobutyric acid	Thr (N-terminal in propeptide)	197.056	Glu-oxazole	Glu-Ser
100.076	N,N-dimethyl-alanine	Ala (N-terminal in propeptide)	197.056	Asp-methyloxazole	Asp-Thr
113.047	monohydroxy-proline	Pro	197.074	Leucine-thiazole	Leu-Cys
125.035	Gly-oxazole	Gly-Ser	197.074	Isoleucine-thiazole	Ile-Cys
127.051	Gly-oxazoline	Gly-Ser	197.128	Leucine-methyloxazoline	Leu-Thr
129.042	dihydroxy-proline	Pro	197.128	Isoleucine-methyloxazoline	Ile-Thr
131.021	β-hydroxy-aspartate	Asp (putative)	198.033	Asn-thiazole	Asn-Cys
137.034	Dha-oxazole	Ser/Cys-Ser	198.087	Asn-methyloxazoline	Asn-Thr
139.050	Ala-oxazole	Ala-Ser	198.088	Gln-oxazoline	Gln-Ser
139.050	Dha-oxazoline	Ser/Cys-Ser	198.124	Lys-oxazoline	Lys-Ser
139.051	Gly-methyloxazole	Gly-Thr	199.017	Asp-thiazole	Asp-Cys
141.012	Gly-thiazole	Gly-Cys	199.053	Met-oxazole	Met-Ser
141.066	Gly-methyloxazoline	Gly-Thr	199.071	Asp-methyloxazoline	Asp-Thr
141.066	Ala-oxazoline	Ala-Ser	199.072	Glu-oxazoline	Glu-Ser
143.028	Gly-thiazoline	Gly-Cys	199.090	Leucine-thiazoline	Leu-Cys
151.050	Dhb-oxazole	Thr-Ser	199.090	Isoleucine-thiazoline	Ile-Cys
151.050	Dha-methyloxazole	Ser/Cys-Thr	200.049	Asn-thiazoline	Asn-Cys
153.011	Dha-thiazole	Ser/Cys-Cys	201.033	Asp-thiazoline	Asp-Cys
153.065	Dha-methyloxazoline	Ser/Cys-Thr	201.069	Met-oxazoline	Met-Ser
153.066	Dhb-oxazoline	Thr-Ser	205.072	His-oxazole	His-Ser
153.066	Ala-methyloxazole	Ala-Thr	207.088	His-oxazoline	His-Ser
155.027	Ala-thiazole	Ala-Cys	210.088	Gln-methyloxazole	Gln-Thr
155.027	Dha-thiazoline	Ser/Cys-Cys	210.124	Lys-methyloxazole	Lys-Thr
155.045	Ser-oxazole	Ser-Ser	211.072	Glu-methyloxazole	Glu-Thr
155.081	Ala-methyloxazoline	Ala-Thr	212.049	Gln-thiazole	Gln-Cys
155.094	dimethylallyl-serine	Ser	212.085	Lys-thiazole	Lys-Cys
157.043	Ala-thiazoline	Ala-Cys	212.103	Gln-methyloxazoline	Gln-Thr
157.061	Ser-oxazoline	Ser-Ser	212.139	Lys-methyloxazoline	Lys-Thr
160.043	N-formyl-methionine	Met (N-terminal)	213.033	Glu-thiazole	Glu-Cys
165.066	Pro-oxazole	Pro-Ser	213.069	Met-methyloxazole	Met-Thr
165.066	Dhb-methyloxazole	Thr-Thr	213.087	Glu-methyloxazoline	Glu-Thr
167.027	Dhb-thiazole	Thr-Cys	214.065	Gln-thiazoline	Gln-Cys
167.081	Dhb-methyloxazoline	Thr-Thr	214.101	Lys-thiazoline	Lys-Cys
167.081	Val-oxazole	Val-Ser	215.030	Met-thiazole	Met-Cys
167.082	Pro-oxazoline	Pro-Ser	215.049	Glu-thiazoline	Glu-Cys
169.043	Dhb-thiazoline	Thr-Cys	215.081	Phe-oxazole	Phe-Ser
169.061	Thr-oxazole	Thr-Ser	215.084	Met-methyloxazoline	Met-Thr
169.061	Ser-methyloxazole	Ser-Thr	217.046	Met-thiazoline	Met-Cys
169.097	Val-oxazoline	Val-Ser	217.097	Phe-oxazoline	Phe-Ser
169.110	dimethylallyl-threonine	Thr	219.088	His-methyloxazole	His-Thr
171.022	Ser-thiazole	Ser-Cys	220.040	5-chlorotryptophan	Trp
171.022	Cys-oxazole	Cys-Ser	221.049	His-thiazole	His-Cys
171.076	Ser-methyloxazoline	Ser-Thr	221.103	His-methyloxazoline	His-Thr
171.077	Thr-oxazoline	Thr-Ser	223.065	His-thiazoline	His-Cys
173.038	Ser-thiazoline	Ser-Cys	224.114	Arg-oxazole	Arg-Ser
173.038	Cys-oxazoline	Cys-Ser	226.130	Arg-oxazoline	Arg-Ser
179.082	Pro-methyloxazole	Pro-Thr	229.097	Phe-methyloxazole	Phe-Thr
181.043	Pro-thiazole	Pro-Cys	231.058	Phe-thiazole	Phe-Cys
181.097	Pro-methyloxazoline	Pro-Thr	231.076	Tyr-oxazole	Tyr-Ser
181.097	Leucine-oxazole	Leu-Ser	231.112	Phe-methyloxazoline	Phe-Thr
181.097	Isoleucine-oxazole	Ile-Ser	231.125	dimethylallyl-tyrosine	Tyr
181.097	Val-methyloxazole	Val-Thr	233.074	Phe-thiazoline	Phe-Cys
182.056	Asn-oxazole	Asn-Ser	233.092	Tyr-oxazoline	Tyr-Ser
183.040	Asp-oxazole	Asp-Ser	238.130	Arg-methyloxazole	Arg-Thr
183.058	Val-thiazole	Val-Cys	240.091	Arg-thiazole	Arg-Cys
183.059	Pro-thiazoline	Pro-Cys	240.145	Arg-methyloxazoline	Arg-Thr
183.077	Thr-methyloxazole	Thr-Thr	242.107	Arg-thiazoline	Arg-Cys
183.112	Val-methyloxazoline	Val-Thr	245.092	Tyr-methyloxazole	Tyr-Thr
183.113	Leucine-oxazoline	Leu-Ser	247.053	Tyr-thiazole	Tyr-Cys
183.113	Isoleucine-oxazoline	Ile-Ser	247.107	Tyr-methyloxazoline	Tyr-Thr
184.072	Asn-oxazoline	Asn-Ser	249.069	Tyr-thiazoline	Tyr-Cys
185.038	Thr-thiazole	Thr-Cys	254.092	Trp-oxazole	Trp-Ser
185.038	Cys-methyloxazole	Cys-Thr	256.108	Trp-oxazoline	Trp-Ser
185.056	Asp-oxazoline	Asp-Ser	265.062	S-glucosyl-cysteine	Cys
185.074	Val-thiazoline	Val-Cys	268.108	Trp-methyloxazole	Trp-Thr
185.092	Thr-methyloxazoline	Thr-Thr	270.069	Trp-thiazole	Trp-Cys
186.999	Cys-thiazole	Cys-Cys	270.123	Trp-methyloxazoline	Trp-Thr
187.053	Cys-methyloxazoline	Cys-Thr	272.085	Trp-thiazoline	Trp-Cys
187.054	Thr-thiazoline	Thr-Cys	287.103	N-succinyltryptophan	Trp (putative, N-terminal in propeptide)
189.015	Cys-thiazoline	Cys-Cys	323.212	Geranyl-tryptophan	Trp
			391.274	Farnesyl-tryptophan	Trp

Supplementary Figure 7 | Mass shifts of NRP monomers (NORINE monomer list excluding lipids) and corresponding NRPS monomers and genome mining accessibility by NRPSpredictor2 (AntiSMASH, * - biosynthetic monomers can be putative due to lack of biosynthetic knowledge)

Mass shift [Da]	Structural monomer	Abbreviation (NORINE)	Biosynthetic monomer *	NRPSpredictor2 accessibility of structural monomer (NRPSpredictor2 monomer code)
43.07252	Ethanolamine	Eta	Eta	no
67.09392	pyrrolidone	Pyr	Pyr	no
69.06674	dehydroalanine	dh-Ala	Dha or Ser	no (Dha), yes (ser)
70.051505	pyruvate	Pya	Pya	no
70.14092	putrescine	Put	Put	no
71.08262	beta-Alanine	b-Ala	b-Ala	yes (ala-b)
71.08262	N-Methyl-Glycine	NMe-Gly	Gly	yes (Gly)
72.06738	D-lactic acid	D-Lac	D-Lac	no
72.06738	Lactic acid	Lac	Lac	no
73.098495	Serinol	Serol	Ser	yes (ser)
83.09332	2,3-dehydro-2-aminobutyric acid	dhAbu	Thr or Dhb	yes (thr), no (Dhb)
83.09332	hydroxy pyrrolidone	OH-Pyr	OH-Pyr	no
83.09332	homoserine lactone	HseL	HseL or Hse	no
83.09332	N-Methyl-dehydroalanine	NMe-Dha	Ser or Dha	yes (Ser), no (Dha)
85.066146	alpha-formylGlycine	aFo-Gly	aFo-Gly	no
85.10921	2-Aminoisobutyric acid	Aib	Aib	yes
85.10921	D-3-methoxyalanine	D-3OMe-Ala	3OMe-Ala or Ser	no (3OMe-Ala), yes (Ser)
85.10921	N-Methyl-Alanine	NMe-Ala	Ala	yes (ala)
85.10921	N-methyl-beta-alanine	NMe-bAla	bAla (Ala-b)	yes (ala-b, beta-ala)
85.10921	D-2-Aminobutyric acid	D-Abu	Abu	yes (abu)
85.10921	2-Aminobutyric acid	Abu	Abu	yes (abu)
85.15226	isovalinol	Ivalol	Ival	yes (iva)
85.15226	valinol	Valol	Val	yes (val)
86.09726	2,3-Diaminopropionic acid	Dpr	Dpr	yes (LDAP)
87.08202	D-Serine	D-Ser	Ser	yes (ser)
87.08202	isoserine	Iser	Iser	no
96.13514	proline carboxamid	ProC	Pro or ProC	yes (pro), no(ProC)
97.1199	norcoronamic acid	norCMA	Val	yes (val)
97.1199	2-methylamino-2-dehydrobutyric acid	2Dh-Mabu	Thr	yes (thr)
97.1199	D-Proline	D-Pro	Pro	yes (pro)
98.14773	3-methylvaleric acid	Me-Vaa	Me-Vaa	no
99.09272	N-formyl-Alanine	NFo-Ala	Ala	yes
99.09272	D-N-formyl-Alanine	D-NFo-Ala	Ala	yes (Ala)
99.13578	N-dimethyl-Alanine	NdMe-Ala	Ala	yes
99.13578	Norvaline	Nva	Nva	no
99.13578	D-Norvaline	D-Nva	Nva	no
99.13578	D-Valine	D-Val	Val	yes (val)
99.13578	D-isovaline	D-Ival	Ival	yes (iva)
99.13578	isovaline	Ival	Ival	yes (iva)
99.13578	2-methyl-3-aminobutanoic acid	Mab	Mab	no
99.17884	Isoleucinol	Ileol	Ile	yes (ile)
99.17884	Leucinol	Leuol	Leu	yes (leu)
100.120536	2-hydroxyisovalerate	Hiv	Hiv	yes (hyv)
100.120536	D-2-hydroxyisovalerate	D-Hiv	D-Hiv	yes (hyv-d)
100.12384	2,4-diaminobutyric acid	Dab	Dab	yes (dab)
100.12384	D-2,4-diaminobutyric acid	D-Dab	D-Dab	yes (dab)
100.12384	(2S)-2,3-diaminobutyric acid	Dbu	Dbu	no
100.12384	(2R,3R)-2,3-diaminobutyric acid	D-Dbu	Dbu	no
100.12384	(2S,3S)-2,3-diaminobutyric acid	L-Dbu	Dbu	no
101.108596	4-amino-3-hydroxybutyric acid	OH-4Abu	OH-4Abu	no
101.108596	D-Threonine	D-Thr	Thr	yes (thr)
101.108596	allo-Threonine	aThr	aThr	yes (allo-thr)
101.108596	D-allo-Threonine	D-aThr	aThr	yes (allo-thr)
101.108596	Homoserine	Hse	Hse	no
101.108596	D-Homoserine	D-Hse	Hse	no
101.108596	N-Methyl-D-Serine	D-NMe-Ser	Ser	yes (ser)
101.108596	N-Methyl-Serine	NMe-Ser	Ser	yes (ser)
101.13174	dehydro-cysteine	dhCys	Cys	yes (cys)
103.14761	D-Cysteine	D-Cys	Cys	yes (cys)
104.110786	benzoic acid	Bz	Bz	no
110.16171	N-methylglutamine	NMe-Gln	Gln	yes (Gln)
110.18486	N-Methyl-Glycine-thiazole	NMe-Gly-Thz	Gly-Cys	yes (Gly, Cys)
111.10342	pyroglutamic acid	pGlu	Glu	yes (Glu)
111.10342	4-oxo-proline	4oxo-Pro	Pro or 4oxo-Pro	yes (pro), no (4oxo-Pro)
111.14648	coronamic acid	CMA	alle	yes (alle)
111.14648	homoproline	Hpr	Hpr	yes (pip)
111.14648	D-homoproline	D-Hpr	Hpr	yes (pip)
111.14648	3-methylproline	3Me-Pro	Pro or 3Me-Pro	yes (pro), no (3Me-Pro)
111.14648	4-methylproline	4Me-Pro	Pro or 4Me-Pro	yes (pro), no (4Me-Pro)
111.14648	5-methylproline	5Me-Pro	Pro or 5Me-Pro	yes (pro), no (5Me-Pro)
112.13124	keto-Leucine	k-Leu	Leu or k-Leu	yes (leu), no (k-Leu)
112.13454	Hydroxy-cycloOrnithine	OH-cOrn	OH-cOrn	no
112.13454	D-Hydroxy-cycloOrnithine	D-OH-cOrn	OH-cOrn	no
113.07624	aziridine dicarboxylic acid	Azd	Azd	no
113.11929	N-formyl-D-aminobyric acid	NFo-D-Abu	Abu	yes (abu)
113.11929	4-Hydroxyproline	4OH-Pro	4OH-Pro	no
113.11929	D-hydroxyproline	D-4OH-Pro	Pro or 4OH-Pro	yes (pro), no (4OH-Pro)
113.11929	3-Hydroxyproline	3OH-Pro	Pro or 3OH-Pro	yes (pro), no (3OH-Pro)
113.16237	D-Isoleucine	D-Ile	Ile	yes (ile)
113.16237	allo-Isoleucine	alle	alle	yes (alle)
113.16237	D-allo-Isoleucine	D-alle	alle	yes (alle)
113.16237	D-tert-Leu	D-t-Leu	t-Leu	no
113.16237	tert-Leu	t-Leu	t-Leu	no
113.16237	D-Leucine	D-Leu	Leu	yes (leu)
113.16237	D-N-methyl-norvaline	D-NMe-Nva	Nva	no
113.16237	D-N-Methylvaline	D-NMe-Val	Val	yes (val)
113.16237	N-Methyl-Valine	NMe-Val	Val	yes (val)
113.16237	2-methyl-3-aminopentanoic acid	Map	Map	no
113.20872	norspermidine	NSpd	NSpd	no
114.10406	hydroxyacetyl propionyl	Hap	Hap	no
114.10406	pentanedioic acid	Pda	Pda	no
114.10736	D-Asparagine	D-Asn	Asn	yes (asn)

Supplementary Figure 7 | Mass shifts of NRP monomers (NORINE monomer list excluding lipids) and corresponding NRPS monomers and genome mining accessibility by NRPSpredictor2 (AntiSMASH, * - biosynthetic monomers can be putative due to lack of biosynthetic knowledge)

Mass shift [Da]	Structural monomer	Abbreviation (NORINE)	Biosynthetic monomer *	NRPSpredictor2 accessibility of structural monomer (NRPSpredictor2 monomer code)
114.10736	N1-formyl-2,3-Diaminopropionic acid	NFo-Dpr	Dpr or NFo-Dpr	yes (LDAP), no (NFo-Dpr)
114.14712	2-hydroxy-3-methyl-pentanoic acid	Hmp	Hmp	no
114.14712	D-2-hydroxy-3-methylpentanoic acid	D-Hmp	D-Hmp	yes (hmp-D)
114.14712	4-Methyl-D-2-hydroxy-valeric acid	4Me-D-Hva	4Me-Hva	no
114.15042	D-ornithine	D-Orn	Orn	yes (orn)
114.15042	Ornithine	Orn	Orn	yes (orn)
115.09213	D-aspartic acid	D-Asp	Asp	yes (asp)
115.09213	N-formyl-isoserine	NFo-Iser	Iser	no
115.13518	D-beta-hydroxyvaline	D-bOH-Val	Val or bOH-Val	yes (val), yes (MeHOval)
115.13518	beta-hydroxyvaline	bOH-Val	Val or bOH-Val	yes (Val), (MeHOval)
115.13518	O-methyl-threonine	OMe-Thr	Thr or OMe-Thr	yes (thr), no (OMe-Thr)
115.13518	N-methylthreonine	NMe-Thr	Thr	yes (thr)
117.108	4-Hydroxythreonine	4OH-Thr	Thr or 4OH-Thr	yes (thr), no (4OH-Thr)
117.17421	alpha-methylcysteine	aMe-Cys	aMe-Cys	no
117.17424	N-methylcysteine	NMe-Cys	Cys	yes (cys)
118.13736	phenylacetic acid	Pha	Pha	no
120.11018	para-hydroxy-benzoic acid	pOH-Bz	pOH-Bz (Sal)	yes (sal)
121.09825	hydroxypicolinic acid	Hpa	Hpa	no
125.13	4-oxo-homoproline	4oxo-Hpr	4oxo-Hpr	no
125.13	N-Formyl-Proline	NFo-Pro	Pro	yes (pro)
125.13	4-oxo-5-methylproline	4oxo-5Me-Pro	Pro or 5Me-Pro	yes (pro), no (4oxo-5Me-Pro)
127.10611	beta-ureido-dehydroAlanine	bU-Ala	Dpr	yes (LDAP)
127.14589	N-Acetyl-2-aminoisobutyric acid	Ac-Aib	Aib	yes (aib)
127.14589	N-formyl-Valine	NFo-Val	Val	yes (val)
127.14589	3-Hydroxy-5-methylproline	3OH-5Me-Pro	Pro or 3OH-5Me-Pro	yes (pro), no (3OH-5Me-Pro)
127.18895	homoisoleucine	Hil	Hil	no
127.18895	D-N-methyl-alloisoleucine	D-NMe-alle	alle	yes (alle)
127.18895	N-Methyl-Isoleucine	NMe-Ile	Ile	yes (ile)
127.18895	beta-methylisoleucine	bMe-Ile	Ile or bMe-Ile	yes (ile), no (bMe-Ile)
127.18895	N-methyl-alloIsoleucine	NMe-alle	alle	yes (alle)
127.18895	D-N-methyl-Leucine	D-NMe-Leu	Leu	yes (leu)
127.18895	N-Methyl-Leucine	NMe-Leu	Leu	yes (leu)
127.18895	alpha-ethylnorvaline	Et-Nva	Et-Nva	no
127.18895	Dolavaline	Dov	Val	yes (val)
127.2353	spermidine	Spd	Spd	no
128.13394	D-N2-methyl-asparagine	D-N2Me-Asn	N2Me-Asn	no
128.13394	beta-methyl-asparagine	bMe-Asn	bMe-Asn	no
128.13394	N-methylasparagine	NMe-Asn	Asn	yes (asn)
128.13394	N1-acetyl-2,3-Diaminopropionic acid	NAc-Dpr	Dpr or NAc-Dpr	yes (LDAP), no (NAc-Dpr)
128.13394	D-Glutamine	D-Gln	Gln	yes (Gln)
128.177	D-Lysine	D-Lys	Lys	yes (lys)
128.177	beta lysine	bLys	bLys	yes (lys-b)
128.177	N-Hydroxy-histamine	N-OH-Hta	N-OH-Hta	no
129.1187	D-beta-methyl-aspartic acid	D-bMe-Asp	bMe-Asp	no
129.1187	beta-methyl-aspartic acid	bMe-Asp	bMe-Asp	no
129.1187	beta-methoxy-aspartic acid	bOMe-Asp	bOMe-Asp	no
129.1187	D-Glutamic Acid	D-Glu	Glu	yes (Glu)
129.1187	O-acetyl-Serine	Ac-Ser	Ser or Ac-Ser	yes (ser), no (Ac-Ser)
129.16177	L-acosamine	Aco		no
129.16177	L-ristosamine	Ria		no
129.16177	3-hydroxyleucine	3OH-Leu	Leu or 3OH-Leu	yes (leu), no (3OH-Leu)
129.16177	gama-hydroxy-N-Methyl-Valine	gOH-NMe-Val	Val or gOH-Val	yes (val), no (gOH-Val)
129.16177	beta-hydroxy-N-Methyl-Valine	bOH-NMe-Val	Val or bOH-Val	yes (val), yes (MeHOval)
130.10676	Hydroxyasparagine	OH-Asn	OH-Asn	yes (hasn)
130.10676	D-HydroxyAsparagine	D-OH-Asn	OH-Asn	yes (hasn)
130.11006	alpha-guanidino Serine	gSer	gSer	no
130.14981	hydroxy-beta lysine	OH-bLys	bLys	yes (lys-b)
130.14981	N5-hydroxy ornithine	OH-Orn	OH-Orn	yes (horn)
130.14981	D-N5-HydroxyOrnithine	D-OH-Orn	OH-Orn	yes (horn)
131.09152	Hydroxyaspartic acid	OH-Asp	OH-Asp	no
131.09152	D-Hydroxyaspartic acid	D-OH-Asp	D-OH-Asp	no
131.20074	N,S-dimethylcysteine	diMe-Cys	Cys	yes (cys)
132.11935	D-arabinose	D-Ara		no
132.11935	Arabinose	Ara		no
132.11935	lyxose	Lyx		no
132.1624	L-Olivose	Oli		no
133.15199	D-PhenylGlycine	D-ph-Gly	Ph-Gly	yes (phg)
133.15199	phenylglycine	Ph-Gly	Ph-Gly	yes (phg)
133.19507	phenylalaninol	Pheol	Phe	yes (phe)
135.55366	4-Chloro-Threonine	4Cl-Thr	Thr	yes (thr)
136.10959	2,3-dihydroxybenzoic acid	diOH-Bz	diOH-Bz	yes (dhb)
138.12877	dehydropyrrolidone	dPyr	dPyr	no
139.15658	2,3-dimethylpyroglutamic acid	2Me-3Me-pGlu	Glu or 2Me-3Me-Glu	yes (Glu), no (2Me-3Me-Glu)
140.19099	arginal	Argal	Arg	yes (arg)
141.17247	4-oxovancosamine	4oxo-Van		no
141.17247	N-formyl-Isoleucine	NFo-Ile	Ile	yes (ile)
141.17247	N-formyl-Leucine	NFo-Leu	Leu	yes (leu)
141.17247	N-acetyl-isovaline	Ac-Ival	Ival	yes (iva)
141.17247	N-Acetylvaline	Ac-Val	Val	yes (val)
141.17247	4-amino-2,2-dimethyl-3-oxopentanoic acid	Ibu	Ibu	no
141.21551	N,O-dimethyl-isoleucine	NMe-OMe-Ile	Ile or OMe-Ile	yes (ile), no (OMe-Ile)
141.21551	O-acetyl-leucinol	OAc-Leuol	Leu	yes (leu)
141.21551	N,beta-dimethylLeucine	NMe-bMe-Leu	bMe-Leu	no
141.21554	N,gamma-alloisoleucine	diMe-alle	alle or Me-alle	yes (alle), no (Me-alle)
141.55978	N-methylchloropyrrole	MCP	N-methylpyrrole-2-carboxylic acid	no
142.16052	D-beta-methylglutamine	D-bMe-Gln	Gln or bMe-Gln	yes (Gln), no (bMe-Gln)
142.16052	beta-methylglutamine	bMe-Gln	Gln or bMe-Gln	yes (Gln), no (bMe-Gln)
143.14528	2-Aminoadipic acid	Aad	Aad	yes (aad)
143.14528	Glutamic Acid methyl ester	MeO-Glu	Glu or MeO-Glu	yes (Glu), no (MeO-Glu)
143.14528	D-Glutamic Acid methyl ester	D-MeO-Glu	Glu or MeO-Glu	yes (Glu), no (MeO-Glu)
143.14528	D-Glutamic Acid methyl ester	MeO-D-Glu	Glu or MeO-Glu	yes (Glu), no (MeO-Glu)
143.14528	3-Methyl-Glutamic acid	3Me-Glu	3Me-Glu	yes (3-me-glu)

Supplementary Figure 7 | Mass shifts of NRP monomers (NORINE monomer list excluding lipids) and corresponding NRPS monomers and genome mining accessibility by NRPSpredictor2 (AntiSMASH, * - biosynthetic monomers can be putative due to lack of biosynthetic knowledge)

Mass shift [Da]	Structural monomer	Abbreviation (NORINE)	Biosynthetic monomer *	NRPSpredictor2 accessibility of structural monomer (NRPSpredictor2 monomer code)
143.18834	L-actinosamine	Act		no
143.18834	L-eremosamine	Ere		no
143.18834	N-methyl-hydroxyisoleucine	NMe-OH-Ile	Ile or OH-Ile	yes (ile), no (OH-Ile)
143.18834	norstatine	Nst	Nst	no
144.13334	D-beta-hydroxy-N2-methyl-asparagine	D-N2Me-bOH-Asn	N2Me-bOH-Asn	no (N2Me-bOH-Asn)
144.13334	beta-hydroxyglutamine	bOH-Gln	bOH-Gln	no
145.1181	methoxyaspartic acid	OMe-Asp	OMe-Asp	no
146.14592	L-rhamnose	Rha		no
147.17857	N-methyl-phenylglycine	NMe-Ph-Gly	Ph-Gly	yes (phg)
147.17857	beta phenylalanine	bPhe	bPhe	no
147.17857	D-Phenylalanine	D-Phe	Phe	yes (phe)
147.17857	D-beta-phenylalanine	D-bPhe	bPhe	no
147.20018	Methionine-S-oxide	O-Met	Met or O-Met	yes (met), no (O-Met)
147.60742	Chloro-Isoleucine	Cl-Ile	Ile or Cl-Ile	yes (ile), no (Cl-Ile)
148.12028	2-hydroxyphenyl-2-oxo-ethanoic acid	Hpoe	Hpoe	no
148.16334	D-Phenyl-lactate	D-Ph-Lac	D-Ph-Lac	no
148.16334	Phenyl-lactate	Ph-Lac	Ph-Lac	no
149.1514	D-HydroxyPhenylGlycine	D-Hpg	Hpg	yes (hpg)
149.1514	HydroxyPhenylGlycine	Hpg	Hpg	yes (hpg)
151.14246	D-4-fluoroPhenylGlycine	D-F-ph-Gly	D-F-ph-Gly	no
151.14581	cysteic acid	CysA	Cys or CysA	yes (cys), no (CysA)
151.14581	D-cysteic acid	D-CysA	Cys or CysA	yes (cys), no (CysA)
153.1434	Hydroxyhistidine	OH-His	His (bOH-His) or OH-His	yes (his), no (OH-His)
153.22621	N-methyl homo vinylogous Valine	NMe-hv-Val	hv-Val	no
154.17451	capreomycidine	Cap	Cap	yes (cap)
154.17451	enduracididine	End	End	no
154.17451	D-enduracididine	D-End	End	no
154.19437	Alanine-thiazole	Ala-Thz	Ala-Cys	yes (ala, cys)
155.19904	O-desmethyldolaproine	dDap	dDap	no
155.19904	N-acetyl-Leucine	NAc-Leu	Leu	yes (leu)
155.2421	methyl-2-aminooctanoic acid	Me-AOA	Me-AOA	no
156.14404	N-formyl-Glutamine	NFo-Gln	Gln	yes (Gln)
156.14554	2-carboxyquinoxaline	COOH-Qui	COOH-Qui	no
156.18381	hydroxyisovalerylpropionyl	Hip	Hip	no
156.1871	3,4-dimethylglutamine	3Me-4Me-Gln	Gln or 3Me-4Me-Gln	yes (Gln), no (3Me-4Me-Gln)
156.25003	N-trimethyl-leucine	NtMe-Leu	Leu	yes (leu)
156.25003	N-dimethyl-leucine	NdMe-Leu	Leu	yes (leu)
157.17517	D-Citrulline	D-Cit	citrulline	no
157.17517	Citrulline	Cit	citrulline	no
157.1904	D-Arginine	D-Arg	Arg	yes (arg)
157.21492	isostatine	Ist	Ist	no
157.21492	statine	Sta	Sta	no
158.15991	N6-formyl-HydroxyOrnithine	Fo-OH-Orn	Fo-OH-Orn	yes (hforn)
158.15991	D-formyl-hydroxyOrnithine	D-Fo-OH-Orn	Fo-OH-Orn	yes (hforn)
159.14468	alpha-amino-hydroxyadipic acid	Ahad	Aad or Ahad	yes (aad)
159.18928	N-methyl-2,3-dehydrophenylalanine	NMe-dPhe	Phe	yes (phe)
160.1725	O-methyl-L-rhamnose	2OMe-Rha		no
161.20517	D-N-Methyl-Phenylalanine	D-NMe-Phe	Phe	yes (phe)
161.20517	3-methylphenylalanine	3Me-Phe	Phe or 3Me-Phe	yes (phe), no (3Me-Phe)
161.20517	N-Methyl-Phenylalanine	NMe-Phe	Phe	yes (phe)
161.20517	Homophenylalanine	Hph	Hph	no
162.14532	D-mannose	D-Man		no
162.14532	beta-D-galactose	bD-Gal		no
162.14532	D-galactose	D-Gal		no
162.14532	D-Glucose	D-Glc		no
162.14532	L-Glucose	Glc		no
163.11627	phosphinothricin	PT	AcDMPT	no
163.13492	N-hydroxy-dehydro-HydroxyPhenylGlycine	OH-dHpg	Hpg	yes (hpg)
163.13492	D-N-hydroxy-dehydro-HydroxyPhenylGlycine	D-OH-dHpg	Hpg	yes (hpg)
163.17798	N-methyl-HydroxyPhenylGlycine	NMe-Hpg	Hpg	yes (hpg)
163.17798	phenylserine	Ph-Ser	Ph-Ser	no
163.17798	D-Tyrosine	D-Tyr	Tyr	yes (tyr)
163.17798	beta-tyrosine	bTyr	bTyr	no
163.19958	Methionine sulfone	O2-Met	Met or O2-Met	yes (met), no (O2-Met)
164.16275	4-hydroxy-D-phenyl-lactate	4OH-D-Ph-Lac	4OH-D-Ph-Lac	no
164.16604	propenoyl-alanyloxazole acid	PALOA	Ala-v-Ser	yes (Ala), no (v-Ser)
165.1508	D-3,5-dihydroxyphenylglycine	D-Dhpg	Dhpg	yes (dpg, dhpg)
165.1508	3,5-dihydroxyphenylglycine	Dhpg	Dhpg	yes (dpg, dhpg)
166.01001	3,4-dichloro-proline	Cl2-Pro	Pro or Cl2-Pro	yes (pro), no (Cl2-Pro)
166.18523	cyclo alpha-ketoarginine	ck-Arg	ck-Arg	no
167.20973	2-carboxy-6-hydroxyoctahydroindole	Choi	Choi	no
167.25281	3-Desoxy-Methyl-4-butenyl-4-methyl threonine	3d-NMe-Bmt	3d-Bmt	no
168.22096	1-methoxy-beta-alanine-thiazole	OMe-bAla-Thz	bAla-Cys	yes (ala-b, cys)
169.22562	4-butenyl-4-methyl threonine	Bmt	Bmt	yes (bmt)
169.22562	Dolaproine	Dap	Dap	no
169.2753	guanylspermidine	GSpd	GSpd	no
170.17392	5-hydroxy-capreomycidine	5OH-Cap	Cap	yes (cap)
170.21039	hydroxysecbutyl acetyl propionyl	Hysp	Hysp	no
170.21699	homoarginine	Har	Har	no
171.19844	N-methoxyacetyl-valine	NOMe-Ac-Val	Val	yes (val)
171.2415	N-desmethyldolaisoleuine	dDil	dDil or Dil	no
172.18651	N-acetyl-HydroxyOrnithine	Ac-OH-Orn	Ac-OH-Orn	yes (haorn)
172.18651	D-N-acetyl-HydroxyOrnithine	D-Ac-OH-Orn	Ac-OH-Orn	yes (haorn)
172.23111	tryptophanol	Trpol	Trp	yes (trp)
175.23174	alpha-amino-phenyl-valeric acid	Apv	Apv	no
175.25334	beta,beta-dimethyl-Methionine-S-oxide	bbMe2-O-Met	bbMe2-O-Met or bbMe2-Met	no
176.00485	N-methyldichloropyrrole-2-carboxylic acid	MdCP	N-methylpyrrole-2-carboxylic acid	no
177.20456	beta-hydroxy-N-Methyl-Phenylalanine	bOH-NMe-Phe	Phe or bOH-Phe	yes (phe), no (bOH-Phe)
177.20456	Homotyrosine	Hty	Hty	no
177.20456	N-methyltyrosine	NMe-Tyr	Tyr	yes (tyr)
178.19261	propenoyl-2-aminobutanoyloxazole acid	PAOA	Abu-v-Ser	yes (abu), no (v-Ser)
179.17738	3,4-dihydroxyphenylalanine	diOH-Phe	Tyr or diOH-Phe	yes (tyr), no (diOH-Phe)

Supplementary Figure 7 | Mass shifts of NRP monomers (NORINE monomer list excluding lipids) and corresponding NRPS monomers and genome mining accessibility by NRPSpredictor2 (AntiSMASH, * - biosynthetic monomers can be putative due to lack of biosynthetic knowledge)

Mass shift [Da]	Structural monomer	Abbreviation (NORINE)	Biosynthetic monomer *	NRPSpredictor2 accessibility of structural monomer (NRPSpredictor2 monomer code)
179.17738	beta-hydroxy-tyrosine	bOH-Tyr	bOH-Tyr	yes (bht)
181.19327	Anticapsin	Aca	Aca	no
182.18462	D-homoarginine	D-Har	Har	no
182.22768	vinylogous arginine	v-Arg	Arg or v-Arg	yes (arg), no (v-Arg)
183.2522	N-methyl-4-butenyl-4-methyl threonine	NMe-Bmt	Bmt	yes (bmt)
183.59647	3-chloro-4-hydroxyphenylglycine	Cl-Hpg	Hpg	yes (hpg)
184.19874	2,3-Dehydro-Tryptophan	dh-Trp	Trp	yes (trp)
184.2005	alpha-ketoarginine	k-Arg	k-Arg	no
185.22656	Dolapyrrolidone	Dpy	Dpy	no
185.26807	Dolaisoleucine	Dil	Dil	no
186.21461	D-Tryptophan	D-Trp	Trp	yes (trp)
186.21638	hydrated alpha-ketoarginine	hk-Arg	hk-Arg	no
186.28082	Dolaphenine	Doe	Phe-Cys	yes (Phe, Cys)
187.19939	dehydro vinylogous tyrosine	dv-Tyr	Tyr or dv-Tyr	yes (tyr), no (dv-Tyr)
188.18919	3,4-dihydroxyArginine	diOH-Arg	Arg or diOH-Arg	yes (arg), no (diOH-Arg)
189.21527	N-acetylphenylalanine	Ac-Phe	Phe	yes (phe)
189.21527	vinylogous tyrosine	v-Tyr	v-Tyr	no
190.20331	D-kynurenine	D-Kyn	Kyn	no
190.20331	kynurenine	Kyn	Kyn	no
190.24638	N-methyl-4-methylamino-phenylalanine	NMe-MeA-Phe	MeA-Phe	no
190.24638	N,O-dimethyl-tyrosinecarboxamid	NMe-OMe-TyrC	OMe-TyrC	no
191.23114	alpha-amino-hydroxyphenyl-valeric acid	Ahv	Ahz	no
191.23114	N-methyl-homotyrosine	NMe-Hty	Hty	no
191.23114	3-methyl-homotyrosine	3Me-Hty	Hty or 3Me-Hty	no
191.23114	D-N,O-dimethyl-tyrosine	D-NMe-OMe-Tyr	Tyr or OMe-Tyr	yes (tyr), no (OMe-Tyr)
191.23114	N,O-dimethyl-tyrosine	NMe-OMe-Tyr	Tyr or OMe-Tyr	yes (tyr), no (OMe-Tyr)
193.20396	beta-methoxy-tyrosine	bOMe-Tyr	Tyr, bOH-Tyr or bOMe-Tyr	yes (bht or tyr), no (bOMe-Tyr)
194.0632	di-chloro-N-methyl-dehydroLeucine	Cl2-NMe-dhLeu	Leu or dhLeu	yes (leu), no (dhLeu)
194.21033	O-sulfate-2-hydroxy-3-methylpentanoic acid	OSu-Hmp	Hmp or OSu-Hmp	no
194.27814	methyloxazoline-isoleucine	MeOx-Ile	Thr-Ile	yes (ile-thr)
196.07907	di-chloro-N-methyl-Leucine	Cl2-NMe-Leu	Leu	yes (leu), no (Cl2-Leu)
197.27878	4-butenyl-4-methyl-N,4-methyl threonine	Me2-Bmt	Me2-Bmt or Bmt	no, yes (bmt)
197.62304	chloro-tyrosine	Cl-Tyr	Tyr or Cl-Tyr	yes (tyr), no (Cl-Tyr)
200.19661	N-acetyl-N6-formyl-N6-hydroxyOrnithine	NAc-Fo-OH-Orn	NAc-Fo-OH-Orn	no
200.24119	N1-methyl-tryptophan	1Me-Trp	Trp or 1Me-Trp	yes (trp), no (1Me-Trp)
202.21402	phototryptophan	pTrp	pTrp	no
202.21402	5-hydroxytryptophan	OH-Trp	Trp or OH-Trp	yes (trp), no (OH-Trp)
204.27295	N-methyl-4-dimethylamino-phenylalanine	NMe-Me2A-Phe	Me2A-Phe	no
205.21467	vinylogous hydroxy tyrosine	v-OH-Tyr	v-OH-Tyr	no
205.25772	alpha-amino-methoxyphenyl-valeric acid	Amv	Amv	no
208.17554	3-nitrotyrosine	3NO2-Tyr	Tyr	yes (tyr)
210.08665	D-PhosphateAsparagine	D-PO-Asn	PO-Asn or OH-Asn	no (PO-Asn), yes (hasn)
210.25761	propenoyl-O-methylserinylthiazole acid	PMST	MeO-Ser-v-Cys	no
211.64961	D-3-chloro-N-methyl-Tyrosine	D-Cl-NMe-Tyr	Tyr or Cl-Tyr	yes (tyr), no (Cl-Tyr)
211.64961	Chloro-N-methyl-tyrosine	Cl-NMe-Tyr	Tyr or Cl-Tyr	yes (tyr), no (Cl-Tyr)
213.62245	beta-hydroxy-chloro-tyrosine	bOH-Cl-Tyr	bOH-Tyr or bOH-Cl-Tyr	yes (bht), no (bOH-Cl-Tyr)
214.22317	3-amino-6-hydroxy-2-piperidone	Ahp	Ahp, incl. Thr	no (Ahp), yes (thr)
216.23904	2,6-diamino-7-hydroxyazelaic acid	Daz	Daz	no
216.2406	N-methyl-5-hydroxytryptophan	NMe-OH-Trp	Trp or OH-Trp	yes (trp), no (OH-Trp)
216.2406	methoxytryptophan	OMe-Trp	Trp or OMe-Trp	yes (trp), yes (s-nmethoxy-trp)
216.24237	4-amino-7-guanidino-2,3-dihydroxyheptanoic aci	Agdha	Arg or Agdha	yes (Arg), no (Agdha)
218.04154	3,5-dichloro-4-hydroxyphenylglycine	Cl2-Hpg	Hpg	yes (hpg)
219.1982	DHP-methyloxazolinyl group	DMOG	diOH-Bz-Thr	yes (dhb, thr)
219.24124	N-methoxyacetyl-D-phenylalanine	NOMe-Ac-D-Phe	Phe	yes (phe)
220.65968	D-6'-chloro-tryptophan	D-Cl-Trp	Trp or Cl-Trp	yes (trp), no (Cl-Trp)
221.2372	dihydroxyphenylthiazol group	DHPT	diOH-Bz-Cys	yes (dhb, cys)
226.07464	bromophenylalanine	Br-Phe	Phe or Br-Phe	yes (phe), no (Br-Phe)
228.2513	N-acetyltryptophan	Ac-Trp	Trp	yes (trp)
228.29437	beta,beta,N-trimethyltryptophan	bbMe-NMe-Trp	Trp or bbMe-Trp	yes (trp), no (bbMe-Trp)
228.50825	tri-chloro-N-methyl-dehydroLeucine	Cl3-NMe-dhLeu	Leu or dhLeu	yes (leu), no (dhLeu)
230.22412	D-2-carboxy-tryptophan	D-COOH-Trp	Trp or COOH-Trp	yes (trp), no (COOH-Trp)
230.29031	thiazolylphenylalanine	Phe-Thz	Phe-Cys	yes (phe, cys)
230.52412	tri-chloro-N-methyl-Leucine	Cl3-NMe-Leu	Leu	yes (leu)
234.68626	N-methyl-6-chloro-tryptophan	NMe-Cl-Trp	Trp or Cl-Trp	yes (trp), no (Cl-Trp)
240.10124	beta-methyl-bromophenylalanine	bMe-Br-Phe	bMe-Br-Phe	no
242.07404	beta-hydroxy-bromophenylalanine	bOH-Br-Phe	Phe or bOH-Br-Phe	yes (phe), no (bOH-Phe)
242.07404	bromotyrosine	Br-Tyr	Tyr or Br-Tyr	yes (tyr), no (Br-Tyr)
242.32094	beta,beta,N1,N-tetramethyltryptophan	bbNMe-NMe-Trp	Trp or bbNMe-Trp	yes (trp), no (bbNMe-Trp)
246.5235	tri-chloro-2-hydroxy-N-methyl-Leucine	Cl3-2OH-NMe-Leu	Leu or 2OH-Leu	yes (leu), no (2OH-Leu)
246.5235	tri-chloro-5-hydroxy-N-methyl-Leucine	Cl3-5OH-NMe-Leu	Leu or 5OH-Leu	yes (leu), no 5OH-Leu
250.68567	N-methyl-6-chloro-5-hydroxytryptophan	NMe-Cl-OH-Trp	Trp or Cl-OH-Trp	yes (trp), no (Cl-OH-Trp)
256.10062	D-3-bromo-N-methyl-Tyrosine	D-Br-NMe-Tyr	Tyr or Br-Tyr	yes (tyr), no (Br-Tyr)
258.27731	N1-carboxy-bichomotryptophan	N1-COOH-bhTrp	N1-COOH-bhTrp or bhTrp	no
259.26534	pyoverdin chromophore	ChrP	Tyr-Dab	yes (tyr), yes (dab)
259.26534	isopyoverdin chromophore	ChrI	Tyr-Dab	yes (tyr), yes (dab)
261.28122	5,6-dihydropyoverdin chromophore	ChrD	Tyr-Dab	yes (tyr), yes (dab)
263.68442	D-6-chloro-N2-formamidotryptophan	D-Cl-CONH2-Trp	Trp or Cl-CONH2-Trp	yes (trp), no (Cl-CONH2-Trp)
265.11069	5-bromo-tryptophan	Br-Trp	Trp or Br-Trp	yes (trp), no (Br-Trp)
266.38054	8,10-Dimethyl-9-hydroxy-7-methoxytridecadienc	DHMDA	DHMDA	no
272.32704	4-propenoyl-2-tyrosylthiazole acid	PTTA	Tyr-v-Cys	yes (tyr), no (v-Cys)
279.13727	N-methyl-2-Bromo-tryptophan	NMe-Br-Trp	Trp or Br-Trp	yes (trp), no (Br-Trp)
281.11008	2-bromo-5-hydroxytryptophan	Br-OH-Trp	Trp or Br-OH-Trp	yes (trp), no (Br-OH-Trp)
285.25958	azotobactins chromophore	ChrA	Tyr-Dab	yes (tyr), yes (dab)
303.10107	D-3-iodo-N-methyl-Tyrosine	D-I-NMe-Tyr	Tyr	yes (tyr)
310.26572	actinomycin chromophore	ChrAct	4-Methyl-3-hydroxy-anthranilic acid	no

Chapter 2, in full, is a reprint of the material as it appears in 'A mass spectrometry guided genome mining approach for natural product peptidogenomics', Kersten, R.D., Yang, Y.L., Xu, Y., Cimermancic, P., Nam, S.J., Fenical, W., Fischbach, M.A., Moore, B.S., Dorrestein, P.C. *Nature Chemical Biology*, 2011, 7, 794-802. The dissertation author was the primary investigator and author of this paper.

R.D.K. designed and carried out experiments, analyzed data and wrote the paper. Y.-L.Y., Y.X. and S.-J.N. carried out experiments and analyzed data. P.C. and M.A.F. carried out the bioinformatic analysis and analyzed data. W.F. analyzed data. B.S.M. and P.C.D. designed experiments, analyzed data and wrote the paper.

Chapter 3 - Bacterial Biosynthesis and Maturation of the Didemnin Anti-cancer Agents

JACS
JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

Article

pubs.acs.org/JACS

Bacterial Biosynthesis and Maturation of the Didemnin Anti-cancer Agents

Ying Xu,[†,#] Roland D. Kersten,[‡,#] Sang-Jip Nam,[‡,¶] Liang Lu,[†] Abdulaziz M. Al-Suwailem,[§] Huajun Zheng,[∥] William Fenical,[‡] Pieter C. Dorrestein,[‡,⊥] Bradley S. Moore,*[,‡,⊥] and Pei-Yuan Qian*[,†]

[†]KAUST Global Collaborative Research, Division of Life Science, School of Science, Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong, China

[‡]Center for Marine Biotechnology and Biomedicine, Scripps Institution of Oceanography, University of California at San Diego, La Jolla, California 92093, United States

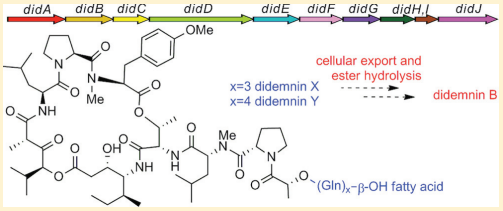
[§]The Coastal and Marine Resources Core Lab, Red Sea Research Center, 4700 King Abdullah University of Science and Technology, Thuwal, Makkah 23955-6900, Kingdom of Saudi Arabia

[∥]Shanghai-MOST Key Laboratory of Health and Disease Genomics, Chinese National Human Genome Center at Shanghai, 250 Bi Bo Road, Shanghai 201203, China

[⊥]Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California at San Diego, La Jolla, California 92093, United States

Supporting Information

ABSTRACT: The anti-neoplastic agent didemnin B from the Caribbean tunicate *Tridemnum solidum* was the first marine drug to be clinically tested in humans. Because of its limited supply and its complex cyclic depsipeptide structure, considerable challenges were encountered during didemnin B's development that continue to limit aplidine (dehydrodidemnin B), which is currently being evaluated in numerous clinical trials. Herein we show that the didemnins are bacterial products produced by the marine α-proteobacteria *Tistrella mobilis* and *Tistrella bauzanensis* via a unique post-assembly line maturation process. Complete genome sequence analysis of the 6,513,401 bp *T. mobilis* strain KA081020-065 with its five circular replicons revealed the putative didemnin biosynthetic gene cluster (*did*) on the 1,126,962 bp megaplasmid pTM3. The *did* locus encodes a 13-module hybrid non-ribosomal peptide synthetase−polyketide synthase enzyme complex organized in a collinear arrangement for the synthesis of the fatty acylglutamine ester derivatives didemnins X and Y rather than didemnin B as first anticipated. Imaging mass spectrometry of *T. mobilis* bacterial colonies captured the time-dependent extracellular conversion of the didemnin X and Y precursors to didemnin B, in support of an unusual post-synthetase activation mechanism. Significantly, the discovery of the didemnin biosynthetic gene cluster may provide a long-term solution to the supply problem that presently hinders this group of marine natural products and pave the way for the genetic engineering of new didemnin congeners.



INTRODUCTION

Sessile marine invertebrates such as sponges and tunicates are intimately linked with microbes that are consumed as food and hosted for their protective chemistry. These marine animals are rich sources of structurally diverse collections of unique bioactive molecules that in some cases have been developed into drugs for the benefit of human health.[1,2] A major challenge in the drug development of marine natural products, however, has historically been the difficulty in supplying sufficient material for pre-clinical and clinical evaluation due to natural limitations of the source organism.[3] Marine products are frequently endowed with complex organic structures that complicate their total syntheses in large scale to meet the demands of a drug development program.[4] Microbes are commonly suggested as the actual producers of many molecules originally isolated from invertebrate tissues, and thus microbial fermentation offers the promise of a renewable supply of important marine drug candidates.[5] However, very few examples exist in which clinically relevant compounds originally extracted from marine invertebrates are produced by cultured microbes. Herein we show that the didemnin family of anti-cancer agents, including didemnin B as the first marine natural product clinically tested in humans,[6] is produced by marine bacteria of the genus *Tistrella* through an unexpected biosynthetic pathway.

Invertebrate-derived natural products, especially those with complex polyketide and non-ribosomal peptide structures, are generally proposed to be microbial in origin since the molecular basis governing their biosyntheses resides exclusively in

Received: February 22, 2012
Published: March 29, 2012

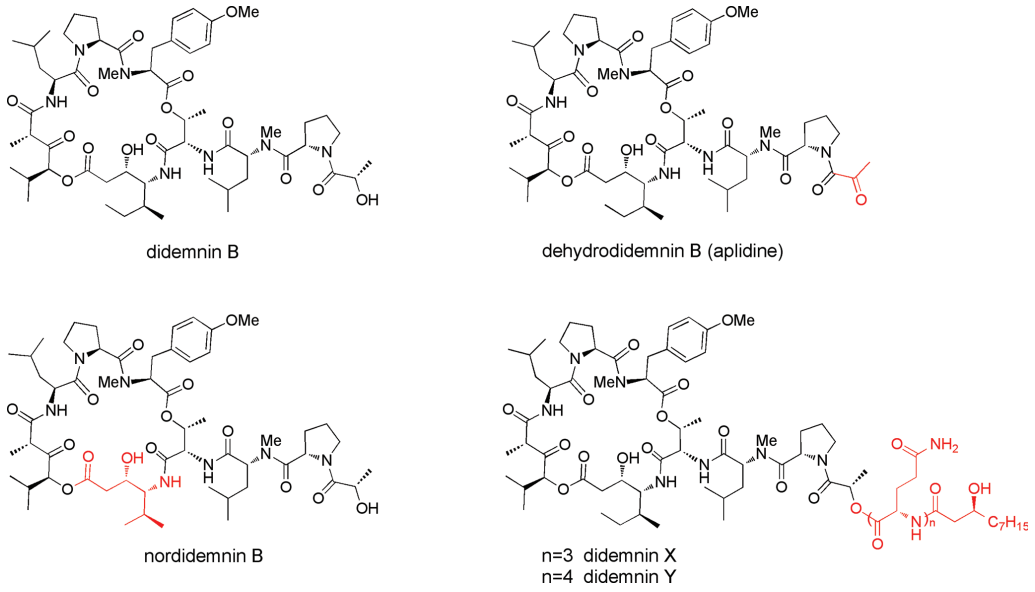


Figure 1. Structures of didemnin B and other representative didemnin analogues in which structural differences are colored red.

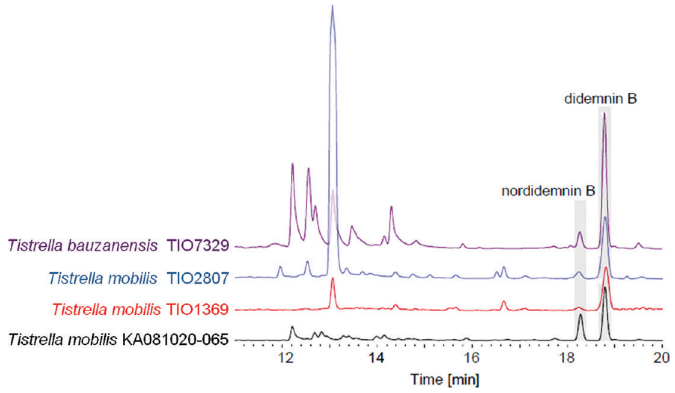


Figure 2. HPLC profiles of didemnin B and nordidemnin B in extracts of four different *Tistrella* strains.

microbes.[7−9] Metagenomic studies have revealed the microbial production of noteworthy bioactive natural products from diverse marine invertebrates, including the pederin group of polyketides from sponges,[10] the patellamides and ecteinascidin-743 (Yondelis) from tunicates,[11,12] and the bryostatin macrolides from bryozoans.[13] In each case, uncultivated bacterial symbionts are thought to be the natural producers. Free-living marine microbes are often not implicated in the synthesis of marine invertebrate natural products, although a field-collected marine cyanobacterium was shown to harbor the sponge macrolide swinholide A.[14] As a consequence, molecular engineering approaches involving whole pathway expression are envisaged to produce complex molecules from marine invertebrate symbionts in laboratory microbial strains in order to develop a renewable supply of the compound and allow for the rational engineering of new chemical entities to probe structure−function relationships.[15] To date, only the cyanobactin group of ribosomal peptides, which includes the patellamides, has been heterologously produced and engineered.[16] Complex polyketides, non-ribosomal peptides, and hybrids thereof, on the other hand, present an unmet technical challenge due to their large biosynthetic gene clusters that often exceed 100 kb.

Didemnins (Figure 1) are cyclic depsipeptides with remarkable anti-tumor, anti-viral, and immunosuppressive properties[6] that were first discovered in the Caribbean tunicate *Tridemnum solidum* in 1981.[17] The most potent analogue, didemnin B, was the first marine natural product to enter clinical trials as an anti-cancer agent, yet it ultimately failed to demonstrate effective anti-tumor activity while displaying

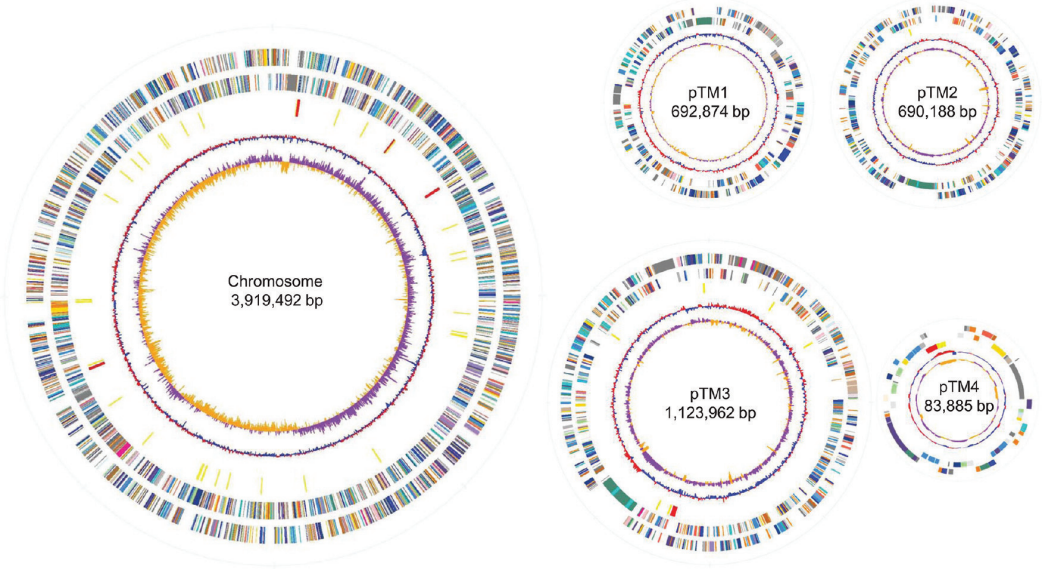


Figure 3. Genome maps of *T. mobilis* KA081020-065. From the outside inward, the first and second circles show predicted coding regions transcribed on the forward and reverse strand, respectively, colored by functional categories according to COG classification. The third circle shows RNAs (yellow for tRNA and red for rRNA). The fourth and fifth circles show G+C content and GC skew (G−C)/(G+C) calculated using a 5 kb window (purple for values >1, orange for values <1), respectively.

cardiac and neuromuscular toxicities.[18,19] A closely related compound, dehydrodidemnin B (aplidine), was isolated from the Mediterranean tunicate *Aplidium albicans* and shown to be more potent and less toxic despite its minor structural change.[20,21] Aplidine has since replaced didemnin B in clinical trials and is currently in multiple phase II and III trials for the treatment of various cancers.[6] Due to their cyclic depsipeptide structures with highly modified amino acid residues, didemnins have long been suspected as microbial products assembled by a hybrid non-ribosomal peptide synthetase−polyketide synthase (NRPS−PKS) enzymatic pathway.[15] In this study, we report that the didemnins are not only the products of the α-proteobacterium *Tistrella mobilis*, which was also recently reported by Tsukimoto and co-workers,[22] but also of *Tistrella bauzanensis*. More importantly, we show through the complete genome sequencing of *T. mobilis* and MALDI-imaging MS technique that didemnin B is biosynthesized by an unanticipated rare mechanism involving a post-assembly activation of lipoglutamine congeners, didemnins X and Y.

RESULTS AND DISCUSSION

Isolation of Didemnins from *Tistrella* Species. We isolated the α-proteobacterium *T. mobilis* KA081020-065 from the Red Sea and measured a very potent cytotoxicity against HeLa cells. Following a bioassay-guided fractionation of the ethyl acetate extract, we isolated two metabolites that, based on high-resolution MS and NMR characterization and comparison with authentic standards, were identical to the cyclic depsipeptides didemnin B (Figure S1) and nordidemnin B (Figure S2) originally isolated from the Caribbean tunicate *Tridemnum solidum*.[17] Since this result showed that the didemnins are bacterial products, we examined additional *Tistrella* strains to explore the extent of didemnin chemistry in this poorly characterized genus. To our surprise, all strains tested, including two Pacific Ocean *T. mobilis* isolates and a *T. bauzanensis* strain, also produced didemnin B and nordidemnin B at varying levels (Figure 2). During the preparation of this manuscript, Tsukimoto and co-workers independently reported their isolation from a *T. mobilis* strain isolated from marine sediments in Japan.[22] Based on these preliminary observations, the didemnins appear to be commonly associated with this recently described bacterial genus whose secondary metabolic potential has not yet been reported.[23] To further explore the natural product biosynthetic capacity of *Tistrella* and to deduce the molecular basis governing didemnin assembly, we sequenced the genome of the Red Sea isolate *T. mobilis* KA081020-065.

Genome Characteristics of *T. moblis* KA081020-065. The complete genome sequence of *T. mobilis* strain KA081020-065 totaling 6,513,401 bp was established by 454 pyrosequencing to reveal five replicons comprising a 3,919,492 bp circular chromosome and four circular plasmids ranging in size from 83,885 bp (pTM4) to 1,126,962 bp (pTM3) (Figure 3). Its genome organization with three megaplasmids greater than 600 kb each is reminiscent of several other α-proteobacteria members such as the rice plant endophyte *Azospirillum* sp. B510 that harbors six plasmids, five of which are in excess of 500 kb.[24] General features of the *T. mobilis* genome, which has a high average G+C content of 68%, are summarized in Table S1.

Based on the chemical structure of the didemnin depsipeptides that contains the two ketide-extended amino acid residues isostatine (Ist) and α-(α-hydroxyisovaleryl)-propionic acid (Hip),[17] we hypothesized their biosynthesis by

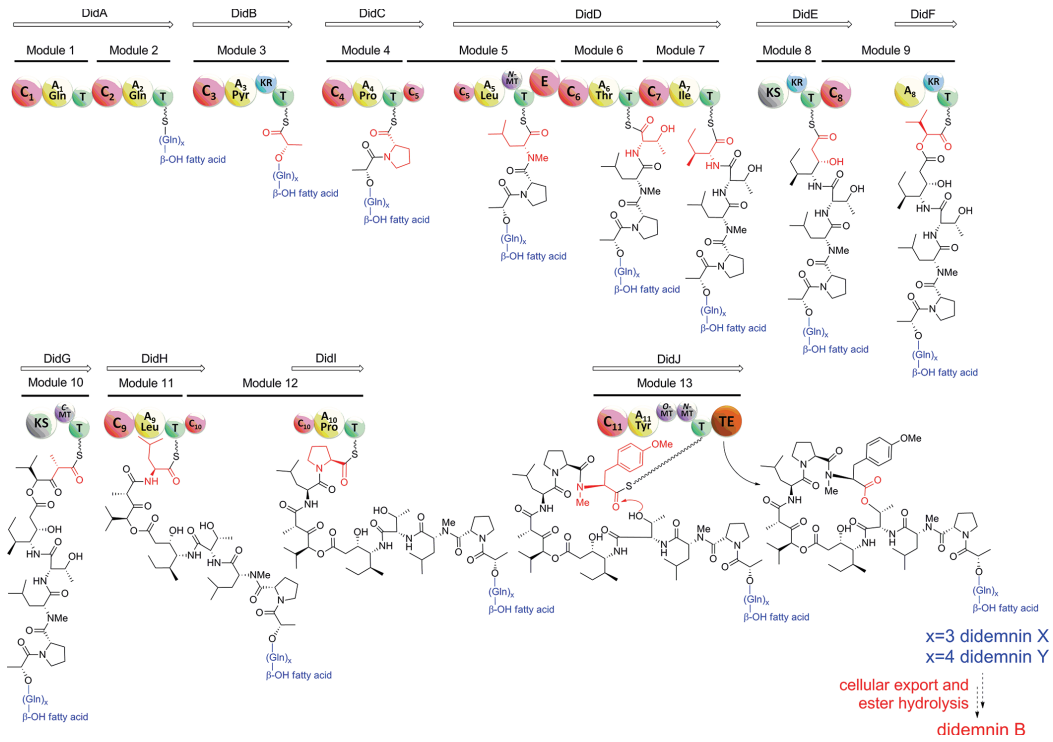


Figure 4. Proposed biosynthetic pathway of didemnins and domain organization of the biosynthetic *did* gene cluster in *T. mobilis* KA081020-065. Domain notation: C, condensation; A, adenylation (substrate included); E, epimerase; KR, ketoreductase; KS, β-ketoacyl synthase; MT, methyltransferase; T, thiolation; TE, thioesterase.

a modular hybrid non-ribosomal peptide synthetase−polyketide synthase (NRPS-PKS) pathway reminiscent of nostopeptolide biosynthesis.[25] Bioinformatic analysis of the *Tistrella* genome revealed just one large NRPS-PKS biosynthetic gene cluster (*did*) located on pTM3 (TMO_c0602−c0611) that was consistent with the proposed biosynthesis of the didemnin family of cyclic depsipeptides. We detected only one other multimodular NRPS-based gene cluster in the genome, a 34,665 bp locus on pTM2 (TMO_b0323−b0325), which rather encodes a putative eight-residue peptide product inconsistent with didemnin assembly (Figure S3).

Bioinformatic Analysis of the Didemnin Biosynthetic Gene Cluster Reveals an Unexpected Post-assembly Processing Mechanism. *In silico* analysis of the putative didemnin biosynthetic gene cluster (*did*) revealed eight NRPS and two PKS-encoding genes spanning from *didA* to *didJ* (Figure S4), comprising 13 modules in total (Figure 4). Based on the enzymatic logic of thiotemplate-mediated assembly line biosynthesis, the projected product of the DidB−J megasynthetase perfectly correlates with the collinear synthesis of didemnin B. The DidB adenylation (A) domain A3 is predicted to load pyruvate and reduce it *in cis* via the adjacent ketoreductase (KR) domain to give lactate bound to the DidB thiolation (T) domain in a reaction reminiscent to that proposed in valinomycin assembly.[26] Further processing by the monomodular DidC and the trimodular DidD NRPSs are next proposed to assemble the Pro-DMeLeu-Thr-Ile tetrapeptide fragment. The *N*-methylated leucine residue is the lone D-amino acid in the didemnin B molecule, which correlates nicely with the epimerase (E) and methyltransferase (MT) domains that reside in the Leu-specific module 5 of DidD. Malonate extension of the DidD product by the PKS DidE, which lacks an integrated acyltransferase (AT) domain reminiscent of *trans*-AT PKSs,[27] yields the ketide-extended β-hydroxy-γ-amino acid residue Ist. Analysis of the genes bordering *didA*−*J* and elsewhere in the *T. mobilis* genome, however, did not reveal a discrete AT from *trans*-AT systems, suggesting that the didemnin polyketide synthase may co-opt the fatty acid synthase AT FabD as reported recently for the biosynthesis of FK228.[28]

Subsequent activation of 2-oxoisovaleric acid by the DidF A domain, whose sequence is similar to the *N*-terminal HetE A domain (43% identity at amino acid level) that is also responsible for its selection in hectochlorin biosynthesis,[29] sets up the incorporation of the α-hydroxy acid 2-hydroxyisovaleric acid and a second round of malonate extension by the DidG PKS. The DidG MT domain putatively incorporates the α-methyl group in the Hip residue. Further chain elongation of the DidG product by the sequential addition of the residues Leu, Pro, and Me_2Tyr via the monomodular DidH−J NRPSs completes the assembly of the linear didemnin B molecule. The *C*-terminal thioesterase (TE)

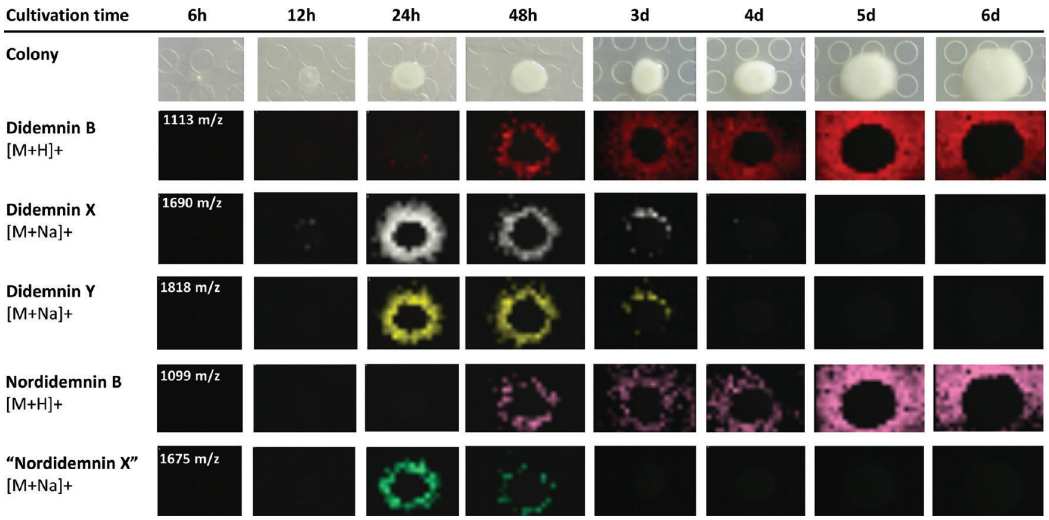


Figure 5. Time-course production of didemnins by *T. mobilis* KA081020-065 detected by MALDI-imaging mass spectrometry.

domain of DidJ is hypothesized to offload didemnin B as the cyclic depsipeptide product. In the case of co-produced nordidemnin B, the DidD A7 domain is anticipated to exhibit substrate flexibility toward aliphatic residues to also activate and incorporate Val. [Hysp]didemnin B[30] is an additionally characterized didemnin B derivative from *T. mobilis* extract (Figure S5) which further indicates substrate promiscuity in DidF domain A9 for α-hydroxy acids 2-hydroxyisovaleric acid (didemnin B) and 2-hydroxy-3-methylvaleric acid ([Hysp]-didemnin B).

While the biosynthetic model described above is clearly consistent with the observed didemnin B and nordidemnin B chemistry of *T. mobilis*, this proposal ignores the DidA bimodular NRPS that should function as the priming synthetase prior to DidB catalysis. Inspection of both DidA A domains revealed their preference for Gln. Furthermore, sequence analysis of the DidA *N*-terminal condensation (C) domain suggested its enzymatic role in the fatty acylation of the first amino acid residue to yield a lipopeptide product (Figure 4) consistent with other important non-ribosomal peptides such as surfactin.[31] Among the natural didemnin congeners previously identified from the tunicate *T. solidum*, only didemnins X and Y have both Gln residues and a lipid component (Figure 1).[32] Interestingly, although these two didemnin derivatives contain Gln residues, they each have more than two residues as predicted by the bimodular DidA if each module were to function just once in peptide elongation. We thus explored the possibility that the immediate product of the didemnin biosynthetic pathway is not (nor)didemnin B, but rather an acylglutamine ester such as (nor)didemnin X or Y in which the ester side chain facilitates the cellular export of the cyclic depsipeptide from the Gram-negative bacterium for extracellular unmasking of the didemnin B byproduct.

Conversion of Didemnins X and Y to Didemnin B Captured by Imaging Mass Spectrometry. Repeated attempts were unsuccessful to isolate additional didemnin congeners by altering the culture medium in order to evaluate the activation hypothesis. We thus turned to MALDI-TOF MS to analyze the bacterial culture extracts, whereupon we successfully detected low-abundance masses consistent with didemnins X and Y in addition to two novel derivatives, nordidemnins X and Y (Figure S6). We confirmed their structures by MS^n analysis and high-resolution Fourier-transform ion cyclotron resonance MS as reported in the Supporting Information (Figure S7). The production of didemnins X and Y not only suggests that they are the fully elaborated biosynthetic product of the didemnin synthetase but also indicates that the DidA bimodular NRPS functions iteratively to install the three to four Gln residues in didemnins X and Y, respectively (Figure 4).

In order to explore the biosynthetic connection between these molecules, we next used MALDI-imaging MS[33,34] to study the temporal and spatial distribution of the didemnin compounds. We interrogated *T. mobilis* colonies over a time course ranging from 6 h to 6 d (Figure 5), which strikingly revealed that didemnins X and Y are produced and excreted from the colony during the early growth phase of the bacterium, peaking at 24 h. Concomitant with their disappearance by day 4 was the emergence of didemnin B as the primary didemnin metabolite that steadily increased over time. This time-lag difference between didemnins X and Y and didemnin B was similarly observed in the nordidemnin series (Figure 5) and is consistent with the proposed model in which the lipoglutamine side chain facilitates the cellular export of the mature didemnin molecule for extracellular processing.

We next set out to determine whether the maturation of didemnins X and Y involved a secreted factor. To this end, we performed a hydrolysis assay of purified didemnin precursors and filtered growth media cultured with and without the bacterium. The assay clearly showed that didemnin X/Y turnover to didemnin B was dependent on the addition of a protein extract from a 1 day *T. mobilis* culture supernatant (Figure S8). While studies are ongoing to identify the secreted hydrolytic factor involved in the peptide activation mechanism, sequence analysis of the open reading frames adjacent to the *didA–J* operon revealed a number of genes encoding

membrane-associated transport proteins (*orf3*, *orf6*) and hydrolytic enzymes (*orf2*, *orf14*) (Figure S4), further supporting the export-hydrolysis model. Recently an activation mechanism was established in the non-ribosomal peptide synthesis of the antibiotic xenocoumacin from the γ-proteobacterium *Xenorhabdus nematophila* in which a precursor molecule is processed by a membrane-bound and D-Asn-specific peptidase.[35] Although the characterized pathway in xenocoumacin processing parallels the proposed pathway for didemnin maturation, the *did* cluster does not harbor a homologue of the xenocoumacin ABC transporter-trans-peptidase XcnG. In addition, the maturation of xenocoumacin takes place in the periplasmic space rather than the extracellular space for didemnins, which may suggest an alternative mechanism in *T. mobilis*.

CONCLUSION

The discovery of the bacterial synthesis of the didemnins raises the question about their origin in the Caribbean tunicate *T. solidum* where they were first characterized over 30 years ago by Rinehart and co-workers.[17] Based on the findings reported here, we suspect that the tunicate-derived didemnins are similarly bacterial in origin. A recent metagenomic profile of the tropical Pacific tunicate *Lissoclinum patella* revealed that the distinctive patellamide ribosomal peptides and the anti-tumor polyketide patellazole are biosynthetically linked to different bacteria within the tunicate microbiome.[36] Similarly, the approved drug Yondelis (ET-743) from the tropical tunicate *Ecteinascidia turbinata* was recently shown to be bacterial in origin.[12] With the *T. mobilis* genomic information now in hand, molecular methods could be used to examine didemnin-associated tunicates for the presence of *Tistrella* bacteria or other microbes that carry the *did* genes. The finding that the didemnin biosynthetic gene cluster resides on a plasmid is also intriguing since other bacteria may acquire the metabolic pathway by horizontal gene transfer. The *did*-containing megaplasmid pTM3 contains several essential elements required for conjugation, including a relaxase (TMO_c0026) needed for DNA transfer. The bacterial chromosome additionally contains a *trb* operon which encodes the components of a bacterial mating apparatus, (TMO_2187−2198), another important component in Gram-negative bacterial conjugation.[37]

With a bacterial production of the didemnins now in place, the supply of these structurally complex and potent anti-cancer agents may no longer be limiting. While clinical trials with didemnin B have been suspended, phase II trials with aplidine (dehydrodidemnin B; plitidepsin) from the Mediterranean tunicate *Aplidium albicans* are actively underway.[38] Presently, multi-step total synthesis is employed to produce aplidine for the clinic due to its limited natural supply.[6,39] Biotechnology now may offer an alternative solution. The *T. mobilis* DidB tetradomain gene product is putatively responsible for the selection and attachment of pyruvate, its ketoreduction to lactate, and its ensuing condensation with the DidA fatty acylglutamine residue (Figure 4). Gene inactivation of the *didB* KR domain would maintain the oxidation state of the pyruvate unit and thereby prevent the esterification reaction, effectively switching the didemnin B biosynthetic pathway to directly produce aplidine. While seemingly straightforward once a genetics system is established in *Tistrella*, production levels may be significantly impacted since the activation and export mechanism that governs didemnin B synthesis from the lipopeptides didemnins X and Y may no longer be effective in such a mutant for aplidine. We are actively pursuing answers to the detailed mechanism of didemnin B biosynthesis and whether we can re-engineer *T. mobilis* to directly produce aplidine. A provisional patent has also been filed on the didemnin biosynthetic gene cluster (U.S. provisional application number 61/537,416).

EXPERIMENTAL SECTION

Strain, Fermentation, and Isolation of Didemnin Compounds. *Tistrella mobilis* KA081020-065 was isolated from seawater collected from the Red Sea during a 2009 research cruise. Its crude extract showed remarkable cytotoxicity on HeLa cells in a bioactive compound screening of Red Sea bacteria. *T. mobilis* KA081020-065 was grown in GYP medium [10 g glucose, 4 g yeast extract, 2 g peptone, and 17 g sea salts per liter of deionized water] using 50-L stirred fermenters at 25 °C. After 72 h, ethyl acetate was added to the culture to extract the metabolites. The crude extract was fractionated by reversed-phase (RP) C18 liquid chromatography and eluted with increasing amounts of methanol in water. The active fraction was further purified by semi-preparative RP HPLC with 63% MeCN in water to give didemnin B and nordidemnin B at 0.2 and 0.1 mg L^{-1}, respectively. For MS^n and FTMS analysis, crude extracts were separated by RP-HPLC (10−100% MeCN in 0.1% trifluoroacetic acid (TFA), 40 min gradient). For hydrolysis assay, didemnin X and Y were purified from crude extract by Sephadex LH20 (GE Healthcare) gel filtration with MeOH as a mobile phase and RP-HPLC (25−75% MeCN in 0.1% TFA, 60 min gradient).

MS Sample Preparation and MS Analysis of Didemnins. For MALDI-TOF MS analysis, extract or peptide samples were dissolved in MeOH, mixed 1:1 with saturated universal MALDI matrix, and spotted on a Bruker MSP96 anchor plate. MALDI-TOF MS analysis was performed on a Microflex Bruker Daltonics mass spectrometer outfitted with Compass 1.3 software suite. For MS^n and FTMS analysis, HPLC-purified samples were redissolved in 50 μL of MeOH/water/formic acid (50:49:1) and injected by a nanomate-electrospray ionization robot (Advion) for consecutive electrospray into the MS inlet of a LTQ 6.4T Fourier transform ion cyclotron resonance (FT-ICR) mass spectrometer (Thermo Finnigan). MS and MS^n data were acquired in the positive ion mode. FTMS data were acquired in 400−2000 m/z scans. Selected peptide mass signals were manually isolated and fragmented by CID. MS^n data were collected either in ion trap or FT detection mode. All data were analyzed using QualBrowser, which is part of the Xcalibur LTQ-FT software package (Thermo Fisher).

Genome Sequencing, Annotation, and Analysis. We used a massively parallel pyrosequencing technology (Roche 454 GS FLX)[40] to sequence the *T. mobilis* KA081020-065 genome. We obtained 315,496 reads with an average length of 334 bp, which provided a 16.2-fold coverage of the 6.4 Mb genome. From these reads, we assembled 112 contigs (>500 bp) using the Newbler software of the 454 suite package. To enhance sequence quality and to construct scaffolds, we sequenced a paired-end Illumina library of an additional 2,625,640 bp with average length of 115 bp, which were then mapped to the genome sequence. All contig relationships within scaffolds were then validated by PCR, while the relationships among scaffolds were determined by multiplex PCR.[41] Gaps were filled by sequencing PCR products. The final sequence assembly was carried out using the phred/phrap/consed package (http://www.genome.washington.edu) in which we resequenced the low sequence quality regions. The final error rate of the genome sequence was 0.28 per 100,000 bases. The sequences of the five replicons were deposited in the GenBank database as accession numbers CP003236−CP003240.

Putative protein-coding sequences (CDS) were determined by combining the prediction results of the Glimmer 3.02[42] and Z-Curve programs.[43] Functional annotations of CDS were performed by searching the KEGG and the NCBI non-redundant protein databases. tRNA genes were predicted with tRNAScan-SE (v1.23).[44] Protein domain predictions and COG assignments[45] were performed by RPS-BLAST using the NCBI CDD library.[46]

***In Silico* Analysis of the Didemnin Biosynthetic Gene Cluster and Other NRPSs in *T. mobilis* KA081020-065.** The putative roles of the proteins in the didemnin gene cluster were assigned using protein−protein BLAST and Pfam analyses. NRPS A domain specificities were predicted using the online program NRPSpredictor2 and antiSMASH.[47−49]

Time-Course Study of Didemnin Production in *T. mobilis* by MALDI-Imaging MS. One microliter of fresh overnight *T. mobilis* culture was spotted onto 1 mm thick GYP agar plates and allowed to grow for 6 h to 6 d. Colonies with different growing times were transferred to Bruker MSP 96 anchor plates and analyzed concurrently using the IMS technique as previously described.[33,50]

Didemnin Precursor Hydrolysis Assay. Samples of 50 μg each of HPLC-purified didemnin X and Y were dissolved in 5 μL of water, mixed with (a) 45 μL of sterile GYP medium (negative control) or (b) 45 μL of protein filtrate (10 kDa cutoff, washed 2× with GYP medium) of 1 mL of a 1 d *T. mobilis* KA081020-065 liquid culture (28 °C), and incubated for 24 h at 25 °C. Hydrolysis samples were taken and spotted on a MALDI-TOF MS plate for subsequent MS analysis at time points 5 min, 1 h, 12 h, and 24 h.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

^{1}H NMR spectra, annotated MS^n spectra of the didemnins, and hydrolysis experiment of didemnin X/Y. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

boqianpy@ust.hk; bsmoore@ucsd.edu

Present Address

[¶]College of Pharmacy, Sunchon National University, Suncheon 540-950, Republic of Korea.

Author Contributions

[#]Y.X. and R.D.K. contributed equally to this work.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We thank Z. Shao for providing the Pacific Ocean *Tistrella* strains and N. Ziemert, J. Yang and P. Lai for assistance with data analysis. Financial support was provided by King Abdullah University of Science and Technology (SA-C0040/UK-C0016 to P.Y.Q.), China Ocean Mineral Resources Research and Development Association (DY125-15-T-02 to P.Y.Q.), and the National Institutes of Health (GM085770 to B.S.M., CA044848 to W.F., GM086283 and S10RR029121 to P.C.D.).

■ REFERENCES

(1) Molinski, T. F.; Dalisay, D. S.; Lievens, S. L.; Saludes, J. P. *Nat. Rev. Drug. Discov.* **2009**, *8*, 69.

(2) Gerwick, W. H.; Moore, B. S. *Chem. Biol.* **2012**, *19*, 85.

(3) Radjasa, O. K.; Vaske, Y. M.; Navarro, G.; Vervoort, H. C.; Tenney, K; Linington, R. G.; Crews, P. *Bioorg. Med. Chem.* **2011**, *19*, 6658.

(4) Florence, G. J.; Gardner, N. M.; Paterson, I. *Nat. Prod. Rep.* **2008**, *25*, 342.

(5) Piel, J. *Curr. Med. Chem.* **2006**, *13*, 39.

(6) Lee, J.; Currano, J. N.; Carroll, P. J.; Joullié, M. M. *Nat. Prod. Rep.* **2012**, *29*, 404.

(7) Faulkner, D. J. *Nat. Prod. Rep.* **2000**, *17*, 1.

(8) Piel, J.; Hui, D.; Wen, G.; Butzke, D.; Platzer, M.; Fusetani, N.; Matsunaga, S. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 16222.

(9) Lane, A. L.; Moore, B. S. *Nat. Prod. Rep.* **2011**, *28*, 411.

(10) Fisch, K. M.; Gurgui, C.; Heycke, N.; van der Sar, S. A.; Anderson, S. A.; Webb, V. L.; Taudien, S.; Platzer, M.; Rubio, B. K.; Robinson, S. J.; Crews, P.; Piel, J. *Nat. Chem. Biol.* **2009**, *5*, 494.

(11) Donia, M. S.; Hathaway, B. J.; Sudek, S.; Haygood, M. G.; Rosovitz, M. J.; Ravel, J.; Schmidt, E. W. *Nat. Chem. Biol.* **2006**, *2*, 729.

(12) Rath, C. M.; Janto, B.; Earl, J.; Ahmed, A.; Hu, F. Z.; Hiller, L.; Dahlgren, M.; Kreft, R.; Yu, F.; Wolff, J. J.; Kweon, H. K.; Christiansen, M. A.; Håkansson, K.; Williams, R. M.; Ehrlich, G. D.; Sherman, D. H. *ACS Chem. Biol.* **2011**, *6*, 1244.

(13) Sudek, S.; Lopanik, N. B.; Waggoner, L. E.; Hildebrand, M.; Anderson, C.; Liu, H.; Patel, A.; Sherman, D. H.; Haygood, M. G. *J. Nat. Prod.* **2007**, *70*, 67.

(14) Andrianasolo, E. H.; Gross, H.; Goeger, D.; Musafija-Girt, M.; McPhail, K.; Leal, R. M.; Mooberry, S. L.; Gerwick, W. H. *Org. Lett.* **2005**, *7*, 1375.

(15) Salomon, C. E.; Magarvey, N. A.; Sherman, D. H. *Nat. Prod. Rep.* **2004**, *21*, 105.

(16) Donia, M. S.; Ravel, J.; Schmidt, E. W. *Nat. Chem. Biol.* **2008**, *4*, 341.

(17) Rinehart, K. L., Jr.; Gloer, J. B.; Hughes, R. G., Jr.; Renis, H. E.; McGovren, J. P.; Swynenberg, E. B.; Stringfellow, D. A.; Kuentzel, S. L. *Science* **1981**, *212*, 933.

(18) Shin, D. M.; Holoye, P. Y.; Murphy, W. K.; Forman, A.; Papasozomenos, S. C.; Hong, W. K.; Raber, M. *Cancer Chemother. Pharmacol.* **1991**, *29*, 145.

(19) Williamson, S. K.; Wolf, M. K.; Eisenberger, M. A.; O'Rourke, M.; Brannon, W.; Crawford, E. D. *Invest. New Drugs* **1995**, *13*, 167.

(20) Ding, X.; Vera, M. D.; Liang, B.; Zhao, Y.; Leonard, M. S.; Joullié, M. M. *Bioorg. Med. Chem. Lett.* **2001**, *11*, 231.

(21) Le Tourneau, C.; Raymond, E.; Faivre, S. *Curr. Pharm. Des.* **2007**, *13*, 3427.

(22) Tsukimoto, M.; Nagaoka, M.; Shishido, Y.; Fujimoto, J.; Nishisaka, F.; Matsumoto, S.; Harunari, E.; Imada, C.; Matsuzaki, T. *J. Nat. Prod.* **2011**, *74*, 2329.

(23) Shi, B. H.; Arunpairojana, V.; Palakawong, S.; Yokota, A. *J. Gen. Appl. Microbiol.* **2002**, *48*, 335.

(24) Kaneko, T.; Minamisawa, K.; Isawa, T.; Nakatsukasa, H.; Mitsui, H.; Kawaharada, Y.; Nakamura, Y.; Watanabe, A.; Kawashima, K.; Ono, A.; Shimizu, Y.; Takahashi, C.; Minami, C.; Fujishiro, T.; Kohara, M.; Katoh, M.; Nakazaki, N.; Nakayama, S.; Yamada, M.; Tabata, S.; Sato, S. *DNA Res.* **2010**, *17*, 37.

(25) Hoffmann, D.; Hevel, J. M.; Moore, R. E.; Moore, B. S. *Gene* **2003**, *311*, 171.

(26) Cheng, Y. Q. *ChemBioChem* **2006**, *7*, 471.

(27) Piel, J. *Nat. Prod. Rep.* **2010**, *27*, 996.

(28) Wesener, S. R.; Potharla, V. Y.; Cheng, Y. Q. *Appl. Environ. Microbiol.* **2011**, *77*, 1501.

(29) Ramaswamy, A. V.; Sorrels, C. M.; Gerwick, W. H. *J. Nat. Prod.* **2007**, *70*, 1977.

(30) Banaigs, B.; Mansour, E. A.; Bonnard, I.; Boulanger, A.; Francisco, C. *Tetrahedron* **1999**, *55*, 9559.

(31) Rausch, C.; Hoof, I.; Weber, T.; Wohlleben, W.; Huson, D. H. *BMC Evol. Biol.* **2007**, *16*, 78.

(32) Sakai, R.; Stroh, J. G.; Sullins, D. W.; Rinehart, K. L. *J. Am. Chem. Soc.* **1995**, *117*, 3734.

(33) Yang, Y. L.; Xu, Y.; Straight, P.; Dorrestein, P. C. *Nat. Chem. Biol.* **2009**, *5*, 885.

(34) Yang, Y. L.; Xu, Y.; Kersten, R. D.; Liu, W. T.; Meehan, M.; Moore, B. S.; Bandeira, N.; Dorrestein, P. C. *Angew. Chem., Int. Ed.* **2011**, *50*, 5839.

(35) Reimer, D.; Pos, K. M.; Thines, M.; Grün, P.; Bode, H. B. *Nat. Chem. Biol.* **2011**, *7*, 888.

(36) Donia, M. S.; Fricke, W. F.; Partensky, F.; Cox, J.; Elshahawi, S. I.; White, J. R.; Phillippy, A. M.; Schatz, M. C.; Piel, J.; Haygood, M. G.; Ravel, J.; Schmidt, E. W. *Proc. Natl. Acad. Sci. U.S.A.* **2011**, *108*, E1423.

(37) Zatyka, M.; Thomas, C. M. *FEMS Microbiol. Rev.* **1998**, *21*, 291.

(38) Muñoz-Alonso, M. J. *Curr. Opin. Investig. Drugs* **2009**, *10*, 536.

(39) Jou, G.; González, I.; Albericio, F.; Lloyd-Williams, P.; Giralt, E. *J. Org. Chem.* **1997**, *62*, 354.
(40) Margulies, M.; Egholm, M.; Altman, W. E.; Attiya, S.; Bader, J. S.; Bemben, L. A.; Berka, J.; Braverman, M. S.; Chen, Y. J.; Chen, Z.; Dewell, S. B.; Du, L.; Fierro, J. M.; Gomes, X. V.; Godwin, B. C.; He, W.; Helgesen, S.; Ho, C. H.; Irzyk, G. P.; Jando, S. C.; Alenquer, M. L.; Jarvie, T. P.; Jirage, K. B.; Kim, J. B.; Knight, J. R.; Lanza, J. R.; Leamon, J. H.; Lefkowitz, S. M.; Lei, M.; Li, J.; Lohman, K. L.; Lu, H.; Makhijani, V. B.; McDade, K. E.; McKenna, M. P.; Myers, E. W.; Nickerson, E.; Nobile, J. R.; Plant, R.; Puc, B. P.; Ronan, M. T.; Roth, G. T.; Sarkis, G. J.; Simons, J. F.; Simpson, J. W.; Srinivasan, M.; Tartaro, K. R.; Tomasz, A.; Vogt, K. A.; Volkmer, G. A.; Wang, S. H.; Wang, Y.; Weiner, M. P.; Yu, P.; Begley, R. F.; Rothberg, J. M. *Nature* **2005**, *437*, 376.
(41) Tettelin, H.; Radune, D.; Kasif, S.; Khouri, H.; Salzberg, S. L. *Genomics* **1999**, *62*, 500.
(42) Delcher, A. L.; Harmon, D.; Kasif, S.; White, O.; Salzberg, S. L. *Nucleic Acids Res.* **1999**, *27*, 4636.
(43) Guo, F. B.; Ou, H. Y.; Zhang, C. T. *Nucleic Acids Res.* **2003**, *31*, 1780.
(44) Lowe, T. M.; Eddy, S. R. *Nucleic Acids Res.* **1997**, *25*, 955.
(45) Tatusov, R. L.; Galperin, M. Y.; Natale, D. A.; Koonin, E. V. *Nucleic Acids Res.* **2000**, *28*, 33.
(46) Marchler-Bauer, A.; Anderson, J. B.; Derbyshire, M. K.; DeWeese-Scott, C.; Gonzales, N. R.; Gwadz, M.; Hao, L.; He, S.; Hurwitz, D. I.; Jackson, J. D.; Ke, Z.; Krylov, D.; Lanczycki, C. J.; Liebert, C. A.; Liu, C.; Lu, F.; Lu, S.; Marchler, G. H.; Mullokandov, M.; Song, J. S.; Thanki, N.; Yamashita, R. A.; Yin, J. J.; Zhang, D.; Bryant, S. H. *Nucleic Acids Res.* **2007**, *35*, D237.
(47) Röttig, M.; Medema, M. H.; Blin, K.; Weber, T.; Rausch, C.; Kohlbacher, O. *Nucleic Acids Res.* **2011**, *39*, W362.
(48) Rausch, C.; Weber, T.; Kohlbacher, O.; Wohlleben, W.; Huson, D. H. *Nucleic Acids Res.* **2005**, *33*, 5799.
(49) Medema, M. H.; Blin, K.; Cimermancic, P.; de Jager, V.; Zakrzewski, P.; Fischbach, M. A.; Weber, T.; Takano, E.; Breitling, R. *Nucleic Acids Res.* **2011**, *39*, W339.
(50) Kersten, R. D.; Yang, Y. L.; Xu, Y.; Cimermancic, P.; Nam, S. J.; Fenical, W.; Fischbach, M. A.; Moore, B. S.; Dorrestein, P. C. *Nat. Chem. Biol.* **2011**, *7*, 794.

Supporting Information

Bacterial biosynthesis and maturation of the didemnin anticancer agents

Ying Xu[†], Roland D. Kersten[‡], Sang-Jip Nam[‡], Liang Lu[†], Abdulaziz M. Al-Suwailem[§], Huajun Zheng[‖], William Fenical[‡], Pieter C. Dorrestein[‡,⊥], Bradley S. Moore[‡,⊥,*], Pei-Yuan Qian[†,*]

[†]KAUST Global Collaborative Research, Division of Life Science, School of Science, Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong, China;

[‡]Center for Marine Biotechnology and Biomedicine, Scripps Institution of Oceanography, University of California San Diego, La Jolla, California 92093, USA;

[§]The Coastal and Marine Resources Core Lab, Red Sea Research Center, 4700 King Abdullah University of Science and Technology, Thuwal, Makkah 23955-6900, Kingdom of Saudi Arabia;

[‖]Shanghai-MOST Key Laboratory of Health and Disease Genomics, Chinese National Human Genome Center at Shanghai, 250 Bi Bo Road, Shanghai 201203, China;

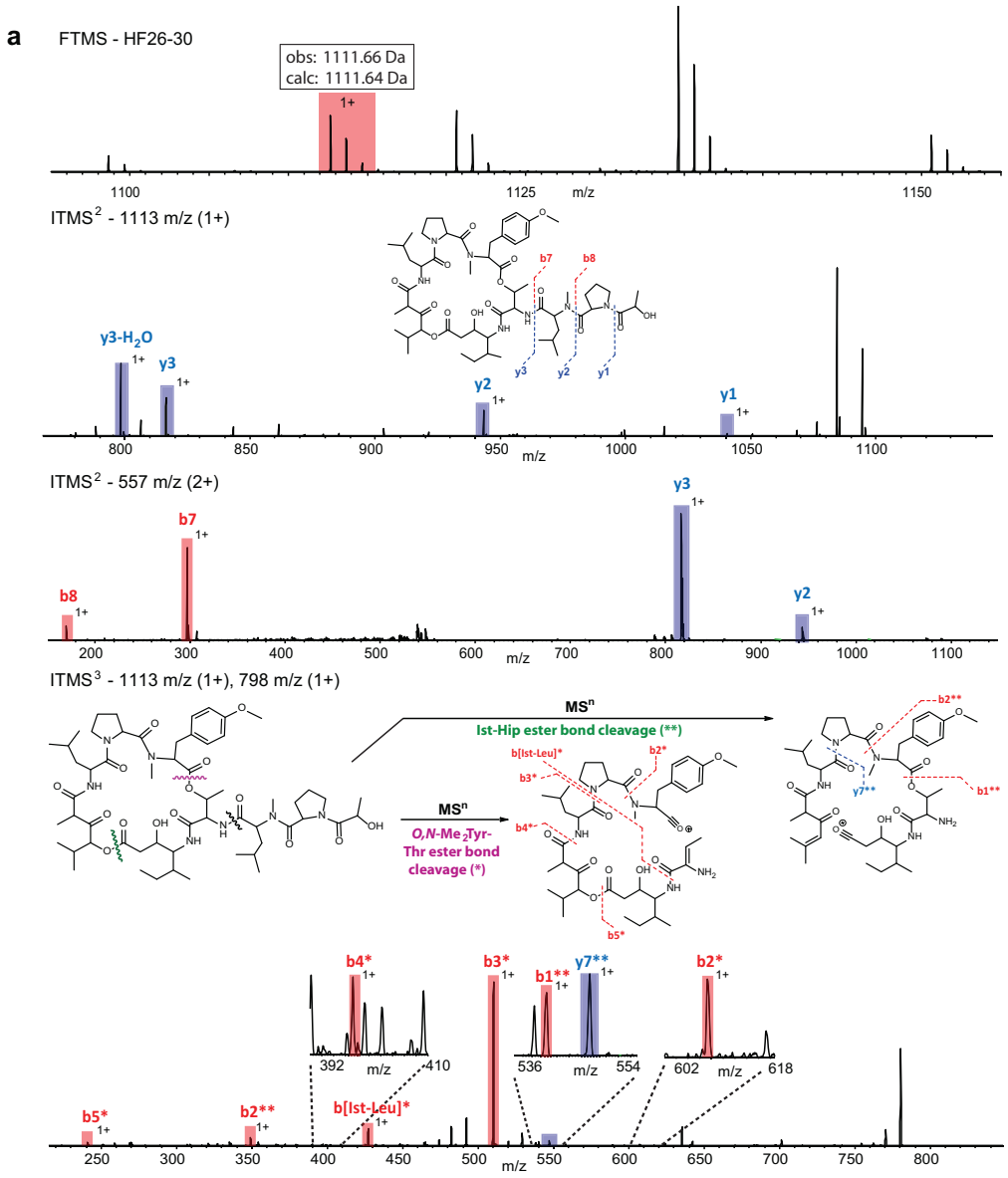
[⊥]Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California at San Diego, La Jolla, California 92093, USA

*Corresponding Authors: Pei-Yuan Qian, Phone: 852-2358-7331, Fax: 852-2358-1559, Email: boqianpy@ust.hk and

Bradley S. Moore, Phone: 858-822-6650, Fax: 858-534-1305, Email: bsmoore@ucsd.edu

Supporting Information

Supplementary Figures



Detector	Precursor ion [m/z]	Observed mass [Da]	Calculated mass [Da]	Error [Da]	Species
IT	1113 (1+)	1040.64	1040.62781	0.01	y1
IT	1113 (1+)	943.55	943.57505	-0.03	y2
IT	1113 (1+)	816.45	816.47533	-0.03	y3
IT	1113 (1+)	798.45	798.46477	-0.01	y3-H_2O
IT	1113 (1+), 798 (1+)	607.27	607.37014	-0.10	b2*
IT	1113 (1+), 798 (1+)	547.36	547.31263	0.05	y7**
IT	1113 (1+), 798 (1+)	540.36	540.30681	0.05	b1**

Detector	Precursor ion [m/z]	Observed mass [Da]	Calculated mass [Da]	Error [Da]	Species
IT	1113 (1+), 798 (1+)	510.36	510.31738	0.04	b3*
IT	1113 (1+), 798 (1+)	427.36	427.28026	0.08	b[Ist-Leu]*
IT	1113 (1+), 798 (1+)	397.27	397.23331	0.04	b4*
IT	1113 (1+), 798 (1+)	349.18	349.21218	-0.03	b2**
IT	1113 (1+)	297.18	297.18088	0.00	b7
IT	1113 (1+), 798 (1+)	241.09	241.15467	-0.06	b5*
IT	1113 (1+)	170	170.08117	-0.08	b8

Fig. S1. Characterization of didemnin B. (**a**) MS^n analysis.

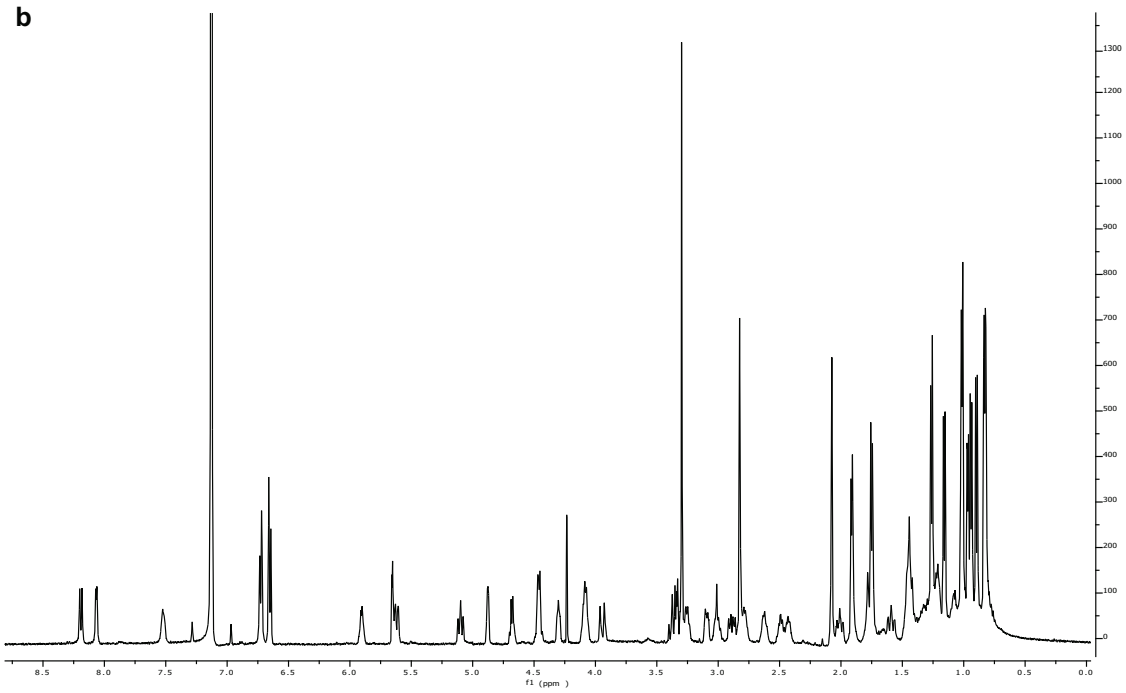
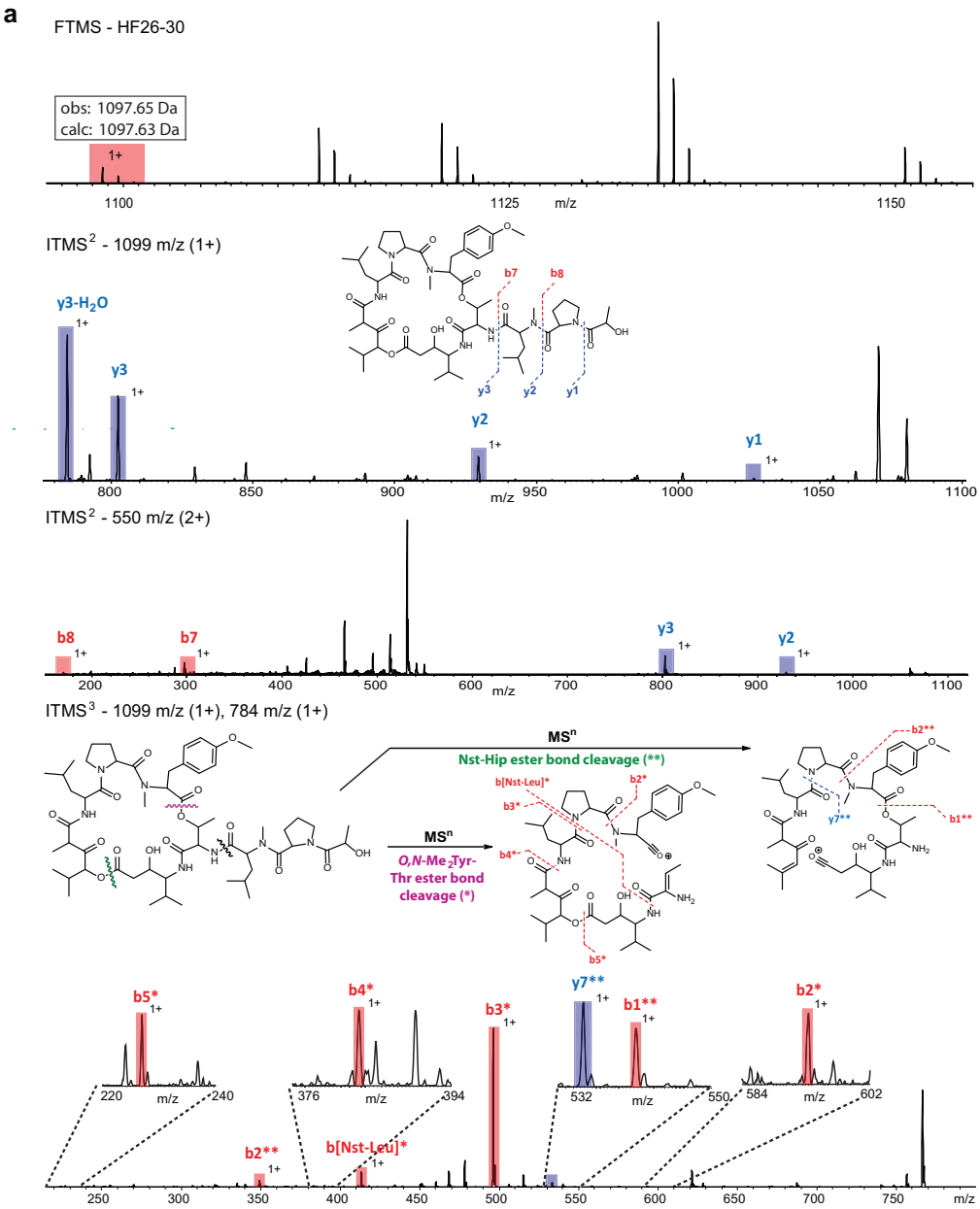


Fig. S1. Characterization of didemnin B. (**b**) ^{1}H NMR spectrum.



Detector	Precursor ion [m/z]	Observed mass [Da]	Calculated mass [Da]	Error [Da]	Species
IT	1099 (1+)	1026.64	1026.61216	0.03	y1
IT	1099 (1+)	929.55	929.5594	-0.01	y2
IT	1099 (1+)	802.45	802.45968	-0.01	y3
IT	1099 (1+)	784.45	784.44912	0.00	y3-H_2O
IT	1099 (1+), 784 (1+)	593.45	593.35449	0.10	b2*
IT	1099 (1+), 784 (1+)	540.18	540.30681	-0.13	b1**
IT	1099 (1+), 784 (1+)	533.36	533.29698	0.06	y7**

Detector	Precursor ion [m/z]	Observed mass [Da]	Calculated mass [Da]	Error [Da]	Species
IT	1099 (1+), 784 (1+)	496.13	496.30173	-0.17	b3*
IT	1099 (1+), 784 (1+)	413.27	413.26461	0.01	b[Nst-Leu]*
IT	1099 (1+), 784 (1+)	383.18	383.21766	-0.04	b4*
IT	1099 (1+), 784 (1+)	349.27	349.21218	0.06	b2**
IT	1099 (1+)	297.18	297.18088	0.00	b7
IT	1099 (1+), 784 (1+)	227.09	227.13902	-0.05	b5*
IT	1099 (1+)	169.91	170.08117	-0.17	b8

Fig. S2. Characterization of nordidemnin B. (**a**) MS^n analysis.

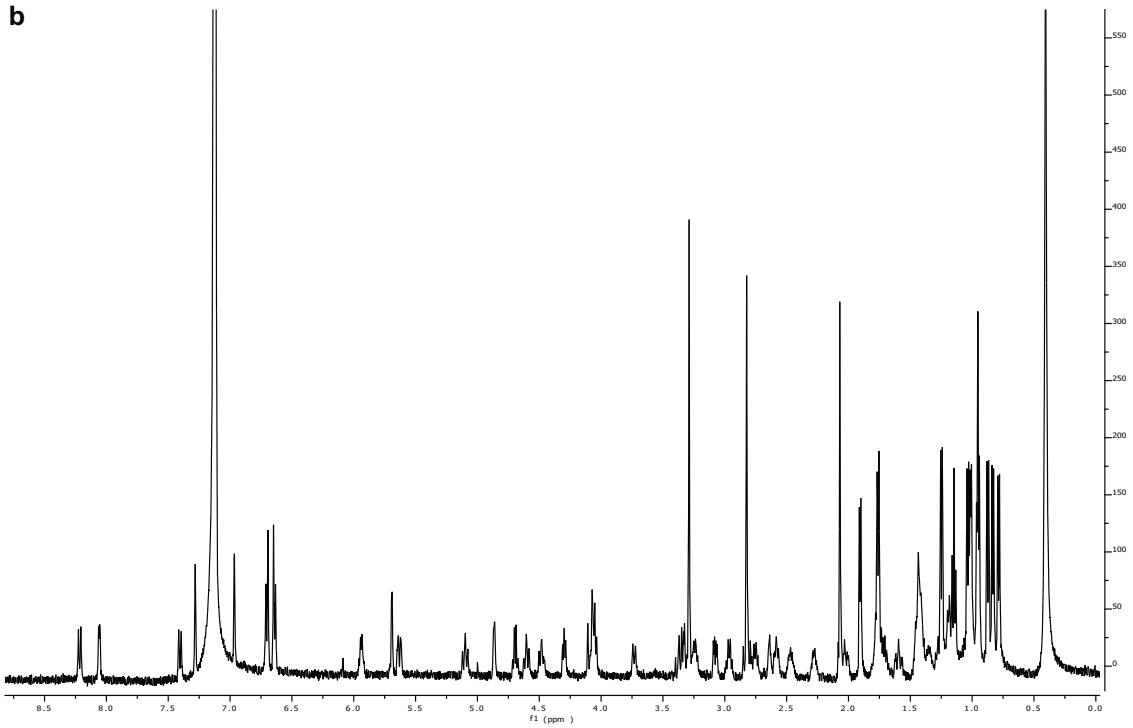


Fig. S2. Characterization of nordidemnin B. (**b**) ^{1}H NMR spectrum.

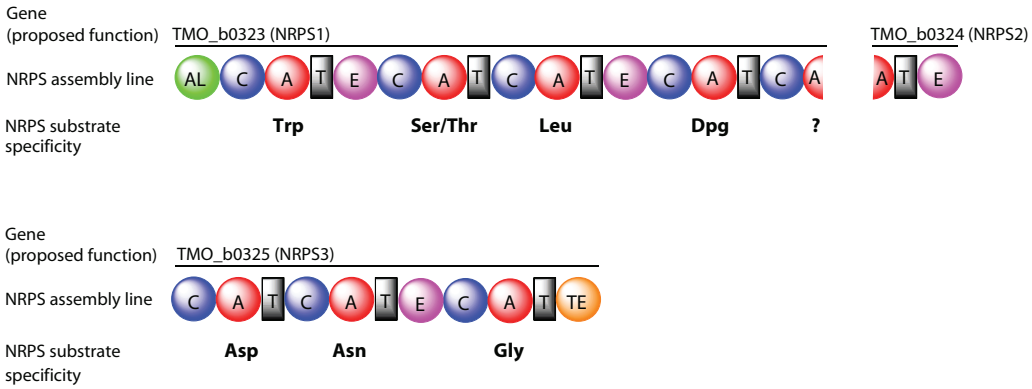
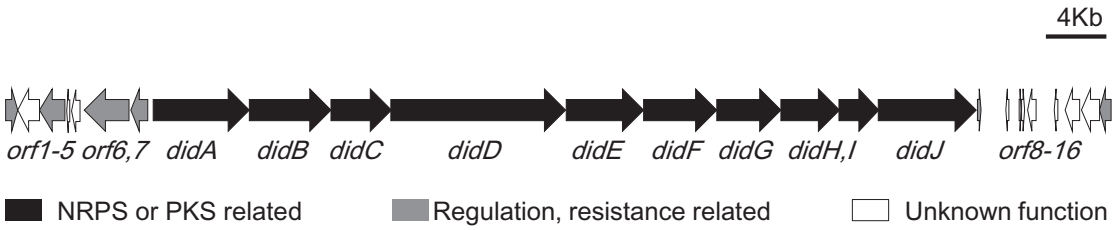


Fig. S3. Bioinformatic analysis of uncharacterized *nrps1* gene cluster on plasmid 2 in the *T. mobilis* KA081020-065 genome. Domain notation: AL, acyl ligase; C, condensation domain; A, adenylation domain; T, thiolation domain; E, epimerization; TE, thioesterase. Substrate abbreviation: Dpg, dihydroxyphenylglycine.

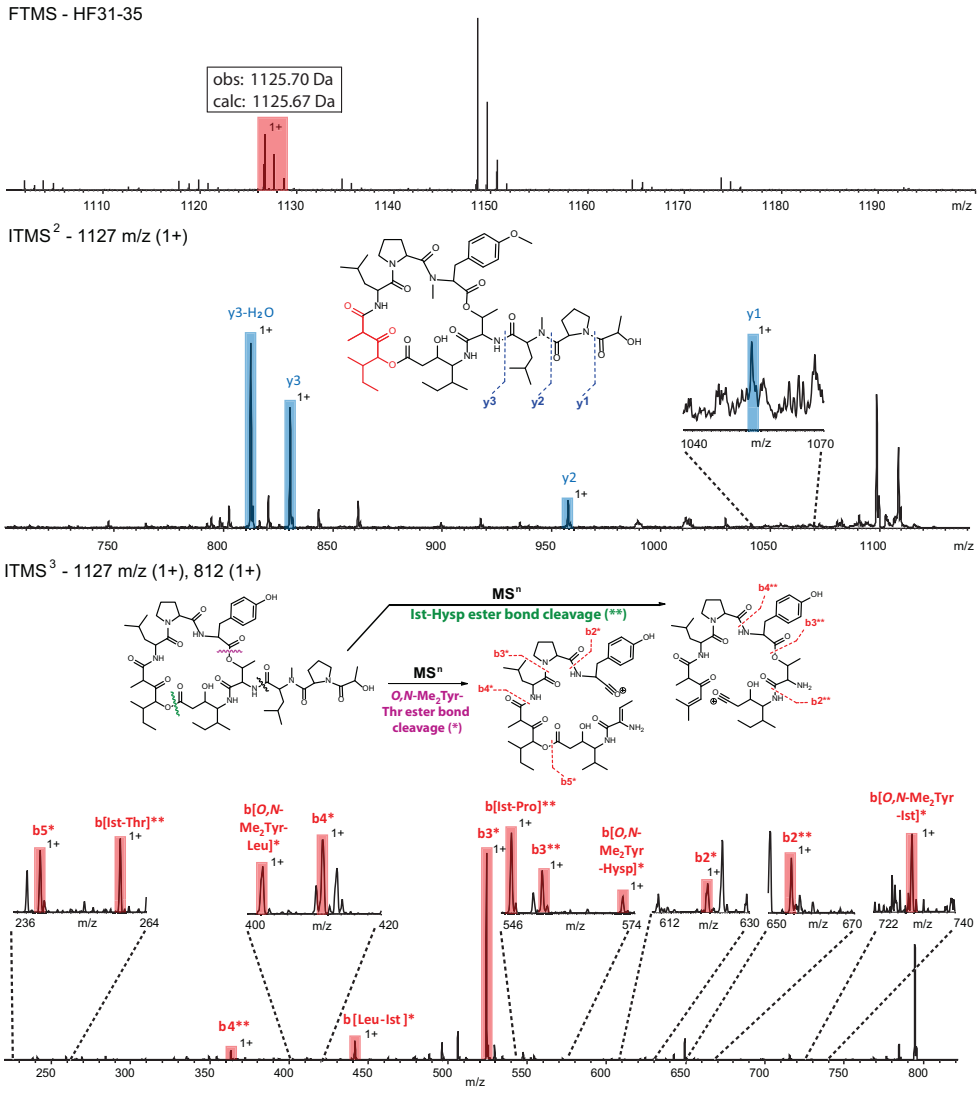


Gene	Size [aa]	Sequence similarity/organism	Proposed function	Identity/ similarity	GenBank accession no.
orf1	261	Thioesterase, *Actinomadura kijaniata*	Thioesterase	41%, 54%	ACB46473
orf2	483	Membrane protease subunit stomatin/prohibitin-like protein, a-proteobacterium BAL199	Putative protease	42%, 59%	ZP_02189733
orf3	560	Cyclic peptide transporter, *Methylobacter tundripaludum* SV96	Transporter	48%, 68%	ZP_07653047
orf4	70	None	Hypothetical		
orf5	190	GTPase domain-containing protein, *Methylobacter tundripaludum*SV96	Hypothetical	39%, 57%	ZP_07653048
orf6	987	Hydrophobic/amphiphilic exporter-1, *Azospirillum* sp. B510	Resistance	38%, 56%	YP_003450508
orf7	377	Secretion protein, *Azospirillum* sp. B510	Resistance	28%, 47%	YP_003450507
didA	2123	OciA protein, *Planktothrix rubescens* NIVA-CYA 98	NRPS	29%, 40%	CAQ48254
didB	1796	Linear gramicidin synthetase subunit D, *Stigmatella aurantiaca* DW4/3-1	NPRS	36%, 48%	ZP_01459555
didC	1330	NRPS, *Myxococcus xanthus* DK 1622	NRPS	41%, 52%	YP_632257
didD	3853	Amino acid adenylation domain protein, *Streptomyces violaceusniger* Tu 4113	NRPS	39%, 50%	ZP_07603194
didE	1705	Amino acid adenylation domain protein, *Acetivibrio cellulolyticus* CD2	NRPS/PKS	36%, 53%	ZP_07325073
didF	1613	HctF, *Lyngbya majuscula*	NRPS	35%, 52%	AAY42398
didG	1413	NRPS/PKS, *Amycolatopsis mediterranei* U32	PKS	46%, 56%	YP_003765866
didH	1286	NRPS/PKS, Myxococcus xanthus DK 1622	NRPS	39%, 53%	YP_631961
didI	873	NRPS, *Myxococcus xanthus* DK 1622	NRPS	39%, 52%	YP_632257
didJ	2163	Amino acid adenylation domain protein, *Lyngbya majuscula* 3L	NRPS	36%, 53%	ZP_08431746

Fig. S4. Organization and deduced functions of the open reading frames within and flanking the didemnin biosynthetic gene cluster.

Gene	Size [aa]	Sequence similarity/organism	Proposed function	Identity/ similarity	GenBank accession no.
orf8	77	MbtH domain-containing protein, *Herpetosiphon aurantiacus* ATCC 23779	MbtH-like protein	80%, 89%	YP_001542806
orf9	68	Hypothetical protein, *Acidovorax* sp. JS42	Hypothetical	79%, 91%	YP_986866
orf10	45	None	Hypothetical		
orf11	60	None	Hypothetical		
orf12	190	Hypothetical protein, *Acidovorax* sp. JS42	Hypothetical	94%, 97%	YP_986861
orf13	75	Hypothetical protein, *Acidovorax* sp. JS42	Hypothetical	98%,100%	YP_004387524
orf14	324	CAAX amino terminal protease family, *Synechococcus* sp. PCC 7335	Putative protease	30%, 49%	ZP_05035401
orf15	398	Cyanate transport system protein, *Pseudomonas syringae* pv. syringae 642	Transport	40%, 52%	ZP_07265073
orf16	255	GntR family transcriptional regulator, *Chromobacterium violaceum* ATCC 12472	Regulation	39%, 55%	NP_903400

Fig. S4, continued. Organization and deduced functions of the open reading frames within and flanking the didemnin biosynthetic gene cluster.



Detector	Precursor ion [m/z]	Observed mass [Da]	Calculated mass [Da]	Error [Da]	Species
IT	1127 (1+)	1054.64	1054.64346	0.00	y1
IT	1127 (1+)	957.55	957.59070	-0.04	y2
IT	1127 (1+)	830.45	830.49098	-0.04	y3
IT	1127 (1+)	812.45	812.48042	-0.03	y3-H_2O
IT	1127 (1+), 812 (1+)	729.45	729.44331	0.01	b[O,N-Me_2Tyr-Ist]*
IT	1127 (1+), 812 (1+)	655.36	655.37069	-0.01	b2**
IT	1127 (1+), 812 (1+)	621.45	621.38634	0.06	b2*
IT	1127 (1+), 812 (1+)	572.36	572.33303	0.03	b[O,N-Me_2Tyr-Hysp]*

Detector	Precursor ion [m/z]	Observed mass [Da]	Calculated mass [Da]	Error [Da]	Species
IT	1127 (1+), 812 (1+)	554.27	554.32301	-0.05	b3**
IT	1127 (1+), 812 (1+)	547.36	547.31263	0.05	b[Ist-Pro]**
IT	1127 (1+), 812 (1+)	524.36	524.33358	0.03	b3*
IT	1127 (1+), 812 (1+)	441.36	441.29591	0.06	b[Leu-Ist]*
IT	1127 (1+), 812 (1+)	411.27	411.24951	0.02	b4*
IT	1127 (1+), 812 (1+)	402.24	402.23873	0.00	b[O,N-Me_2Tyr-Leu]*
IT	1127 (1+), 812 (1+)	363.18	363.22838	-0.05	b4**
IT	1127 (1+), 812 (1+)	259.18	259.16523	0.01	b[Ist-Thr]**
IT	1127 (1+), 812 (1+)	241.18	241.15522	0.02	b5*

Fig. S5. Characterization of didemnin B derivative [Hysp]didemnin B produced by *Tistrella mobilis* KA081020-065 due to putative NRPS substrate promiscuity in DidF A8 domain. MS^n analysis of [Hysp]didemnin B. Hysp, α-(α-hydroxy sec-butylacetyl) propionic acid.

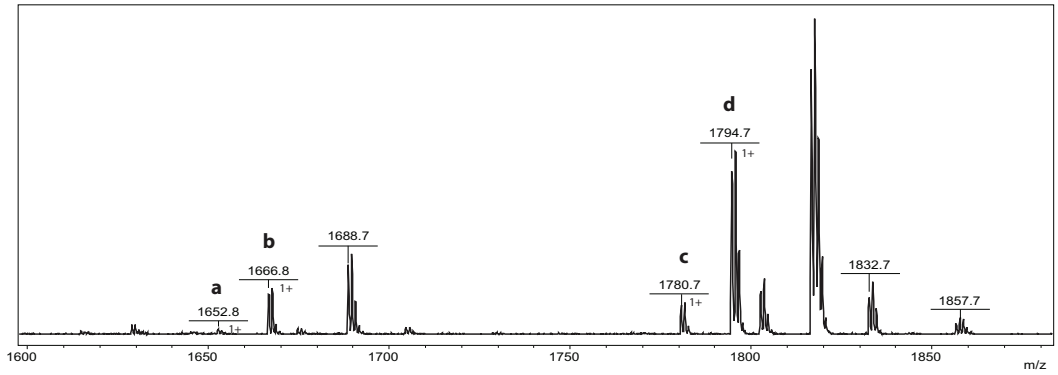
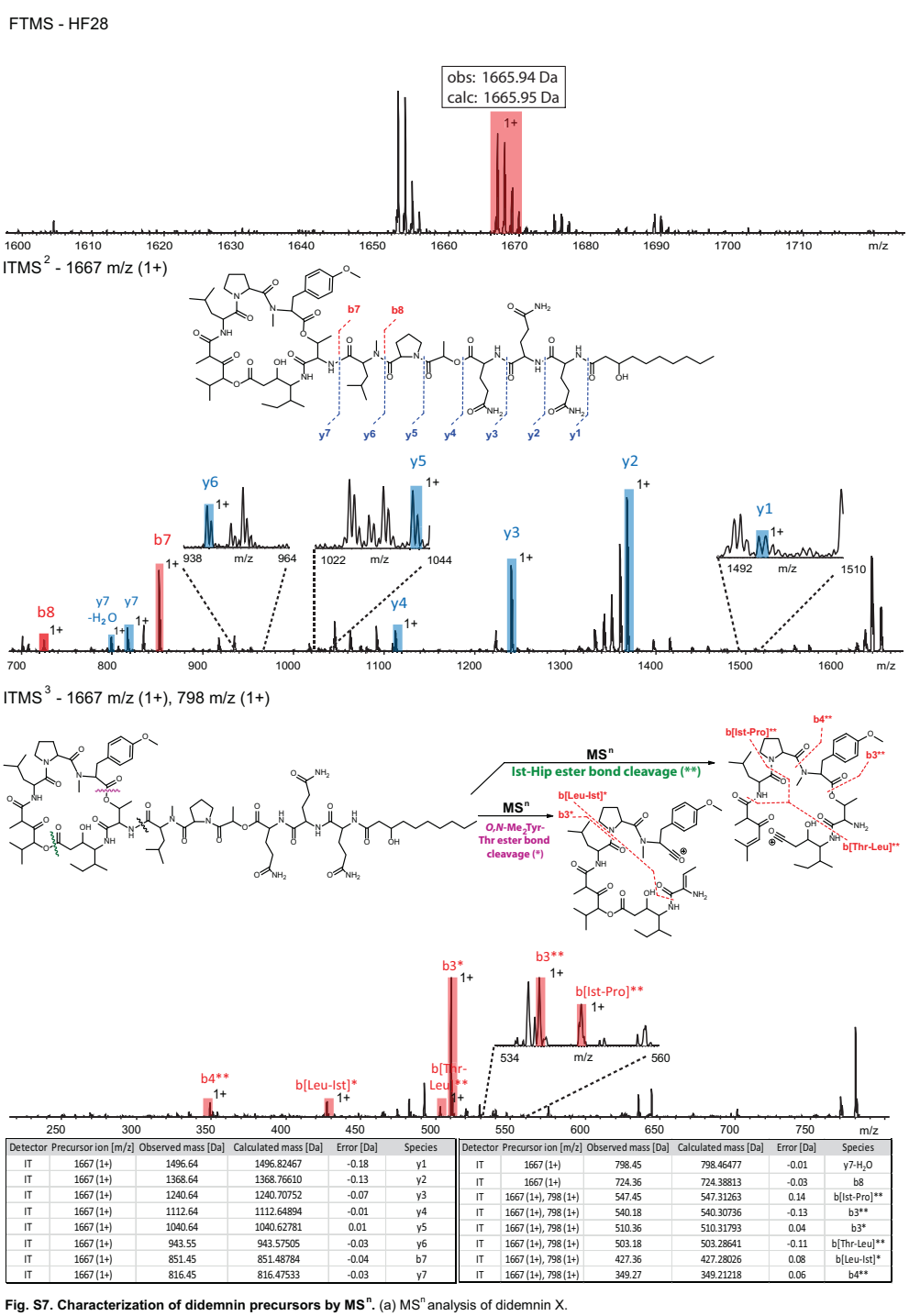


Fig. S6. Detection of didemnin precursors in the extract of *Tistrella mobilis* KA081020-065 by MALDI-TOF MS. a – Nordidemnin X ($[M+H]^+$, obs 1651.8 Da, calc 1651.9 Da), b – Didemnin X ($[M+H]^+$, obs 1665.8 Da, calc 1666.0), c – Nordidemnin Y ($[M+H]^+$, obs 1779.7 Da, calc 1780.0 Da), d – Didemnin Y ($[M+H]^+$, obs 1793.7 Da, calc 1794.0 Da).

a

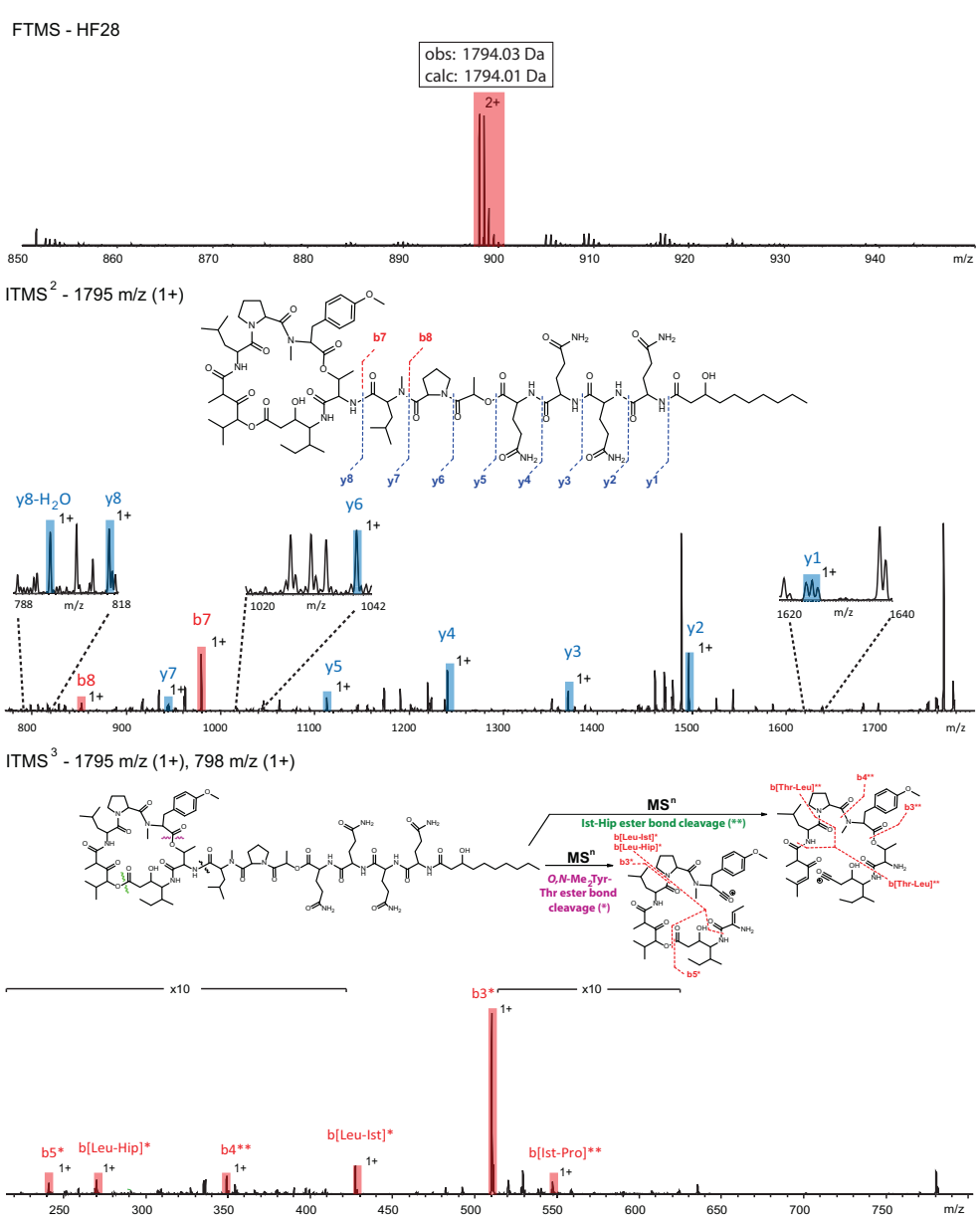


Detector	Precursor ion [m/z]	Observed mass [Da]	Calculated mass [Da]	Error [Da]	Species
IT	1667 (1+)	1496.64	1496.82467	-0.18	y1
IT	1667 (1+)	1368.64	1368.76610	-0.13	y2
IT	1667 (1+)	1240.64	1240.70752	-0.07	y3
IT	1667 (1+)	1112.64	1112.64894	-0.01	y4
IT	1667 (1+)	1040.64	1040.62781	0.01	y5
IT	1667 (1+)	943.55	943.57505	-0.03	y6
IT	1667 (1+)	851.45	851.48784	-0.04	b7
IT	1667 (1+)	816.45	816.47533	-0.03	y7

Detector	Precursor ion [m/z]	Observed mass [Da]	Calculated mass [Da]	Error [Da]	Species
IT	1667 (1+)	798.45	798.46477	-0.01	y7-H_2O
IT	1667 (1+)	724.36	724.38813	-0.03	b8
IT	1667 (1+), 798 (1+)	547.45	547.31263	0.14	b[Ist-Pro]**
IT	1667 (1+), 798 (1+)	540.18	540.30736	-0.13	b3**
IT	1667 (1+), 798 (1+)	510.36	510.31793	0.04	b3*
IT	1667 (1+), 798 (1+)	503.18	503.28641	-0.11	b[Thr-Leu]**
IT	1667 (1+), 798 (1+)	427.36	427.28026	0.08	b[Leu-Ist]*
IT	1667 (1+), 798 (1+)	349.27	349.21218	0.06	b4**

Fig. S7. Characterization of didemnin precursors by MS^n. (a) MS^n analysis of didemnin X.

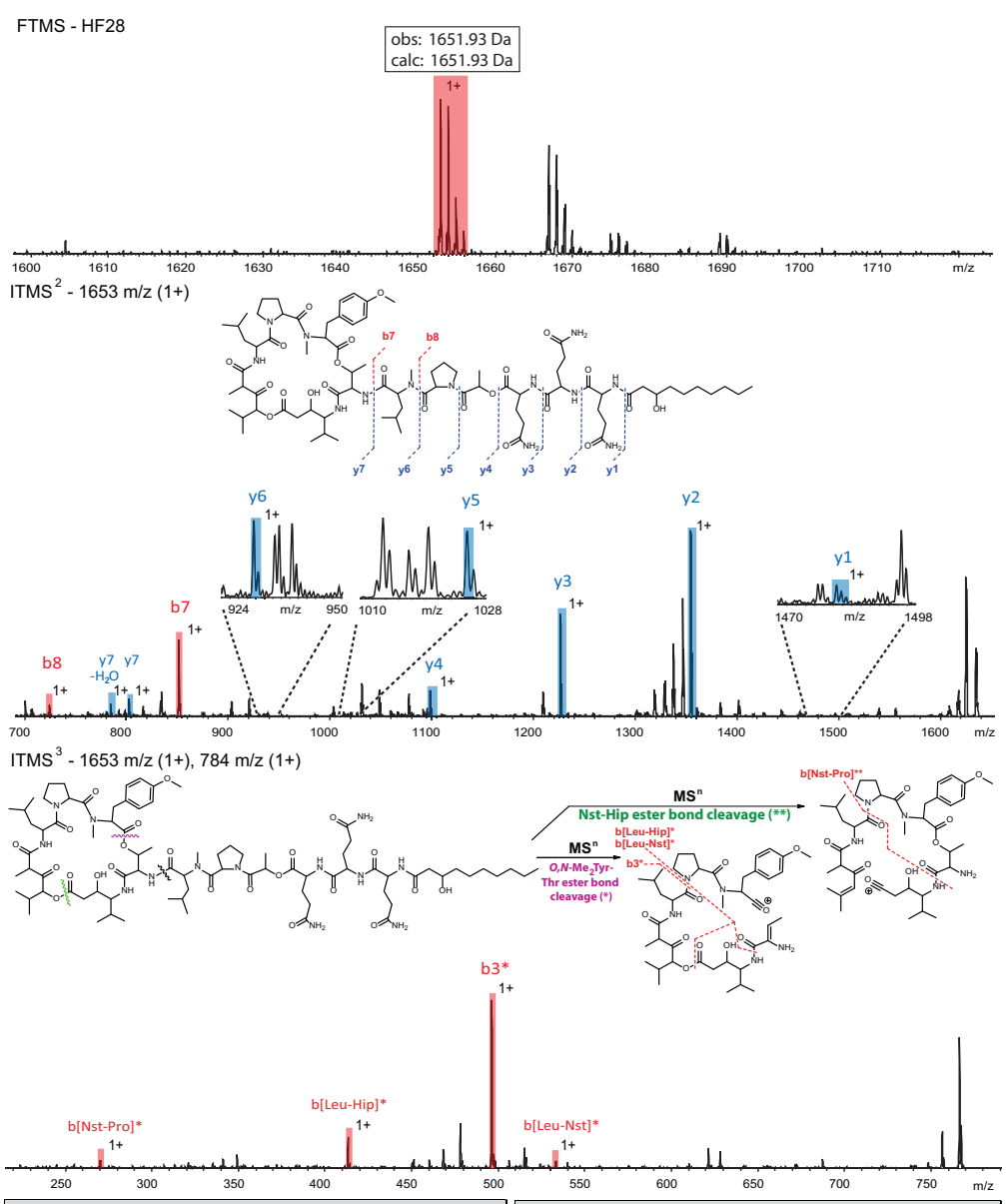
b



Detector	Precursor ion [m/z]	Observed mass [Da]	Calculated mass [Da]	Error [Da]	Species
IT	1795 (1+)	1624.7	1624.88325	-0.18	y1
IT	1795 (1+)	1496.73	1496.82467	-0.09	y2
IT	1795 (1+)	1368.64	1368.76610	-0.13	y3
IT	1795 (1+)	1240.64	1240.70752	-0.07	y4
IT	1795 (1+)	1112.64	1112.64894	-0.01	y5
IT	1795 (1+)	1040.64	1040.62781	0.01	y6
IT	1795 (1+)	979.55	979.54642	0.00	b7
IT	1795 (1+)	943.55	943.57505	-0.03	y7
IT	1795 (1+)	852.45	852.44671	0.00	b8

Detector	Precursor ion [m/z]	Observed mass [Da]	Calculated mass [Da]	Error [Da]	Species
IT	1795 (1+)	816.45	816.47533	-0.03	y8
IT	1795 (1+)	798.45	798.46477	-0.01	y8-H_2O
IT	898 (2+), 798 (1+)	547.36	547.31263	0.05	b[Ist-Pro]**
IT	898 (2+), 798 (1+)	510.36	510.31793	0.04	b3*
IT	898 (2+), 798 (1+)	427.36	427.28026	0.08	b[Leu-Ist]*
IT	898 (2+), 798 (1+)	349.18	349.21218	-0.03	b4**
IT	898 (2+), 798 (1+)	270.27	270.16998	0.10	b[Leu-Hip]*
IT	898 (2+), 798 (1+)	241.18	241.15522	0.02	b5*

Fig. S7. Characterization of didemnin precursors by MS^n. (b) MS^n analysis of didemnin Y.

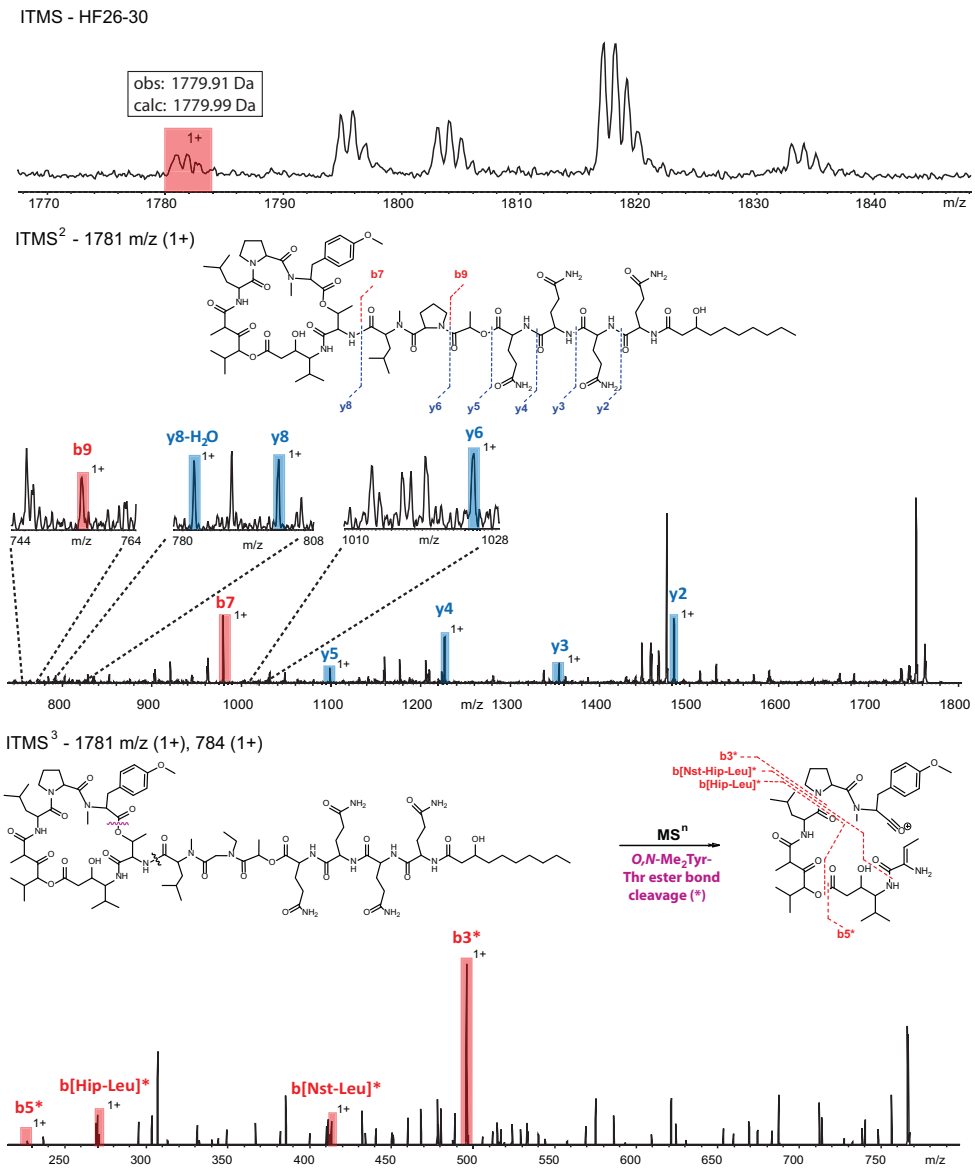


Detector	Precursor ion [m/z]	Observed mass [Da]	Calculated mass [Da]	Error [Da]	Species
IT	1653 (1+)	1482.73	1482.80902	-0.08	y1
IT	1653 (1+)	1354.64	1354.75045	-0.11	y2
IT	1653 (1+)	1226.64	1226.69187	-0.05	y3
IT	1653 (1+)	1098.55	1098.63329	-0.08	y4
IT	1653 (1+)	1026.64	1026.61216	0.03	y5
IT	1653 (1+)	929.55	929.55940	-0.01	y6
IT	1653 (1+)	802.45	802.45968	-0.01	y7

Detector	Precursor ion [m/z]	Observed mass [Da]	Calculated mass [Da]	Error [Da]	Species
IT	1653 (1+)	784.45	784.44912	0.00	y7-H_2O
IT	1653 (1+)	851.45	851.48784	-0.04	b7
IT	1653 (1+)	724.36	724.38813	-0.03	b8
IT	1653 (1+), 784 (1+)	533.45	533.29698	0.15	b[Nst-Pro]*
IT	1653 (1+), 784 (1+)	496.36	496.30228	0.06	b3*
IT	1653 (1+), 784 (1+)	413.36	413.26461	0.10	b[Leu-Nst]*
IT	1653 (1+), 784 (1+)	270.27	270.16998	0.10	b[Leu-Hip]*

Fig. S7. Characterization of didemnin precursors by MS^n. (c) MS^n analysis of nordidemnin X.

d



Detector	Precursor ion [m/z]	Observed mass [Da]	Calculated mass [Da]	Error [Da]	Species
IT	1781 (1+)	1482.73	1482.80902	0.08	y2
IT	1781 (1+)	1354.64	1354.75045	0.11	y3
IT	1781 (1+)	1226.64	1226.69187	0.05	y4
IT	1781 (1+)	1098.64	1098.63329	-0.01	y5
IT	1781 (1+)	1026.64	1026.61216	-0.03	y6
IT	1781 (1+)	979.45	979.54642	0.10	b7
IT	1781 (1+)	802.45	802.45968	0.01	y8

Detector	Precursor ion [m/z]	Observed mass [Da]	Calculated mass [Da]	Error [Da]	Species
IT	1781 (1+)	784.45	784.44912	0.00	y8-H2O
IT	1781 (1+)	755.27	755.39394	0.12	b9
IT	1781 (1+), 784 (1+)	496.36	496.30228	-0.06	b3*
IT	1781 (1+), 784 (1+)	413.36	413.26461	-0.10	b[Nst-Leu]*
IT	1781 (1+), 784 (1+)	270.18	270.16998	-0.01	b[Hip-Leu]*
IT	1781 (1+), 784 (1+)	227.13	227.13957	0.01	b5*

Fig. S7. Characterization of didemnin precursors by MS^n. (d) MS^n analysis of nordidemnin Y.

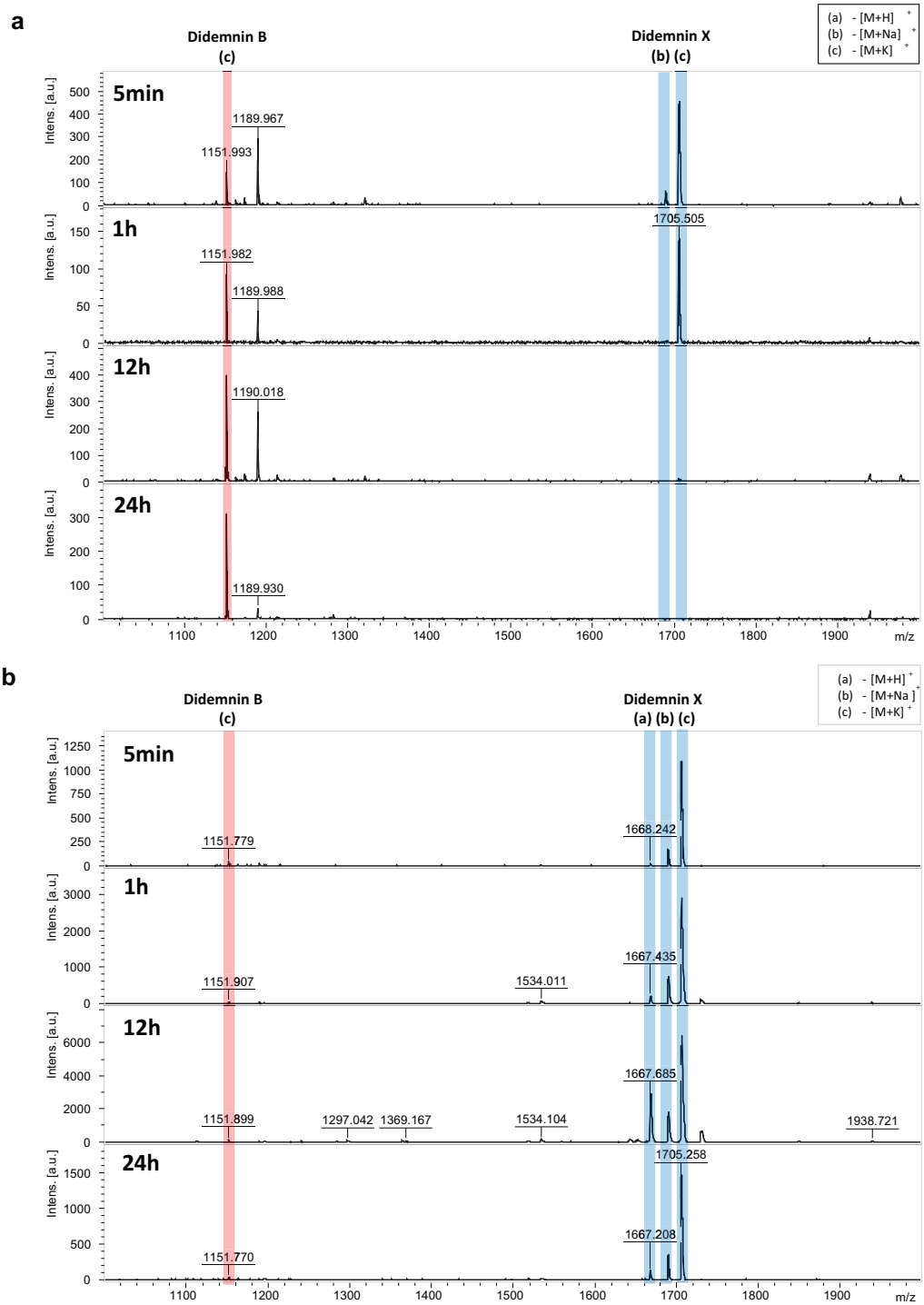


Fig. S8. Didemnin precursor hydrolysis assay. MALDI-TOF MS analysis of didemnin X hydrolysis time course in presence of *Tistrella mobilis* secreted proteome ((**a**), 1d growth in liquid culture, >10kDa protein cutoff filter) and in presence of sterile GYP medium ((**b**), negative control).

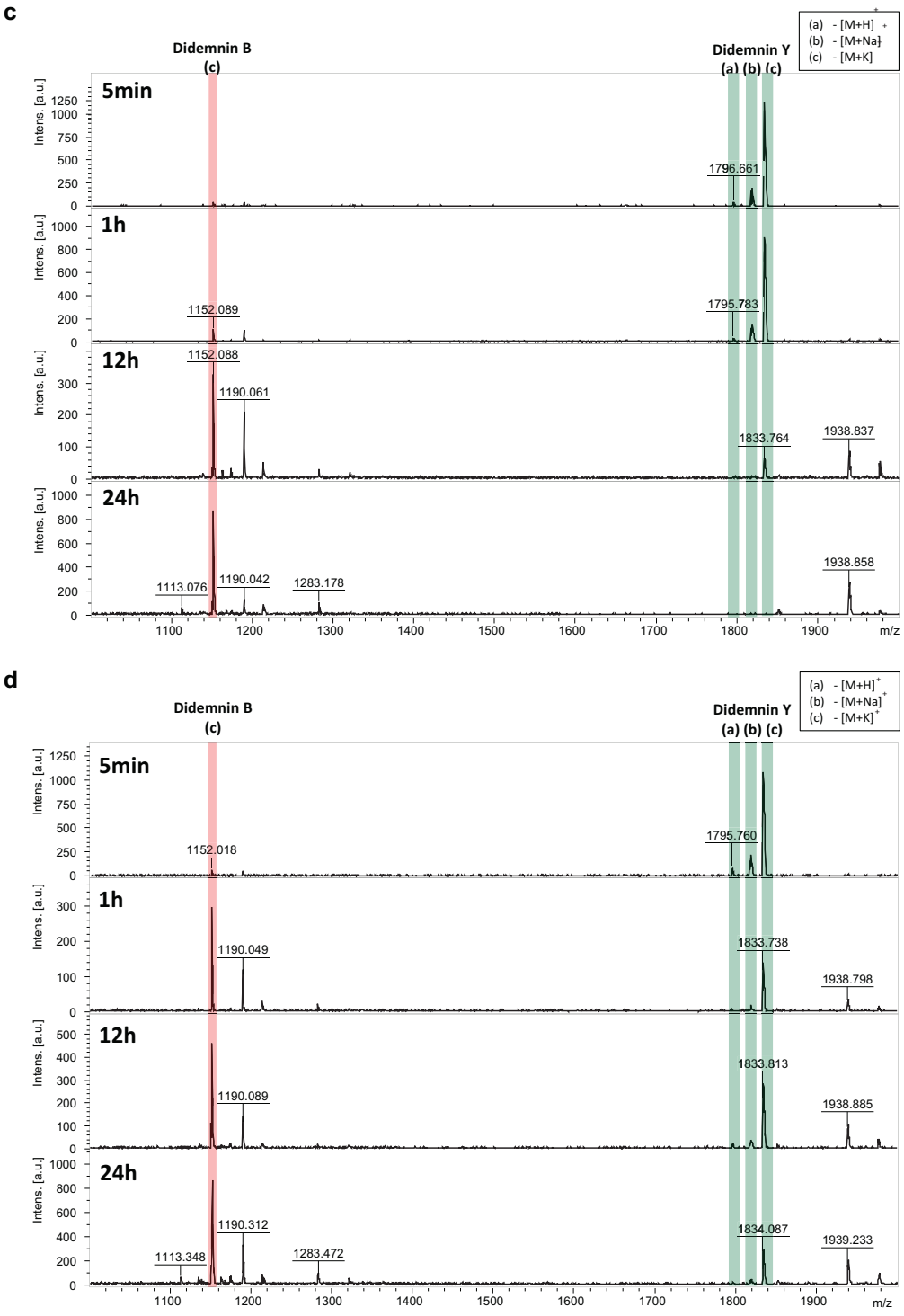


Fig. S8. Didemnin precursor hydrolysis assay. MALDI-TOF MS analysis of didemnin Y hydrolysis time course in presence of *Tistrella mobilis* secreted proteome ((**c**), 1d growth in liquid culture, >10kDa protein cutoff filter) and in presence of sterile GYP medium ((**d**), negative control).

Table S1. General genome features of *Tistrella mobilis* KA081020-065

Features	Chromosome	pTM1	pTM2	pTM3	pTM4
Topology	Circular	Circular	Circular	Circular	Circular
Genome size (bp)	3919492	692874	690188	1126962	83885
G+C content	68.15%	68.26%	67.61%	68.08%	67.26%
CDs number	3565	605	602	942	72
Coding density	89.08%	90.38%	88.92%	88.55%	85.06%
Average CDs length (bp)	979	1035	1019	1059	991
Assigned function	2852	432	501	726	60
Conserved hypothetical	472	101	71	165	9
Hypothetical	244	72	30	51	3
rRNA operons	2	0	0	1	1
tRNA operons	41	0	1	9	3

Chapter 3, in full, is a reprint of the material as it appears in 'Bacterial biosynthesis and maturation of the didemnin anti-cancer agents', Xu, Y., Kersten, R.D., Nam, S.J., Lu, L., Al-Suwailem, A.M., Zheng, H., Fenical, W., Dorrestein, P.C., Moore, B.S., Qian, P.Y. *Journal of the American Chemical Society*, 2012, 134, 8625-8632. The dissertation author was one of two equally contributing primary investigators and authors of this paper.

Y.X. and R.D.K. designed and carried out chemical isolation, MS structure elucidation and imaging MS experiments, analyzed data and wrote the paper, S.-J.N. and L.L. carried out chemical isolation and NMR structure elucidation experiments and analyzed data, A.M.A.-S. carried out bacterial strain isolation experiments, H.Z. carried out genome sequencing experiments, W.F. and P.C.D. designed experiments and analyzed data, B.S.M. and P.-Y.Q. designed experiments, analyzed data and wrote the paper. Y.X. and R.D.K. contributed equally to the paper.

Chapter 4 - Glycogenomics, a mass spectrometry-guided genome mining method to connect biosynthesis genes to glycosylated natural products

4.1 Introduction

Glycosylated natural products (GNPs) produced by microbes comprise many compounds with therapeutic and agrochemical applications such as the antibiotic erythromycin [1] and the insecticide avermectin [2], respectively. A glycosylated natural product consists of an aglycone moiety and one or multiple glycosyl units (Figure 14) [3] which often directly mediate the bioactivity of the compound [4]. In microbial genomes, the genes for biosynthesis and attachment of these glycosyl groups are usually clustered with the biosynthetic genes of the aglycone (Figure 15).

In this chapter, a new experiment-guided genome mining approach is introduced that rapidly connects GNP chemotypes with their genotypes in microbial genomes via sugar moieties. Herein, *O*- and *N*-glycosyl groups are characterized in their sugar monomers by tandem mass spectrometry and matched to corresponding glycosylation genes in secondary metabolic pathways by a MS-glycogenetic code. The aglycone biosynthetic genes of the GNP genotype then classify the natural product and guide further structure elucidation. The glycogenomic approach extends the concept of MS-guided genome mining to non-peptidic natural products [5] and has the potential for automation. The glycogenomic strategy is highlighted by characterization of glycosylated aromatic polyketides, cinerubin B from *Streptomyces* sp. SPB74, and a new arenimycin derivative from *Salinispora arenicola* CNB-527.

4.2 Results

4.2.1 A MS-glycogenetic code connecting microbial GNP chemo- and genotypes

In order to connect GNP chemotypes by tandem MS with GNP genotypes, a template had to be established that would link *de novo* MS^n fragmentation data of each sugar with the corresponding biosynthetic genes from characterized microbial GNP pathways. This MS-glycogenetic code comprises 83 microbial sugar monomers including the most common microbial sugars from the Bacterial Carbohydrate Structure Data Base (BCSDB) [6] and most known deoxysugars involved in natural product glycosylation [3]. For each sugar, calculated masses of an *O*-/*N*-glycosidic neutral loss from the parent ion (Y-ion) and of B/C-ions in CID-based tandem MS experiments are listed together with the common and specific biosynthetic genes of the corresponding verified or predicted sugar pathway (Table 2).

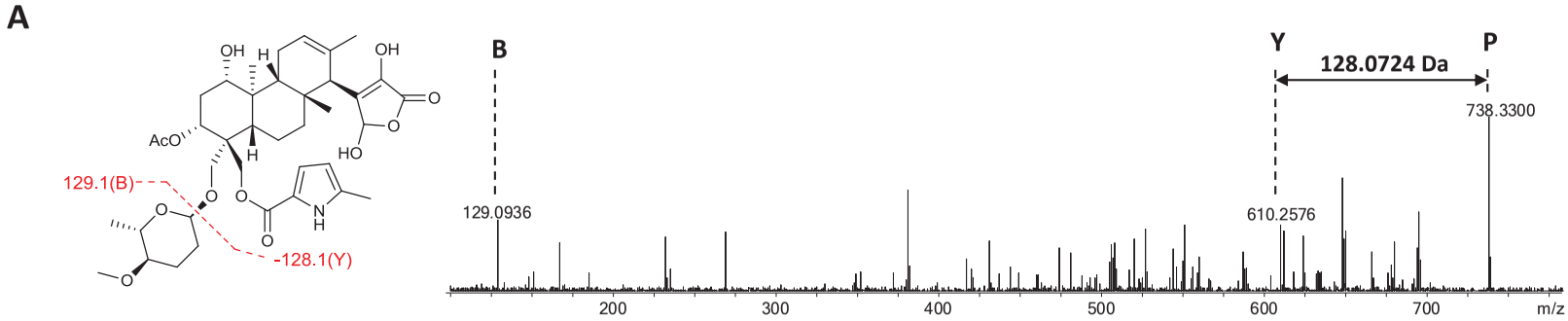
The MS-glycogenetic code was first tested whether it could connect MS^n data of known GNP chemotypes with their corresponding GNP genotypes from GenBank. The analyzed GNPs were selected based on the availability of MS^n data in databases, the literature or this study and the availability of nucleotide sequences associated with biosynthetic gene clusters (Tables 3 and 4). 17 of 19 analyzed GNPs could be connected successfully with their biosynthetic gene cluster by the MS-glycogenetic code. 14 of the 17 GNPs with observed sugar losses showed sugar-specific B-ions, while 15 of 17 showed sugar-specific Y-ion neutral losses. The first negative GNP result, nystatin [7] showed no expected sugar fragmentation in the database MS^n spectrum [8] and, thus, could not be matched to its gene cluster. The second negative GNP result, glycosylated thiopeptide Sch40832 [9], rather showed dideoxysugar-specific Y-ion neutral losses. However, the putative thiopeptide gene cluster in the *Micromonospora aurantiaca* ATCC 27029 genome comprised a glycosyltransferase but not the specific genes involved in dideoxysugar biosynthesis (Table 3 and 4) [10]. Thus, Sch40832 could not be connected with the gene cluster using a MS-glycogenetic approach. Among the analyzed GNPs were no C-glycosides. It is anticipated that C-glycosylated natural products won't be accessible by the MS-glycogenetic code

as C-glycosides don't fragment in a B/Y-ion pathway at the C-glycosidic bond under low-energy CID conditions but in an A/X-pathway at [0,2]-intraglycosidic bonds [11].

Exemplifying the MS-glycogenetic analysis, phenalinolactone A, a glycosylated terpene from genome-sequenced *Streptomyces* sp. Tu6071 [12], shows a neutral loss of 128.072 Da from the parent ion (738.345 *m/z*, $[M+Na]^+$) and a complementary 129.094 Da B-ion in its MS^2 spectrum. This putative Y-ion mass shift and B-ion correspond to isomeric *O*-methyl-*L*-amicetose or 4-*O*-methyl-*L*-rhodinose as MS^n candidate sugars (Table 1). BLAST analysis of the phenalinolactone gene cluster predicted three common glycosylation genes encoding a nucleotidylyltransferase, a 4,6-dehydratase and a glycosyltransferase, plus six specific glycosylation genes, i.e. a 2,3-dehydratase, a 3,4-dehydratase, a 3-ketoreductase, a 4-ketoreductase, an epimerase and an *O*-methyltransferase. In the MS-glycogenetic code, these specific genes match to the biosynthetic pathway of the two MS^n candidate sugars, *O*-methyl-*L*-amicetose and 4-*O*-methyl-*L*-rhodinose, thus connecting MS^n data of phenalinolactone A with its gene cluster (Figure 23).

4.2.2 MS-guided genome mining of a GNP from *Streptomyces* sp. SPB74

The MS-glycogenetic code was integrated into a workflow of MS-guided genome mining of glycosylated microbial natural products (Figure 24). This glycogenomic strategy starts with the LC-MS^n analysis of a metabolic extract of a genome-sequenced bacterium (Figure 24A). Candidate GNP fractions can be identified in the chromatogram by peaks in extracted ion chromatograms (EIC) of sugar-specific B/C-ion masses or Y/Z-ion neutral losses (Table 3, Fig. 24B). Candidate GNPs are then characterized in their putative glycosyl groups by neutral losses and B/C-ions in the corresponding MS^n spectra (Figure 24C). In the next step, the secondary metabolic gene cluster that contains biosynthetic genes to produce the MS^n candidate sugars is characterized (Figure 24D). First, all gene clusters with glycosylation genes are identified in the genome. Then, the GNP gene cluster with biosynthetic genes corresponding to any MS^n candidate sugars is characterized. Finally, the analysis of the aglycone genes enables the



B

Y-Mass shift [Da]	B-mass shift [Da]	C-mass shift [Da]	Sugar	Common genes	Specific BS genes																
					DH			KR		IM			AT			MT					
				NT	4,6-DH	2,3-DH	3,4-DH	3-KR	4-KR	(3,5-)E	3,4-IM	FuPyIM	2-AT	3-AT	4-AT	C3-MT	C5-MT	N-MT	N,N-dMT	O-MT
110.03678	111.04398	129.05454	L-aculose																	
112.05243	113.05963	131.07019	L-cinerulose A																	
114.06808	115.07528	133.08584	L-rhodinose																	
125.04768	126.05488	144.06544	rednose																	
128.04735	129.05455	147.06511	L-cinerulose B																	
128.08373	129.09093	147.10149	O-methyl-L-amicetose																	
128.08373	129.09093	147.10149	4-O-methyl-L-rhodinose																	
129.07898	130.08618	148.09674	L-daunosamine																	
129.07898	130.08618	148.09674	L-acosamine	N/A																
129.07898	130.08618	148.09674	L-ristosamine																	

C

Glycosylation and aglycone gene prediction

Gene	Reference	Size [aa]	Predicted function
PlaA1	ABB69735.1	434	3,4-dehydratase
PlaA2	ABB69736.1	463	2,3-dehydratase
PlaA3	ABB69737.1	323	4,6-dehydratase
PlaA4	ABB69738.1	301	nucleotidylyltransferase
PlaM1	ABB69739.1	303	O-methyltransferase
PlaT1	ABB69742.1	449	Squalene epoxidase
PlaT2	ABB69743.1	571	Putative squalene cyclase
PlaT3	ABB69744.1	337	Putative prenyltransferase
PlaA5	ABB69760.1	335	3-ketoreductase
PlaA6	ABB69761.1	396	glycosyltransferase
PlaA7	ABB69762.1	311	4-ketoreductase
PlaA8	ABB69763.1	191	epimerase

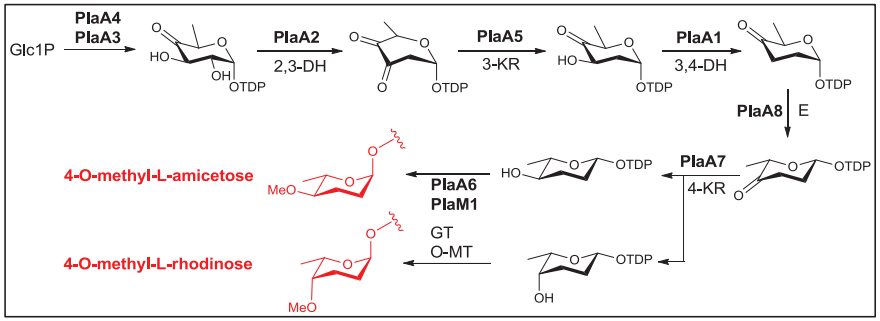


Figure 23: Connection of glycosylated terpene phenalinolactone A from *Streptomyces* sp. Tu6071 with its gene cluster by the MS-glycogenetic code. (A) Tandem MS spectrum of phenalinolactone A. A putative B-ion and Y-ion mass shift of a methyltrideoxysugar was detected using the sugar mass list of the MS-glycogenetic code. (B) MS-glycogenetic code with observed putative sugar fragments (red) and the corresponding biosynthetic genes of these sugars (red box). (C) Prediction of glycosylation genes from the phenalinolactone gene cluster (GenBank: DQ230532, [Durr, *et al. Chem Biol*, 2006]). Predicted gene functions match the pathways of the observed MS/MS sugars, 4-O-methyl-L-rhodinose and 4-O-methyl-L-amicetose.

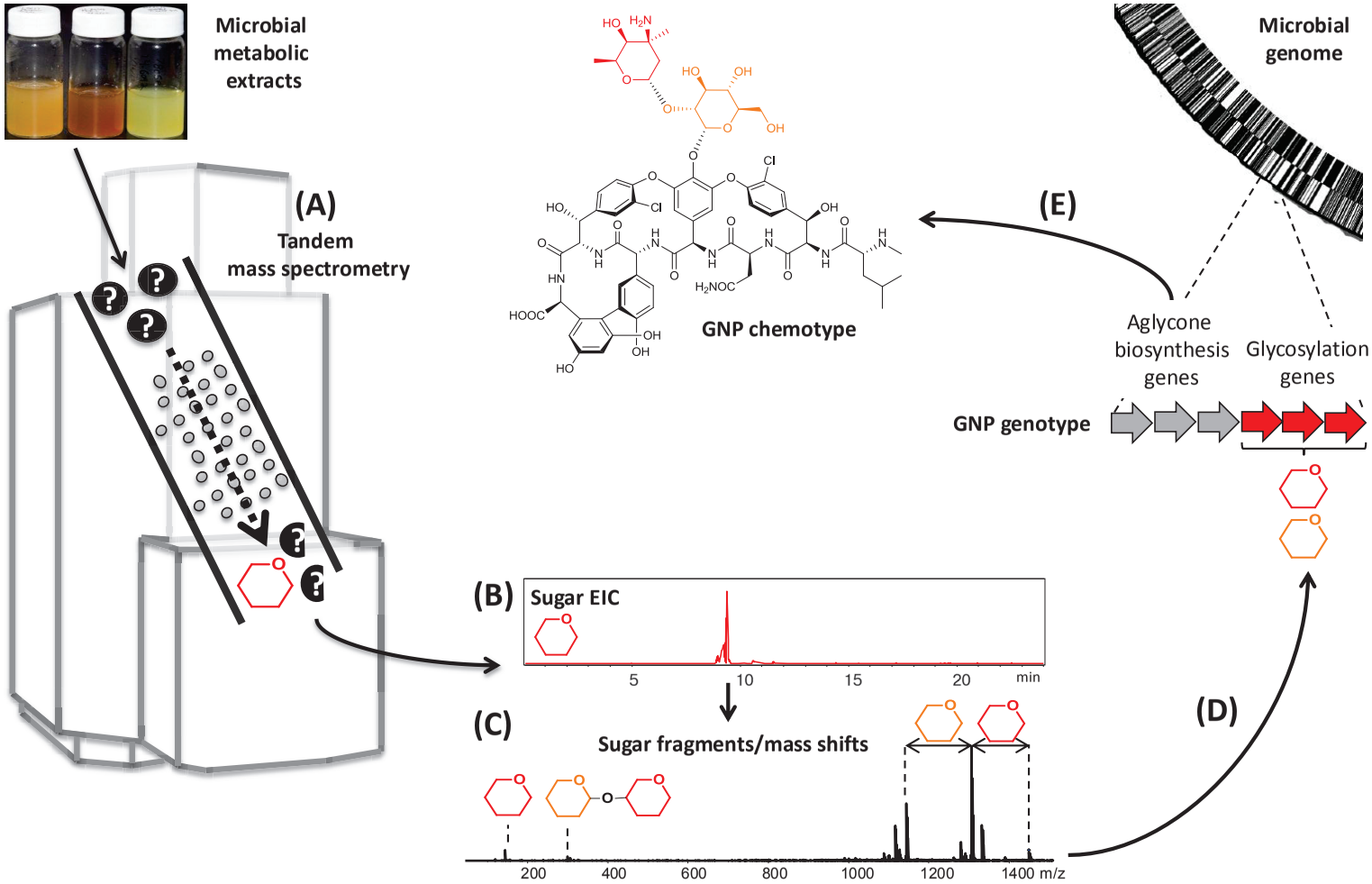


Figure 24: The glycogenomic workflow for characterization of glycosylated natural products from genome-sequenced microbes. (A) Tandem mass spectrometric analysis of microbial metabolic samples can reveal biosynthetic building blocks such as amino acids and sugar monomers of natural products via tandem MS fragment ions. (B) Identification of putative glycosylated natural products in a liquid chromatography-tandem mass spectrometry analysis as peaks in extracted ion chromatograms of known sugar fragment masses. (C) Verification of putative GNPs by characterization of candidate sugar monomers by sugar neutral losses and corresponding sugar fragment ions in tandem MS spectra. (D) Connection of putative GNP chemotype with corresponding GNP genotype in target microbial genome by genome mining of GNP pathway with glycosylation genes matching observed sugar fragments. (E) Characterization of GNP chemotype by analysis of aglycone biosynthetic genes of candidate GNP pathway and further structure elucidation.

classification of the GNP chemotype (Figure 24E). The connection of GNP structure and genes is verified by iterative analysis of MS^n and genetic data. NMR structure elucidation finally characterizes the GNP chemotype.

As a proof-of-concept experiment of the glycogenomic approach, we characterized the glycosylated anthracycline cinerubin B from a previously unknown producer, *Streptomyces* sp. SPB74 [13] (Figure 25 and 26). An organic extract of this genome-sequenced actinobacterium was analyzed by LC-MS^n to give a putative GNP with a parent mass of 825.317 Da (Figure 25A). Fragmentation of this molecule resulted in two mass shifts and two low-*m/z* fragment ions that corresponded to MS^n candidate sugars. The observed mass shifts of 110.0 Da and 130.1 Da matched *L*-aculose and a collection of six dideoxyhexose isomers, respectively, while the putative B-ion 158.12 *m/z* suggested the additional presence of one of eight aminodeoxysugars (Figure 25B). We next interrogated the genome sequence of *S.* sp. SPB74 for the biosynthesis of a natural product adorned with at least three sugar monomers. Of the 16 secondary metabolic gene clusters we identified by AntiSMASH analysis [14], just two harbored glycosylation genes and only one contained specific glycosylation genes (Figure 25C and 26). Among these specific genes were six associated with the biosynthesis of *L*-aculose (Table 1), a derivative of the trideoxy sugar *L*-rhodinose found in the polyketide antibiotic aclacinomycin Y [15]. Additional genes associated with deoxysugar biosynthesis were consistent with the predicted pathways for the four candidate aminodeoxysugars – megosamine, nogalamine, rhodosamine and angolosamine – and all candidate dideoxysugars, excluding the biosynthetically uncharacterized esperamicin A1 sugar [16]. Co-clustered with the deoxysugar biosynthesis gene locus were genes predicted for aglycone biosynthesis comprising an aromatic type II polyketide synthase. Further analysis of the gene set revealed its similarity to the aclacinomycin gene cluster from *Streptomyces galileus* [17,18]. This enabled the classification of the target compound as a glycosylated anthracycline polyketide. Structure elucidation of the purified compound by NMR identified cinerubin B (calculated mass = 825.32079 Da, Δm = 4.6 ppm), a highly bioactive polyketide first characterized from *Streptomyces antibioticus* [19] (Figure 27, Table 5). The fast

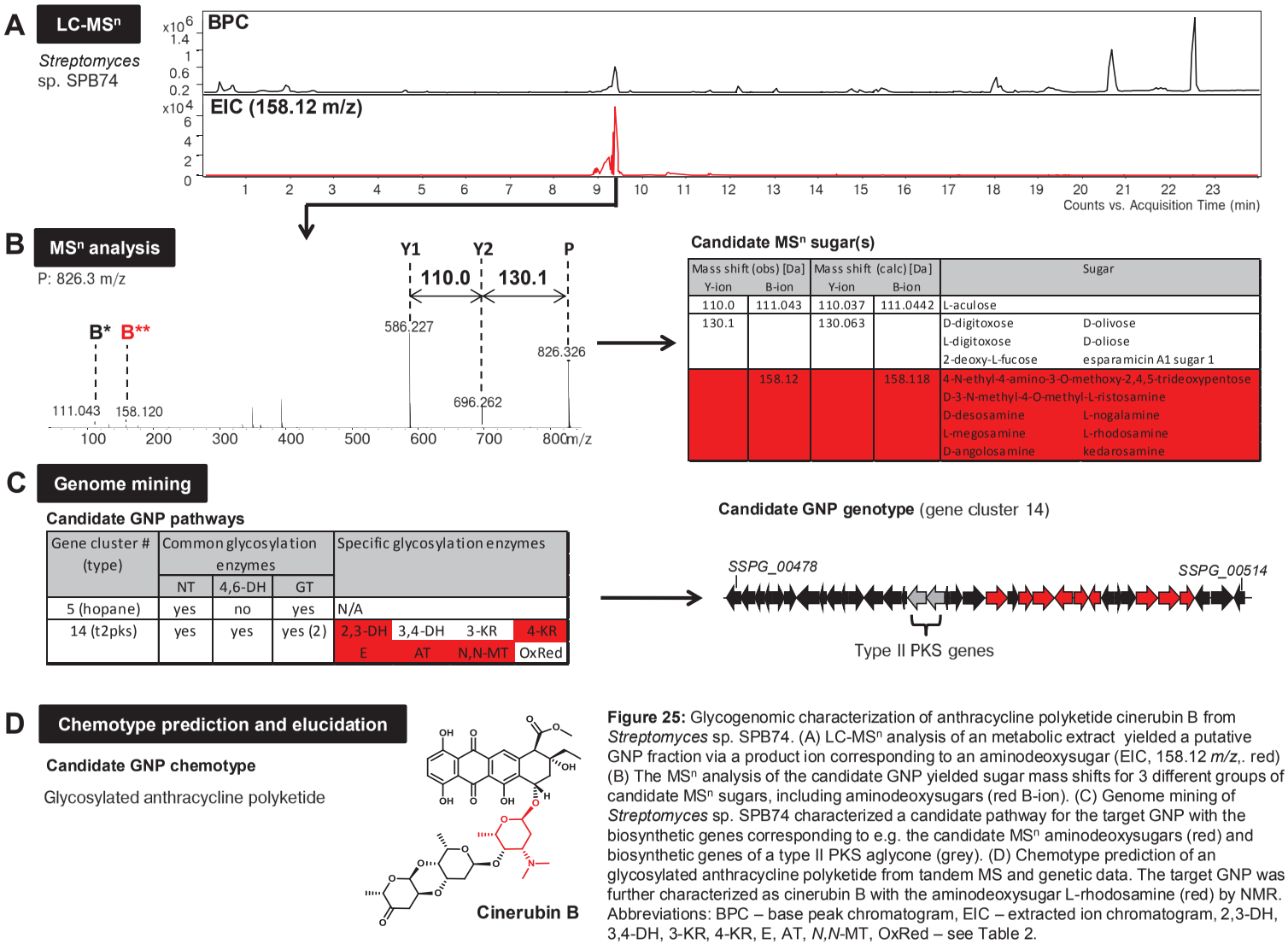
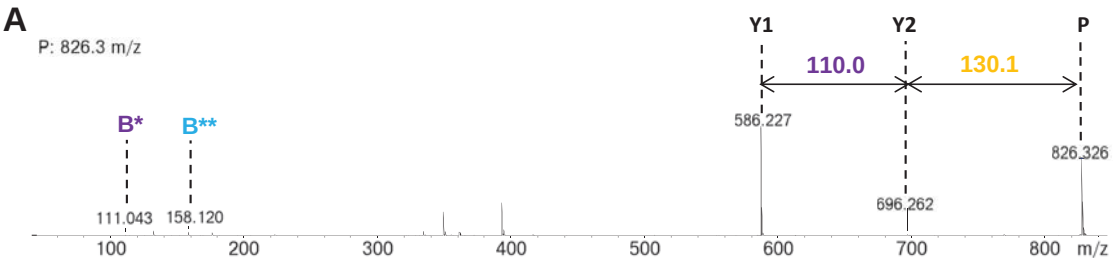


Figure 25: Glycogenomic characterization of anthracycline polyketide cinerubin B from *Streptomyces* sp. SPB74. (A) LC-MSⁿ analysis of an metabolic extract yielded a putative GNP fraction via a product ion corresponding to an aminodeoxysugar (EIC, 158.12 *m/z*,. red) (B) The MSⁿ analysis of the candidate GNP yielded sugar mass shifts for 3 different groups of candidate MSⁿ sugars, including aminodeoxysugars (red B-ion). (C) Genome mining of *Streptomyces* sp. SPB74 characterized a candidate pathway for the target GNP with the biosynthetic genes corresponding to e.g. the candidate MSⁿ aminodeoxysugars (red) and biosynthetic genes of a type II PKS aglycone (grey). (D) Chemotype prediction of an glycosylated anthracycline polyketide from tandem MS and genetic data. The target GNP was further characterized as cinerubin B with the aminodeoxysugar L-rhodosamine (red) by NMR. Abbreviations: BPC – base peak chromatogram, EIC – extracted ion chromatogram, 2,3-DH, 3,4-DH, 3-KR, 4-KR, E, AT, *N,N*-MT, OxRed – see Table 2.



B

Y-Mass shift [Da]	B-mass shift [Da]	C-mass shift [Da]	Sugar	Glycosylation genes: NT	4,6-DH	2,3-DH	3,4-DH	3-KR	4-KR	(3,5-)E	3-AT	4-AT	N-MT	N,N-MT	O-MT	N-ET	oxDA	OxRed
110.0368	111.0440	129.0545	L-aculose															
130.0630	131.0702	149.0808	D-digitoxose															
130.0630	131.0702	149.0808	L-digitoxose															
130.0630	131.0702	149.0808	2-deoxy-L-fucose															
130.0630	131.0702	149.0808	D-olivose															
130.0630	131.0702	149.0808	D-oliose															
130.0630	131.0702	149.0808	L-boivinose	N/A														
157.1103	158.1175	176.1280	4-N-ethyl-4-amino-3-O-methoxy-2,4,5-trideoxypentose															
157.1103	158.1175	176.1280	D-3-N-methyl-4-O-methyl-L-ristosamine															
157.1103	158.1175	176.1280	D-desosamine															
157.1103	158.1175	176.1280	L-megosamine															
157.1103	158.1175	176.1280	nogalamine															
157.1103	158.1175	176.1280	L-rhodosamine															
157.1103	158.1175	176.1280	D-angolosamine															
157.1103	158.1175	176.1280	kedarosamine	N/A														

C

Gene	Reference	Size [aa]	Predicted function	Closest homolog [sim/id, %]
SSBG_00478	EDY42516.1	284	daunorubicin resistance protein	BAB72060.1| AcrW [Streptomyces galilaeus] (93/89)
SSBG_00479	EDY42517.2	326	daunorubicin resistance protein DrrA	BAB72059.1| AcrV [Streptomyces galilaeus] (93/86)
SSBG_00481	EDY42519.2	206	transcriptional regulator	BAB72058.1| AclS [Streptomyces galilaeus] (80/74)
SSBG_00482	EDY42520.2	144	aglycone C-1 hydroxylase	BAB72057.1| AclR [Streptomyces galilaeus] (95/88)
SSBG_00483	EDY42521.1	269	putative hydroxylase	BAB72056.1| AclQ [Streptomyces galilaeus] (83/74)
SSBG_00484	EDY42522.1	242	N-methyltransferase	BAB72055.1| AclP [Streptomyces galilaeus] (86/80)
SSBG_00485	EDY42523.2	519	oxidoreductase (OxRed)	ABI15166 | AknOx [Streptomyces galilaeus] (corrected from AclO)
SSBG_00486	EDY42524.1	159	hypothetical protein	BAB72053.1| AclJ [Streptomyces galilaeus] (87/81)
SSBG_00487	EDY42525.1	281	transcriptional regulatory protein	BAB72052.1| AclI [Streptomyces galilaeus] (92/88)
SSBG_00488	EDY42526.1	144	CYC, cyclase	BAB72051.1| AknH [Streptomyces galilaeus] (96/91)
SSBG_00489	EDY42527.1	286	MET, methyltransferase	BAB72050.1| AknG [Streptomyces galilaeus] (90/87)
SSBG_00490	EDY42528.2	365	Propionate starter unit, AT	BAB72049.1| AknF [Streptomyces galilaeus] (86/82)
SSBG_00491	EDY42529.1	354	Propionate starter unit, KSIII	BAB72048.1| AknE2 [Streptomyces galilaeus] (88/83)
SSBG_00492	EDY42530.1	92	minPKS, ACP	BAB72047.1| AknD [Streptomyces galilaeus] (88/81)
SSBG_00493	EDY42531.1	412	minPKS, AT	BAB72046.1| AknC [Streptomyces galilaeus] (88/81)
SSBG_00494	EDY42532.1	424	minPKS, KS	BAB72045.1| AknB [Streptomyces galilaeus] (95/89)
SSBG_00495	EDY42533.2	118	OXY, oxygenase	BAB72044.1| AknX [Streptomyces galilaeus] (77/75)
SSBG_00496	EDY42534.1	261	KR, ketoreductase	BAB72043.1| AknA [Streptomyces galilaeus] (93/87)
SSBG_00497	EDY42535.1	450	ARO, aromatase	BAB72042.1| AknE1 [Streptomyces galilaeus] (89/83)
SSBG_00498	EDY42536.1	440	glycosyltransferase (GT)	BAB72041.1| AclK [Streptomyces galilaeus] (94/88)
SSBG_00499	EDY42537.1	205	epimerase (E)	BAB72040.1| AclL [Streptomyces galilaeus] (93/88)
SSBG_00500	EDY42538.2	311	4-ketoreductase (4-KR)	BAB72039.1| AclM [Streptomyces galilaeus] (81/75)
SSBG_00501	EDY42539.2	448	2,3-dehydratase (2,3-DH)	BAB72038.1| AclN [Streptomyces galilaeus] (86/81)
SSBG_00502	EDY42540.1	369	aminotransferase (AT)	BAB72037.1| AclZ [Streptomyces galilaeus] (91/87)
SSBG_00503	EDY42541.1	291	TDP-glucose synthase (NT)	BAB72036.1| AclY [Streptomyces galilaeus] (96/91)
SSBG_00504	EDY42542.1	239	N,N-dimethyltransferase (N,N-MT)	AAF73460.1|AF264025_11 AknX2 [Streptomyces galilaeus] (92/86)
SSBG_00505	EDY42543.1	259	CYC, cyclase	AAF73459.1|AF264025_10 AknW [Streptomyces galilaeus] (96/90)
SSBG_00507	EDY42544.1	144	putative oxidoreductase	AAF73458.1|AF264025_9 AknV [Streptomyces galilaeus] (85/81)
SSBG_00508	EDY42546.1	276	C-7 KR, ketoreductase	AAF73457.1|AF264025_8 AknU [Streptomyces galilaeus] (87/82)
SSBG_00509	EDY42547.2	445	P450-like, GTF activation	AAF73456.1|AF264025_7 AknT [Streptomyces galilaeus] (79/72)
SSBG_00510	EDY42548.2	438	glycosyltransferase (GT)	AAF73455.1|AF264025_6 AknS [Streptomyces galilaeus] (89/83)
SSBG_00511	EDY42549.1	328	4,6-dehydratase (4,6-DH)	AAF73454.1|AF264025_5 AknR [Streptomyces galilaeus] (89/84)
SSBG_00512	EDY42550.2	300	3-ketoreductase (3-KR)	AAF73453.1|AF264025_4 AknQ [Streptomyces galilaeus] (90/85)
SSBG_00513	EDY42551.2	440	3,4-dehydratase (3,4-DH)	AAF73452.1|AF264025_3 AknP [Streptomyces galilaeus] (97/92)
SSBG_00514	EDY42552.1	319	transcriptional regulatory protein	AAF73451.1|AF264025_2 AknO [Streptomyces galilaeus] (93/87)

Figure 26: Glycogenomic characterization of cinerubin B, a glycosylated anthracycline polyketide, from *Streptomyces sp.* SPB74. (A) Tandem MS spectrum of cinerubin B with Y-ion mass shifts (purple, orange) and B-ions (blue) corresponding to putative sugar monomers. (B) Characterization of candidate MS^n sugars from cinerubin B with corresponding glycosylation genes by the MS-glycogenetic code (Table 2). (C) Gene cluster analysis of candidate cinerubin B pathway with highlighted glycosylation genes (red) and aglycone biosynthetic genes (grey).

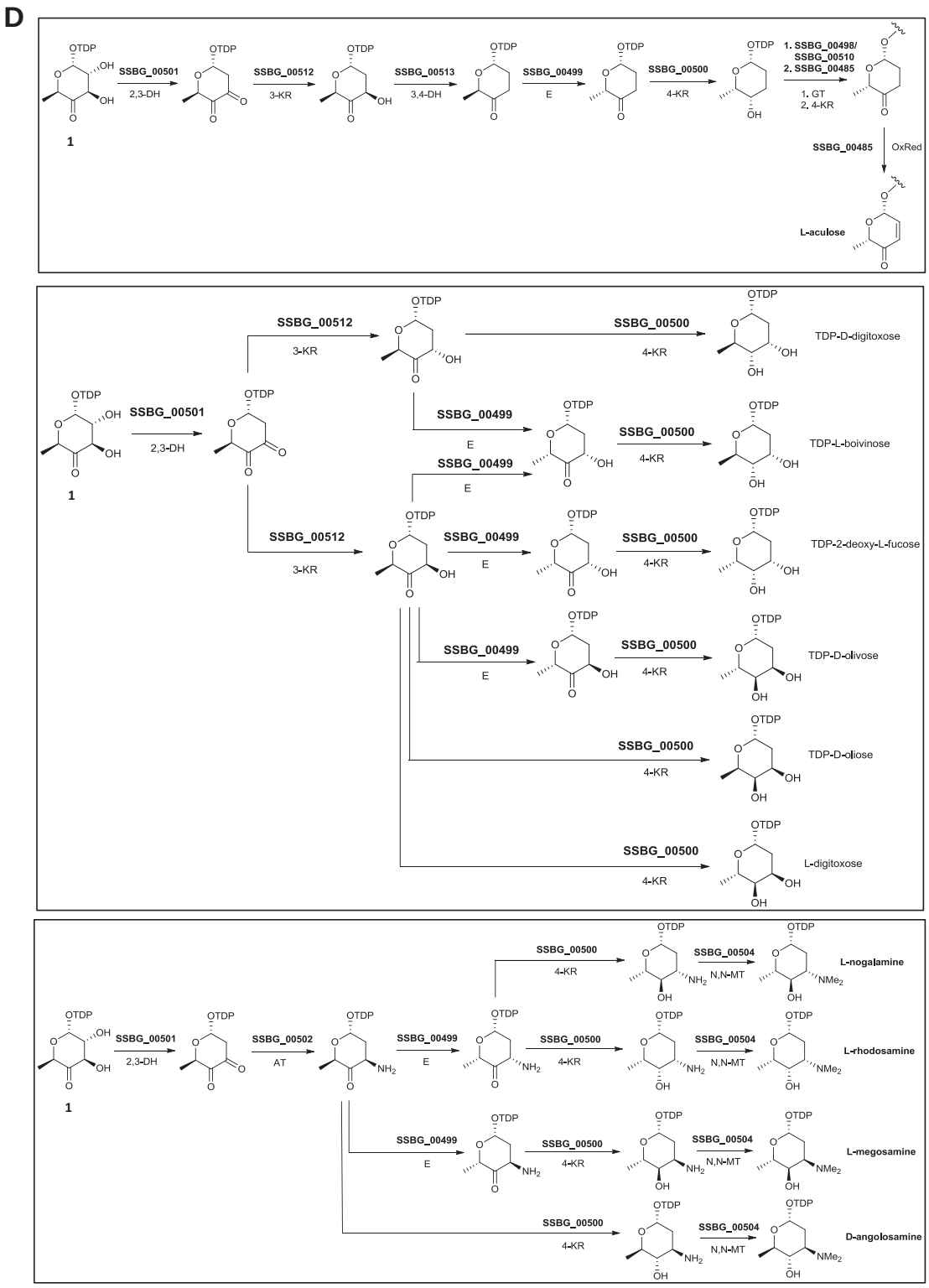


Figure 26: MS-glycogenomic characterization of cinerubin B, a glycosylated anthracycline polyketide, from *Streptomyces* sp. SPB74. (D) Matching pathways of specific glycosylation genes with candidate MS/MS sugars from cinerubin B, each starting at deoxysugar biosynthetic intermediate TDP-4-keto-6-deoxy-α-D-glucose (1).

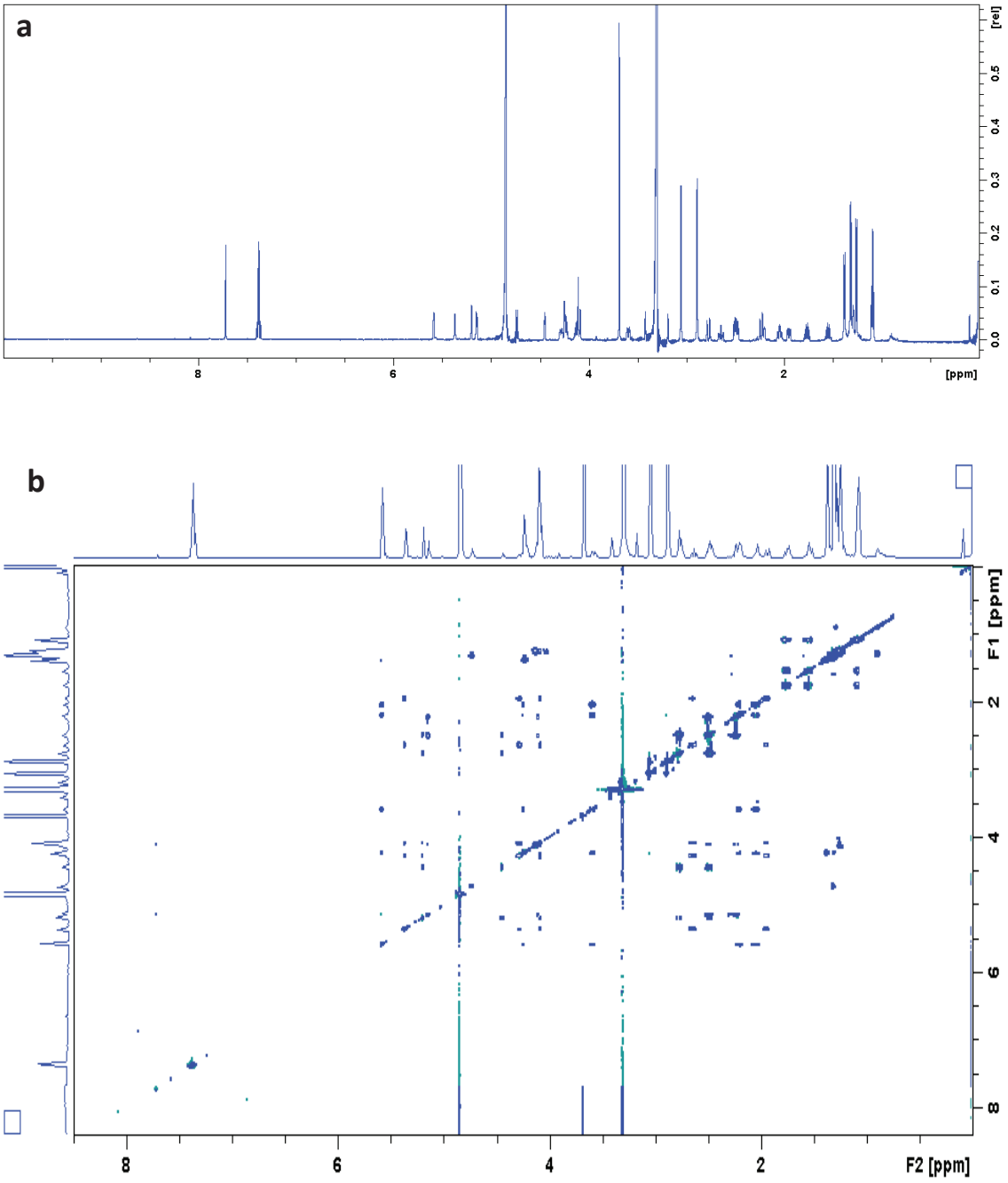


Figure 27: NMR spectra of cinerubin B (1-hydroxyaclacinomycin A). 1D and 2D NMR analysis of cinerubin B could verify the candidate deoxysugars of the glycogenomic analysis as rhodosamine (Rhn), 2'-deoxyfucose (dFuc) and cinerulose B (CinB) which is attached via a 1''',2'''-*O,O*-di-glycosidic bond to 2'-deoxyfucose. The sugar stereochemistry was assigned based on a ^{1}H-^{1}H NOESY experiment. (a) ^{1}H NMR spectrum. The spectrum was observed in MeOD-d4, 600 MHz. The detailed annotations were listed in Table 5. (b),(c) ^{1}H-^{1}H TOCSY spectra and annotations. The spectrum was observed in MeOD-d4, 600 MHz, with a mixing time of 90 ms. Subfigure b is a full spectrum, subfigure c is a zoom in the spectrum with annotations of the sugar spin systems. (d),(f) ^{1}H-^{13}C HMBC spectrum and annotations. The spectrum was observed in MeOD-d4, 600 MHz with $^{2,3}J_{1H/13C}$ = 7Hz. Subfigure d is a full spectrum and subfigure f shows HMBC annotations. (e),(g) ^{1}H-^{13}C HSQC spectrum with annotations. The spectrum was observed in MeOD-d4, 600 MHz, with $^{1}J_{1H/13C}$ = 145 Hz. Subfigure e is a full spectrum and subfigure g shows HMBC annotations. (h) ^{1}H-^{1}H NOESY spectrum and annotations. The spectrum was observed in MeOD-d4, 600 MHz, with a Bruker NMR .

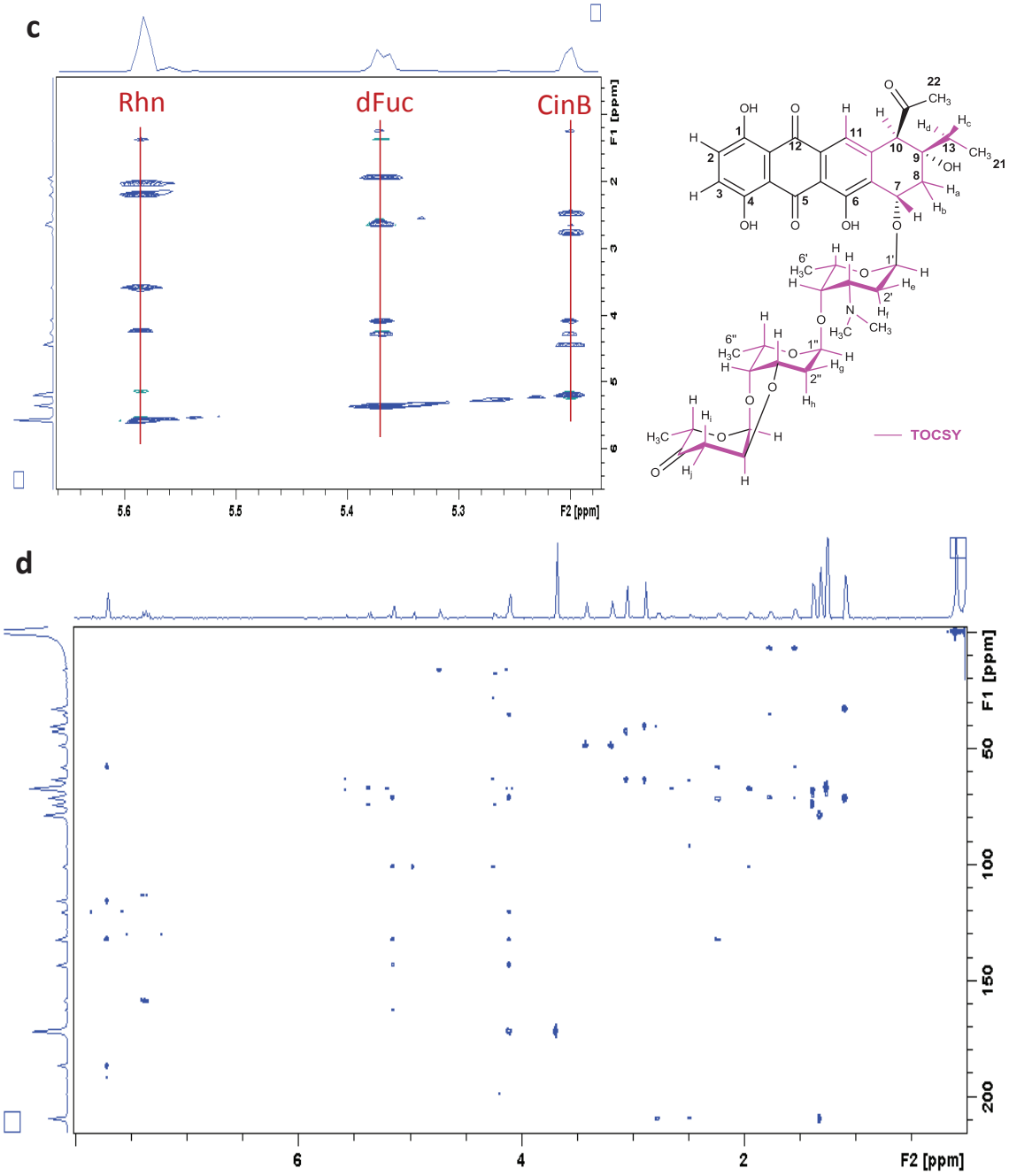


Figure 27: NMR spectra of cinerubin B (1-hydroxyaclacinomycin A). 1D and 2D NMR analysis of cinerubin B could verify the candidate deoxysugars of the glycogenomic analysis as rhodosamine (Rhn), 2'-deoxyfucose (dFuc) and cinerulose B (CinB) which is attached via a 1''',2'''-*O,O*-di-glycosidic bond to 2'-deoxyfucose. The sugar stereochemistry was assigned based on a ^{1}H-^{1}H NOESY experiment. (a) ^{1}H NMR spectrum. The spectrum was observed in MeOD-d4, 600 MHz. The detailed annotations were listed in Table 5. (b),(c) ^{1}H-^{1}H TOCSY spectra and annotations. The spectrum was observed in MeOD-d4, 600 MHz, with a mixing time of 90 ms. Subfigure b is a full spectrum, subfigure c is a zoom in the spectrum with annotations of the sugar spin systems. (d),(f) ^{1}H-^{13}C HMBC spectrum and annotations. The spectrum was observed in MeOD-d4, 600 MHz with $^{2,3}J_{1H/13C}$ = 7Hz. Subfigure d is a full spectrum and subfigure f shows HMBC annotations. (e),(g) ^{1}H-^{13}C HSQC spectrum with annotations. The spectrum was observed in MeOD-d4, 600 MHz, with $^{1}J_{1H/13C}$ = 145 Hz. Subfigure e is a full spectrum and subfigure g shows HMBC annotations. (h) ^{1}H-^{1}H NOESY spectrum and annotations. The spectrum was observed in MeOD-d4, 600 MHz, with a Bruker NMR .

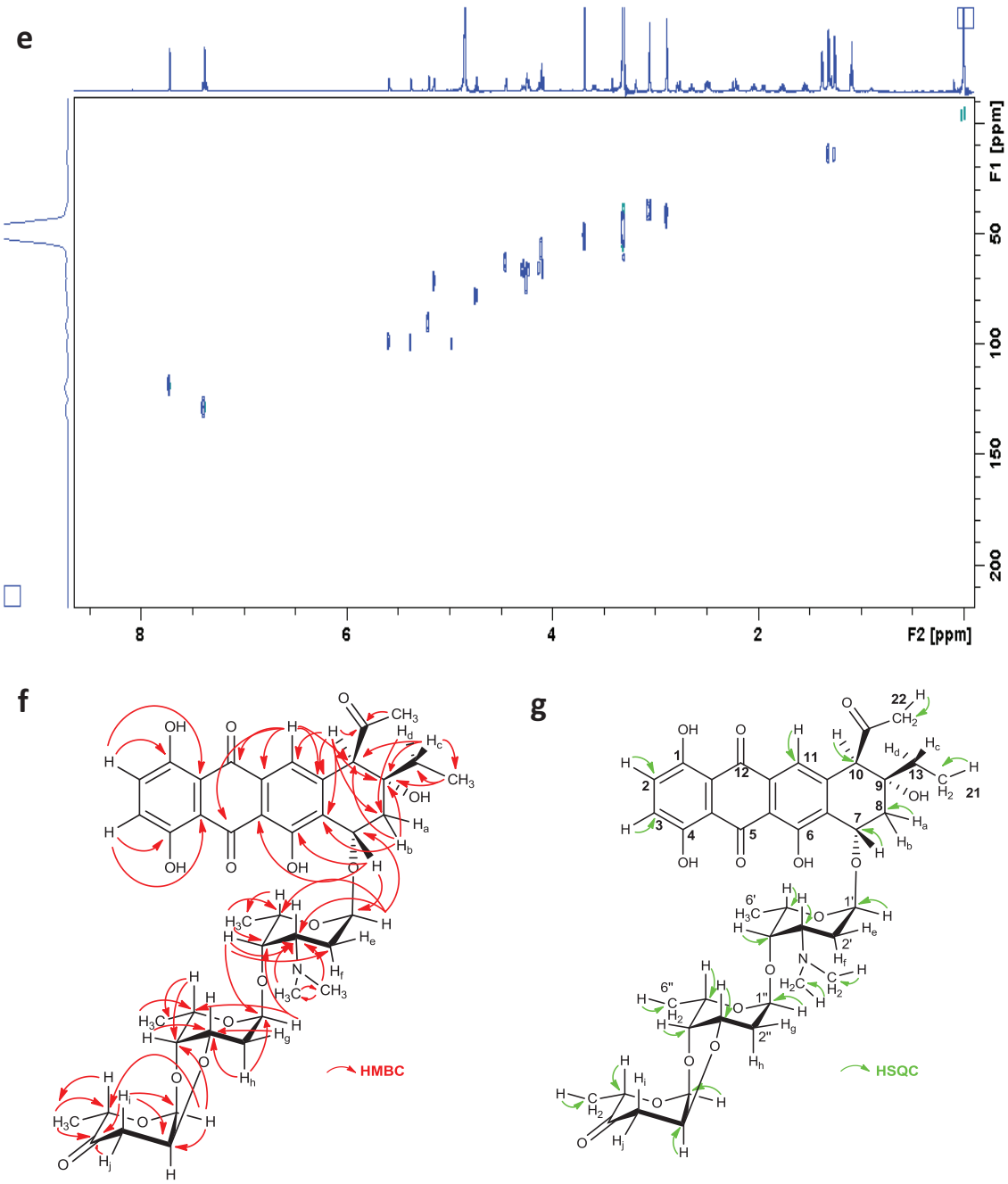


Figure 27: NMR spectra of cinerubin B (1-hydroxyaclacinomycin A). 1D and 2D NMR analysis of cinerubin B could verify the candidate deoxysugars of the glycogenomic analysis as rhodosamine (Rhn), 2'-deoxyfucose (dFuc) and cinerulose B (CinB) which is attached via a 1''',2'''-*O,O*-di-glycosidic bond to 2'-deoxyfucose. The sugar stereochemistry was assigned based on a ^{1}H-^{1}H NOESY experiment. (a) ^{1}H NMR spectrum. The spectrum was observed in MeOD-d4, 600 MHz. The detailed annotations were listed in Table 5. (b),(c) ^{1}H-^{1}H TOCSY spectra and annotations. The spectrum was observed in MeOD-d4, 600 MHz, with a mixing time of 90 ms. Subfigure b is a full spectrum, subfigure c is a zoom in the spectrum with annotations of the sugar spin systems. (d),(f) ^{1}H-^{13}C HMBC spectrum and annotations. The spectrum was observed in MeOD-d4, 600 MHz with $^{2,3}J_{1H/13C}$ = 7Hz. Subfigure d is a full spectrum and subfigure f shows HMBC annotations. (e),(g) ^{1}H-^{13}C HSQC spectrum with annotations. The spectrum was observed in MeOD-d4, 600 MHz, with $^{1}J_{1H/13C}$ = 145 Hz. Subfigure e is a full spectrum and subfigure g shows HMBC annotations. (h) ^{1}H-^{1}H NOESY spectrum and annotations. The spectrum was observed in MeOD-d4, 600 MHz, with a Bruker NMR .

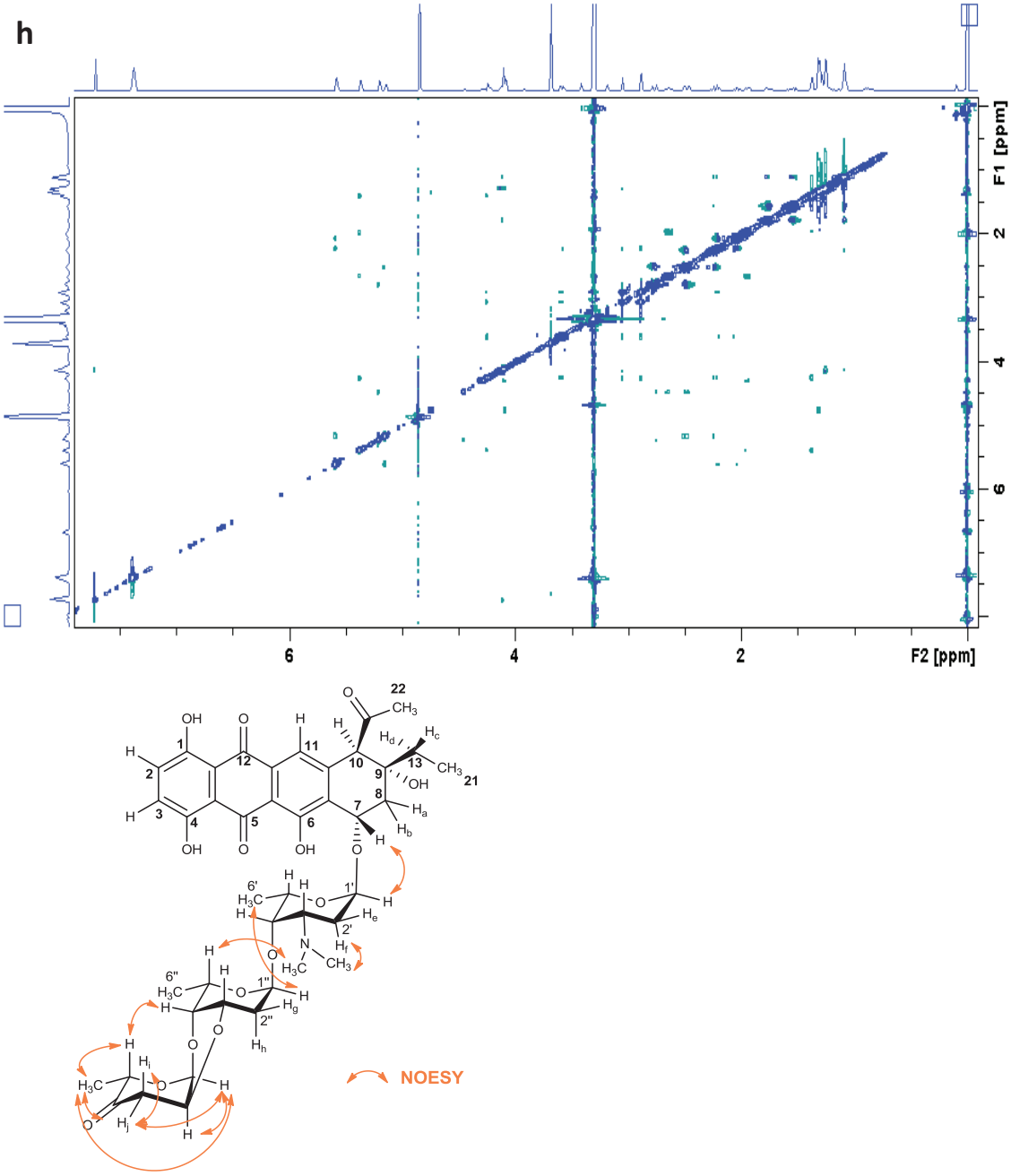


Figure 27: NMR spectra of cinerubin B (1-hydroxyaclacinomycin A). 1D and 2D NMR analysis of cinerubin B could verify the candidate deoxysugars of the glycogenomic analysis as rhodosamine (Rhn), 2'-deoxyfucose (dFuc) and cinerulose B (CinB) which is attached via a 1''',2'''-*O,O*-di-glycosidic bond to 2'-deoxyfucose. The sugar stereochemistry was assigned based on a ^{1}H-^{1}H NOESY experiment. (a) ^{1}H NMR spectrum. The spectrum was observed in MeOD-d4, 600 MHz. The detailed annotations were listed in Table 5. (b),(c) ^{1}H-^{1}H TOCSY spectra and annotations. The spectrum was observed in MeOD-d4, 600 MHz, with a mixing time of 90 ms. Subfigure b is a full spectrum, subfigure c is a zoom in the spectrum with annotations of the sugar spin systems. (d),(f) ^{1}H-^{13}C HMBC spectrum and annotations. The spectrum was observed in MeOD-d4, 600 MHz with $^{2,3}J_{1H/13C}$ = 7Hz. Subfigure d is a full spectrum and subfigure f shows HMBC annotations. (e),(g) ^{1}H-^{13}C HSQC spectrum with annotations. The spectrum was observed in MeOD-d4, 600 MHz, with $^{1}J_{1H/13C}$ = 145 Hz. Subfigure e is a full spectrum and subfigure g shows HMBC annotations. (h) ^{1}H-^{1}H NOESY spectrum and annotations. The spectrum was observed in MeOD-d4, 600 MHz, with a Bruker NMR .

characterization of cinerubin B as a glycosylated anthracycline polyketide from a standard LC-MS^n run of a crude microbial extract of a genome sequenced microbe showed the feasibility of the glycogenomic approach in connecting a GNP chemotype with its genotype.

4.2.3 Glycogenomic characterization of a new arenimycin chemotype and genotype from *Salinispora arenicola* CNB-527

LC-MS^n analysis of an organic extract of marine actinobacterium *Salinispora arenicola* CNB-527, a well-studied and prolific secondary metabolite producer [20], yielded a compound of mass 808 Da that showed a forosamine sugar B-ion EIC (142.12 *m/z*, Figure 28A). The MS^2 spectrum of the compound also showed a forosamine Y-ion mass shift (141.12 m/z) and another putative methyldeoxysugar mass shift in the B- and Y-ion series (Figure 28B). The candidate MS^n methyldeoxysugars were digitalose, *O*-methylrhamnose and 6-deoxy-3-*C*-methylmannose (Figure 28B). The AntiSMASH analysis of the *Salinispora* arenicola CNB-527 genome revealed four gene clusters with putative glycosylated products – a type II PKS pathway, an indole pathway, an enediyne PKS pathway and a type I PKS pathway. The type II PKS gene cluster and the indole gene cluster both had the specific glycosylation genes for digitalose or *O*-methylrhamnose biosynthesis, i.e. a 4-ketoreductase, an epimerase and an *O*-methyltransferase (Figure 28C, blue). However, only the type II PKS cluster had the specific genes for a forosamine glycosylation, i.e. a 2,3-dehydratase, 3,4-dehydratase, 3-ketoreductase, an aminotransferase, and an *N,N*-dimethyltransferase (Figure 29). The target type II PKS gene cluster from *S. arenicola* CNB-527 had not been yet been characterized (Figure 29). The presence of type II PKS genes and the characterized glycosylation genes and MS^n data again indicated an aromatic polyketide product with a diglycosyl group. It was hypothesized that the characterized compound is a new arenimycin derivative, 'arenimycin B', with an additional D-forosamine glycosyl group attached at the 4-hydroxyl group of the 2-*O*-methyl-L-rhamnose unit of arenimycin (Figure 28D*)*, which was previously characterized from another *Salinispora arenicola* strain, CNR-647 [21]. We further identified arenimycin during $LCMS^n$ analysis that showed a similar tandem MS spectrum

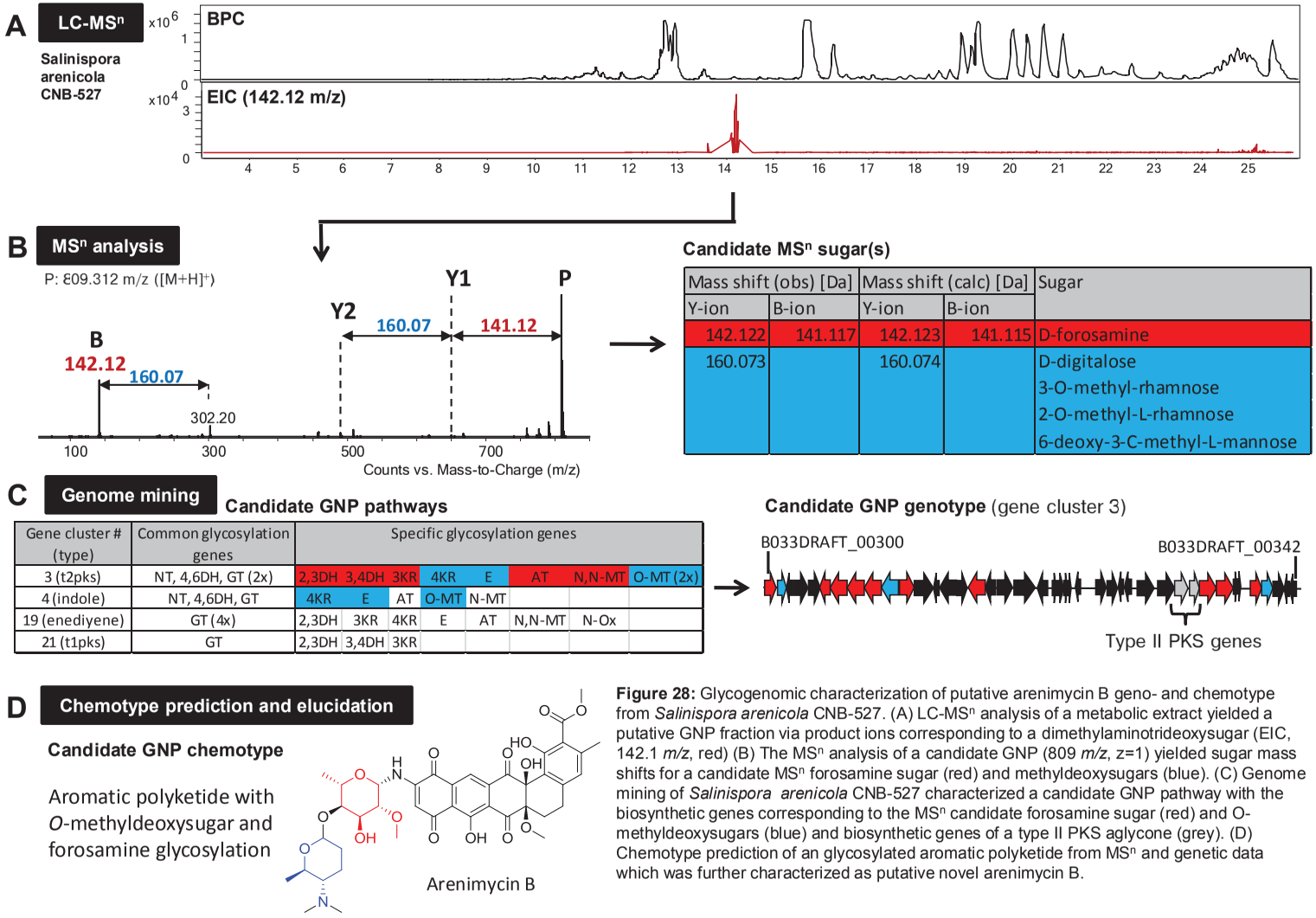


Figure 28: Glycogenomic characterization of putative arenimycin B geno- and chemotype from *Salinispora arenicola* CNB-527. (A) LC-MS^n analysis of a metabolic extract yielded a putative GNP fraction via product ions corresponding to a dimethylaminotrideoxysugar (EIC, 142.1 *m/z*, red) (B) The MS^n analysis of a candidate GNP (809 *m/z*, z=1) yielded sugar mass shifts for a candidate MS^n forosamine sugar (red) and methyldeoxysugars (blue). (C) Genome mining of *Salinispora arenicola* CNB-527 characterized a candidate GNP pathway with the biosynthetic genes corresponding to the MS^n candidate forosamine sugar (red) and O-methyldeoxysugars (blue) and biosynthetic genes of a type II PKS aglycone (grey). (D) Chemotype prediction of an glycosylated aromatic polyketide from MS^n and genetic data which was further characterized as putative novel arenimycin B.

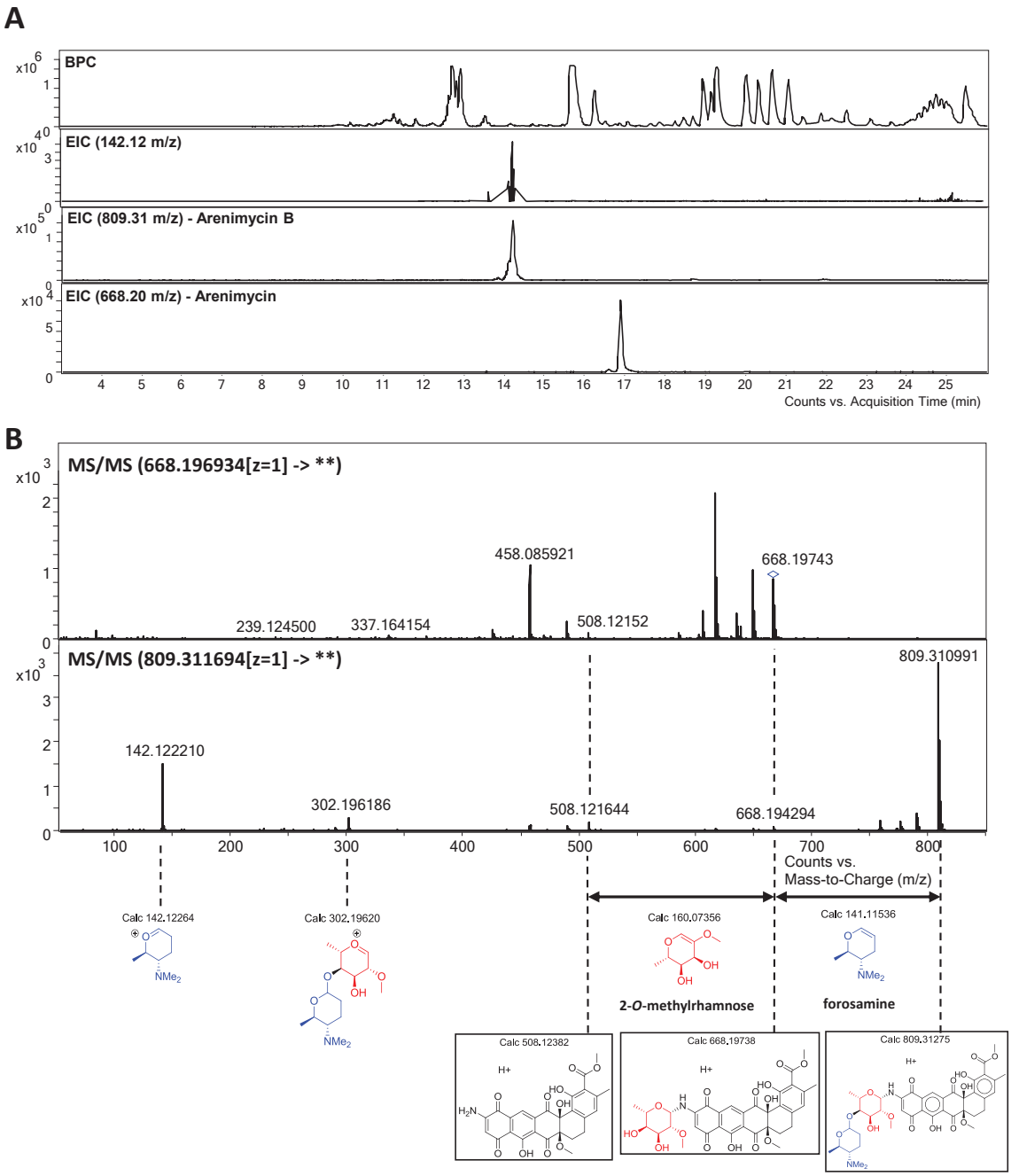


Figure 29: LCMS and MS/MS characterization of arenimycins. (A) LCMS profiles of a crude ethylacetate extract of *S. arenicola* CNB-527. Abbreviations: BPC – base peak chromatogram, EIC – extracted ion chromatogram. (B) MS/MS spectra of arenimycin (top) and arenimycin B (bottom) with structural peak assignments.

to ‘arenimycin B’ except for the putative forosamine Y- and B-ion shifts (Figure 29). Purification, NMR structure elucidation and bioactivity tests are ongoing.

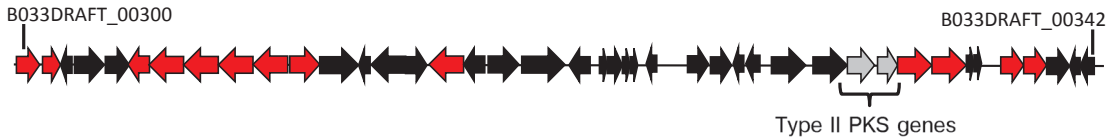
Arenimycin is a benzo[α]naphthacene quinone natural product first isolated from *Salinispora arenicola* CNR-647 [21]. It is structurally closely related to SF2446B1 from *Streptomyces* sp. SF2446 [21] and has antibacterial activity against rifampin- and methicillin-resistant *Staphylococcus aureus* [22]. The detection of arenimycin A and ‘B’ indicate that the corresponding type II PKS pathway codes for the biosynthesis of arenimycins in general. The arenimycin gene cluster has not been characterized to date but a biosynthetic gene cluster of structurally similar pradimicin was characterized [23]. Pradimicin shares the benzo[α]naphthacene quinone core with arenimycins but has different hydroxylation regiochemistry, quinone patterns and glycosylation sites and groups [24]. However, the putative biosynthetic genes of the pradimicin benzo[α]naphtacene quinone core are present in the arenimycin gene cluster (Figure 30). In contrast to the pradimicin cluster, the arenimycin pathway comprises a flavin-dependent monooxygenase that has homology to TcmG, a monooxygenase from the tetrenomycin pathway [25], and that may explain the different hydroxylation chemistry at positions 6a and 14a. Ultimately, the characterization of a new putative bioactive glycosylated compound and its biosynthetic pathway from *Salinispora arenicola* CNB-527 highlights how targeting glycosylation on small molecules as a genome mining approach can lead to a fast discovery of bioactive molecules and their biosynthetic pathways.

4.3 Discussion

In this study, a new experiment-guided genome mining strategy was introduced to characterize glycosylated natural products with their biosynthetic gene clusters in microbial genome sequences. The glycogenomic approach is based on a MS-glycogenetic code that connects predictable glycosylation fragments from MS^n experiments of GNPs with their glycosylation genes in microbial genomes. The approach led to the rapid characterization of

cinerubin B, a glycosylated anthracycline antibiotic, from *Streptomyces sp.* SPB74, and to the discovery of

A



Gene	Size [aa]	Predicted function	Closest homolog (Sim/Id, %/%)	Reference	Pradimicin pathway homolog
B033DRAFT_00300	330	NT	dTDP-glucose synthase [Micromonospora griseorubida] (82/73)	BAC57039	
B033DRAFT_00301	218	E	dTDP-4-dehydrorhamnose 3,5-epimerase [Micromonospora lupini str. Lupac 08] (82/68)	ZP_21030398	
B033DRAFT_00302	169	regulator	MarR family transcriptional regulator [Stackebrandtia nassauensis DSM 44728] (59/46)	YP_003512399	
B033DRAFT_00303	316	transporter	daunorubicin ABC transporter ATPase [Actinosynnema mirum DSM 43827] (74/64)	YP_003098925	
B033DRAFT_00304	280	transporter	ABC transporter [Actinosynnema mirum DSM 43827] (71/53)	YP_003098924	
B033DRAFT_00305	247	N,N-MT	probable N-dimethyltransferase [Saccharopolyspora spinosa] (75/60)	AAG23280	
B033DRAFT_00306	386	AT	probable aminotransferase [Saccharopolyspora spinosa] (79/65)	AAG23279	
B033DRAFT_00307	448	3,4-DH	putative 3-dehydratase [Streptomyces galilaeus] (86/77)	AAF73452	
B033DRAFT_00308	401	GT-like (P450)	AknT [Streptomyces galilaeus] (44/36)	AAF73456	
B033DRAFT_00309	376	O-MT	putative sugar O-methyltransferase [Streptomyces sp. SANK 60405] (56/43)	BAJ05903	
B033DRAFT_00310	411	GT	glycosyl transferase [Saccharopolyspora erythraea NRRL 2338] (63/45)	YP_001102993	
B033DRAFT_00311	490	deacetylase	acetylornithine deacetylase [Microvirga sp. WSM3557] (56/36)	ZP_10180744	
B033DRAFT_00312	114	NK	nucleoside diphosphate kinase [Listeria ivanovii FSL F6-596] (54/30)	ZP_07874356	
B033DRAFT_00313	411	cytochrome P450	cytochrome P450 [Streptomyces davawensis JCM 4913] (57/40)	CCK25399	
B033DRAFT_00314	243	hypothetical protein	LmbE family protein [Acidothermus cellulolyticus 11B] (68/58)	YP_872134	
B033DRAFT_00315	383	GT	ChaGT1 protein [Streptomyces chartreusis] (56/41)	CAH10164	
B033DRAFT_00316	325	reductase	aldo/keto reductase [Streptomyces flavogriseus ATCC 33331] (64/51)	YP_004926245	
B033DRAFT_00317	346	methyltransferase	PdmT [Actinomadura hibisca] (60/47)	ABK58689	
B033DRAFT_00318	540	monooxygenase	putative FAD-dependent monooxygenase GrhO8 [Streptomyces sp. JP95] (69/54)	AAM33675	
B033DRAFT_00319	276	regulator	DNA-binding transcriptional activator of the SARP family [Frankia sp. QA3] (55/42)	ZP_10309867	
B033DRAFT_00320	117	cyclase	polyketide cyclase [Streptomyces sp. SANK 61196] (74/59)	ADG86313	PdmK
B033DRAFT_00321	131	cyclase	Cupin 2 barrel domain-containing protein [Micromonospora aurantiaca ATCC 27029] (69/53)	YP_003836600	PdmL
B033DRAFT_00322	161	cyclase	Aromatase WhiE VI [Streptomyces gancidicus BKS 13-15] (78/67)	EMF26754	PdmD
B033DRAFT_00323	120	oxygenase	antibiotic biosynthesis monooxygenase [Amycolatopsis mediterranei U32] (68/47)	YP_003767590	PdmI
B033DRAFT_00324	120	biotin carboxylase	acetyl-CoA carboxylase, carboxyl transferase, beta subunit [Frankia sp. EUN1f] (80/67) (PrmP1)	ZP_06417970	
B033DRAFT_00325	286	regulator	SARP family regulator [Streptomyces lavendulae] (68/50)	BAG74714	
B033DRAFT_00326	280	methyltransferase	Chal protein [Streptomyces chartreusis] (56/42)	CAH10176	
B033DRAFT_00327	139	hypothetical protein	hypothetical protein RER_29830 [Rhodococcus erythropolis PR4] (55/40)	YP_002766430	
B033DRAFT_00328	151	oxygenase	RubQ [Streptomyces collinus] (64/49)	AAM97373	
B033DRAFT_00329	412	cytochrome P450	cytochrome P450 [Streptomyces avermitilis MA-4680] (57/43)	NP_823237	
B033DRAFT_00331	422	KS	PdmA [Actinomadura hibisca] (83/74)	ABM21747	PdmA
B033DRAFT_00332	405	CLF	PdmB [Actinomadura hibisca] (74/65)	ABM21748	PdmB
B033DRAFT_00333	241	reductase	putative 3-oxoacyl-ACP reductase [Streptomyces sp. TA-0256] (75/61)	ABM21753	PdmG
B033DRAFT_00334	458	2,3DH	NDP-hexose 2,3-dehydratase [Frankia sp. CcI3] (75/62)	YP_483211	
B033DRAFT_00335	332	3KR	NDP-hexose-3-ketoreductase [Streptomyces lydicus] (68/57)	CBA11563	
B033DRAFT_00336	84	ACP	acyl carrier protein [Streptomyces arenae] (71/54)	AAD20269	PdmC
B033DRAFT_00337	108	oxygenase	PdmH [Actinomadura hibisca] (72/58)	ABM21754	PdmH
B033DRAFT_00338	339	4,6DH	dTDP-glucose 4,6-dehydratase [Streptomyces sp. SirexAA-E] (83/73)	YP_004802450	
B033DRAFT_00339	283	4KR	putative dTDP-4-keto-6-deoxyhexose reductase [Actinoplanes missouriensis 431] (80/74)	YP_005462169	
B033DRAFT_00340	206	regulator	Transcriptional regulator ArsR family [Patulibacter sp. I11] (61/42)	ZP_09470860	
B033DRAFT_00341	69	carboxylase	carboxylase [Streptomyces ghanaensis ATCC 14672] (65/54)	ZP_06581973	
B033DRAFT_00342	243	reductase	putative NADPH-dependent FMN reductase [Actinoplanes missouriensis 431] (80/72)	YP_005464808	

Figure 30: Glycogenomic connection of putative arenimycin B with its biosynthetic gene cluster from *Salinispora arenicola* CNB-527. (A) Gene cluster analysis of candidate arenimycin B pathway, with highlighted glycosylation genes (red) and aglycone biosynthetic genes (grey).

B

Y-Mass shift [Da]	B-mass shift [Da]	C-mass shift [Da]	Sugar	Common genes	Specific BS genes										
					DH			KR		IM	AT	MT		
				NT	4,6-DH	2,3-DH	3,4-DH	3-KR	4-KR	(3,5-)E	4-AT	C3-MT	N,N-dMT	O-MT
141.115363	142.122639	160.133204	D-forosamine											
160.07356	161.080836	179.091401	D-digitalose											
160.07356	161.080836	179.091401	3-O-methyl-rhamnose											
160.07356	161.080836	179.091401	2-O-methyl-L-rhamnose											
160.07356	161.080836	179.091401	6-deoxy-3-C-methyl-L-mannose											

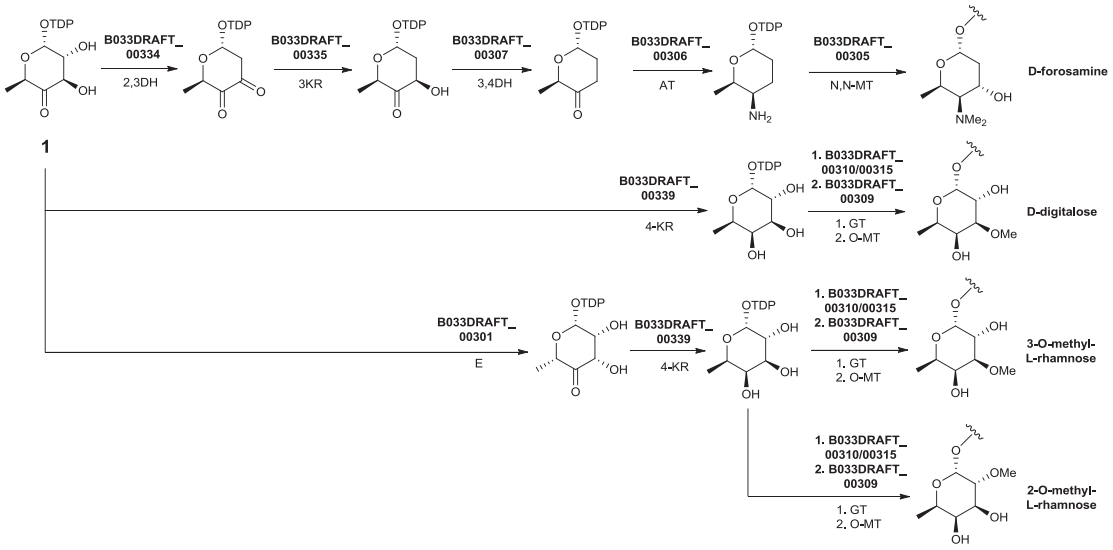


Figure 30: Glycogenomic connection of putative arenimycin B with its biosynthetic gene cluster from *Salinispora arenicola* CNB-527. (B) Matching pathways of specific glycosylation genes with candidate MS/MS forosamine sugar and O-methyldeoxysugars from arenimycin B starting at deoxysugar biosynthetic intermediate TDP-4-keto-6-deoxy-α-*D*-glucose (1).

putative ‘arenimycin B’, a new glycosylated aromatic polyketide with bioactive potential [21], and its biosynthetic gene cluster from *Salinispora arenicola* CNB-527.

Genome sequences are becoming a standard resource in microbial research [26]. In the analysis of microbial secondary metabolism, genome sequences have revealed a large pool of uncharacterized or so-called ‘cryptic’ natural product pathways as potential sources of new chemical entities that may have therapeutic potential [27]. Harvesting these orphan pathways has been mainly done by *in silico*-guided approaches in which predictions of biosynthetic genes in these pathways select the experiments to isolate the target cryptic natural products [28]. This workflow allows for the characterization of only one pathway per experiment. In light of an exponential growth of genome sequences, a one-by-one connection of an unknown chemotype with its genotype won’t match the pace of sequencing new cryptic pathways.

Experiment-guided genome mining such as the glycogenomic approach starts at the chemotype level to characterize biosynthetic building blocks of an unknown natural product that can be connected to the corresponding biosynthetic genes in the genome sequence based on current biosynthetic knowledge. This chemotype-to-genotype direction should enable a characterization of multiple cryptic pathways by initial parallel analysis of unknown secondary metabolites and subsequent genome mining of their pathways [5]. Implementation of this concept into new metabolomic [29] and metagenomic [30] approaches can facilitate studies of more complex microbiome systems where parallel characterization of metabolomes and metagenomes will require connections of expressed chemotypes with present genotypes in a more automated fashion.

The first steps of glycogenomics rely on tandem mass spectrometric identification of *O*- and *N*-glycosyl groups from microbial GNPs. These sugars can be often characterized as B-ion fragments of CID experiments in the low m/z region. *O*- and *N*-glycosylation is found in an estimated >95% of glycosylated natural products (Table 2). We implemented this fragmentation behavior in our analysis by creating EICs from LC-MS^n data for all 46 B-ion masses of the 71 known sugars involved in natural product glycosylation (Table 1). Putative GNP fractions based

on sugar EIC peaks can then be verified by identification of corresponding Y-ion neutral losses or B/C-ion fragments of the observed sugar and/or other sugar fragments. The result of the LC-MS analysis is a list of MS^n candidate sugars of a putative GNP which are used for finding the GNP genotype by genome mining their corresponding glycosylation genes in a secondary metabolic gene cluster. A limitation of GNP characterization by MS^n is a certain variability in ionization, fragmentation and fragment stability of structurally diverse GNPs. This variability in spectral outcome can be due to compound-inherent properties, e.g. a better ionization of aminosugars versus non-aminosugars, or instrument- and experiment-based differences. For example, more B/C-ions and less Y/Z-ions are observed in experiments with higher CID energies. Different instruments can also yield differences in MS^n fragment intensities or fragmentation patterns. However, the general B/Y- and C/Z-fragmentation of *O*- and *N*-glycosylated natural products applies across different mass spectrometers with CID capabilities (Figure 20). For our glycogenomic approach, Q-TOF MS instruments are better suited than ion trap MS instruments because more low-*m/z* information such as B/C-ions is observed by Q-TOF mass spectrometers. Based on tandem MS studies on C-glycosides [11], no B/Y-fragmentation is expected for C-glycosylated natural products which will prevent their connection to biosynthetic genes by the glycogenomic approach.

Genome mining of GNP pathways and connecting observed glycosyl groups with glycosyl genes in these pathways are the next steps in the MS-glycogenomic approach. First, all secondary metabolic gene clusters are predicted from a target genome by AntiSMASH [14] and analyzed for the presence of common and specific glycosylation genes. For functional prediction of glycosylation genes by BLAST, only some glycosylation genes enable a reliable sequence-based regioselectivity prediction of their enzymatic products [3], e.g. 3- vs. 4-ketoreductases and 2,3- vs. 3,4-dehydratases, whereas aminotransferase, epimerase and methyltransferase regioselectivity were not predicted in our analysis. Among methyltransferases, the methylation site was predicted in terms of its element, i.e. *N*-, *C*- or *O*-methylation. It is generally difficult to accurately predict a glycosyl group *de novo* from a set of common and specific glycosylation

genes in a GNP gene cluster because these enzymes are often promiscuous in substrates and even catalyzed reactions [31]. In addition, crosstalk with primary and secondary metabolic pathways involved in carbohydrate biosynthesis, such as cell wall formation, can sometimes lead to lack of important genes in a GNP pathway and, thus, lead to a false or no sugar prediction. For example, in a putative gene cluster of glycosylated thiopeptide Sch40832 [9,10] from *Micromonospora carbonacea* ATCC 39149, only one glycosyltransferase gene can be predicted to be involved in the glycosylation (Table 3 and 4). The other remaining common and specific genes are most likely encoded in other sugar pathways in the genome. In glycogenomic analysis, the glycosylation genes in candidate GNP pathways are used to test several natural product glycosylation hypotheses made based on the MS^n candidate sugars rather than to do *de novo* sugar prediction. A match of a putative GNP with its biosynthetic genes is made by reanalysis of MS^n data and glycosylation genes, and, ultimately, by genetic knockout or NMR structure elucidation.

Here, we introduce a new genome mining approach that can characterize unknown GNP chemotypes and their genotypes in microbial genomes by iterative identification of *O*-/*N*-glycosyl groups in tandem MS spectra and of their glycosylation genes in secondary metabolic gene clusters. This work extends the concept of experiment-guided genome mining to more natural product classes such as glycosylated polyketides and, therefore, sets another blueprint for future automated characterization of complex secondary metabolomes by a combined application of tandem mass spectrometry and genomics. The implementation of the MS-glycogenetic code and glycogenomic workflow in data acquisition and processing programs could lead to a faster characterization of new GNP chemistry, biochemistry and bioactivity from the increasing microbial genome resources. This would also enable accelerated access and understanding of cryptic GNP pathways in microbial communities and as a therapeutic source.

4.4 Materials and methods

4.4.1 Cultivation and extraction of actinobacteria

A liquid ISP2 starter culture of *Streptomyces* sp. SPB74 (BROAD genome sequence strain, [13]) was inoculated from a spore suspension and incubated at 28 ºC, 225 rpm for 6 days. A 50 ml ISP2 culture was inoculated with 1% of the starter culture and incubated at 28 ºC, 225 rpm for 7 days. The supernatant and cells were extracted with ethyl acetate. The crude extract was dried by rotovaporation and analyzed by LCMS for presence of glycosylated natural products.

A liquid A1 starter culture of *Salinispora arenicola* CNB-527 was inoculated from a spore suspension and incubated at 28 ºC, 225 rpm for 6 days. A 50 ml A1 culture was inoculated with 1% of the starter culture and incubated at 28 ºC, 225 rpm for 7 days. The supernatant was extracted with ethyl acetate, the cells were resuspended in methanol and stirred for 30 min. Ethyl acetate and methanol extracts were combined and dried by rotovaporation. The crude extract was analyzed by LCMS for presence of glycosylated natural products.

4.4.2 MS analysis of microbial metabolic extracts

Crude microbial extracts were dissolved in methanol and filtered through Acrodisc MS Syringe Filter (PTFE membrane, 25 mm, 0.2 µm, PALL Life Sciences). The samples were adjusted to a concentration of 200µg/ml and injected into an Agilent 1260 LC system (injection volume: 5 µl) with an Agilent Extend-C18 RP UPLC column (2.1x100 mm, 1.8 µm) connected to an Agilent 6530 Accurate-Mass Q-TOF LC/MS. For analysis of the *Salinispora arenicola* extract, the LC gradient was as follows: 10% acetonitrile (0.1% TFA, 0-3 min), 10-100% acetonitrile (0.1% TFA)/0.1% TFA (3-23 min), 100% acetonitrile (0.1% TFA, 23-25 min), 10% acetonitrile (0.1% TFA, 25-30 min). The column compartment temperature was 25 ºC. For *Streptomyces* sp. SPB74 extract analysis, the LC gradient was as follows: 10-100% acetonitrile (0.1% TFA)/0.1% TFA (0-20 min), 100% acetonitrile (0.1% TFA, 20-24 min), 10% acetonitrile (0.1% TFA, 24-30 min). For

Salinispora arenicola extract analysis, the Q-TOF settings were as follows: acquisition mode auto-MS2 - MS range: 125-1500 m/z, MS scan rate: 1 spectrum/s, MS/MS scan rate: 2 spectra/s, isolation width: 4 m/z, CID energy: 20 eV, precursor selection static exclusion: 100-500 m/z, ESI source – gas temperature: 300ºC, gas flow: 11 L/min, nebulizer: 45psig, positive ion polarity, scan source parameters: VCap 3000 V, fragmentor 100 V. For *Streptomyces* sp. SPB74 extract analysis, the Q-TOF settings were as follows: acquisition mode auto-MS2 - MS range: 100-3000 m/z, MS scan rate: 1 spectrum/s, MS/MS scan rate: 3 spectra/s, isolation width: 4 m/z, CID energy: 30+0.1(x[m/z]) eV, ESI source – gas temperature: 350ºC, gas flow: 11 l/min, nebulizer: 45psig, positive ion polarity, scan source parameters: VCap 4000 V, fragmentor 200 V. LCMS/MS data was analyzed with Qualitative analysis software of MassHunter software B5 (Agilent). LC-MS/MS data was searched for sugar footprints in extracted ion chromatograms (EIC) of B/C-ion fragments of Table 1 and/or Y-ion neutral loss chromatograms (NLC). Peaks in EICs or NLCs were verified or discarded as candidate glycosylated natural products by reanalysis of MS/MS spectra for corresponding sugar B/C-ions and Y/Z-ion neutral losses. From a candidate GNP MS/MS spectrum, a list of candidate MS/MS sugars was generated by including all sugars from Table 1 that matched observed sugar mass shifts.

4.4.3 Genome mining of glycosylated natural products

Genome sequences of *Streptomyces* sp. SPB74 (GenBank files: GG770539 and GG770540) and *Salinispora arenicola* CNB-527 (Paul Jensen, GenBank project ID PRJNA169705) were analyzed by AntiSMASH [14] for prediction of secondary metabolic gene clusters. Each predicted gene cluster was analyzed for presence of common and specific glycosylation genes (a) based on gene annotation in “Genes and detection info overview” of each cluster and (b) based on BLAST-analysis of putative glycosylation genes. Glycosylation gene functions were assigned based on gene annotation and closest functional BLAST homologs. Specific glycosylation genes were differentiated (if possible) into: 2,3DH, 3,4DH, 3KR, 4KR, 3,4IM, E, FuPyIM, AT, *O*-MT, *N,N*-MT, *N*-MT, *C*-MT, *N*-ET, AcT, CarbT, PyT, oxDA, OxRed,

Dhg, ThiS, *N*-Ox (see Table 2 for abbreviations). A reference list of all gene clusters with reported glycosylation genes was compiled.

Each gene cluster was tested if the specific glycosylation genes match any of the observed MS/MS candidate sugars based on Table 2, i.e. if the biosynthetic genes of an observed sugar are present in a candidate GNP gene cluster. A putative match was confirmed by matching of additional candidate MS/MS sugars to genes in the candidate gene cluster. Next, the candidate GNP gene cluster was fully analyzed by BLAST-analysis of closest functional homologs and a natural product class was assigned based on non-glycosylation biosynthetic genes.

4.4.4 Purification of glycosylated natural products

Cinerubin B was isolated from a 1 L ISP2 medium culture (4 g yeast extract, 10 g malt extract, 4 g D-glucose, ad 1000 ml Millipore-filtered water) which was inoculated with a 10 ml ISP2 starter culture (6 d, 28 ºC, 225 rpm) from spore suspension inoculation and incubated for 7 days at 28 ºC and 225 rpm. The liquid culture was extracted with ethyl acetate (3x). The crude extracts were combined and dried completely by rotovaporation. The crude extract was resuspended in methanol and separated by gel filtration chromatography (solid phase: Sephadex LH20, GE Life Sciences, mobile phase: methanol). Gel filtration fractions were analyzed by dried-droplet MALDI-TOF MS for the presence of cinerubin B. Gel filtration fractions with cinerubin B were further purified by semi-preparative reversed-phase HPLC (Luna reversed phase C18 column, 10x250 mm, 5 micron, 100 Å, Phenomenex). HPLC fractions were analyzed by dried-droplet MALDI-TOF MS for the presence and purity of cinerubin B for subsequent NMR analysis.

Arenimycin B was isolated from a 6 l A1 medium culture of *Salinispora arenicola* CNB-527 after 9 day growth at 28°C and 225 rpm. At the end of the cultivation, washed Amberlite XAD-7 resin (20 g/L, Sigma-Aldrich) was added to the *S. arenicola* culture for 2 h. Subsequently, the resin and cells were separated from the supernatant by cheesecloth filtration. Half of the resin and cells were soaked in acetone and the other half was soaked in methanol for 2 h. The crude extracts were filtered, rotary evaporated, resuspended in methanol and subjected to flash column

liquid-chromatography (reverse phase C18 silica gel, Sigma Aldrich). 20-100% methanol/water fractions were collected and analyzed by LC-MS to identify fractions with arenimycin B. Arenimycin B eluted in 80% methanol. Arenimycin fractions were combined, concentrated by rotovaporation and subjected to semi-preparative HPLC (Luna reversed phase C18 column, 10x250 mm, 5 micron, 100 Å, Phenomenex) with a 10-100% acetonitrile/water gradient (45 min, 3 mL/min flow rate). HPLC fractions were analyzed for arenimycin B by LC-MS or dry droplet MALDI-MS. Arenimycin HPLC fractions were combined and subjected to analytical HPLC (Luna reversed phase C18 column, 4.6x250 mm, 5 micron, 100 Å, Phenomenex) with a 60 min gradient (25-75% acetonitrile (0.1%TFA)). HPLC fractions were analyzed for arenimycin B by LC-MS. <1mg of arenimycin B were isolated from 6 L of liquid culture of *S. arenicola* CNB-527.

4.4.5 NMR analysis of glycosylated natural products

NMR data were acquired at the UCSD Skaggs School of Pharmacy and Pharmaceutical Sciences NMR Facility on a 600 MHz Varian NMR spectrometer (Topspin 2.1.6 software, Bruker) with a 1.7 mm cryoprobe. Each purified glycosylated natural products was dissolved in MeOD-d4 and subjected to NMR structure elucidation (^{1}H, DQF-COSY, ^{1}H-^{13}C HMBC, NOESY). NMR data were analyzed with Topspin 2.1.6 software (Bruker).

4.5 References

1. Staunton, J., Weissman, K.J. Polyketide biosynthesis: a millennium review. *Nat. Prod. Rep.* **18**, 380–416 (2001).

2. Ikeda, H., Nonomiya, T., Usami, M., Ohta, T., Omura, S. Organization of the biosynthetic gene cluster for the polyketide anthelmintic macrolide avermectin in *Streptomyces avermitilis*. *Proc. Natl. Acad. Sci. U. S. A.* **96**, 9509-9514 (1999).

3. Thibodeaux, C.J., Melançon, C.E. 3rd, Liu, H.W. Natural-product sugar biosynthesis and enzymatic glycodiversification. *Angew. Chem. Int. Ed. Engl.* **47**, 9814-9859 (2008).

4. La Ferla, B., Airoldi, C., Zona, C., Orsato, A., Cardona, F., Merlo, S., Sironi, E., D'Orazio, G., Nicotra, F. Natural glycoconjugates with antitumor activity. *Nat. Prod. Rep.* **28**, 630-648 (2011).

5. Kersten, R.D., Yang, Y.L., Xu, Y., Cimermancic, P., Nam, S.J., Fenical, W., Fischbach, M.A., Moore, B.S., Dorrestein, P.C. A mass spectrometry-guided genome mining approach for natural product peptidogenomics. *Nat. Chem. Biol.* 7, 794-802 (2011).

6. Herget, S., Toukach, P.V., Ranzinger, R., Hull, W.E., Knirel, Y.A., von der Lieth, C.W. Statistical analysis of the Bacterial Carbohydrate Structure Data Base (BCSDB): characteristics and diversity of bacterial carbohydrates in comparison with mammalian glycans. *BMC Struct. Biol.* **8**, 35 (2008).

7. Brautaset, T., Sekurova, O.N., Sletta, H., Ellingsen, T.E., StrŁm, A.R., Valla, S., Zotchev, S.B. Biosynthesis of the polyene antifungal antibiotic nystatin in Streptomyces noursei ATCC 11455: analysis of the gene cluster and deduction of the biosynthetic pathway. *Chem. Biol.* 7, 395-403 (2000).

8. Smith, C.A., O'Maille, G., Want, E.J., Qin, C., Trauger, S.A., Brandon, T.R., Custodio, D.E., Abagyan, R., Siuzdak, G. METLIN: A Metabolite Mass Spectral Database. *Ther. Drug Monit.* **7**, 747-751 (2005).

9. Puar, M.S., Chan, T.M., Hegde, V., Patel, M., Bartner, P., Ng, K.J., Pramanik, B.N., MacFarlane, R.D. Sch 40832: a novel thiostrepton from *Micromonospora carbonacea*. J. Antibiot. **51**, 221-224 (1998).

10. Li, J., Qu, X., He, X., Duan, L., Wu, G., Bi, D., Deng, Z., Liu, W., Ou, H.Y. ThioFinder: A Web-Based Tool for the Identification of Thiopeptide Gene Clusters in DNA Sequences. *PLoS One* **7** (2012).

11. Vukics V, Guttman A. Structural characterization of flavonoid glycosides by multi-stage mass spectrometry. *Mass Spectrom. Rev*. **29**, 1-16 (2010).

12. Dürr, C., Schnell, H.J., Luzhetskyy, A., Murillo, R., Weber, M., Welzel, K., Vente, A., Bechthold, A. Biosynthesis of the terpene phenalinolactone in Streptomyces sp. Tü6071: analysis of the gene cluster and generation of derivatives. *Chem. Biol.***13**, 365-377 (2006).

13. Scott, J.J., Oh, D.C., Yuceer, M.C., Klepzig, K.D., Clardy, J., Currie, C.R. Bacterial protection of beetle-fungus mutualism. *Science.* **322**, 63 (2008).

14. Medema, M.H., Blin, K., Cimermancic, P., de Jager, V., Zakrzewski, P., Fischbach, M.A., Weber, T., Takano, E., Breitling R. antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters. *Nucleic Acids Res.* **39**, W339–W346 (2011).

15. Alexeev, I., Sultana, A., Mäntsälä, P., Niemi, J., Schneider, G. Aclacinomycin oxidoreductase (AknOx) from the biosynthetic pathway of the antibiotic aclacinomycin is an unusual flavoenzyme with a dual active site. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 6170-6175 (2007).

16. Konishi, M., Ohkuma, H., Saitoh, K., Kawaguchi, H., Golik, J., Dubay, G., Groenewold, G., Krishnan, B., Doyle, T.W. Esperamicins, a novel class of potent antitumor antibiotics. I. Physico-chemical data and partial structure. *J. Antibiot. (Tokyo)* **38**, 1605-1609 (1985).

17. Räty, K., Kantola, J., Hautala, A., Hakala, J., Ylihonko, K., Mäntsälä, P. Cloning and characterization of *Streptomyces galilaeus* aclacinomycins polyketide synthase (PKS) cluster. *Gene* **293**, 115-122 (2002).

18. Räty, K., Kunnari, T., Hakala, J., Mäntsälä, P., Ylihonko, K. A gene cluster from *Streptomyces galilaeus* involved in glycosylation of aclarubicin. *Mol. Gen. Genet*. **264**, 164-172 (2000).

19. Ettlinger, L., Gäumann, E., Hutter, R., Keller-Schierlein, W., Kradolfer, F., Neipp, L., Prelog, V. Reusser, P., Zahner, H. Stoffwechselprodukte von Actinomyceten, XVI. Cinerubine. *Chem. Ber.* **92**, 1867-1879 (1959).

20. Jensen, P.R., Mafnas, C. Biogeography of the marine actinomycete Salinispora. *Environ. Microbiol*. **8**, 1881-1888 (2006).

21. Asolkar, R.N., Kirkland, T.N., Jensen, P.R., Fenical, W. Arenimycin, an antibiotic effective against rifampin- and methicillin-resistant Staphylococcus aureus from the marine actinomycete Salinispora arenicola. *J Antibiot (Tokyo).* **63**, 37-39 (2010).

22. Gomi, S., Sasaki, T., Itoh, J., Sezaki, M. SF2446, new benzo[a]naphthacene quinone antibiotics II. The structural elucidation. *J. Antibiot.* **41**, 425–432 (1988).

23. Kim, B.C., Lee, J.M., Ahn, J.S., Kim, B.S. Cloning, sequencing, and characterization of the pradimicin biosynthetic gene cluster of Actinomadura hibisca P157-2. *J. Microbiol. Biotechnol*. **17**, 830-839 (2007).

24. Tsunakawa, M., Nishio, M., Ohkuma, H., Tsuno, T., Konishi, M., Naito, T., Oki, T., Kawaguchi, H. The structure of pradimicins A, B and C: a novel family of antifungal antibiotics. *J. Org. Chem.* **54**, 2532–2536 (1989).

25. Rafanan, E.R. Jr., Hutchinson, C.R., Shen, B. Triple hydroxylation of tetracenomycin A2 to tetracenomycin C involving two molecules of O(2) and one molecule of H(2)O. *Org. Lett.* **2**, 3225-3227 (2000).

26. Pagani, I., Liolios, K., Jansson, J., Chen, I.M., Smirnova, T., Nosrat, B., Markowitz, V.M., Kyrpides, N.C. The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata *Nucleic Acids Res.* **40**, D571–D579 (2012).

27. Corre, C., Challis, G.L. New natural product biosynthetic chemistry discovered by genome mining. *Nat. Prod. Rep.* **26**, 977-986 (2009).

28. Winter, J.M., Behnken, S., Hertweck, C. Genomics-inspired discovery of natural products. *Curr. Opin. Chem. Biol.* **15**, 22–31 (2011).

29. Watrous, J., Roach, P., Alexandrov, T., Heath, B.S., Yang, J.Y., Kersten, R.D., van der Voort, M., Pogliano, K., Gross, H., Raaijmakers, J.M., Moore, B.S., Laskin, J., Bandeira, N., Dorrestein, P.C. Mass spectral molecular networking of living microbial colonies. Proc. Natl. Acad. Sci. U. S. A. 109, 1743-1752 (2012).

30. Allen, E.E., Banfield, J.F. Community genomics in microbial ecology and evolution. *Nat. Rev. Microbiol.* **3**, 489-498 (2005).

31. Dürr, C., Hoffmeister, D., Wohlert, S.E., Ichinose, K., Weber, M., Von Mulert, U., Thorson, J.S., Bechthold, A. The glycosyltransferase UrdGT2 catalyzes both C- and O-glycosidic sugar transfers. *Angew. Chem. Int. Ed. Engl*. **43**, 2962-2965 (2004).

Table 1: Prediction of gene clusters of glycosylated natural products in finished actinobacterial genomes (Oct 2012, JGI database) by AntiSMASH analysis of GenBank genome files and subsequent analysis of glycosylation genes in predicted gene clusters. Predicted GNP pathways were differentiated by presence or absence of specific glycosylation genes. Predicted GNP pathways are highlighted in grey and corresponding strain genomes and families in yellow.

Strain	Genbank	Genus	Family	Putative GNP pathway - no specific genes (# - AntiSMASH gene cluster)	Putative GNP pathway - with specific genes (# - AntiSMASH gene cluster)
Acidothermus cellulolyticus 11B	NC_008578.1	*Acidothermus*	*Acidothermaceae*	other - 1GT	none
Arcanobacterium haemolyticum CCM, DSM 20595	CP002045.1	*Arcanobacterium*	*Actinomycetaceae*	none	none
Mobiluncus curtisii ATCC 43063	CP001992.1	*Mobiluncus*	*Actinomycetaceae*	none	none
Actinosynnema mirum 101, DSM 43827	CP001630.1	*Actinosynnema*	*Actinosynnemataceae*	10 - nrps, 1GT 12 - nrps-t1pks - 1GT 14 - nrps, 1GT 15 - t1pks, 1GT	5 - nucleoside, 1GT, 1 spec gene 22 - oligosaccharide-t1pks, 2GT, 1 spec gene
Beutenbergia cavernae HKI 0122, DSM 12333	CP001618.1	*Beutenbergia*	*Beutenbergiaceae*	none	none
Catenulispora acidiphila ID139908, DSM 44928	CP001700.1	*Catenulispora*	*Catenulisporaceae*	6 - nrps, 1GT 7 - nrps-lant-t1pks, 1GT	8 - terpene, 2GT, 2 spec genes
Cellulomonas flavigena 134, DSM 20109	CP001964.1	*Cellulomonas*	*Cellulomonadaceae*	t2pks - 1GT	none
Cellulomonas fimi NRS 133, ATCC 484	CP002666.1	*Cellulomonas*	*Cellulomonadaceae*	none	none
Cellvibrio gilvus ATCC 13127	CP002665.1	*Cellulomonas*	*Cellulomonadaceae*	terpene - 1GT	none
Corynebacterium aurimucosum CN-1, ATCC 700975	CP001601.1	*Corynebacterium*	*Corynebacteriaceae*	t1pks - 1GT	none
Corynebacterium diphtheriae gravis NCTC 13129	BX248353.1	*Corynebacterium*	*Corynebacteriaceae*	none	none
Corynebacterium efficiens YS-314	NC_004369.1	*Corynebacterium*	*Corynebacteriaceae*	terpene - 1GT	none
Corynebacterium glutamicum Kalinowski ATCC 13032	NC_006958.1	*Corynebacterium*	*Corynebacteriaceae*	terpene - 1GT t1pks - 1GT	none
Corynebacterium glutamicum Nakagawa ATCC 13032	NC_003450.3	*Corynebacterium*	*Corynebacteriaceae*	terpene - 1GT t1pks - 1GT	none
Corynebacterium glutamicum R	NC_009342.1	*Corynebacterium*	*Corynebacteriaceae*	none	none
Corynebacterium jeikeium K411	NC_007164.1	*Corynebacterium*	*Corynebacteriaceae*	none	none
Corynebacterium kroppenstedtii DSM 44385	CP001620.1	*Corynebacterium*	*Corynebacteriaceae*	none	none
Corynebacterium pseudotuberculosis 1002	CP001809.2	*Corynebacterium*	*Corynebacteriaceae*	4 -t1pks, 1GT	none
Corynebacterium pseudotuberculosis C231	CP001829.1	*Corynebacterium*	*Corynebacteriaceae*	4 -t1pks, 1GT	none
Corynebacterium pseudotuberculosis FRC41	CP002097.1	*Corynebacterium*	*Corynebacteriaceae*	4 -t1pks, 1GT	none
Corynebacterium pseudotuberculosis I19	CP002251.1	*Corynebacterium*	*Corynebacteriaceae*	none	none
Corynebacterium resistens DSM 45100	CP002857.1	*Corynebacterium*	*Corynebacteriaceae*	none	none
Corynebacterium ulcerans 809	CP002790.1	*Corynebacterium*	*Corynebacteriaceae*	5 -t1pks, 1GT	none
Corynebacterium ulcerans BR-AD22	CP002791.1	*Corynebacterium*	*Corynebacteriaceae*	5 -t1pks, 1GT	none
Corynebacterium urealyticum DSM 7109	NC_010545.1	*Corynebacterium*	*Corynebacteriaceae*	none	none
Corynebacterium diphtheriae 31A	CP003206.1	*Corynebacterium*	*Corynebacteriaceae*	4 - nrps, 1GT 6 - t1pks, 1GT	none
Corynebacterium diphtheriae BH8	CP003209.1	*Corynebacterium*	*Corynebacteriaceae*	5 -t1pks, 1GT	none
Corynebacterium diphtheriae C7 (beta)	NC_016801.1	*Corynebacterium*	*Corynebacteriaceae*	5 - nrps, 1GT 6 - t1pks, 1GT	none
Corynebacterium diphtheriae CDCE 8392	CP003211.1	*Corynebacterium*	*Corynebacteriaceae*	5 - nrps, 1GT 7 - t1pks, 1GT	none
Corynebacterium diphtheriae HC01	CP003212.1	*Corynebacterium*	*Corynebacteriaceae*	4 -t1pks, 1GT	none
Corynebacterium diphtheriae HC02	CP003213.1	*Corynebacterium*	*Corynebacteriaceae*	6 - t1pks, 1GT	none
Corynebacterium diphtheriae HC03	CP003214.1	*Corynebacterium*	*Corynebacteriaceae*	6 - t1pks, 1GT	none
Corynebacterium diphtheriae HC04	CP003215.1	*Corynebacterium*	*Corynebacteriaceae*	7 - t1pks, 1GT	none
Corynebacterium diphtheriae INCA 402	CP003208.1	*Corynebacterium*	*Corynebacteriaceae*	2 - nrps, 1GT 6 - t1pks, 1GT	none
Corynebacterium diphtheriae PW8	CP003216.1	*Corynebacterium*	*Corynebacteriaceae*	5 - t1pks, 1GT	none
Corynebacterium diphtheriae VA01	CP003217.1	*Corynebacterium*	*Corynebacteriaceae*	6 - t1pks, 1GT	none
Corynebacterium pseudotuberculosis 1/06-A	CP003082.1	*Corynebacterium*	*Corynebacteriaceae*	3 - t1pks, 1GT	none
Corynebacterium pseudotuberculosis 267	CP003407.1	*Corynebacterium*	*Corynebacteriaceae*	none	none
Corynebacterium pseudotuberculosis 3/99-5	CP003152.1	*Corynebacterium*	*Corynebacteriaceae*	4 - t1pks, 1GT	none
Corynebacterium pseudotuberculosis 316	CP003077.1	*Corynebacterium*	*Corynebacteriaceae*	4 - t1pks, 1GT	none
Corynebacterium pseudotuberculosis 42/02-A	CP003062.1	*Corynebacterium*	*Corynebacteriaceae*	4 - t1pks, 1GT	none
Corynebacterium pseudotuberculosis CIP52.97	CP003061.1	*Corynebacterium*	*Corynebacteriaceae*	4 - t1pks, 1GT	none
Corynebacterium pseudotuberculosis P54B96	CP003385.1	*Corynebacterium*	*Corynebacteriaceae*	4 - t1pks, 1GT	none
Corynebacterium pseudotuberculosis PAT10	CP002924.1	*Corynebacterium*	*Corynebacteriaceae*	4 - t1pks, 1GT	none
Corynebacterium ulcerans 0102	AP012284.1	*Corynebacterium*	*Corynebacteriaceae*	5 - t1pks, 1GT	none
Corynebacterium variabile DSM 44702	CP002917.1	*Corynebacterium*	*Corynebacteriaceae*	none	none
Brachybacterium faecium 6-10, DSM 4810	CP001643.1	*Brachybacterium*	*Dermabacteraceae*	none	none
Kytococcus sedentarius 541, DSM 20547	CP001686.1	*Kytococcus*	*Dermacoccaceae*	2 - t1pks, 1GT	none
Frankia alni ACN14a	CT573213.2	*Frankia*	*Frankiaceae*	GC1 - t1pks - 1GT GC4 - terpene - 1GT GC5 - t1pks - 1GT GC6 - t1pks - 1GT GC7 - nrps - 1GT GC11 - t1pks -1GT GC13 - t1pks - 1GT GC18 - other - 1GT	none
Frankia sp. CcI3	CP000249.1	*Frankia*	*Frankiaceae*	GC2 - terpene - 1GT GC3 - t1pks - 1GT GC4 - t1pks - 1GT GC8 - other - 5GTs GC19 - lant - 1GT GC24 - terpene - 1GT	none
Frankia sp. EAN1pec	CP000820.1	*Frankia*	*Frankiaceae*	none	none
Frankia sp. EuI1c	CP002299.1	*Frankia*	*Frankiaceae*	2 - t1pks, 1GT 4 - t1pks, 1GT	none
Frankia symbiont of *Datisca glomerata*	NC_015656.1	*Frankia*	*Frankiaceae*	2 - t2pks, 2GTs 5 - t1pks, 1GT 11 - t2pks, 1GT	none
Blastococcus saxobsidens DD2	FO117623.1	*Blastococcus*	*Geodermatophilaceae*	none	none
Geodermatophilus obscurus G-20, DSM 43160	CP001867.1	*Geodermatophilus*	*Geodermatophilaceae*	GC2 - other, 6GTs	none
Stackebrandtia nassauensis LLR-40K-21, DSM 44728	CP001778.1	*Stackebrandtia*	*Glycomycetaceae*	GC2 - nrps-t1pks - 1GT GC15 - other - 1GT	none
Gordonia bronchialis 3410, DSM 43247	CP001802.1	*Gordonia*	*Gordoniaceae*	1 - t1pks, 1GT 6 -terpene, 1GT	none
Gordonia polyisoprenivorans VH2, DSM 44266	CP003119.1 CP003120.1	*Gordonia*	*Gordoniaceae*	6 - butyrolactone, 1GT 7 - terpene, 5GTs 9 - nrps, 1GT 11 - nrps, 3GTs 13 - terpene, 1GT	4 - nrps, 5GT, 1NT, spec genes
Intrasporangium calvum 7KIP, DSM 43043	CP002343.1	*Intrasporangium*	*Intrasporangiaceae*	none	none
Jonesia denitrificans 55134, DSM 20603	CP001706.1	*Jonesia*	*Jonesiaceae*	none	none
Kineococcus radiotolerans SRS30216	NC_009664.2	*Kineococcus*	*Kineosporiaceae*	3 - other, 1GT	none
Clavibacter michiganensis michiganensis NCPPB 382	NC_009480.1 NC_009478.1 NC_009479.1	*Clavibacter*	*Microbacteriaceae*	none	none
Leifsonia xyli xyli CTCB07	NC_006087.1	*Leifsonia*	*Microbacteriaceae*	none	none
Microbacterium testaceum StLB037	NC_015125.1	*Microbacterium*	*Microbacteriaceae*	1- t3pks, 1GT	none
Tropheryma whipplei TW08/27	BX072543.1	*Tropheryma*	*Microbacteriaceae*	none	none
Tropheryma whipplei Twist	NC_004572.3	*Tropheryma*	*Microbacteriaceae*	none	none
Arthrobacter arilaitensis re117, CIP108037	NC_014550.1	*Arthrobacter*	*Micrococcaceae*	1 - siderophore, 2GT	none
Arthrobacter aurescens TC1	NC_008711.1 NC_008712.1 NC_008713.1	*Arthrobacter*	*Micrococcaceae*	4 - t3pks, 1GT	none
Arthrobacter chlorophenolicus A6	NC_011886.1 NC_011881.1 NC_011879.1	*Arthrobacter*	*Micrococcaceae*	none	none
Arthrobacter phenanthrenivorans Sphe3	NC_015145.1 NC_015146.1 NC_015147.1	*Arthrobacter*	*Micrococcaceae*	2 - t3pks, 1GT	none
Arthrobacter sp. FB24	NC_008541.1 NC_008537.1 NC_008538.1 NC_008539.1	*Arthrobacter*	*Micrococcaceae*	none	none
Kocuria rhizophila DC2201	NC_010617.1	*Kocuria*	*Micrococcaceae*	none	none
Micrococcus luteus Fleming NCTC 2665	CP001628.1	*Micrococcus*	*Micrococcaceae*	terpene - 1GT	none
Renibacterium salmoninarum ATCC 33209	CP000910.1	*Renibacterium*	*Micrococcaceae*	4 - t1pks, 1GT	none
Rothia dentocariosa ATCC 17931	CP002280.1	*Rothia*	*Micrococcaceae*	none	none
Rothia mucilaginosa DY-18	NC_013715.1	*Rothia*	*Micrococcaceae*	none	none
Actinoplanes missouriensis NBRC 102363	AP012319.1	*Actinoplanes*	*Micromonosporaceae*	GC2 - nrps - 1GT GC7 - t1pks-nrps-terpene - 1GT GC8 - other - 1GT	GC6 - t2pks - 1GT, 1NT, 14,6DH, 5 spec genes
Actinoplanes sp. SE50/110	CP003170.1	*Actinoplanes*	*Micromonosporaceae*	GC12 - t1pks - 1GT	GC4 - nrps(LPS) - 2GT, spec genes GC5 - amglycycl - 2GTs, 1NT, 1 4,6DH, 1 spec genes

Table 1 (continued): Prediction of gene clusters of glycosylated natural products in finished actinobacterial genomes (Oct 2012, JGI database) by AntiSMASH analysis of GenBank genome files and subsequent analysis of glycosylation genes in predicted gene clusters. Predicted GNP pathways were differentiated by presence or absence of specific glycosylation genes. Predicted GNP pathways are highlighted in grey and corresponding strain genomes and families in yellow.

Strain	Genbank	Genus	Family	Putative GNP pathway - no specific genes (# - AntiSMASH gene cluster)	Putative GNP pathway - with specific genes (# - AntiSMASH gene cluster)
Micromonospora aurantiaca ATCC 27029	CP002162.1	*Micromonospora*	*Micromonosporaceae*	GC4 - NRPS-PKS - 2GT GC6 - hybrid - 1GT GC10 - oligosaccharide - 3GT	GC12 - oligosaccharide-type II PKS - 7GT, 1NT, 1 4,6DH, 5 spec genes
Micromonospora sp. L5	CP002399.1	*Micromonospora*	*Micromonosporaceae*	GC10 - nrps-oligosaccharide-terpene - 3 GT GC14 - lant-nrps-t1pks - 1GT GC17 - NRPS-PKS - 2GT	GC8 - oligosaccharide-type II PKS - 5GT, 1 4,6DH, 1NT, 6 spec genes
Salinispora arenicola CNS-205	CP000850.1	*Salinispora*	*Micromonosporaceae*	GC12 - amglyccycl	GC4 - type I pks-nrps - 2GT, 1NT, spec genes GC7 - oligosaccharide-t1pks - 4GT, 5 spec genes GC10 - indole - 1GT, 1NT, 1 4,6DH, 4 spec genes
Salinispora tropica CNB-440	NC_009380.1	*Salinispora*	*Micromonosporaceae*	none	GC4 - type II pks -2GT, 1NT, 1 4,6-DH, 4 spec genes
Verrucosispora maris AB-18-032	CP002638.1	*Verrucosispora*	*Micromonosporaceae*	none	none
Amycolicicoccus subflavus DQS3-9A1	NC_015564.1 NC_015560.1 NC_015561.1	*Amycolicicoccus*	*Mycobacteriaceae*	6 terpene, 1GT 8 - t1pks, 1GT	2 - nrps, 1GT, 1NT, 1spec gene
Mycobacterium africanum GM041182	NC_015758.1	*Mycobacterium*	*Mycobacteriaceae*	12 - t1pks-nrps, 2GTs	6 - t1pks, 5GTs, 1spec gene
Mycobacterium avium 104	CP000479.1	*Mycobacterium*	*Mycobacteriaceae*	11 - nrps-oligosaccharide, 4GTs	13 - terpene, 1GT, 1spec gene
Mycobacterium leprae Br4923	NC_011896.1	*Mycobacterium*	*Mycobacteriaceae*	2 - oligosaccharide-t1pks, 3GTs	5 - t1pks, 1GT
Mycobacterium avium paratuberculosis K-10	NC_002944.2	*Mycobacterium*	*Mycobacteriaceae*	none	none
Mycobacterium bovis AF2122/97	NC_002945.3	*Mycobacterium*	*Mycobacteriaceae*	12 - t1pks-nrps, 3GTs	6 - t1pks, 5GTs, 1spec gene
Mycobacterium bovis BCG Pasteur 1173P2	NC_008769.1	*Mycobacterium*	*Mycobacteriaceae*	12 - t1pks-nrps, 2GTs	6 - t1pks, 5GTs, 1spec gene
Mycobacterium bovis BCG Tokyo 172	NC_012207.1	*Mycobacterium*	*Mycobacteriaceae*	12 - t1pks-nrps, 2GTs	6 - t1pks, 5GTs, 1spec gene
Mycobacterium canettii CIPT 140010059	HE572590.1	*Mycobacterium*	*Mycobacteriaceae*	6 - t1pks, 5GTs 12 - t1pks-nrps-oligosaccharide, 3GTs 14 - t1pks, 1GT	none
Mycobacterium gilvum PYR-GCK	NC_009338.1 NC_009339.1 CP000658.1 CP000659.1	*Mycobacterium*	*Mycobacteriaceae*	2 - nrps, 1GT 7 - terpene, 1GT 11 - t13pks, 1GT	4 - t1pks, spec gene
Mycobacterium leprae TN	NC_002677.1	*Mycobacterium*	*Mycobacteriaceae*	2 - oligosaccharide-t1pks, 2GTs 5 - t1pks, 1GT	none
Mycobacterium marinum M, ATCC BAA-535	NC_010612.1 NC_010604.1	*Mycobacterium*	*Mycobacteriaceae*	10 - t1pks-nrps, 1GT 14 - t1pks-nrps, 4GTs	none
Mycobacterium smegmatis MC2 155	NC_008596.1	*Mycobacterium*	*Mycobacteriaceae*	2 - nrps-t1pks, 3GTs, 1 4,6DH 10 - t1pks-nrps, 2GTs 11 - nrps, 1GTs 14 - nrps, 3GTs	none
Mycobacterium gilvum Spyr1	NC_014814.1 NC_014811.1 NC_014812.1	*Mycobacterium*	*Mycobacteriaceae*	1 - nrps, 1GT 4 - terpene, 1GT 7 - t3pks, 1GT 13 - t1pks, 1GT	none
Mycobacterium sp. JDM601	CP002329.1	*Mycobacterium*	*Mycobacteriaceae*	none	7 - other, 1GT, 1spec gene
Mycobacterium sp. JLS	CP000580.1	*Mycobacterium*	*Mycobacteriaceae*	12 - t1pks, 1GT	none
Mycobacterium sp. KMS	CP000518.1 CP000519.1 CP000520.1	*Mycobacterium*	*Mycobacteriaceae*	7 - nrps-t1pks, 3GTs	none
Mycobacterium sp. MCS	CP000384.1 CP000385.1	*Mycobacterium*	*Mycobacteriaceae*	3 - nrps-t1pks, 3GTs	none
Mycobacterium tuberculosis CCDC5079	CP001641.1	*Mycobacterium*	*Mycobacteriaceae*	6 - t1pks, 5GTs 12 - t1pks-nrps-oligosaccharide, 2GTs	none
Mycobacterium tuberculosis CCDC5180	CP001642.1	*Mycobacterium*	*Mycobacteriaceae*	6 - t1pks, 5GTs	none
Mycobacterium tuberculosis CDC1551	NC_002755.2	*Mycobacterium*	*Mycobacteriaceae*	12 - t1pks-nrps-oligosaccharide, 2GTs	
Mycobacterium tuberculosis F11 (ExPEC)	NC_009525.1	*Mycobacterium*	*Mycobacteriaceae*	6 - t1pks, 5GTs 12 - t1pks-nrps-oligosaccharide, 3GTs	none
Mycobacterium tuberculosis H37Ra	CP000611.1	*Mycobacterium*	*Mycobacteriaceae*	6 - t1pks, 5GT 12 - t1pks-nrps-oligosaccharide, 2GTs	none
Mycobacterium tuberculosis KZN 1435 (MDR)	CP001658.1	*Mycobacterium*	*Mycobacteriaceae*	4 - oligosaccharide-t1pks-nrps, 3GTs 11 - t1pks, 5GTs	none
Mycobacterium ulcerans Agy99	NC_008611.1 NC_005916.1	*Mycobacterium*	*Mycobacteriaceae*	none	3 - nrps, 3GTs, 2spec genes
Mycobacterium vanbaalenii PYR-1	NC_008726.1	*Mycobacterium*	*Mycobacteriaceae*	7 - other, 4GTs 13 - t1pks, 1GT	none
Mycobacterium bovis BCG Mexico	CP002095.1	*Mycobacterium*	*Mycobacteriaceae*	6 - t1pks, 4GTs 12 - t1pks-nrps, 3GTs	none
Mycobacterium chubuense NBB4	CP003053.1 CP003054.1 CP003055.1	*Mycobacterium*	*Mycobacteriaceae*	12 - t1pks, 2GT	11 - other, 1GT, 1 spec gene
Mycobacterium intracellulare MOTT-02	CP003323.1	*Mycobacterium*	*Mycobacteriaceae*	3 - nrps, 1GT	15 - terpene, 1GT, 2 spec genes
Mycobacterium intracellulare MOTT-64	CP003324.1	*Mycobacterium*	*Mycobacteriaceae*	none	15 - terpene, 1GT, 2 spec genes
Mycobacterium massiliense GO 06	CP003699.1	*Mycobacterium*	*Mycobacteriaceae*	1 - t1pks, 5GT 13 - nrps, 2GTs	12 - nrps, 3GT, 1NT, 1 spec gene
Mycobacterium rhodesiae NBB3	CP003169.1	*Mycobacterium*	*Mycobacteriaceae*	4 - lant, 1GT 6 - t1pks, 1GT 13 - t2pks, 1GT	2 - t1pks-nrps, 5GTs, 1 4,6DH, spec genes
Mycobacterium tuberculosis CTRI-2	CP002992.1	*Mycobacterium*	*Mycobacteriaceae*	6 - t1pks, 3GTs 13 - t1pks-nrps, 3GTs	none
Mycobacterium tuberculosis KZN 4207 (DS)	CP001662.1	*Mycobacterium*	*Mycobacteriaceae*	4 - oligosaccharide-t1pks-nrps, 4GTs 11 - t1pks, 5GTs	none
Mycobacterium tuberculosis RGTB327	CP003233.1	*Mycobacterium*	*Mycobacteriaceae*	none	none
Mycobacterium tuberculosis RGTB423	CP003234.1	*Mycobacterium*	*Mycobacteriaceae*	8 - t1pks-nrps, 1GT	none
Mycobacterium tuberculosis UT205	NC_016934.1	*Mycobacterium*	*Mycobacteriaceae*	6 - t1pks, 4GTs	none
Nakamurella multipartita Y-104, DSM 44233	CP001737.1	*Nakamurella*	*Nakamurellaceae*	terpene - 1GT	none
Nocardia farcinica IFM 10152	AP006618.1 AP006619.1 AP006620.1	*Nocardia*	*Nocardiaceae*	GC1 - t1pks - 5GTs GC17 - nrps - 1GT	none
Nocardia cyriacigeorgica GUH-2	NC_016887.1	*Nocardia*	*Nocardiaceae*	1 - t1pks, 3GTs 6 - nrps, 3GTs 15 - nrps, 1GT, 1 4,6DH 19 - nrps, 1GT	none
Rhodococcus equi 103S	FN563149.1	*Rhodococcus*	*Nocardiaceae*	GC2 - terpene - 1GT GC7 - terpene - 1GT GC10 - other - 1GT GC11 - t1pks - 4GTs	GC14 - nrps - 2GT, 1NT, 1 4,6DH, 1 spec gene
Rhodococcus erythropolis PR4	AP008957.1 AP008931.1 AP008932.1 AP008933.1	*Rhodococcus*	*Nocardiaceae*	GC1 - t1pks - 6GTs GC2 - t1pks - 1GT GC5 - nrps - 2GTs GC8 - nrps - 1GT GC10 - nrps - 1GT GC12 - terpene - 1GT GC15 - amglyccycl - 1GT	none
Rhodococcus opacus B4	AP011115.1 AP011116.1 AP011117.1	*Rhodococcus*	*Nocardiaceae*	GC1 - nrps - 1GT GC3 - terpene - 1GT GC12 - t1pks - 12GTs GC19 - nrps - 4GTs GC20 - nrps - 3GTs	GC13 - t1pks - 1GT, 1 spec gene
Rhodococcus jostii RHA1	CP000431.1 CP000432.1 CP000433.1 CP000434.1	*Rhodococcus*	*Nocardiaceae*	GC11 - t1pks - 5GTs	GC17 - nrps - 2GTs, 2NTs, 1 4,6DH, spec genes GC18 - nrps - 1GT, 2NTs, 1 4,6DH, spec genes GC20 - nrps - 1GT, 1NT, spec gene
Kribbella flavida IFO 14399, DSM 17836	CP001736.1	*Kribbella*	*Nocardioidaceae*	none	none
Nocardioides sp. JS614	NC_008699.1 NC_008697.1	*Nocardioides*	*Nocardioidaceae*	none	none
Nocardiopsis dassonvillei dassonvillei DSM 43111	NC_014210.1 NC_014211.1	*Nocardiopsis*	*Nocardiopsaceae*	none	13 - lant-oligosaccharide, 8GTs, 1NT, 1 4,6DH, spec genes
Thermobifida fusca YX	CP000088.1	*Thermobifida*	*Nocardiopsaceae*	3 - bcin, 1GT	none
Isoptericola variabilis 225	CP002810.1	*Isoptericola*	*Promicromonosporaceae*		
Xylanimonas cellulosilytica XIL07, DSM 15894	CP001821.1 CP001822.1	*Xylanimonas*	*Promicromonosporaceae*	none	none
Microlunatus phosphovorus NM-1	AP012204.1	*Microlunatus*	*Propionibacteriaceae*	none	none
Propionibacterium acnes 266	CP002409.1	*Propionibacterium*	*Propionibacteriaceae*	none	none
Propionibacterium acnes 6609	CP002815.1	*Propionibacterium*	*Propionibacteriaceae*	none	none
Propionibacterium acnes KPA171202	NC_006085.1	*Propionibacterium*	*Propionibacteriaceae*	none	none

Table 1 (continued): Prediction of gene clusters of glycosylated natural products in finished actinobacterial genomes (Oct 2012, JGI database) by AntiSMASH analysis of GenBank genome files and subsequent analysis of glycosylation genes in predicted gene clusters. Predicted GNP pathways were differentiated by presence or absence of specific glycosylation genes. Predicted GNP pathways are highlighted in grey and corresponding strain genomes and families in yellow.

Strain	Genbank	Genus	Family	Putative GNP pathway - no specific genes (# - AntiSMASH gene cluster)	Putative GNP pathway - with specific genes (# - AntiSMASH gene cluster)
Propionibacterium acnes SK137	CP001977.1	*Propionibacterium*	*Propionibacteriaceae*	none	none
Propionibacterium freudenreichii shermanii CIRM-BIA1	NC_014215.1	*Propionibacterium*	*Propionibacteriaceae*	none	none
Propionibacterium acnes ATCC 11828	CP003084.1	*Propionibacterium*	*Propionibacteriaceae*	none	none
Propionibacterium acnes TypeIA2 P.acn17	CP003196.1	*Propionibacterium*	*Propionibacteriaceae*	none	none
Propionibacterium acnes TypeIA2 P.acn31	CP003197.1	*Propionibacterium*	*Propionibacteriaceae*	none	none
Propionibacterium acnes TypeIA2 P.acn33	CP003195.1	*Propionibacterium*	*Propionibacteriaceae*	none	none
Amycolatopsis mediterranei U32	CP002000.1	*Amycolatopsis*	*Pseudonocardiaceae*	GC13 - nrps-t1pks-terpene - 1GT GC14 - t1pks - 1GT GC15 - t1pks - 1GT	GC1 - type I PKS - 1GT, 2 spec genes GC22 - nrps-oligosaccharide-t2pks - 9GT, 1NT, spec genes
Amycolatopsis mediterranei S699	CP003729.1	*Amycolatopsis*	*Pseudonocardiaceae*	14 - t1pks, 1GT	22 - nrps-oligosaccharide-t1pks - 9GTs, 1NT, 3 spec genes
Pseudonocardia dioxanivorans CB1190	CP002593.1 CP002594.1 CP002595.1 CP002596.1 CP002597.1 CP002598.1	*Pseudonocardia*	*Pseudonocardiaceae*	none none none none none none	none none none none none none
Saccharomonospora viridis P101, DSM 43017	CP001683.1	*Saccharomonospora*	*Pseudonocardiaceae*	none	none
Saccharopolyspora erythraea NRRL 2338 white	AM420293.1	*Saccharopolyspora*	*Pseudonocardiaceae*	GC7 - t1pks - 1GT GC14 - terpene - 1GT GC25 - t1pks - 1GT GC27 - terpene - 1GT	GC3 - t1pks - 2GT, 11 spec genes
Sanguibacter keddieii ST-74, DSM 10542	CP001819.1	*Sanguibacter*	*Sanguibacteraceae*	none	none
Segniliparus rotundus CDC 1076, DSM 44985	CP001958.1	*Segniliparus*	*Segniliparaceae*	8 -t1pks, 1GT	none
Kitasatospora setae KM-6054, NBRC 14216	NC_016109.1	*Kitasatospora*	*Streptomycetaceae*	8 - terpene, 1GT, 1NT 10 - t3pks, 1GT, 1NT, 1 4,6DH	7 - siderophore, 1GT, 1 spec gene
Streptomyces avermitilis MA-4680	NC_003155.4 NC_004719.1	*Streptomyces*	*Streptomycetaceae*	GC5 - NRPS-PKS - 1GT GC27 - melanin - 2GT, polysaccharide? GC31 - type I PKS - 1GT GC33 - type I PKS - 1GT	GC6 -type I PKS - 1GT, 1NT, 1 4,6DH, 5 spec genes GC7 - terpene - 1GT, 1NT, 2 spec genes GC11 - hopane - 1GT, 1NT, 2 spec genes
Streptomyces bingchenggensis BCW-1	NC_016582.1	*Streptomyces*	*Streptomycetaceae*	GC7- type I PKS - 1GT, 1 4,6DH GC19 - terpene - 1GT GC26 - NRPS-PKS - 1GT GC27 - NRPS - 1GT	GC34 - type II PKS - 5GT, 1NT, 1 4,6DH, 5 spec genes GC39 - type I PKS - 3GT, 1NT, 4,6DH - 3 spec genes
Streptomyces coelicolor A3(2)	NC_003888.3 NC_003904.1	*Streptomyces*	*Streptomycetaceae*	GC8 - melanin - 1GT GC21 - hopane - 1GT, 1NT	none
Streptomyces griseus griseus NBRC 13350	NC_010572.1	*Streptomyces*	*Streptomycetaceae*	GC6 - t1pks - 1GT GC33 - nrps - 1GT	GC7 - oligosaccharide - 1GT, 1NT, 1 4,6DH, 4 spec genes GC12 - terpene - 2GT, 1NT, 1 spec gene GC27 - amglyccycl - 2GTs, 1 NT, 1 4,6DH, spec genes
Streptomyces scabiei 87.22	NC_013929.1	*Streptomyces*	*Streptomycetaceae*	none	GC7- terpene - 1GT, 1NT, 1 spec genes GC26 - type I PKS - 1GT, 1 4,6DH, 5 spec genes
Streptomyces cattleya NRRL 8057	CP003219.1 CP003229.1	*Streptomyces*	*Streptomycetaceae*	GC13 - terpene - 1GT GC19 - hopene - 1GT, 1NT GC20 - t3pks - 1GT none	GC24 - butyrolactone-t1pks-nrps - 1GT, 1NT, 2 spec genes none
Streptomyces sp. Tu6071	CM001165.1	*Streptomyces*	*Streptomycetaceae*	GC5 - hopene - 1GT, 1NT	GC8 - terpene - 1GT, 1NT, 1 4,6DH, spec genes
Streptomyces flavogriseus ATCC 33331	CP002475.1 CP002476.1 CP002477.1	*Streptomyces*		GC16 - terpene - 1GT	GC6- terpene - 1GT, 1NT, 1 spec genes
Streptomyces venezuelae ATCC 10712	FR845719.1	*Streptomyces*	*Streptomycetaceae*	none	GC18 - t2pks - 1GT, 1NT, 1 4,6DH, 3 spec genes
Streptomyces hygroscopicus jinggangensis 5008	NC_017765.1 NC_017766.1 NC_016972.1	*Streptomyces*	*Streptomycetaceae*	none	2 - aminoglycoside, 1GT, 1NT, spec genes 32 - terpene - 1GT, 1NT, 1spec gene
Streptosporangium roseum NI 9100, DSM 43021	NC_013595.1 NC_013596.1	*Streptosporangium*	*Streptosporangiaceae*	none none	none none
Thermomonospora curvata DSM 43183	NC_013510.1	*Thermomonospora*	*Thermomonosporaceae*	3 - nrps, 1GT 9 - t1pks, 1GT	4 - t2pks-t1pks, 2GT, 1NT, 1 4,6DH, 6 spec genes
Tsukamurella paurometabola 33, DSM 20162	NC_014158.1 NC_014159.1	*Tsukamurella*	*Tsukamurellaceae*	5 - t3pks, 1GT	8 - t1pks, 1NT, 1GT, 2 spec genes
Thermobispora bispora R51, DSM 43833	CP001874.1	*Thermobispora*	*unclassified*	none	none
Bifidobacterium adolescentis ATCC 15703	NC_008618.1	*Bifidobacterium*	*Bifidobacteriaceae*	none	none
Bifidobacterium animalis lactis AD011	CP001213.1	*Bifidobacterium*	*Bifidobacteriaceae*	none	none
Bifidobacterium animalis lactis BB-12	CP001853.1	*Bifidobacterium*	*Bifidobacteriaceae*	none	none
Bifidobacterium animalis lactis Bl-04, ATCC SD5219	CP001515.1	*Bifidobacterium*	*Bifidobacteriaceae*	none	none
Bifidobacterium animalis lactis DSM 10140	NC_017834.1	*Bifidobacterium*	*Bifidobacteriaceae*	none	none
Bifidobacterium animalis lactis V9	CP001892.1	*Bifidobacterium*	*Bifidobacteriaceae*	none	none
Bifidobacterium animalis subsp. lactis CNCM I-2494	CP002915.1	*Bifidobacterium*	*Bifidobacteriaceae*	none	none
Bifidobacterium bifidum PRL2010	CP001840.1	*Bifidobacterium*	*Bifidobacteriaceae*	none	none
Bifidobacterium bifidum S17	CP002220.1	*Bifidobacterium*	*Bifidobacteriaceae*	none	none
Bifidobacterium breve ACS-071-V-Sch8b	CP002743.1	*Bifidobacterium*	*Bifidobacteriaceae*	none	none
Bifidobacterium breve UCC2003	CP000303.1	*Bifidobacterium*	*Bifidobacteriaceae*	none	none
Bifidobacterium dentium Bd1	CP001750.1	*Bifidobacterium*	*Bifidobacteriaceae*	none	none
Bifidobacterium longum DJO10A	NC_010816.1	*Bifidobacterium*	*Bifidobacteriaceae*	none	none
Bifidobacterium longum NCC2705	NC_004307.2	*Bifidobacterium*	*Bifidobacteriaceae*	none	none
Bifidobacterium longum infantis 157F-NC	NC_015052.1	*Bifidobacterium*	*Bifidobacteriaceae*	none	none
Bifidobacterium longum infantis ATCC 15697	NC_017219.1	*Bifidobacterium*	*Bifidobacteriaceae*	none	none
Bifidobacterium longum longum BBMN68	CP002286.1	*Bifidobacterium*	*Bifidobacteriaceae*	none	none
Bifidobacterium longum longum JDM301	CP002010.1	*Bifidobacterium*	*Bifidobacteriaceae*	none	none
Bifidobacterium longum subsp. longum KACC 91563	NC_017221.1	*Bifidobacterium*	*Bifidobacteriaceae*	none	none
Bifidobacterium animalis animalis ATCC 25527	CP002567.1	*Bifidobacterium*	*Bifidobacteriaceae*	none	none
Bifidobacterium animalis lactis BLC1	CP003039.1	*Bifidobacterium*	*Bifidobacteriaceae*	none	none
Gardnerella vaginalis ATCC 14019	NC_014644.1	*Gardnerella*	*Bifidobacteriaceae*	none	none
Gardnerella vaginalis 409-05	NC_013721.1	*Gardnerella*	*Bifidobacteriaceae*	none	none
Gardnerella vaginalis HMP9231	NC_017456.1	*Gardnerella*	*Bifidobacteriaceae*	none	none
Atopobium parvulum IPP 1246, DSM 20469	CP001721.1	*Atopobium*	*Coriobacteriaceae*	none	none
Coriobacterium glomerans PW2, DSM 20642	NC_015389.1	*Coriobacterium*	*Coriobacteriaceae*	none	none
Cryptobacterium curtum 12-3, DSM 15641	NC_013170.1	*Cryptobacterium*	*Coriobacteriaceae*	none	none
Eggerthella lenta VPI 0255, DSM 2243	CP001726.1	*Eggerthella*	*Coriobacteriaceae*	none	none
Eggerthella sp. YY7918	NC_015738.1	*Eggerthella*	*Coriobacteriaceae*	none	none
Gordonibacter pamelaeae 7-10-1-bT, DSM 19378	FP929047.1	*Gordonibacter*	*Coriobacteriaceae*	none	none
Olsenella uli VPI, DSM 7084	CP002106.1	*Olsenella*	*Coriobacteriaceae*	none	none
Slackia heliotrinireducens RHS 1, DSM 20476	CP001684.1	*Slackia*	*Coriobacteriaceae*	none	none
Rubrobacter xylanophilus DSM 9941	NC_008148.1	*Rubrobacter*	*Rubrobacteraceae*	none	none
Conexibacter woesei ID131577, DSM 14684	NC_013739.1	*Conexibacter*	*Conexibacteraceae*	none	none

Table 2: MS-glycogenetic code. Grey square indicates enzyme is present in corresponding sugar pathway. White square indicates enzyme is not present in corresponding sugar pathway.

Y-Mass shift [Da]	B-mass shift [Da]	C-mass shift [Da]	Sugar	Common genes	Specific BS genes												
					DH			KR		IM			AT			MT	
				NT	4,6-DH	2,3-DH	3,4-DH	3-KR	4-KR	(3,5-)E	3,4-IM	FuPyIM	2-AT	3-AT	4-AT	C3-MT	C5-MT
110.03678	111.044056	129.054621	L-aculose														
112.05243	113.059706	131.070271	L-cinerulose A														
114.06808	115.075356	133.085921	L-rhodinose														
125.047679	126.054955	144.065520	rednose														
128.047345	129.054621	147.065186	L-cinerulose B														
128.08373	129.091006	147.101571	O-methyl-L-amicetose														
128.08373	129.091006	147.101571	4-O-methyl-L-rhodinose														
129.078979	130.086255	148.096820	L-daunosamine														
129.078979	130.086255	148.096820	L-acosamine	N/A													
129.078979	130.086255	148.096820	L-ristosamine														
130.062995	131.070271	149.080836	D-digitoxose														
130.062995	131.070271	149.080836	L-digitoxose														
130.062995	131.070271	149.080836	2-deoxy-L-fucose														
130.062995	131.070271	149.080836	D-olivose														
130.062995	131.070271	149.080836	D-oliose														
130.062995	131.070271	149.080836	L-boivinose	N/A													
141.078979	142.086255	160.096820	4-oxo-L-vancosamine														
141.115363	142.122639	160.133204	D-forosamine														
143.094628	144.101904	162.112469	L-actinosamine														
143.094628	144.101904	162.112469	3-epi-L-vancosamine														
143.094628	144.101904	162.112469	L-vancosamine														
143.094628	144.101904	162.112469	L-vicenisamine														
144.078645	145.085921	163.096486	D-chalcose														
144.078645	145.085921	163.096486	D-mycarose														
144.078645	145.085921	163.096486	3-O-methyl-L-olivose/L-oleandrose														
144.078645	145.085921	163.096486	olivomose														
145.073894	146.081170	164.091735	D-mycosamine														
146.040152	147.047428	165.057993	4-deoxy-4-thio-D-digitoxose														
146.05791	147.065186	165.075751	D-fucofuranose														
146.05791	147.065186	165.075751	D-fucose														
146.05791	147.065186	165.075751	L-rhamnose														
146.05791	147.065186	165.075751	L-quinovose	N/A													
146.05791	147.065186	165.075751	D-quinovose	N/A													
157.110279	158.117555	176.128120	4-N-ethyl-4-amino-3-O-methoxy-2,4,5-trideoxypentose														
157.110279	158.117555	176.128120	D-3-N-methyl-4-O-methyl-L-ristosamine														
157.110279	158.117555	176.128120	N,N-dimethyl-L-pyrrolosamine	N/A													
157.110279	158.117555	176.128120	D-desosamine														
157.110279	158.117555	176.128120	L-megosamine														
157.110279	158.117555	176.128120	nogalamine														
157.110279	158.117555	176.128120	L-rhodosamine														
157.110279	158.117555	176.128120	D-angolosamine														
157.110279	158.117555	176.128120	kedarosamine	N/A													
158.081719	159.088995	177.099560	L-noviose														
158.094295	159.101571	177.112136	L-cladinose														
159.089544	160.096820	178.107385	2'-N-methyl-D-fucosamine														
160.055802	161.063078	179.073643	4-deoxy-4-methylthio-a-D-digitoxose	N/A													
160.07356	161.080836	179.091401	D-digitalose														
160.07356	161.080836	179.091401	3-O-methyl-rhamnose														
160.07356	161.080836	179.091401	2-O-methyl-L-rhamnose														
160.07356	161.080836	179.091401	6-deoxy-3-C-methyl-L-mannose														
161.068809	162.076085	180.086650	4,6-dideoxy-4-hydroxylamino-D-glucose														
162.052825	163.060101	181.070666	hexose														
171.125929	172.133205	190.143770	3-N,N-dimethyl-L-eremosamine														
171.125929	172.133205	190.143770	namenamicin sugar C	N/A													
172.07356	173.080836	191.091401	chromose D (4-O-acetyl-β-D-oliose)														
173.068809	174.076085	192.086650	4-O-carbamoyl-D-olivose														
173.105194	174.112470	192.123035	D-ravidosamine														
173.105194	174.112470	192.123035	3-N,N-dimethyl-D-mycosamine														
174.08921	175.096486	193.107051	2,3-O-dimethyl-L-rhamnose														
174.08921	175.096486	193.107051	2,4-O-dimethyl-L-rhamnose														
174.08921	175.096486	193.107051	3,4-O-dimethyl-L-rhamnose														
178.029982	179.037258	197.047823	2-thioglucose														
186.08921	187.096486	205.107051	olivomycose (L-chromose B)														
187.120844	188.128120	206.138685	4-N,N-dimethylamino-4-deoxy-5-C-methyl-l-rhamnose														
188.10486	189.112136	207.122701	2,3,4-tri-O-methylrhamnose														
200.10486	201.112136	219.122701	4-O-acetyl-L-arcanose														
201.100109	202.107385	220.117950	3-N-acetyl-D-ravidosamine														
201.100109	202.107385	220.117950	3-O-carbamoyl-L-noviose														
202.12051	203.127786	221.138351	L-nogalose														
215.115759	216.123035	234.133600	4-O-acetyl-D-ravidosamine														
217.095024	218.102300	236.112865	3-O-carbamoyl-4-O-methyl-L-noviose														
243.110674	244.117950	262.128515	3-N-acetyl-4-O-acetyl-D-ravidosamine														
281.126324	282.133600	300.144165	3-(5'-methyl-2'-pyrrolylcarbonyl-)4-O-methyl-L-noviose														
				NT	4,6-DH	2,3-DH	3,4-DH	3-KR	4-KR	(3,5-)E	3,4-IM	FuPyIM	2-AT	3-AT	4-AT	C3-MT	C5-MT
					DH			KR		IM			AT			AT	

Common bacterial sugar monomers

Y-Mass shift [Da]	B-mass shift [Da]	C-mass shift [Da]	Sugar
132.04244	133.04964	151.0602	D-arabinose
146.05744	147.06464	165.0752	L-rhamnose
146.05744	147.06464	165.0752	L-fucose
146.05744	147.06464	165.0752	D-rhamnose
161.06844	162.07564	180.0862	D-glucosamine
162.05244	163.05964	181.0702	D-glucose
162.05244	163.05964	181.0702	D-galactose
162.05244	163.05964	181.0702	D-mannose
176.03244	177.03964	195.0502	glucuronic acid
176.03244	177.03964	195.0502	D-galacturonic acid
176.06844	177.07564	195.0862	6-methoxy-D-glucose
176.06844	177.07564	195.0862	3-methoxy-D-mannose
203.079	204.08664	222.0972	N-acetyl-D-glucosamine
203.07944	204.08664	222.0972	N-acetyl-D-galactose
220.05844	221.06564	239.0762	ketodeoxyoctonic acid
291.09542	292.10262	310.11318	N-5-acetylneuraminic acid

Legend	
NT	nucleotidylyltransferase
4,6-DH	4,6-dehydratase
2,3-DH	2,3-dehydratase
3,4-DH	3,4-dehydratase
3-KR	3-ketoreductase
4-KR	4-ketoreductase
E	epimerase
3,4-IM	3,4-isomerase
FuPyIM	fucopyranose isomerase
2-AT	2-aminotransferase
3-AT	3-aminotransferase
4-AT	4-aminotransferase
C3-MT	C3-methyltransferase
C5-MT	C5-methyltransferase
N-MT	N-methyltransferase
N,N-MT	N,N-dimethyltransferase
O-MT	O-methyltransferase
N-ET	N-ethyltransferase
AcT	acetyltransferase
CarbT	carbamoyltransferase
PyT	pyrrolyltransferase
OxRed	oxidoreductase
oxDA	oxidative deaminase
Dhg	dehydrogenase
N-Ox	N-oxidase

Table 2 (continued): MS-glycogenetic code. Grey square indicates enzyme is present in corresponding sugar pathway. White square indicates enzyme is not present in corresponding sugar pathway.

Sugar	MT												Pathway reference	
	N-MT	N,N-dMT	O-MT	N-ET	AcT	CarbT	PyT1/2	OxRed	oxDA	Dhg	ThiS	N-Ox	(Pubmed ID)	(Genbank ID)
L-aculose								■					17395717, 12137949, 11016846	AF264025, AF257324
L-cinerulose A													17395717, 12137949, 11016846	AF264025, AF257324
L-rhodinose													17395717, 12137949, 11016846	AF264025, AF257324
rednose													17395717, 12137949, 11016846	AF264025, AF257324
L-cinerulose B													17395717, 12137949, 11016846	AF264025, AF257324
O-methyl-L-amicetose			■										16632249	DQ230532
4-O-methyl-L-rhodinose			■										12770825	AF521085
L-daunosamine													19777229	STMDNRLM
L-acosamine													12596194	N/A
L-ristosamine													12617516	AB088119
D-digitoxose													16723573	DQ176871
L-digitoxose													17985890	EU301739
2-deoxy-L-fucose													12137949, 11016846	AF264025, AF257324
D-olivose													10770761	AF055579
D-oliose													10660060, 11254130, 15712316	X89899
L-boivinose													N/A	N/A
4-oxo-L-vancosamine													10390204	Y16952
D-forosamine		■											11358695, 11544225	AY007564
L-actinosamine													12596194	N/A
3-epi-L-vancosamine													9545426	AJ223998, AJ223999
L-vancosamine													12596194	HE589771
L-vicenisamine	■												15112997	AB086653
D-chalcose			■						■				15561847	AY509120
D-mycarose													10220165, 10658660	AF055922, AF147704, SFU08223
3-O-methyl-L-olivose/L-oleandrose			■										10449723	AB032523
olivomose			■										15112992	AJ578458
D-mycosamine													10873841	AF263912
4-deoxy-4-thio-D-digitoxose											■		12183629	AF497482
D-fucofuranose													12822997	AY233211
D-fucose													15911378	AJ786382, AJ786383
L-rhamnose													11358695, 11544225	AY007564
L-quinovose													N/A	N/A
D-quinovose													N/A	N/A
4-N-ethyl-4-amino-3-O-methoxy-2,4,5-trideoxypentose				■						■			12183629	AF497482
D-3-N-methyl-4-O-methyl-L-ristosamine	■		■										12617516	AB088119
N,N-dimethyl-L-pyrrolosamine													N/A	N/A
D-desosamine		■							■				9353926	SEU77459
L-megosamine		■											10972798	AF263245
nogalamine		■											9349712	AF187532
L-rhodosamine		■											12137949, 11016846	AF264025, AF257324
D-angolosamine		■											15271354	AY196994
kedarosamine													N/A	N/A
L-noviose													10770754	AF170880
L-cladinose			■										9353926	SEU77459
2'-N-methyl-D-fucosamine	■												15797213	AY117439
4-deoxy-4-methylthio-a-D-digitoxose													14528002	N/A
D-digitalose			■										15911378	AJ786382, AJ786383
3-O-methyl-rhamnose			■										12183629	AF497482
2-O-methyl-L-rhamnose			■										11358695, 11544225	AY007564
6-deoxy-3-C-methyl-L-mannose													9349712	AF187532
4,6-dideoxy-4-hydroxylamino-D-glucose												■	12183629	AF497482
hexose													12596194	HE589771
3-N,N-dimethyl-L-eremosamine		■											15271354	AY196994
namenamicin sugar C													N/A	N/A
chromose D (4-O-acetyl-β-D-oliose)					■								15112992	AJ578458
4-O-carbamoyl-D-olivose						■							16207901	DQ149987
D-ravidosamine		■											21264378	FN565485
3-N,N-dimethyl-D-mycosamine		■											10220165, 10658660	AF055922, AF147704, SFU08223
2,3-O-dimethyl-L-rhamnose			■										11358695, 11544225	AY007564
2,4-O-dimethyl-L-rhamnose			■										11358695, 11544225	AY007564
3,4-O-dimethyl-L-rhamnose			■										11358695, 11544225	AY007564
2-thioglucose											■		20443562	HM055942
olivomycose (L-chromose B)					■								15112992	AJ578458
4-N,N-dimethylamino-4-deoxy-5-C-methyl-l-rhamnose		■			■								12183628	AY048670
2,3,4-tri-O-methylrhamnose			■										11358695, 11544225	AY007564
4-O-acetyl-L-arcanose			■		■								10767533	AB088224
3-N-acetyl-D-ravidosamine		■			■								21264378	FN565485
3-O-carbamoyl-L-noviose						■							10770754	AF170880
L-nogalose			■										9349712	AF187532
4-O-acetyl-D-ravidosamine		■			■								21264378	FN565485
3-O-carbamoyl-4-O-methyl-L-noviose			■			■							10770754	AF170880
3-N-acetyl-4-O-acetyl-D-ravidosamine		■			■								21264378	FN565485
3-(5'-methyl-2'-pyrrolylcarbonyl-)4-O-methyl-L-noviose			■				■						12480894	AF329398
	N-MT	N,N-dMT	O-MT	N-ET	AcT	CarbT	PyT1/2	OxRed	oxDA	Dhg	ThiS	N-Ox		
	MT													

Legend	
NT	nucleotidylyltransferase
4,6-DH	4,6-dehydratase
2,3-DH	2,3-dehydratase
3,4-DH	3,4-dehydratase
3-KR	3-ketoreductase
4-KR	4-ketoreductase
E	epimerase
3,4-IM	3,4-isomerase
FuPyIM	fucopyranose isomerase
2-AT	2-aminotransferase
3-AT	3-aminotransferase
4-AT	4-aminotransferase
C3-MT	C3-methyltransferase
C5-MT	C5-methyltransferase
N-MT	N-methyltransferase
N,N-MT	N,N-dimethyltransferase
O-MT	O-methyltransferase
N-ET	N-ethyltransferase
AcT	acetyltransferase
CarbT	carbamoyltransferase
PyT	pyrrolyltransferase
OxRed	oxidoreductase
oxDA	oxidative deaminase
Dhg	dehydrogenase
N-Ox	N-oxidase

Table 3: Connection of known GNP chemo- and genotypes by the MS-glycogenetic code. Reference GNP chemotypes were analyzed in sugar-specific MS^n neutral losses or B-/C-ion fragments. MS/MS candidate sugars were identified based on observed sugar masses (see Table 4). Corresponding reference GNP genotypes from GenBank were then analyzed in predicted common and specific glycosylation genes (see Table 4). Predicted glycosylation genes were analyzed in matching the biosynthesis of the MS^n candidate sugars utilizing Table 2. A GNP chemotype-to-genotype connection by the MS-glycogenetic code was a match of MS^n and genetic candidate sugars (green). GNPs with an unsuccessful chemotype-genotype connection were highlighted in red. Abbreviations: see Table 2.

#	Reference GNP chemotype	Instrument	Reference	Observed MS/MS sugar footprint [Da] (see Table 4)		MS/MS candidate sugars	Genetic candidate sugars	Matching reference pathway based on Table 2	BLAST analysis of gene clusters (see Table 4)		Gene cluster (Genbank #)
				Y-ion neutral loss	B-ion				Specific glycosylation genes	Common glycosylation genes	
1	phenalinolactone	ESI-LTQ-FT-MS	this study	128.072 (Y)	129.094 (B)	O-methyl-L-amicetose 4-O-methyl-L-rhodinose	O-methyl-L-amicetose 4-O-methyl-L-rhodinose	2,3-DH, 3,4-DH, 3-KR, 4-KR, E, O-MT	2,3-DH, 3,4-DH, 3-KR, 4-KR, E, O-MT	NT, 4,6-DH, GT	DQ230532
2	daunomycin	ESI-Q-TOF-MS	Metlin, ID 550		130.0852 (B)	L-daunosamine L-acosamine	L-daunosamine	2,3-DH, AT, E, 4-KR	2,3-DH, AT, E, 4-KR	NT, 4,6-DH, GT	STMDNRLM STMDNRQ SPU77891 STMDNRI
						L-ristosamine L-ristosamine	L-ristosamine	2,3-DH, AT, E, 4-KR			
3	staurosporine	ESI-Q-TOF-MS	Metlin, ID 3307	129.080 (Y)	130.086 (B)	L-daunosamine L-acosamine	L-daunosamine	2,3-DH, AT, E, 4-KR	2,3-DH, AT, E, 4-KR, O-MT, MT	NT, 4,6-DH, GT (2x)	AB088119
						L-ristosamine L-ristosamine	L-ristosamine	2,3DH, AT, E, 4KR			
4	oleandomycin	ESI-Q-TOF-MS	Metlin, ID 44103	144.084 (Y)	145.086 (B)	D-chalcose D-mycarose			2,3-DH, 3,4-DH, 3-KR, E, 4-KR, E, 3,4-DH/AT, O-MT, N, N-MT, oxDA	NT, 4,6-DH, GT (4x)	AF055579 AJ002638
						L-oleandrose olivomose	L-oleandrose olivomose	2,3-DH, 3-KR, E, 4-K-R, O-MT 2,3-DH, 3-KR, 4-KR, O-MT			
					158.116 (B)	4-N-ethyl-4-amino-3-O-methoxy-2,4,5-trideoxypentose D-3-N-methyl-4-O-methyl-L-ristosamine					
						D-desosamine	D-desosamine	3,4-DH, oxDA, AT, N,N-MT			
						N,N-dimethyl-L-pyrrolosamine		N/A			
						L-megosamine	L-megosamine	2,3-DH, AT, E, 4-KR, N,N-MT			
						L-nogalamine	nogalamine	2,3-DH, AT, E, 4-KR, N,N-MT			
						L-rhodosamine	L-rhodosamine	2,3-DH, AT, E, 4-KR, N,N-MT			
						D-angolosamine	D-angolosamine	2,3-DH, AT, 4-KR, N,N-MT			
						kedarosamine		N/A			
5	spinosyn A	ESI-TOF-MS	Ferrer, I. *et al. Trends Anal. Chem.* **24**, 671-682 (2005)	188.105 (Y)		2,3,4-tri-O-methylrhamnose			2,3-DH, 3,4-DH, AT, 3-KR, O-MT (4x), N,N-MT	GT	AY007564
					142.123 (B)	D-forosamine	D-forosamine	2,3-DH, 3-KR, 3,4-DH, AT, N,N-MT			
6	vancomycin	ESI-Q-TOF-MS	Metlin, ID 582	143.082 (Y1)	144.100 (B)	L-actinosamine			2,3-DH, 4-KR, E, C-MT	GT (2x)	
						3-epi-L-vancosamine L-vancosamine	3-epi-L-vancosamine L-vancosamine	2,3-DH, 3-KR, E, 4-KR, C-MT			
						L-vicenisamine calicheamicin deoxyaminopentose L-actinosamine					
				162.061 (Y2)		hexose					
7	tylosin	ESI-Q-TOF-MS	Metlin, ID 40964	144.076 (Y)	145.085 (B)	D-chalcose D-mycarose	D-mycarose	2,3-DH, 3-KR, C-MT, E, 4-KR	2,3-DH, 3-KR, 4-KR (2x), E (2x), AT, C-MT, O-MT	NT, 4,6-DH, GT (2x)	AF055922 AF147704
						L-oleandrose olivomose	L-oleandrose olivomose	2,3-DH, 3-KR, O-MT, E, 4-KR 2,3-DH, 3-KR, O-MT, 4-KR			SFU08223
8	avermectin B1a	ESI-Q-TOF-MS	Metlin, ID 40988	144.075 (Y1)		D-chalcose			2,3-DH, 3-KR, 4-KR, E, O-MT	NT, 4,6-DH, GT	AB032523
				144.077 (Y2)		D-mycarose					
						L-oleandrose olivomose	L-oleandrose olivomose	2,3-DH, 3-KR, O-MT, E, 4-KR 2,3-DH, 3-KR, O-MT, 4-KR			
9	nystatin	ESI-Q-TOF-MS	Metlin, ID 40970	no sugar fragmentation		no sugar observed					
10	chartreusin	ESI-MS	Xu, Z. *et al.* *Chem. Biol.* (2005)	160(Y)		4-deoxy-4-methylthio-a-D-digitoxose	4-deoxy-4-methylthio-a-D-digitoxose		4-KR, E/KR, MT	NT, 4,6-DH, GT(2x)	AJ786382 AJ786383
						D-digitalose 3-O-methyl-rhamnose 2-O-methyl-L-rhamnose	D-digitalose 3-O-methyl-rhamnose 2-O-methyl-L-rhamnose	4-KR, O-MT 4-KR, E, O-MT 4-KR, E, O-MT			
						6-deoxy-3-C-methyl-L-mannose	6-deoxy-3-C-methyl-L-mannose				

Table 3 (continued): Connection of known GNP chemo- and genotypes by the MS-glycogenetic code. Reference GNP chemotypes were analyzed in sugar-specific MS^n neutral losses or B-/C-ion fragments. MS/MS candidate sugars were identified based on observed sugar masses (see Table 4). Corresponding reference GNP genotypes from GenBank were then analyzed in predicted common and specific glycosylation genes (see Table 4). Predicted glycosylation genes were analyzed in matching the biosynthesis of the MS^n candidate sugars utilizing Table 2. A GNP chemotype-to-genotype connection by the MS-glycogenetic code was a match of MS^n and genetic candidate sugars (green). GNPs with an unsuccessful chemotype-genotype connection were highlighted in red. Abbreviations: see Table 2.

#	Reference GNP chemotype	Instrument	Reference	Observed MS/MS sugar footprint [Da] (see Table 4)		MS/MS candidate sugars	Genetic candidate sugars	Matching reference pathway based on Table 2	BLAST analysis of gene clusters (see Table 4)		Gene cluster (Genbank #)
				Y-ion neutral loss	B-ion				Specific glycosylation genes	Common glycosylation genes	
11	erythromycin A	ESI-Q-TOF-MS	Metlin, ID 2573	158.0928 (Y)		L-cladinose	L-cladinose	2,3-DH, 3-KR, E, C-MT, 4-KR, MT	2,3-DH, 3,4-DH, 3-KR, 4-KR, E, 3,4-IM, 3,4-DH/AT, C-MT,	GT (2x)	AM420293 SEU77459
					158.1168 (B)	4-N-ethyl-4-amino-3-O-methoxy-2,4,5-trideoxypentose					
						D-3-N-methyl-4-O-methyl-L-ristosamine					
						D-desosamine	D-desosamine	3,4-DH, oxDA, AT, N, N-MT			
						N,N-dimethyl-L-pyrrolosamine		N/A			
						L-megosamine	L-megosamine	2,3-DH, AT, E, 4-KR, N,N-MT			
						L-nogalamine	nogalamine	2,3-DH, AT, E, 4-KR, N,N-MT			
						L-rhodosamine	L-rhodosamine	2,3-DH, AT, E, 4-KR, N,N-MT			
						D-angolosamine	D-angolosamine	2,3-DH, AT, 4-KR, N, N-MT			
						kedarosamine		N/A			
12	megalomicin	N/A	Urseglio, M. *et al. Appl. Env. Microbiol.* **76**, 3869-3877 (2010)	144.08 (Y1)		D-chalcose			2,3-DH (2x), 3-KR, 4-KR (4x), 3,4-IM, E, AT, N,N-MT, AcT	GT (4x)	AF263245
						D-mycarose					
						L-oleandrose	L-oleandrose	2,3-DH, 3-KR, O-MT, E, 4-KR			
						olivomose	olivomose	2,3-DH, 3-KR, O-MT, 4-KR			
				156.91 (Y2)	158.12 (B), 174.11 (C)	4-N-ethyl-4-amino-3-O-methoxy-2,4,5-trideoxypentose					
						D-3-N-methyl-4-O-methyl-L-ristosamine					
						D-desosamine					
						N,N-dimethyl-L-pyrrolosamine					
						L-megosamine	L-megosamine	2,3-DH, 4-KR, E, AT, N,N-MT			
						L-nogalamine	nogalamine	2,3-DH, 4-KR, E, AT, N,N-MT			
						L-rhodosamine	L-rhodosamine	2,3-DH, 4-KR, E, AT, N,N-MT			
						D-angolosamine	D-angolosamine	2,3-DH, 4-KR, AT, N,N-MT			
						kedarosamine					
13	aclacinomycin A	ESI-Q-TOF-MS	Metlin, ID 551	112.0495(Y2)	113.0598(B) 131.0697 (C)	L-cinerulose A	L-cinerulose A	2,3-DH, 3,4-DH, 3-KR, 4-KR, E	DH, 3,4-DH, 3-KR, 4-KR, E, AT, N,N-MT	NT, 4,6-DH, GT (3x)	AF264025 AF257324
				130.0675 (Y1)		D-digitoxose	D-digitoxose	2,3-DH, 3-KR, 4-KR			
						L-digitoxose	L-digitoxose	2,3-DH, 3-KR, 4-KR, E			
						2-deoxy-L-fucose	2-deoxy-L-fucose	2,3-DH, 3-KR, 4-KR, E			
						D-olivose	D-olivose	2,3-DH, 3-KR, 4-KR, E			
						D-oliose	D-oliose	2,3-DH, 3-KR, 4-KR			
						esparamicin A1 sugar 1					
14	novobiocin	ESI-Q-TOF-MS	Metlin, ID 43281	217.0938(Y)	218.1035(B)	3-O-carbamoyl-4-O-methyl-L-noviose	3-O-carbamoyl-4-O-methyl-L-noviose	E, 4-KR, C-MT, O-MT, CarbT	E, 4-KR, C-MT, O-MT, CarbT, MT	NT, 4,6-DH, GT	AF170880
15	neocarzinostatin	FAB-MS	Edo, K., Mizugaki, M. *Tetrahedron Lett.* **26**, 331-334 (1985)		160 (B)	2'-N-methyl-D-fucosamine	2'-N-methyl-D-fucosamine	2,3-DH, 4-KR, AT, N-MT	2,3-DH, 4-KR, AT, N-MT, O-MT	NT, 4,6-DH, GT	AY117439
16	amphotericin B	ESI-Q-TOF-MS	this study (Supplementary Table 4)	163.084(Z)		D-mycosamine	D-mycosamine	3,4-IM (CytP450), AT	3,4-IM (CytP450), AT	4,6DH, GT	AF357202
17	lankamycin	CI-MS	**59**, 1886-1894 (1976)	200 (Y1)	201 (B)	4-O-acetyl-L-arcanose	4-O-acetyl-L-arcanose	2,3-DH, 3-KR, 4-KR, E, C-MT, O-MT, AcT	2,3-DH, 3-KR (2x), 4-KR, 3,4-IM, E, 3,4-DH/AT, C-MT, oxDA, AcT, O-MT (2x),	NT, 4,6DH, GT (3x)	AB088224
				144 (Y2)	145 (B)	D-chalcose					
						D-mycarose					
						L-oleandrose	L-oleandrose	2,3-DH, 3-KR, O-MT, E, 4-KR			
						olivomose	olivomose	2,3-DH, 3-KR, O-MT, 4-KR			
18	chalcomycin	ESI-Q-TOF-MS	this study (Supplementary Table 4)	174.062 (Y1)		2,3-O-dimethyl-L-rhamnose	2,3-O-dimethyl-L-rhamnose	4-KR, E, O-MT	3-KR, 4-KR, E, 3,4-DH/AT, O-MT (3x), oxDA	NT, 4,6-DH, GT (2x)	AY509120
						2,4-O-dimethyl-L-rhamnose	2,4-O-dimethyl-L-rhamnose				
						3,4-O-dimethyl-L-rhamnose	3,4-O-dimethyl-L-rhamnose				
				144.072 (Y2)	145.071 (B)	D-chalcose	D-chalcose	AT, oxDA, 3-KR, O-MT			
						D-mycarose					
						L-oleandrose					
						olivomose					
19	Sch40832	ESI-MS	Puar, M. S. *et al. J Antibiotics* (1998)	130 (Y)		D-digitoxose			none	GT	GG657738.1
						L-digitoxose					
						2-deoxy-L-fucose					
						D-olivose					
						D-oliose					
						L-boivinose					

Table 4: MS/MS-sugar fragmentation and glycosylation gene prediction from chemotypes and genotypes of characterized glycosylated natural products (GNPs) from databases (Table 3) or self-acquired MS/MS data.

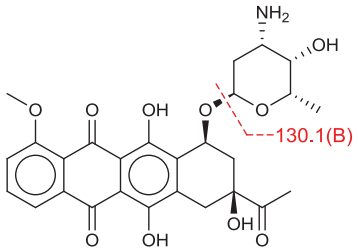
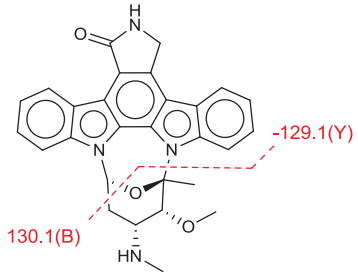
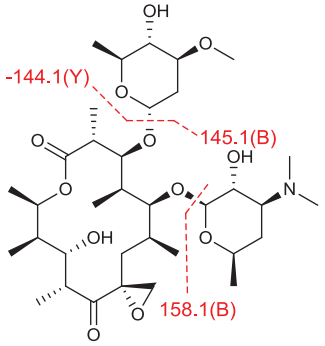
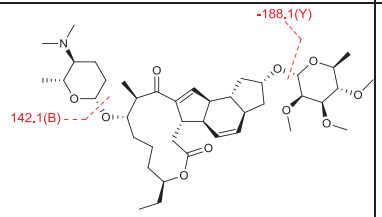
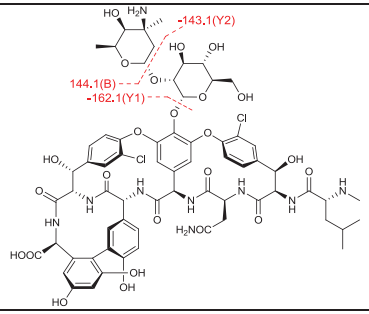
<table>
<tr><th>GNP</th><th>MS/MS glycosylation footprints of characterized GNP</th><th>Glycosylation genes in gene cluster of characterized GNP</th></tr>
<tr><td>daunomycin</td><td> 
</td><td>
<table>
<tr><th>Gene</th><th>Predicted function</th><th>Closest functional homolog by BLAST (similiarity/identity) [%/%]</th></tr>
<tr><td>dnmT</td><td>2,3DH</td><td>BAD08363.1| dTDP-4-keto-6-deoxyglucose 2,3-dehydratase [Streptomyces halstedii] (73/57)</td></tr>
<tr><td>dnrH</td><td>GT</td><td>AAB08020.1| glycosyltransferase [Streptomyces sp. C5] (93/93)</td></tr>
<tr><td>dnrS</td><td>GT</td><td>ABC00729.1| CosG (glycosyltransferase) [Streptomyces olindensis] (75/61)</td></tr>
<tr><td>dnmU</td><td>E</td><td>AF257324_2 AknL (epimerase) [Streptomyces galilaeus] (83/73)</td></tr>
<tr><td>dnmV</td><td>4KR</td><td>CAA12010.1| SnogG [Streptomyces nogalater] (63/53)</td></tr>
<tr><td>dnrM</td><td>4,6DH</td><td>ZP_06913915.1| dTDP-glucose 4,6-dehydratase [Streptomyces pristinaespiralis ATCC 25486] (62/57)</td></tr>
<tr><td>dnmL</td><td>NT</td><td>ZP_06913914.1| dTDP-glucose synthase [Streptomyces pristinaespiralis ATCC 25486] (89/79)</td></tr>
<tr><td>dnrJ</td><td>AT</td><td>CCD33157.1| putative C-3 aminotransferase [Amycolatopsis orientalis] (85/74)</td></tr>
</table>
</td></tr>
<tr><td>staurosporine</td><td> 
</td><td>
<table>
<tr><th>Gene</th><th>Predicted function</th><th>Closest functional homolog by BLAST (similiarity/identity) [%/%]</th></tr>
<tr><td>staB</td><td>4,6DH</td><td>ZP_05008524.1| dTDP-glucose 4,6-dehydratase [Streptomyces clavuligerus ATCC 27064] (89/81)</td></tr>
<tr><td>staA</td><td>NT</td><td>ZP_05008523.1| glucose-1-phosphate thymidyltransferase [Streptomyces clavuligerus ATCC 27064] (87/78)</td></tr>
<tr><td>staG</td><td>GT</td><td>CAD58668.1| putative glycosyltransferase [Streptomyces longisporoflavus]</td></tr>
<tr><td>staN</td><td>GT</td><td>CAD58669.1| putative P450 protein [Streptomyces longisporoflavus] (99/98)</td></tr>
<tr><td>staMA</td><td>O-MT</td><td>ZP_06776292.1| Staurosporine biosynthesis (O-)methyltransferase StaMA [Streptomyces clavuligerus ATCC 27064] (81/71)</td></tr>
<tr><td>staI</td><td>AT</td><td>CCD33157.1| putative C-3 aminotransferase [Amycolatopsis orientalis] (92/85)</td></tr>
<tr><td>staJ</td><td>2,3DH</td><td>CAC48374.1| putative NDP-hexose 2,3-dehydratase [Amycolatopsis balhimycina DSM 5908] (85/75)</td></tr>
<tr><td>staK</td><td>4KR</td><td>ZP_06776290.1| Staurosporine biosynthesis 4-ketoreductase StaK [Streptomyces clavuligerus ATCC 27064] (84/76)</td></tr>
<tr><td>staE</td><td>E</td><td>ZP_06776288.1| Staurosporine biosynthesis 3,5-epimerase StaE [Streptomyces clavuligerus ATCC 27064] (92/82)</td></tr>
<tr><td>staMB</td><td>MT</td><td>YP_001537181.1| type 11 methyltransferase [Salinispora arenicola CNS-205] (92/81)</td></tr>
</table>
</td></tr>
</table>

Table 4: MS/MS-sugar fragmentation and glycosylation gene prediction from chemotypes and genotypes of characterized glycosylated natural products (GNPs) from databases (Table 3) or self-acquired MS/MS data.

GNP	MS/MS glycosylation footprints of characterized GNP	Glycosylation genes in gene cluster of characterized GNP
oleando-mycin		
spinosyn A		
vancomycin		

Gene	Predicted function	Closest functional homolog by BLAST (similiarity/identity) [%/%]
oleW	3KR	CCH33151.1| NDP-hexose 3-ketoreductase [Saccharothrix espanaensis DSM 44229] (69/55)
oleV	2,3DH	ZP_10457036.1| NDP-hexose 2,3-dehydratase [Streptomyces acidiscabies 84-104] (74/60)
oleL	E	ADI50280.1| sugar 3,5-epimerase [Streptomyces sp. MK730-62F2] (66/56)
oleS	NT	ZP_07308385.1| glucose-1-phosphate thymidylyltransferase [Streptomyces viridochromogenes DSM 40736] (85/75)
oleU	4KR	ZP_07308387.1| dTDP-4-dehydrorhamnose reductase [Streptomyces viridochromogenes DSM 40736]
oleE	4,6DH	ZP_07308386.1| dTDP-glucose 4,6-dehydratase [Streptomyces viridochromogenes DSM 40736] (84/76)
oleNI	3,4DH/AT	YP_001102983.1| eryCIV NDP-6-deoxyhexose 3,4-dehydratase [Saccharopolyspora erythraea NRRL 2338] (79/66)
oleT	ox.DA	YP_001102982.1| EryCV NDP-4,6-dideoxyhexose 3,4-enoyl reductase [Saccharopolyspora erythraea NRRL 2338] (76/65)
oleNII	AT	YP_001103001.1| erythromycin biosynthesis transaminase EryCI [Saccharopolyspora erythraea NRRL 2338]
oleI	GT	ZP_07289948.1| oleandomycin glycosyltransferase [Streptomyces sp. C] (59/44)
oleD	GT	YP_006248311.1| oleandomycin glycosyltransferase [Streptomyces hygroscopicus subsp. jinggangensis 5008] (92/87)
oleY	O-MT	BAC57026.1| methyltransferase [Micromonospora griseorubida] (58/41)
oleM1	N,N-MT	YP_001102985.1| TDP-desosamine-N-dimethyltransferase [Saccharopolyspora erythraea NRRL 2338] (77/68)
oleG2	GT	YP_001102993.1| glycosyl transferase [Saccharopolyspora erythraea NRRL 2338] (71/54)
oleG1	GT	YP_001102993.1| glycosyl transferase [Saccharopolyspora erythraea NRRL 2338] (70/54)

Gene	Predicted function	Closest functional homolog by BLAST (similiarity/identity) [%/%]
spnS	N,N-MT	CBH32796.1| putative N,N-dimethyltransferase [Streptomyces ravidus] (67/50)
spnR	AT	ADM72812.1| putative NDP-hexose aminotransferase [Streptomyces aureofaciens] (78/64)
spnQ	3,4DH	AF264025_3 putative 3-dehydratase [Streptomyces galilaeus] (85/72)
spnP	GT	ZP_06826086.1| glycosyltransferase family 28 domain-containing protein [Streptomyces sp. SPB74] (61/45)
spnO	2,3DH	AF324838_29 putative dNDP-4-keto-6-deoxy-glucose-2,3-dehydratase SimB3 [Streptomyces antibioticus] (67/53)
spnN	3KR	AF264025_4 putative 3-ketoreductase [Streptomyces galilaeus] (68/53)
spnK	O-MT	BAJ05901.1| putative sugar O-methyltransferase [Streptomyces sp. SANK 60405] (69/53)
spnI	GT	YP_004965109.1| Gene info linked to YP_004965109.1glycosyl transferase [Streptomyces bingchenggensis BCW-1] (54/41)
spnI	O-MT	ZP_05000469.1| O-methyltransferase [Streptomyces sp. Mg1] (59/42)
spnH	O-MT	AEP40941.1| sugar O-methyltransferase [Nocardiopsis sp. FU40] (77/62)
spnF	O-MT	ACY01395.1| O-methyl transferase [Streptomyces platensis subsp. rosaceus] (55/39)

Gene	Predicted function	Closest functional homolog by BLAST (similiarity/identity) [%/%]
gtfD	GT	CAA76553.1| glycosyltransferase [Amycolatopsis balhimycina DSM 5908] (81/71)
gtfE	GT	CAA76552.1| glycosyltransferase [Amycolatopsis balhimycina DSM 5908] (88/83)
vcaC	C-MT	CAC48364.1| putative C-3 methyl transferase [Amycolatopsis balhimycina DSM 5908] (96/93)
vcaA	2,3DH	CAC48374.1| putative NDP-hexose 2,3-dehydratase [Amycolatopsis balhimycina DSM 5908] (92/87)
vcaE	4KR	AEI58885.1| 4-ketoreductase [Amycolatopsis orientalis HCCB10007] (85/76)
vcaD	E	CAC48377.1| putative 3,5 epimerase [Amycolatopsis balhimycina DSM 5908] (94/88)
vcaB	AT	CAC48376.1| putative C-3 amino transferase [Amycolatopsis balhimycina DSM 5908] (96/90)

Table 4: MS/MS-sugar fragmentation and glycosylation gene prediction from chemotypes and genotypes of characterized glycosylated natural products (GNPs) from databases (Table 3) or self-acquired MS/MS data.

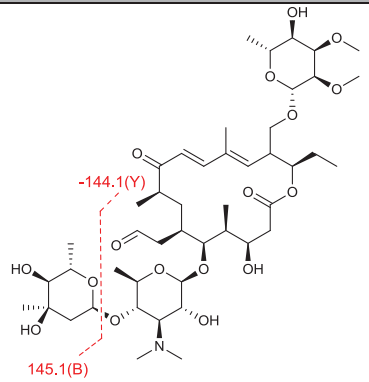
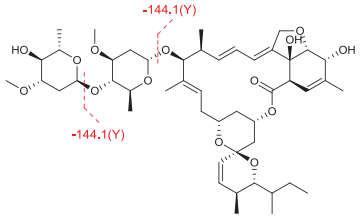
<table>
<tr><th>GNP</th><th>MS/MS glycosylation footprints of characterized GNP</th><th>Glycosylation genes in gene cluster of characterized GNP</th></tr>
<tr><td>tylosin</td><td> 
</td><td>
<table>
<tr><th>Gene</th><th>Predicted function</th><th>Closest functional homolog by BLAST (similiarity/identity) [%/%]</th></tr>
<tr><td>tylN</td><td>GT</td><td>ABV49604.1| glycosyltransferase [Streptomyces eurythermus] (83/75)</td></tr>
<tr><td>tylE</td><td>O-MT</td><td>ABV49603.1| O-methyltransferase [Streptomyces eurythermus] (83/75)</td></tr>
<tr><td>tylD</td><td>4KR</td><td>ABV49602.1| NDP-4-ketoreductase [Streptomyces eurythermus] (74/62)</td></tr>
<tr><td>tylJ</td><td>E</td><td>ABV49598.1| NDP-hexose 3,(5)-epimerase [Streptomyces eurythermus] (84/72)</td></tr>
<tr><td>tylCI</td><td>3KR</td><td>YP_001102994.1| TDP-4-keto-6-deoxyhexose 2,3-reductase [Saccharopolyspora erythraea NRRL 2338] (84/74)</td></tr>
<tr><td>tylCIV</td><td>4KR</td><td>ABW91155.1| NDP-hexose 4-ketoreductase [Streptomyces eurythermus] (70/60)</td></tr>
<tr><td>tylCII</td><td>C-MT</td><td>ABW91157.1| NDP-hexose 3-C-methyltransferase [Streptomyces eurythermus] (88/78)</td></tr>
<tr><td>tylCV</td><td>GT</td><td>ABW91158.1| glycosyltransferase [Streptomyces eurythermus] (78/68)</td></tr>
<tr><td>tylCVII</td><td>E</td><td>ABW91159.1| NDP-hexose 3,5-epimerase [Streptomyces eurythermus] (80/74)</td></tr>
<tr><td>tylB</td><td>AT</td><td>AAF59939.1| aminotransferase-like protein [Streptomyces antibioticus] (73/63)</td></tr>
<tr><td>tylA1</td><td>NT</td><td>ABV49608.1| NDP-hexose synthase [Streptomyces eurythermus] (83/76)</td></tr>
<tr><td>tylA2</td><td>4,6DH</td><td>ABV49607.1| NDP-hexose 4,6-dehydratase [Streptomyces eurythermus] (83/74)</td></tr>
<tr><td>tylCVI</td><td>2,3DH</td><td>ABV49606.1| NDP-hexose 2,3-dehydratase/thioesterase [Streptomyces eurythermus] (77/66)</td></tr>
</table>
</td></tr>
<tr><td>avermectin B1a</td><td> 
</td><td>
<table>
<tr><th>Gene</th><th>Predicted function</th><th>Closest functional homolog by BLAST (similiarity/identity) [%/%]</th></tr>
<tr><td>aveBI</td><td>GT</td><td>YP_004967697.1|glycosyl transferase family protein [Streptomyces bingchenggensis BCW-1] (59/47)</td></tr>
<tr><td>aveBII</td><td>4,6DH</td><td>YP_003132507.1| dTDP-glucose 4,6-dehydratase [Saccharomonospora viridis DSM 43017] (84/73)</td></tr>
<tr><td>aveBIII</td><td>NT</td><td>CAJ88197.1| putative glucose-1-phosphate thymidyltransferase [Streptomyces ambofaciens ATCC 23877] (85/75)</td></tr>
<tr><td>aveBIV</td><td>4KR</td><td>BAC55215.1| 4-ketoreductase [Streptomyces sp. TP-A0274] (64/54)</td></tr>
<tr><td>aveBV</td><td>E</td><td>NP_851468.1| putative NDP-4-keto-6-deoxyhexose 3,5-epimerase [Streptomyces rochei] (73/65)</td></tr>
<tr><td>aveBVI</td><td>2,3DH</td><td>AAD13549.1| NDP-hexose 2,3-dehydratase homolog [Streptomyces cyanogenus] (79/71)</td></tr>
<tr><td>aveBVII</td><td>O-MT</td><td>NP_851467.1| putative NDP-hexose 3-O-methyltransferase [Streptomyces rochei] (92/85)</td></tr>
<tr><td>aveBVIII</td><td>3KR</td><td>ZP_06594449.1| dTDP-4-keto-6-deoxy-L-hexose 2,3-reductase [Streptomyces albus J1074] (88/81)</td></tr>
</table>
</td></tr>
<tr><td>nystatin</td><td>No sugar fragmentation</td><td>-</td></tr>
</table>

Table 4: MS/MS-sugar fragmentation and glycosylation gene prediction from chemotypes and genotypes of characterized glycosylated natural products (GNPs) from databases (Table 3) or self-acquired MS/MS data.

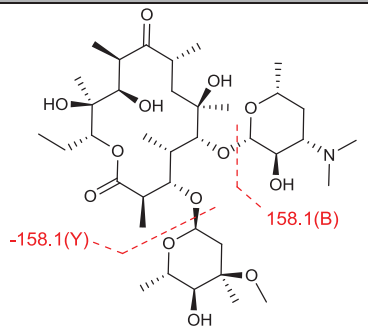
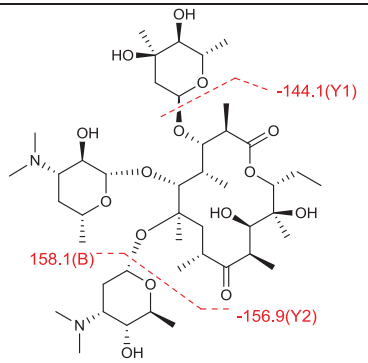
<table>
<tr><th>GNP</th><th>MS/MS glycosylation footprints of characterized GNP</th><th>Glycosylation genes in gene cluster of characterized GNP</th></tr>
<tr><td>erythromycin A</td><td> 
</td><td>
<table>
<tr><th>Gene</th><th>Predicted function</th><th>Closest functional homolog by BLAST (similiarity/identity) [%/%]</th></tr>
<tr><td>eryBVII</td><td>E</td><td>ZP_10551370.1| dTDP-4-deoxyglucose 3,5-epimerase [Streptomyces auratus AGR0001] (82/76)</td></tr>
<tr><td>eryCV</td><td>ox.DA</td><td>NP_851461.1| putative NDP-4,6-dideoxyhexose 3,4-enoyl reductase [Streptomyces rochei] (77/66)</td></tr>
<tr><td>eryCIV</td><td>3,4DH/AT</td><td>NP_851460.1| putative NDP-6-deoxyhexose 3,4-dehydratase [Streptomyces rochei] (81/71)</td></tr>
<tr><td>eryBVI</td><td>2,3DH</td><td>ZP_10457036.1| NDP-hexose 2,3-dehydratase [Streptomyces acidiscabies 84-104] (68/53)</td></tr>
<tr><td>eryCVI</td><td>N,N-MT</td><td>CAA05643.1| methyltransferase [Streptomyces antibioticus] (76/68)</td></tr>
<tr><td>eryBV</td><td>GT</td><td>AAG13915.1|AF263245_11 TDP-mycarose glycosyltransferase [Micromonospora megalomicea subsp. nigra] (86/75)</td></tr>
<tr><td>eryBIV</td><td>4KR</td><td>AAG13916.1|AF263245_12 TDP-4-keto-6-deoxyhexose 4-ketoreductase [Micromonospora megalomicea subsp. nigra] (81/70)</td></tr>
<tr><td>eryCII</td><td>3,4IM</td><td>AAG13920.1|AF263245_16 TDP-4-keto-6-deoxyglucose 3,4-isomerase [Micromonospora megalomicea subsp. nigra] (79/73)</td></tr>
<tr><td>eryCIII</td><td>GT</td><td>AAG13921.1|AF263245_17 TDP-desosamine glycosyltransferase [Micromonospora megalomicea subsp. nigra] (90/84)</td></tr>
<tr><td>eryBII</td><td>3KR</td><td>AAG13914.1|AF263245_10 TDP-4-keto-6-deoxyhexose 2,3-reductase [Micromonospora megalomicea subsp. nigra] (89/81)</td></tr>
<tr><td>eryG</td><td>MT</td><td>AAU93801.1| methylase [Aeromicrobium erythreum] (83/73)</td></tr>
<tr><td>eryBIII</td><td>C-MT</td><td>AAD41823.1|AF147704_3 NDP-hexose 3-C-methyltransferase TylCIII [Streptomyces fradiae] (85/72)</td></tr>
<tr><td>eryCI</td><td>AT</td><td>AAC68680.1| transaminase [Streptomyces venezuelae] (76/67)</td></tr>
</table>
</td></tr>
<tr><td>megalomicin</td><td> 
</td><td>
<table>
<tr><th>Gene</th><th>Predicted function</th><th>Closest functional homolog by BLAST (similiarity/identity) [%/%]</th></tr>
<tr><td>megT</td><td>2,3DH</td><td>YP_001102984.1| NDP-4-keto-6-deoxy-glucose 2,3-dehydratase [Saccharopolyspora erythraea NRRL 2338] (63/56)</td></tr>
<tr><td>megDVI</td><td>3,4IM</td><td>YP_001102992.1| TDP-4-keto-6-deoxy-glucose 3,4-isomerase [Saccharopolyspora erythraea NRRL 2338] (61/51)</td></tr>
<tr><td>megDI</td><td>GT</td><td>YP_001102993.1| glycosyl transferase [Saccharopolyspora erythraea NRRL 2338] (82/68)</td></tr>
<tr><td>megY</td><td>AcT</td><td>ZP_09183446.1| acyltransferase 3 [Streptomyces sp. S4] (62/43)</td></tr>
<tr><td>megDIII</td><td>N,N-MT</td><td>YP_005464796.1| putative methyltransferase [Actinoplanes missouriensis 431] (75/66)</td></tr>
<tr><td>megDIV</td><td>E</td><td>YP_006881544.1| dTDP-4-dehydrorhamnose 3,5-epimerase [Streptomyces venezuelae ATCC 10712] (72/58)</td></tr>
<tr><td>megDVI</td><td>4KR</td><td>AAL14256.1| NDP-4-keto-6-deoxyhexose 4-ketoreductase [Streptomyces venezuelae ATCC 10712] (67/55)</td></tr>
<tr><td>megDVII</td><td>2,3DH</td><td>YP_001102994.1| TDP-4-keto-6-deoxyhexose 2,3-reductase [Saccharopolyspora erythraea NRRL 2338] (89/81)</td></tr>
<tr><td>megBV</td><td>GT</td><td>YP_001102986.1| 6-DEB TDP-mycarosyl glycosyltransferase [Saccharopolyspora erythraea NRRL 2338] (86/75)</td></tr>
<tr><td>megBIV</td><td>4KR</td><td>YP_001102987.1| dTDP-4-keto-6-deoxy-L-hexose 4-reductase [Saccharopolyspora erythraea NRRL 2338] (81/70)</td></tr>
<tr><td>megDII</td><td>AT</td><td>ACB46490.1| sugar 3-aminotransferase [Actinomadura kijaniata] (85/74)\</td></tr>
<tr><td>megCI</td><td>GT</td><td>AAB84066.1| EryCII [Saccharopolyspora erythraea NRRL 2338] (79/72)</td></tr>
<tr><td>megCII</td><td>GT</td><td>YP_001102993.1|glycosyl transferase [Saccharopolyspora erythraea NRRL 2338] (90/84)</td></tr>
<tr><td>megBII</td><td>3-KR</td><td>YP_001102994.1| TDP-4-keto-6-deoxyhexose 2,3-reductase [Saccharopolyspora erythraea NRRL 2338] (73/63)</td></tr>
</table>
</td></tr>
</table>

Table 4: MS/MS-sugar fragmentation and glycosylation gene prediction from chemotypes and genotypes of characterized glycosylated natural products (GNPs) from databases (Table 3) or self-acquired MS/MS data.

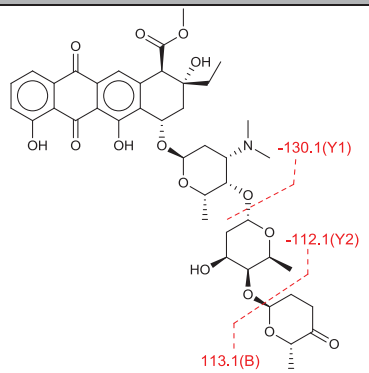
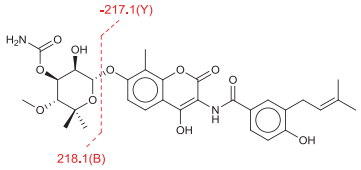
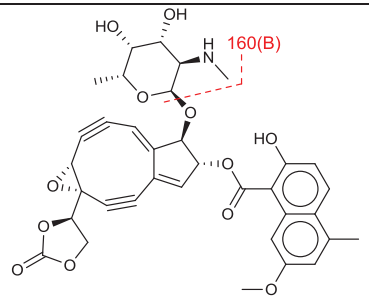
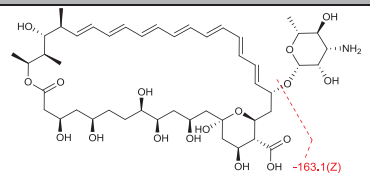
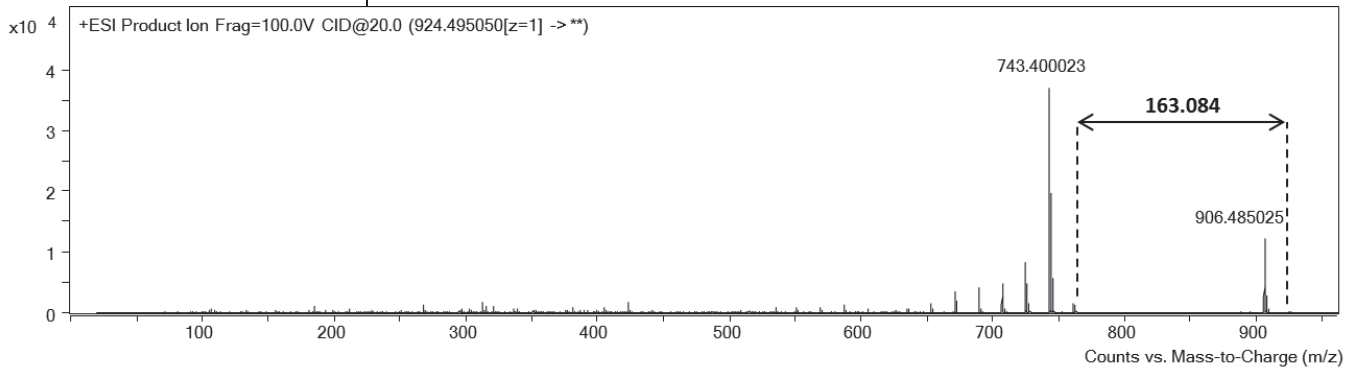
<table>
<tr><th>GNP</th><th>MS/MS glycosylation footprints of characterized GNP</th><th>Glycosylation genes in gene cluster of characterized GNP</th></tr>
<tr><td>aclacino-mycin A</td><td> 
</td><td>
<table>
<tr><th>Gene</th><th>Predicted function</th><th>Closest functional homolog by BLAST (similiarity/identity) [%/%]</th></tr>
<tr><td>aknP</td><td>3,4DH</td><td>AAL24451.1| RdmI [Streptomyces purpurascens] (87/76)</td></tr>
<tr><td>aknQ</td><td>3KR</td><td>AAA83425.2| RdmF [Streptomyces purpurascens] (79/64)</td></tr>
<tr><td>aknR</td><td>4,6DH</td><td>ZP_06826098.1| dTDP-glucose 4,6-dehydratase [Streptomyces sp. SPB74] (90/85)</td></tr>
<tr><td>aknS</td><td>GT</td><td>ZP_06826097.1| glycosyltransferase family 28 domain-containing protein [Streptomyces sp. SPB74] (89/83)</td></tr>
<tr><td>aknT</td><td>GT</td><td>ABC00727.1| CosT [Streptomyces olindensis] (55/45)</td></tr>
<tr><td>aknX2</td><td>N,N-MT</td><td>ZP_06826092.1| methyltransferase type 11 [Streptomyces sp. SPB74] (92/86)</td></tr>
<tr><td>aknY</td><td>NT</td><td>ZP_06826091.1| glucose-1-phosphate thymidylyltransferase [Streptomyces sp. SPB74] (96/91)</td></tr>
<tr><td>aknZ</td><td>AT</td><td>CCD33157.1| putative C-3 aminotransferase [Amycolatopsis orientalis] (88/80)</td></tr>
<tr><td>aknM</td><td>4KR</td><td>CAA12010.1| SnogG [Streptomyces nogalater] (63/53)</td></tr>
<tr><td>aknL</td><td>E</td><td>ZP_06826087.1| dTDP-4-dehydrorhamnose 3,5-epimerase [Streptomyces sp. SPB74] (94/88)</td></tr>
<tr><td>aknK</td><td>GT</td><td>ZP_06826086.1| glycosyltransferase family 28 domain-containing protein [Streptomyces sp. SPB74] (95/90)</td></tr>
<tr><td>aknN</td><td>DH</td><td>AEM44313.1| goadsporin biosynthetic protein [uncultured bacterium] (64/56)</td></tr>
</table>
</td></tr>
<tr><td>novobiocin</td><td> 
</td><td>
<table>
<tr><th>Gene</th><th>Predicted function</th><th>Closest functional homolog by BLAST (similiarity/identity) [%/%]</th></tr>
<tr><td>novM</td><td>GT</td><td>AAN65229.1|AF329398_19 glycosyltransferase [Streptomyces roseochromogenes subsp. oscitans] (89/84)</td></tr>
<tr><td>novN</td><td>CarbT</td><td>AAO06921.1| GdmN [Streptomyces hygroscopicus] (74/62)</td></tr>
<tr><td>novO</td><td>MT</td><td>AAG29793.1|AF235050_16 putative methyltransferase [Streptomyces rishiriensis]</td></tr>
<tr><td>novP</td><td>O-MT</td><td>AAG29794.1|AF235050_17 O-methyltransferase [Streptomyces rishiriensis] (97/95)</td></tr>
<tr><td>novS</td><td>4KR</td><td>AAN65241.1|AF329398_31 dTDP-4-keto-6-deoxyhexose reductase [Streptomyces roseochromogenes subsp. oscitans] (91/84)</td></tr>
<tr><td>novT</td><td>4,6DH</td><td>AAG29802.1| dTDP-glucose 4,6-dehydratase [Streptomyces rishiriensis] (93/89)</td></tr>
<tr><td>novU</td><td>C-MT</td><td>AAN65243.1|AF329398_33 C-methyltransferase [Streptomyces roseochromogenes subsp. oscitans] (93/89)</td></tr>
<tr><td>novV</td><td>NT</td><td>AAG29804.1| dTDP-glucose synthase [Streptomyces rishiriensis] (96/91)</td></tr>
<tr><td>novW</td><td>E</td><td>AAG29805.1| dTDP-4-keto-6-deoxyglucose 3,5-epimerase [Streptomyces rishiriensis] (93/88)</td></tr>
</table>
</td></tr>
<tr><td>neocarzino-statin</td><td> 
</td><td>
<table>
<tr><th>Gene</th><th>Predicted function</th><th>Closest functional homolog by BLAST (similiarity/identity) [%/%]</th></tr>
<tr><td>AY117439.1:10097..10894</td><td>4-KR</td><td>CAJ88311.1| putative NDP-hexose 4-ketoreductase [Streptomyces ambofaciens ATCC 23877] (80/75)</td></tr>
<tr><td>AY117439.1:14674..15672</td><td>O-MT</td><td>YP_001539692.1| O-methyltransferase family protein [Salinispora arenicola CNS-205] (77/66)</td></tr>
<tr><td>AY117439.1:26475..27179</td><td>NT</td><td>CCK25159.1| mannose-1-phosphate guanyltransferase [Streptomyces davawensis JCM 4913] (82/70)</td></tr>
<tr><td>AY117439.1:27203..28198</td><td>4,6DH</td><td>ZP_11210703.1| NAD-dependent epimerase/dehydratase [Streptomyces somaliensis DSM 40738] (78/67)</td></tr>
<tr><td>AY117439.1:28341..29060</td><td>N-MT</td><td>ZP_10580916.1| methylase involved in ubiquinone/menaquinone biosynthesis [Bradyrhizobium sp. YR681] (62/</td></tr>
<tr><td>AY117439.1:29285..30493</td><td>GT</td><td>CCK29032.1| glycosyl transferase [Streptomyces davawensis JCM 4913] (64/52)</td></tr>
<tr><td>AY117439.1:67831..68811</td><td>2,3DH</td><td>ZP_05000564.1| NDP-hexose-2,3-dehydratase [Streptomyces sp. Mg1] (73/59)</td></tr>
<tr><td>AY117439.1:11170..12750</td><td>AT</td><td>ZP_05001621.1| histidine ammonia-lyase [Streptomyces sp. Mg1] (69/54)</td></tr>
</table>
</td></tr>
</table>

Table 4: MS/MS-sugar fragmentation and glycosylation gene prediction from chemotypes and genotypes of characterized glycosylated natural products (GNPs) from databases (Table 3) or self-acquired MS/MS data.

<table>
<tr><th>GNP</th><th>MS/MS glycosylation footprints of characterized GNP</th><th>Glycosylation genes in gene cluster of characterized GNP</th></tr>
<tr><td>amphotericin B</td><td> 



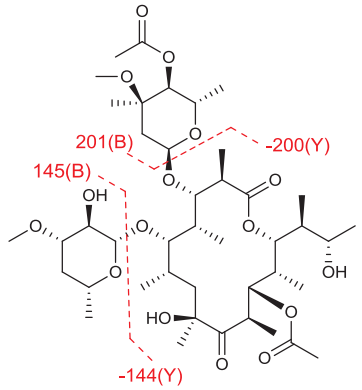
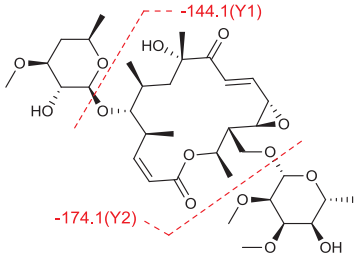
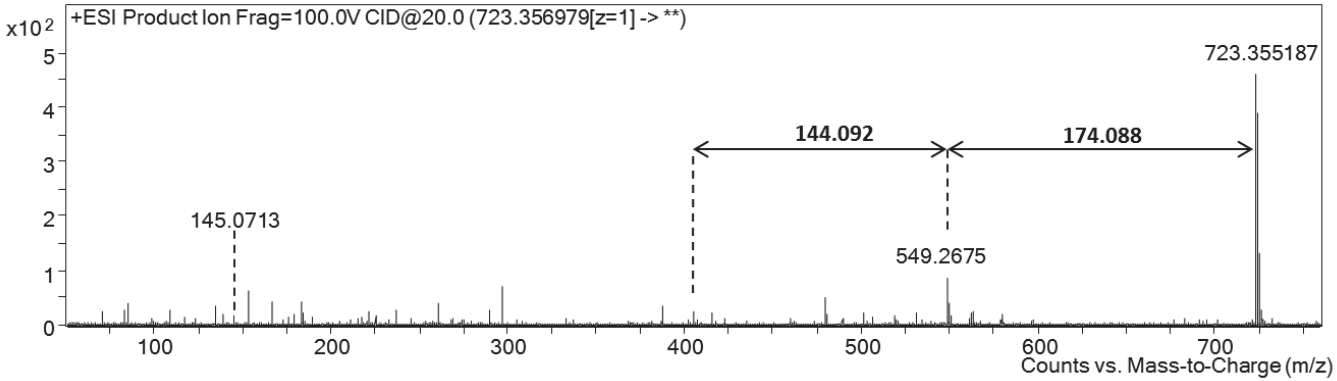
<table>
<tr><th>Gene</th><th>Predicted function</th><th>Closest functional homolog by BLAST (similiarity/identity) [%/%]</th></tr>
<tr><td>amphDIII</td><td>4,6DH</td><td>AAF71765.1|AF263912_4 NysDIII [Streptomyces noursei ATCC 11455] (94/89)</td></tr>
<tr><td>amphDII</td><td>AT</td><td>AAF71772.1|AF263912_11 NysDII [Streptomyces noursei ATCC 11455] (96/90)</td></tr>
<tr><td>amphDI</td><td>GT</td><td>AAF71773.1|AF263912_12 NysDI [Streptomyces noursei ATCC 11455] (93/85)</td></tr>
<tr><td>amphN</td><td>3,4IM (CytP450)</td><td>AAF71771.1|AF263912_10 NysN [Streptomyces noursei ATCC 11455] (91/84)</td></tr>
</table></td></tr>
<tr><td>lankamycin</td><td> 
</td><td>
<table>
<tr><th>Gene</th><th>Predicted function</th><th>Closest functional homolog by BLAST (similiarity/identity) [%/%]</th></tr>
<tr><td>lkmD</td><td>4,6DH</td><td>YP_003103995.1| dTDP-glucose 4,6-dehydratase [Actinosynnema mirum DSM 43827] (79/67)</td></tr>
<tr><td>lkmBIII</td><td>C-MT</td><td>AAD41823.1|AF147704_3 NDP-hexose 3-C-methyltransferase TylCIII [Streptomyces fradiae] (82/71)</td></tr>
<tr><td>lkmG</td><td>O-MT</td><td>ADU56358.1| putative D-glucose O-methyltransferase [Streptomyces tacrolimicus] (58/43)</td></tr>
<tr><td>lkmBII</td><td>3KR</td><td>YP_001102994.1| TDP-4-keto-6-deoxyhexose 2,3-reductase [Saccharopolyspora erythraea NRRL 2338] (84/75)</td></tr>
<tr><td>lkmI</td><td>GT</td><td>YP_001102993.1| glycosyl transferase [Saccharopolyspora erythraea NRRL 2338] (78/64)</td></tr>
<tr><td>lkmCII</td><td>3,4IM</td><td>AAG13907.1|AF263245_3 TDP-4-keto-6-deoxyhexose 3,4-isomerase [Micromonospora megalomicea subsp. nigra] (56/46)</td></tr>
<tr><td>lkmJ</td><td>AcT</td><td>P_06711012.1| acyltransferase MdmB [Streptomyces sp. e14] (59/45)</td></tr>
<tr><td>lkmCIV</td><td>3,4DH/AT</td><td>YP_001102983.1| eryCIV NDP-6-deoxyhexose 3,4-dehydratase [Saccharopolyspora erythraea NRRL 2338] (81/71)</td></tr>
<tr><td>lkmCV</td><td>ox.DA</td><td>AAB84076.1| EryCV [Saccharopolyspora erythraea NRRL 2338] (76/65)</td></tr>
<tr><td>lkmL</td><td>GT</td><td>CAA05642.1| glycosyltransferase [Streptomyces antibioticus] (68/56)</td></tr>
<tr><td>lkmBIV</td><td>4KR</td><td>ABW91155.1| NDP-hexose 4-ketoreductase [Streptomyces eurythermus] (63/54)</td></tr>
<tr><td>lkmCVI</td><td>3KR</td><td>ADI71457.1| putative sugar 3-ketoreductase [Amycolatopsis orientalis subsp. vinearia] (64/54)</td></tr>
<tr><td>lkmM</td><td>GT</td><td>YP_004818444.1| MGT family glycosyltransferase [Streptomyces violaceusniger Tu 4113] (82/73)</td></tr>
<tr><td>lkmO</td><td>O-MT</td><td>ZP_06594450.1| dTDP-6-deoxy-L-hexose 3-O-methyltransferase [Streptomyces albus J1074] (95/88)</td></tr>
<tr><td>lkmBVII</td><td>E</td><td>NP_822124.1| dTDP-4-keto-6-deoxyhexose 3,5-epimerase [Streptomyces avermitilis MA-4680] (73/65)</td></tr>
<tr><td>lkmU</td><td>NT</td><td>ADO32770.1| putative dTDP-1-glucose synthase [Streptomyces vietnamensis] (74/62)</td></tr>
<tr><td>lkmBVI</td><td>2,3DH</td><td>AAD55451.1| 2,3-dehydratase [Streptomyces antibioticus] (61/49)</td></tr>
</table></td></tr>
</table>

Table 4: MS/MS-sugar fragmentation and glycosylation gene prediction from chemotypes and genotypes of characterized glycosylated natural products (GNPs) from databases (Table 3) or self-acquired MS/MS data.

<table>
<tr><th>GNP</th><th>MS/MS glycosylation footprints of characterized GNP</th><th>Glycosylation genes in gene cluster of characterized GNP</th></tr>
<tr><td>chalcomycin</td><td> 



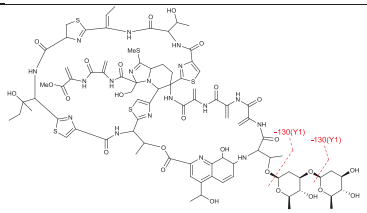
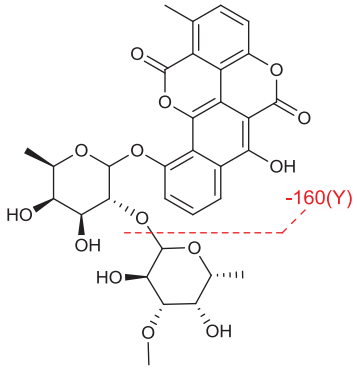
<table>
<tr><th>Gene</th><th>Predicted function</th><th>Closest functional homolog by BLAST (similiarity/identity) [%/%]</th></tr>
<tr><td>chmCIV</td><td>3,4DH/AT</td><td>ABB52533.1| 3,4-dehydratase-like protein [Streptomyces sp. KCTC 0041BP] (98/93)</td></tr>
<tr><td>chmCI</td><td>O-MT</td><td>ZP_05000456.1| sugar-O-methyltransferase [Streptomyces sp. Mg1] (93/87)</td></tr>
<tr><td>chmCV</td><td>ox.DA</td><td>ZP_05000457.1| NDP-4,6-dideoxyhexose-3,4-enoyl reductase [Streptomyces sp. Mg1] (92/89)</td></tr>
<tr><td>chmAII</td><td>4,6DH</td><td>ZP_05000462.1| dTDP-glucose-4,6-dehydratase [Streptomyces sp. Mg1] (97/97)</td></tr>
<tr><td>chmAI</td><td>NT</td><td>ABB52525.1| alpha-D-glucose-1-phosphate thymidylyltransferase [Streptomyces sp. KCTC 0041BP] (99/97)</td></tr>
<tr><td>chmJ</td><td>E</td><td>ZP_05000464.1| NDP-hexose-3-epimerase [Streptomyces sp. Mg1] (98/96)</td></tr>
<tr><td>chmMII</td><td>O-MT</td><td>ABB52523.1| 3-O-methyltransferase [Streptomyces sp. KCTC 0041BP] (98/96)</td></tr>
<tr><td>chmD</td><td>4-KR</td><td>ABB52541.1| hexose-4-ketoreductase [Streptomyces sp. KCTC 0041BP] (96/93)</td></tr>
<tr><td>chmMI</td><td>O-MT</td><td>ABB52542.1| O-methyltransferase [Streptomyces sp. KCTC 0041BP] (97/94)</td></tr>
<tr><td>chmN</td><td>GT</td><td>ZP_05000470.1| 6-deoxy-D-allosyltransferase [Streptomyces sp. Mg1] (97/95)</td></tr>
<tr><td>chmCII</td><td>3,4IM</td><td>ZP_05001883.1| NDP-hexose-3,4-isomerase [Streptomyces sp. Mg1]</td></tr>
<tr><td>chmCIII</td><td>GT</td><td>ABB52547.1| chalcosyltransferase [Streptomyces sp. KCTC 0041BP] (98/94)</td></tr>
<tr><td>chmU</td><td>3KR</td><td>ABB52548.1| 3-oxoacyl-(acyl-carrier-protein)-reductase [Streptomyces sp. KCTC 0041BP] (96/90)</td></tr>
</table></td></tr>
<tr><td>Sch40832</td><td>  </td><td>Predicted gene cluster in Micromonospora carbonacea ATCC 39149: MCAG_03933 - MCAG_03962 (based on THIOBASE (Li, J. et al., PLOS ONE (2012)).
<table>
<tr><th>Gene</th><th>Predicted function</th><th>Closest functional homolog by BLAST (similiarity/identity) [%/%]</th></tr>
<tr><td>MCAG_03952</td><td>GT</td><td>YP_003102514.1 Sterol 3-beta-glucosyltransferase [Actinosynnema mirum DSM 43827] (60/49)</td></tr>
</table></td></tr>
</table>

Table 4: MS/MS-sugar fragmentation and glycosylation gene prediction from chemotypes and genotypes of characterized glycosylated natural products (GNPs) from databases (Table 3) or self-acquired MS/MS data.

GNP	MS/MS glycosylation footprints of characterized GNP	Glycosylation genes in gene cluster of characterized GNP
chartreusin		

Gene	Predicted function	Closest functional homolog by BLAST (similiarity/identity) [%/%]
ChaGT2	GT	ZP_09566415.1| glycosyl transferase family 2 [Singulisphaera acidiphila DSM 18658] (72/54)
ChaM	MT	YP_002518358.1| methyltransferase [Caulobacter crescentus NA1000] (57/41)
ChaGT1	GT	CAE17548.1| glycosyltransferase [Streptomyces griseus subsp. griseus] (58/44)
ChaS1	NT	ADO32770.1| putative dTDP-1-glucose synthase [Streptomyces vietnamensis] (72/54)
ChaS3	4KR	YP_003301686.1| GtrA family protein [Thermomonospora curvata DSM 43183] (68/52)
ChaS2	4,6DH	YP_005464806.1| putative dTDP-glucose-4,6-dehydratase [Actinoplanes missouriensis 431] (82/69)
ChaS4	E/KR	ZP_04607399.1| nucleoside-diphosphate-sugar epimerase [Micromonospora sp. ATCC 39149] (56/44)

Table 5: ^{1}H and ^{13}C NMR Data for cinerubin B (1-hydroxyaclacinomycin A) in MeOD-d4.

Site		δC [ppm]	δH [ppm]	Signal	*J*(H-H) [Hz]	Site		δC [ppm]	δH [ppm]	Signal	*J*(H-H) [Hz]
1		158.9				**rhodosamine**					
2		130.3	7.39	dd	9.5, 13.4	1'		101.3	5.59	d	4
3			7.39	dd	9.5, 13.4	2'	(e)	28.6	2.04	dt	4, 12.8
4		158.9					(f)		2.22	cm	4
4-OH			N/A			3'		63.8	12.8	d	12.8
4a		113.9				4'		74.5		bs	
5		192.2				5'		68.4	4.24	q	6.6
5a						6'	(Me)	17.8	1.38	d	6.6
6		163.1				3'-NMe_2	(Me)	2.89	43	s	
6-OH			N/A					3.05	40.7	s	
6a		143.9				**2'-deoxyfucose**					
7		72.2	5.15	d	4.4	1''		101.3	5.38	d	4.4
8	(a)	35.6	2.5	cm		2''	(g)	26.5	2.65	dt	4.4, 18.9
	(b)		2.23	cm			(h)		1.95	dd	4.8, 12.3
9		71.7				3''		67.5	4.29	dt	4.0, 12.3
9-OH			N/A			4''		67.5	4.09	bs	
10		58.1	4.11	s		5''		67.5	4.13	q	6.6
10a		132.7				6''	(Me)	16.4	1.26	d	6.6
11		120.5	7.72	s		3''-OH			N/A		
11a		116.3				**cinerulose B**					
12		187				1'''		92.4	5.2	d	3
12a		113.9				2'''		64.2	4.45	dd	3
13	(c)	33.3	1.77	dt	21.6, 7.0	3'''	(i)	53	2.48	cm	
	(d)		1.55	dt	21.8, 7.0		(j)		2.77	dd	2.6, 14.5
15						4'''		210			
21	(Me)	6.6	1.09	t	7.3	5'''		79.2	4.74	q	6.6
22	(MeO)	51.7	3.69	s		6'''	(Me)	16.4	1.31	d	6.6

Abbreviations: Me – methyl, MeO – methoxy, N/A – not annotated, the hydroxyl-protons were solvent exchanged and, thus, not observed. s – singlet, d – duplet, m – multiplet, q – quartet, t – triplet, c – complex.

Chapter 4, in full, is currently being prepared for submission for publication of the material. Kersten, R.D., Ziemert, N., Crüsemann, M., Duggan, B.M., Jensen, P.R., Dorrestein, P.C., Moore, B.S. The dissertation author was the primary investigator and author of this material.

R.D.K. designed and carried out experiments, analyzed data and wrote the paper. N.Z. carried out bioinformatic experiments and analyzed data. M.C. carried out MS experiments and analyzed data B.M.D. carried out NMR experiments. P.R.J. provided genome information and analyzed data. B.S.M. and P.C.D. designed experiments, analyzed data and wrote the paper.

Chapter 5 - Bioactivity-guided genome mining reveals the lomaiviticin biosynthetic gene cluster in *Salinispora tropica*

DOI: 10.1002/cbic.2001300147

Bioactivity-guided genome mining reveals the lomaiviticin biosynthetic gene cluster in *Salinispora tropica*

Roland D. Kersten[a], Amy L. Lane[b], Markus Nett[c], Taylor K.S. Richter[a], Brendan M. Duggan[d], Pieter C. Dorrestein[a],[d],[e], Bradley S. Moore*[a],[d]

The use of genome sequences has become routine in guiding the discovery and identification of microbial natural products and their biosynthetic pathways. *In silico* prediction of molecular features, such as metabolic building blocks, physico-chemical properties or biological functions, from orphan gene clusters has opened up the characterization of many new chemo- and genotypes in genome mining approaches. Here, we guided our genome mining of two predicted enediyne pathways in *Salinispora tropica* CNB-440 by a DNA interference bioassay to isolate DNA-targeting enediyne polyketides. An organic extract of *S. tropica* showed DNA-interference activity that surprisingly was not abolished in genetic mutants of the targeted enediyne pathways, *ST_pks1* and *spo*. Instead we showed that the product of the orphan type II polyketide synthase pathway, *ST_pks2*, is solely responsible for the DNA-interfering activity of the parent strain. Subsequent comparative metabolic profiling revealed the lomaiviticins, glycosylated diazofluorene polyketides, as the *ST_pks2* products. This study marks the first report of the 59 open reading frame lomaiviticin gene cluster (*lom*) and supports the biochemical logic of their dimeric construction via a pathway related to the kinamycin monomer.

Introduction

Microbial genome sequencing has increased dramatically over the past decade, providing access to a tremendous amount of information about the genetic capability of microorganisms to produce natural products.[1] Bioinformatics analyses have revealed that the number of putative natural product biosynthetic gene clusters often greatly exceeds the number of reported metabolites,[2] suggesting that a large portion of nature's chemical potential remains unexplored. Realization of the chemistry encoded by orphan gene clusters may afford access to the next generations of drugs as well as insights into evolutionary benefits of the plethora of biosynthetic genes maintained within microbial genomes.

Researchers have developed a variety of approaches to discover natural products associated with orphan gene clusters.[3,4] Such genome mining strategies include inactivation of selected gene clusters followed by comparative metabolic profiling of mutant and wild-type strains,[5] prediction of physico-chemical properties from biosynthetic gene sequences,[6] manipulation of genes regulating biosynthetic pathway expression,[7] isotope labeling of predicted biosynthetic precursors in combination with isotope-guided fractionation,[8] heterologous expression of orphan biosynthetic genes,[9] and mass spectrometry-guided genome mining.[10]

The discovery of novel natural products that impede DNA replication or function is of interest since DNA-interfering molecules have potential cancer chemotherapeutic properties,[11,12] especially as drug conjugates linked to monoclonal antibodies.[13] Hallmark examples of DNA-targeting natural products include the enediyne antibiotics.[14] Their biosynthesis involves an iterative acting type I polyketide synthase, which assembles the distinctive core structure common to all members of this natural product family.[15] Enediynes feature two acetylenic groups that are linked via a conjugated double bond. They are differentiated into two subclasses possessing either 9-membered rings, as exemplified by neocarzinostatin, or 10-membered rings, such as calicheamicin and dynemicin (Figure 1A).[15] These cytotoxic agents intercalate into chromosomal DNA in a sequence-specific manner and cause strand scission via a radical mechanism.[16]

[a] *Roland D. Kersten, Taylor K.S. Richter, Prof. Dr. Pieter C. Dorrestein, Prof. Dr. Bradley S. Moore*
Center for Marine Biotechnology and Biomedicine, Scripps Institution of Oceanography, University of California at San Diego, La Jolla, CA 92093 (USA)
E-mail: bsmoore@ucsd.edu
[b] *Prof. Dr. Amy Lane*
University of North Florida, Chemistry Department, 1 UNF Dr., Jacksonville, FL 32224 (USA)
[c] *Dr. Markus Nett*
Leibniz Institute for Natural Product Research and Infection Biology, Hans-Knöll-Institute, Jena (Germany)
[d] *Dr. Brendan M. Duggan, Prof. Dr. Pieter C. Dorrestein, Prof. Dr. Bradley S. Moore*
Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California at San Diego, La Jolla, CA 92093 (USA)
[e] *Prof. Dr. Pieter C. Dorrestein*
Department of Pharmacology and Department of Chemistry and Biochemistry, University of California at San Diego, La Jolla, California (USA)

Actinomycetes are prevalent sources of DNA-interfering natural products, including enediynes.[12,14] Genome sequenced *Salinispora* species devote approximately 10% of their genetic material to natural product biosynthesis, which includes several orphan enediyne polyketide synthase (ePKS) gene clusters.[17,18] Although the 5.2-Mbp genome of the *Salinispora tropica* strain CNB-440 harbors at least 19 putative biosynthetic gene clusters,[1] only five of these loci are linked with characterized natural products.[18-22] Its untapped metabolic potential thus makes *S. tropica* a promising candidate for genomics-guided drug discovery. The *S. tropica* CNB-440 genome contains two ePKS gene clusters, the *ST_pks1* and *spo* loci. While the product of the former is not known, the *spo* gene set is associated with the biosynthesis of two chlorinated cyclopenta[*a*]indenes. Sporolides A and B were proposed to be degradation products of a reactive 9-membered enediyne precursor molecule (Figure 1A, Supplementary Figure 1A).[22-24] While the potent proteasome inhibitor salinosporamide A was isolated from this strain and is currently in clinical trials for treatment of multiple myeloma,[25] no DNA-interfering anticancer natural products have been reported from *S. tropica*.

Herein, we report the identification of DNA-targeting natural products from *S. tropica* CNB-440 via bioactivity-guided genome mining. Allelic-exchange mutagenesis of select biosynthesis genes, DNA interference assays and comparative metabolite profiling tracked the observed biological activity not to one of the anticipated enediyne pathways but exclusively to *ST_pks2*, an orphan biosynthetic gene cluster predicted to yield an anthracycline-type antibiotic. The metabolites associated with the *ST_pks2* locus were subsequently identified as the lomaiviticins, glycosylated diazofluorene polyketides with potent anticancer activity originally discovered from *Salinispora pacifica* strain DPJ-0019 (formerly *Micromonospora lomaivitiensis*).[26-29] The discovery of the lomaiviticin pathway provides the molecular logic for the construction of these complex organic molecules and, in the long term, may help generate new bioactive derivatives by genetic engineering. Furthermore, the described genome mining approach may offer an effective strategy to mine for new DNA-interference natural products and their pathways with potential applicability as antitumor chemotherapies.

Results and Discussion

Bioinformatic prediction and biochemical detection of DNA-damaging chemotypes in *Salinispora tropica* CNB-440

The first step in the search for DNA-interfering chemotypes from the *S. tropica* CNB-440 genome was the bioinformatic analysis of candidate gene clusters. Initially, the two ePKS pathways in the *S. tropica* CNB-440 genome, *spo* and *ST_pks1*, were targeted. The *ST_pks1* locus was uncharacterized and not related to known enediyne molecules from *S. tropica*.

Proteins encoded by the *ST_pks1* genes exhibit strongest homology to enzymes required for the biosynthesis of dynemicin, a 10-membered enediyne.[29] Interestingly, *ST_pks1* (Figure 1B, Supplementary Table 1) lacks ORFs for the synthesis of anthraquinone- or sugar-based groups that are typically clustered with *epks* genes and account for the extensive functionalization of the enediyne core. In dynemicin biosynthesis, genes encoding tailoring enzymes are chromosomally distinct from enediyne-associated genes,[15] and an analogous scenario is plausible for the proposed *ST_pks1*-encoded enediyne as well.

The *spo* cluster (Figure 1C) includes all genes required for enediyne biosynthesis and is expected to yield a 9-membered enediyne product (Figure 1A). Most 9-membered enediynes are highly unstable relative to their 10-membered counterparts,[23] and the *spo* enediyne product was proposed to readily undergo nonenzymatic Bergman-type cyclization in the presence of chloride anions, yielding sporolides A and B (Supplementary Figure 1A).[22,23,30] Even though the enediyne origin of the sporolides has not been experimentally confirmed, previous *in vitro* studies supported an involvement of *spo* genes in the biosynthesis of their cyclohexenone residue.[22]

To determine whether *S. tropica* produces DNA-damaging antibiotics, as predicted by bioinformatic analysis, the strain was fermented in a commonly employed seawater-based medium, and chemical extracts from multiple time points were evaluated for DNA-interfering natural products using the well-established biochemical prophage induction assay (BIA).[31] DNA-damaging activity was most pronounced for methanol/ethyl acetate extracts of *S. tropica* cultures harvested 6-10 days after inoculation (Supplementary Figure 2A), suggesting that at least one of the biosynthetic clusters predicted to yield DNA-interfering enediyne compounds was expressed.

***S. tropica* enediyne genotypes are not responsible for biosynthesis of DNA-interfering chemotypes**

To explore whether the candidate gene clusters *ST_pks1* and *spo* are associated with the observed DNA-damaging activity, a key biosynthetic gene from each cluster was inactivated and replaced with a gene for apramycin resistance (*aprR*) via PCR-targeted gene replacement.[32] BIA activities were then compared between chemical extracts from individual mutants and the wild type strain (Supplementary Figure 2).

Elimination of the ePKS gene *spoE* from the *spo* cluster yielded a *spoE::apr*^{R} mutant incapable of sporolide production (Supplementary Figure 1B), confirming that the sporolides are indeed ePKS-derived products as originally suspected. However, chemical extracts from the *spoE::apr*^{R} mutant exhibited the same activity profile as the wild type strain in the BIA (Supplementary Figure 2C), indicating neither sporolides A and B nor their proposed 9-membered enediyne precursor (Supplementary Figure 1A) were responsible for the observed DNA interference in the BIA. We suspect that this enediyne may decompose too rapidly for BIA detection due to the lack of a protective carrier. Most characterized gene clusters for 9-membered enediynes encode an apoprotein that sequesters the enediyne to enhance its stability and thus facilitate biological activity.[15,23] However, no proteins with homology to characterized enediyne-stabilizing proteins were found within the *S. tropica* genome, suggesting that this organism lacks mechanisms for 9-membered enediyne stabilization and offering an explanation for the apparent instability of the proposed sporolide precursor. A similar transformation was hypothesized for the cyanosporasides from *Salinispora pacifica* CNS-143, which also have a cyclopenta[*a*]indene core structure derived from a putative 9-membered enediyene precursor.[33]

From the *ST_pks1* gene cluster, PCR-targeted gene elimination of the putative 10-membered ePKS encoded by *strop0598* yielded a mutant with BIA activity equivalent to that of the wild type (Supplementary Figure 2B), thus eliminating *ST_pks1* as a candidate for biosynthesis of the observed DNA-interfering active metabolite. Further, LC-MS profiling of the mutant and wild type extracts revealed no evidence for enediyne natural products (data not shown), suggesting that the orphan *ST_pks1* gene cluster is either inactive under the selected fermentation conditions or expressed at low levels.

Type II PKS pathway *ST_pks2* produces DNA-interfering compounds in S. tropica CNB-440

As no enediyne was responsible for causing DNA-interference in *S. tropica* extracts as originally suspected, we evaluated the other orphan biosynthetic pathways for the responsible agent. Another prominent class of DNA-targeting natural products from actinomycetes are aromatic polyketides such as the anthracycline anticancer agents.[34] Aromatic polyketides are commonly biosynthesized by type II PKSs and are often glycosylated.[35] Glycosylated anthracycline anticancer agents, such as daunomycin, can intercalate into chromosomal DNA with their polyketide moiety, while the glycosyl moiety binds the phosphoribose DNA backbone.[34] The *S. tropica* CNB-440 genome contains two orphan type II PKS pathways, *ST_pks2* and *ST_pks3*. The *ST_pks3* locus (Supplementary Figure 3A, Supplementary Table 2) encodes the synthesis of a putative spore pigment polyketide related to the *whiE* dodecaketide product of unknown structure from *Streptomyces coelicolor* A3(2).[36] We explored its function by gene elimination of the predicted *ST_pks3* β-ketosynthase (*strop2500*) and sporulation phenotyping on A1 agar. The *ST_pks3* mutant showed loss of spore pigmentation compared to wild type *S. tropica* confirming a predicted spore pigment product (Supplementary Figure 3B). The *ST_pks2* gene cluster, on the other hand, is predicted to encode production of a glycosylated aromatic polyketide, thereby making it a more promising alternative target in the search for a DNA-interfering molecule than the *ST_pks3* spore pigment.

To examine the function of the *ST_pks2* gene cluster (Figure 1D), we eliminated the putative β-ketosynthase gene *strop2223* to yield a *strop2223::apr*^{R} mutant. The organic extract of this mutant was inactive in the BIA (Supplementary Figure 2D), thereby correlating the DNA-interference activity observed from wild type *S. tropica* exclusively to the *ST_pks2* gene cluster and not to either of the enediyne-encoding gene clusters (*ST_pks1* and *spo*). In light of prior research correlating the observation of ePKS-encoding genes with BIA activity,[37] we found this result surprising. This earlier work, however, did not confirm their observations with gene deletions or compound isolation, thus leaving open the possibility that non-enediyne polyketides were responsible for the observed biological activity.

Characterization of *ST_pks2* as the lomaiviticin biosynthetic gene cluster (lom)

To identify the DNA-interfering glycosylated polyketide predicted by bioinformatics analysis of the orphan *ST_pks2* gene cluster, we employed an LCMS^n-based comparative metabolite profiling strategy. In LCMS^n traces of the wild type extract, several masses were detected that were absent in the corresponding mutant sample (Figure 2A/B, Supplementary Figure 4). One compound with an exact mass of 1338.5319 Da showed candidate deoxysugar fragments (158.12 m/z – putative *N,N*-dimethyldideoxysugar; 145.09 m/z – putative *O*-methyldideoxysugar) in the MS/MS spectrum (Figure 2C), indicating a putative glycosylated natural product. The compound was isolated as a burgundy-red solid by MS-guided fractionation via reversed phase-flash column chromatography from a methanol/ethyl acetate extract of a culture of *S. tropica* CNB-440. NMR analysis enabled the characterization of the molecule as lomaiviticin C (Figure 2D, Supplementary Figure 5, Supplementary Table 3), a glycosylated aromatic polyketide recently described from *S. pacifica* strain DPJ-0019.[26] In addition, other known derivatives, lomaiviticins A, D and E, were identified by LCMS^n (Supplementary Figure 6).[27,28] In the β-ketosynthase mutant, no NDP-deoxysugar accumulation was detected (data not shown) as these NDP-species should have a generally short lifetime in cell metabolism due to their high energy content.

Lomaiviticin A was first isolated by Carter[27] and is the derivative with most potent anticancer activity.[26] Its C–C dimeric diazofluorene core is also found in its monomeric form in kinamycin natural products from *Streptomyces murayamaensis*.[28,38] While the kinamycin gene cluster (*kin*) was previously reported by Gould in 1998,[39] the lomaiviticin biosynthetic gene cluster (*lom*) remained unknown until now. The *lom* locus is predicted to comprise 59 ORFs, which includes all biosynthetic genes putatively involved in the construction of the diazofluorene core in kinamycin (Table 1, Supplementary Figure 7).[28,39] The *lom* cluster contains three additional PKS genes adjacent to the minimal type II PKS genes for the incorporation of a propionate starter unit in contrast to a starter acetate unit in kinamycin (Supplementary Figure 7). Putative biosynthetic gene products include two glycosyltransferases and several deoxysugar biosynthetic enzymes that nicely correlate with the glycosylation pattern of the lomaiviticins (Supplementary Figure 8). In addition, the gene cluster contains a FAD-dependent monooxygenase gene, *strop2191*, that is homologous to ActVA-Orf4 from the actinorhodin biosynthetic gene cluster in *S. coelicolor* A3(2). ActVA-Orf4 catalyzes the C–C dimerization reaction of two benzoisochromanequinone monomers in actinorhodin biosynthesis.[40] We thus suspect that its homolog, Strop2191, catalyzes the dimerization of kinamycin-like diazofluorene monomers in lomaiviticin biosynthesis. The dimerization could occur from C2 after a dihydroxylation at C3 and C4 and a hydroxylation at C10 as enzymatic hydroxylations might be sterically hindered in the dimer (Supplementary Figure 7). A homolog of Strop2191 is not present in the homologous *kin* cluster. These diazo-forming and dimerizing enzymes await characterization.

Conclusion

In this study, we identified the lomaiviticin biosynthetic gene cluster (*ST_pks2* = *lom*) in *Salinispora tropica* CNB-440 via bioactivity-guided genome mining. Motivated by *ST_pks1* and *spo*, two *S. tropica* enediyne gene clusters without characterized enediyne products, we analyzed organic extracts with the BIA, an assay for rapid screening of DNA-interfering chemotypes, to detect a *S. tropica* DNA-targeting chemotype. Although we did indeed measure a BIA-based activity, gene elimination experiments showed that this observed bioactivity was not caused by an enediyne pathway product. Instead we correlated the BIA bioactivity to the products of the type II PKS pathway, *ST_pks2*. Comparative metabolic profiling yielded the lomaiviticins as *ST_pks2* products and, thus, as the observed BIA-positive compounds. Our approach of using the BIA in combination with genome mining and gene elimination may prove an effective strategy to identify new DNA-targeting chemo- and genotypes beyond enediyne scaffolds. The identified biosynthetic gene cluster for a highly complex and stereospecific natural product such as lomaiviticin A can open up semisynthetic and heterologous expression routes as alternatives for their challenging production via total synthesis.

Experimental Section

General

All chemicals were acquired from Fisher Scientific, Sigma Aldrich and Honeywell Burdick & Jackson, and solvents were of HPLC grade or higher. HPLC analyses were conducted with an Agilent 1200 HPLC system with diode array detection. LC-MS analyses were conducted with an Agilent 6530 Accurate-Mass Q-TOF MS (MassHunter software, Agilent) equipped with a Dual electrospray ionization source and with an Agilent 1260 LC system (ChemStation software, Agilent) with diode array detector. NMR data were acquired at the UCSD Skaggs School of Pharmacy and Pharmaceutical Sciences NMR Facility on a 600MHz Varian NMR spectrometer (Topspin 2.1.6 software, Bruker) with a 1.7 mm cryoprobe.

Gene inactivation

Targeted biosynthetic genes within *S. tropica* gene clusters (*ST_pks1, ST_pks3, lom*, and *spo*) predicted to yield DNA-interfering natural products were inactivated by PCR-based mutagenesis following previously established methods.[32] Briefly, the apramycin resistance (*aac(3)IV*) cassette from pIJ773 was PCR amplified and extended using primer sequences (Supplementary Table 4). Extended antibiotic resistance cassettes were introduced by electroporation into *Escherichia coli* BW25113/pKD20 carrying appropriate pCCFOS-based fosmids BHXS2039 (for *ST_pks1* genes), BHXS0939 (for *ST_pks3* genes), BHXS5407 (for *lom* genes), and BHXS4676 (for *spo* genes). The mutated fosmid was then transferred into *S. tropica* CNB-440 by conjugation from *E. coli* S17-1, and gene replacement confirmed by colony PCR and sequencing of PCR products. Gene replacement experiments were minimally carried out in duplicate.

***S. tropica* fermentation and biochemical prophage induction assay (BIA)**

Wild type *S. tropica* CNB-440 as well as *spoE::apr*^{R}, *strop2223::apr*^{R}, *strop2500::apr*^{R}, and *strop0598::apr*^{R} mutants were cultured at 28 °C in Fernbach flasks containing A1 medium (1 L, 1% starch, 0.4% yeast extract, 0.2% peptone, 0.1% calcium carbonate, 3% InstantOcean® sea salt). Inoculation occurred with A1 medium starter cultures (10 mL) in falcon tubes (50 mL). Flasks were shaken at 225 rpm, and time point samples (50 mL) were drawn every two days starting on day 4. Each sample (50 mL) was evenly divided into two portions. The first portion was clarified by centrifugation at 5000 rpm for 10 min to provide a supernatant sample. The remaining mycelial pellet from this portion was suspended in methanol (MeOH, 25 mL), stirred for 30 min at room temperature, and cell debris subsequently removed by centrifugation. The second portion of *S. tropica* fermentation (25 mL) was extracted three times with an equal volume of ethyl acetate. Supernatant samples, as well as MeOH and EtOAc extracts, were concentrated to dryness *in vacuo*. Samples were re-suspended at 1 mg/mL in their respective extraction solvents. The biochemical prophage induction assay (BIA) was performed as previously described.[31] Briefly, extract (10 µL) was applied to agar plates seeded with *E. coli* ATCC 33312 and incubated for 5 h at 37 °C. Soft agar containing Fast Blue RR salt (1.2 mg/mL) and 6-bromo-2-naphthyl-β-D-galactopyranoside (0.4 mg/mL) was added onto the plate and color development was observed within 15 min. Cisplatin was included as a positive reference (test concentration: 5 mg/mL and 2-fold serial dilutions thereof). The fermentation medium and solvents were also assayed alone and negative for BIA activity. Every assay was run at least in duplicate. For sporulation phenotyping, wild type *S. tropica* CNB-440 and *strop2500::apr*^{R} mutant were inoculated on an A1 agar plate (18 g agar/L A1 medium) and incubated for 14 days at 28°C.

Comparative metabolite profiling of wild type and mutant *S. tropica* chemical extracts by LC-MS

Crude MeOH/EtOAc extracts of wild type and mutant *S. tropica* strains were filtered through Acrodisc MS PTFE Syringe filters (Pall Inc.) and adjusted in concentration (0.2 mg/mL). For HPLC analysis, 25 µl was injected on a Luna reversed phase C18 (5 µm, 4.6 x 100 mm) HPLC column and subjected to a 10-100% MeCN (0.1% TFA) / 0.1% TFA gradient (15 min, 0.7 mL/min). For LCMS^n analysis, crude extract (2 µg) was injected onto a Phenomenex Luna C18 reverse phase HPLC column (5µm, 150x4.6mm) and was analyzed with an Agilent 6530 Accurate-Mass LC/MS with an Agilent 1260 LC system under the following LC conditions (1-3 min – 10% MeCN (0.1% TFA), 3-23 min - 10-100% MeCN (0.1% TFA), 23-25 min – 100% MeCN (0.1% TFA), 0.7 mL/min). Q-TOF MS settings during the LC gradient were as follows: Acquisition -Mass range 300-1500 m/z, MS scan rate 1/sec, MS/MS scan rate 2/sec, fixed collision energy 20eV, Source – Gas Temperature 300°C, gas flow 11 L/min, Nebulizer 45 psig, Ion polarity positive, Scan source parameters – VCap 3000, Fragmentor 100, Skimmer1 65, OctopoleRFPeak 750. The MS was autotuned with Agilent tuning solution in positive mode before each measurement. LC(DAD) data were analyzed with ChemStation software (Agilent), and MS data were analyzed with MassHunter software (Agilent).

Isolation and characterization of *ST_pks2* product, lomaiviticin C

Lomaiviticin C was isolated by MS-guided fractionation from an A1 medium culture (1L) of wild type *S. tropica* CNB-440 after 10 d incubation. The cells of 500 ml of the culture were extracted with MeOH for 30 min under stirring, and the remaining supernatant and culture were extracted twice with EtOAc. The crude organic extract was concentrated *in vacuo,* resuspended in MeOH (2 mL), and loaded on a reversed phase C18 silica gel column for flash-column chromatography. Lomaiviticin C was further purified as previously described[26] using LC-MS for identification of lomaiviticin C-containing fractions. Purified lomaiviticin C was dissolved in MeOD-d4 and subjected to NMR structure elucidation (^{1}H, DQF-COSY, ^{1}H-^{13}C HMBC, NOESY). NMR data were analyzed with Topspin 2.1.6 software (Bruker).

Acknowledgements

We thank Paul R. Jensen and William Fenical for providing S. tropica CNB-440. This work was supported by a postdoctoral fellowship from the German Academic Exchange Service (DAAD) to M.N. and an NIH IRACDA postdoctoral fellowship (GM068524) to A.L.L. This work was supported by grants from the NIH (GM085770 to B.S.M., GM097509 to P.C.D. and instrument grant 1-S10-RR031562-01A1).

Keywords: polyketides · biosynthesis · genome mining · lomaiviticin · *Salinispora tropica*

[1] M. Nett, H. Ikeda, B. S. Moore, *Nat. Prod. Rep.* **2009**, *26*, 1362-1384.
[2] G. L. Challis, *J. Med. Chem.* **2008**, *51*, 2618-2628.
[3] C. Corre, G. L. Challis, *Nat. Prod. Rep.* **2009**, *26*, 977-986.
[4] J. M. Winter, S. Behnken, C. Hertweck, *Curr. Opin. Chem. Biol.* **2011**, *15*, 22–31.
[5] L. J. Song, F. Barona-Gomez, C. Corre, L. K. Xiang, D. W. Udwary, M. B. Austin, J. P. Noel, B. S. Moore, G. L. Challis, *J. Am. Chem. Soc.* **2006**, *128*, 14754-14755.
[6] A. S. Eustaquio, S.-J. Nam, K. Penn, A. Lechner, M. C. Wilson, W. Fenical, P. R. Jensen, B. S. Moore, *Chembiochem* **2011**, *12*, 61-64.
[7] a) S. Lautru, R. J. Deeth, L. M. Bailey, G. L. Challis, *Nat. Chem. Biol.* **2005**, *1*, 265-269; b) K. Ishida, T. Lincke, S. Behnken, C. Hertweck, *J. Am. Chem. Soc.* **2010**, *132*, 13966-13968; c) L. Laureti, L. Song, S. Huang, C. Corre, P. Leblond, G. L. Challis, B. Aigle, *Proc. Natl. Acad. Sci. USA*, **2011**, *108*, 6258-6263.
[8] H. Gross, V. O. Stockwell, M. D. Henkels, B. Nowak-Thompson, J. E. Loper, W. H. Gerwick, *Chem. Biol.* **2007**, *14*, 53-63.
[9] L. C. Blasiak, J. Clardy, *J. Am. Chem. Soc.* **2010**, *132*, 926-927.

[10] R. D. Kersten, Y. L. Yang, Y. Xu, P. Cimermancic, S. J. Nam, W. Fenical, M. A. Fischbach, B. S. Moore, P. C. Dorrestein, *Nat. Chem. Biol.* **2011**, *7*, 794-802.

[11] B. La Ferla, C. Airoldi, C. Zona, A. Orsato, F. Cardona, S. Merlo, E. Sironi, G. D'Orazio, F. Nicotra, *Nat. Prod. Rep.* **2011**, *28*, 630-648.

[12] a) D. J. Newman, G. M. Cragg, 2009. *Curr. Opin. Invest. Dr.*, **2009**, *10*, 1280-1296; b) C. Olano, C. Mendez, J. A. Salas, *Nat. Prod. Rep.* **2009**, *26*, 628-660.

[13] a) P. R. Hamann, L. M. Hinman, I. Hollander, C. F. Beyer, D. Lindh, R. Holcomb, W. Hallett, H. R. Tsou, J. Upeslacis, D. Shochat, A. Mountain, D. A. Flowers, I. Bernstein, *Bioconjug. Chem.* **2002**, *13*, 47-58; b) L. Ducry, B. Stump, *Bioconjug. Chem*., *21*, 5–13 (2010).

[14] U. Galm, M. H. Hager, S. G. Van Lanen, J. H. Ju, J. S. Thorson, B. Shen, *Chem. Rev*., **2005**, *105*, 739-758.

[15] Z. X. Liang, *Nat. Prod. Rep*. **2010**, *27*, 499-528.

[16] A. L. Smith, K. C. Nicolaou, *J. Med. Chem.*, **1996**, *39*, 2103–2117.

[17] K. Penn, C. Jenkins, M. Nett, D. W. Udwary, E. A. Gontang, R. P. McGlinchey, B. Foster, A. Lapidus, S. Podell, E. E. Allen, B. S. Moore, P. R. Jensen, *ISME J.*, **2009**, *3*, 1193-1203.

[18] D. W. Udwary, L. Zeigler, R. N. Asolkar, V. Singan, A. Lapidus, W. Fenical, P. R. Jensen, B. S. Moore, *Proc. Natl. Acad. Sci. USA* **2007**, *104*, 10376-10381.

[19] A. A. Roberts, A. W. Schultz, R. D. Kersten, P. C. Dorrestein, B. S. Moore, *FEMS Microbiol. Lett.* **2012**, *335*, 95-103.

[20] A. S. Eustáquio, R. P. McGlinchey, Y. Liu, C. Hazzard, L. L. Beer, G. Florova, M. M. Alhamadsheh, A. Lechner, A. J. Kale, Y. Kobayashi, K. A. Reynolds, B. S. Moore, *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106*, 12295-12300 (2009).

[21] A. Miyanaga, J. E. Janso, L. McDonald, M. He, H. Liu, L. Barbieri, A. S. Eustáquio, E. N. Fielding, G. T. Carter, P. R. Jensen, X. Feng, M. Leighton, F. E. Koehn, B. S. Moore, *J. Am. Chem. Soc.* **2011**, *133*, 13311-13313.

[22] R. P. McGlinchey, M. Nett, B. S. Moore, *J. Am. Chem. Soc.* **2008**, *130*, 2406-2407.

[23] M. Jean, S. Tomasi, P. Van de Weghe, *Org. Biomol. Chem.* **2012**, *10*, 7453-7456.

[24] G. O. Buchanan, P. G. Williams, R. H. Feling, C. A. Kauffman, P. R. Jensen, W. Fenical, *Org. Lett*. **2005**, *7*, 2731-2734.

[25] a) T. A. Gulder, B. S. Moore, *Angew. Chem. Int. Ed.* **2010**, *49*, 9346-9367; b) R. H. Feling, G. O. Buchanan, T. J. Mincer, C. A. Kauffman, P. R. Jensen, W. Fenical, *Angew. Chem. Int. Ed.* **2003**, *42*, 355-357.

[26] C. M. Woo, N. E. Beizer, J. E. Janso, S. B. Herzon, *J. Am. Chem. Soc*. **2012**, *134*, 15285-15288.

[27] H. He, W. D. Ding, V. S. Bernan, A. D. Richardson, C. M. Ireland, M. Greenstein, G. A. Ellestad, G. T. Carter, *J. Am. Chem. Soc.* **2001**, *123*, 5362-5363.

[28] S. B. Herzon, C. M. Woo, *Nat. Prod. Rep.* **2012**, *29*, 87-118.

[29] a) M. Konishi, H. Ohkuma, T. Tsuno, T. Oki, G. D. VanDuyne, J. Clardy, *J. Am. Chem. Soc.* **1990**, *112*, 3715-3716; b) Q. Gao, J. S. Thorson, *FEMS Microbiol. Lett.* **2008**, *282*, 105-114.

[30] C. L. Perrin, B. L. Rodgers, J. M. O'Connor, *J. Am. Chem. Soc.* **2007**, *129*, 4795-4799.

[31] R. K. Elespuru, R. J. White, *Cancer Res*., **1983**, *43*, 2819-2830.

[32] a) B. Gust, G. L. Challis, K. Fowler, T. Kieser, K. F. Chater, *Proc. Natl. Acad. Sci. USA* **2003**, *100*, 1541-1546. b) A. S. Eustáquio, F. Pojer, J. P. Noel, B. S. Moore. *Nat. Chem. Biol*. **2008**, *4*, 69-74.

[33] A. L. Lane, S. J. Nam, T. Fukuda, K. Yamanaka, C. A. Kauffman, P. R. Jensen, W. Fenical, B. S. Moore, *J. Am. Chem. Soc.* **2013**, *135*, 4171-4174.

[34] A. Rabbani, R. M. Finn, J. Ausio, *Bioessays* **2005**, *27*, 50–56.

[35] I. Fujii, Y. Ebizuka, *Chem. Rev.* **1997**, *97*, 2511-2523.

[36] T. W. Yu, Y. Shen, R. McDaniel, H. G. Floss, C. Khosla, D. A. Hopwood and B. S. Moore, *J. Am. Chem. Soc.* **1998**, *120*, 7749-7759.

[37] E. Zazopoulos, K. X. Huang, A. Staffa, W. Liu, B. O. Bachmann, K. Nonaka, J. Ahlert, J. S. Thorson, B. Shen, C. M. Farnet, *Nat. Biotechnol*. **2003**, *21*, 187-190.

[38] a) S. Ito, T. Matsuya, S. Omura, M. Otani, A. Nakagawa, *J. Antibiot*. **1970**, *23*, 315–317. b) S. J. Gould, N. Tamayo, C. R. Melville, M. C. Cone, *J. Am. Chem. Soc.* **1994**, *116*, 2207–2208.

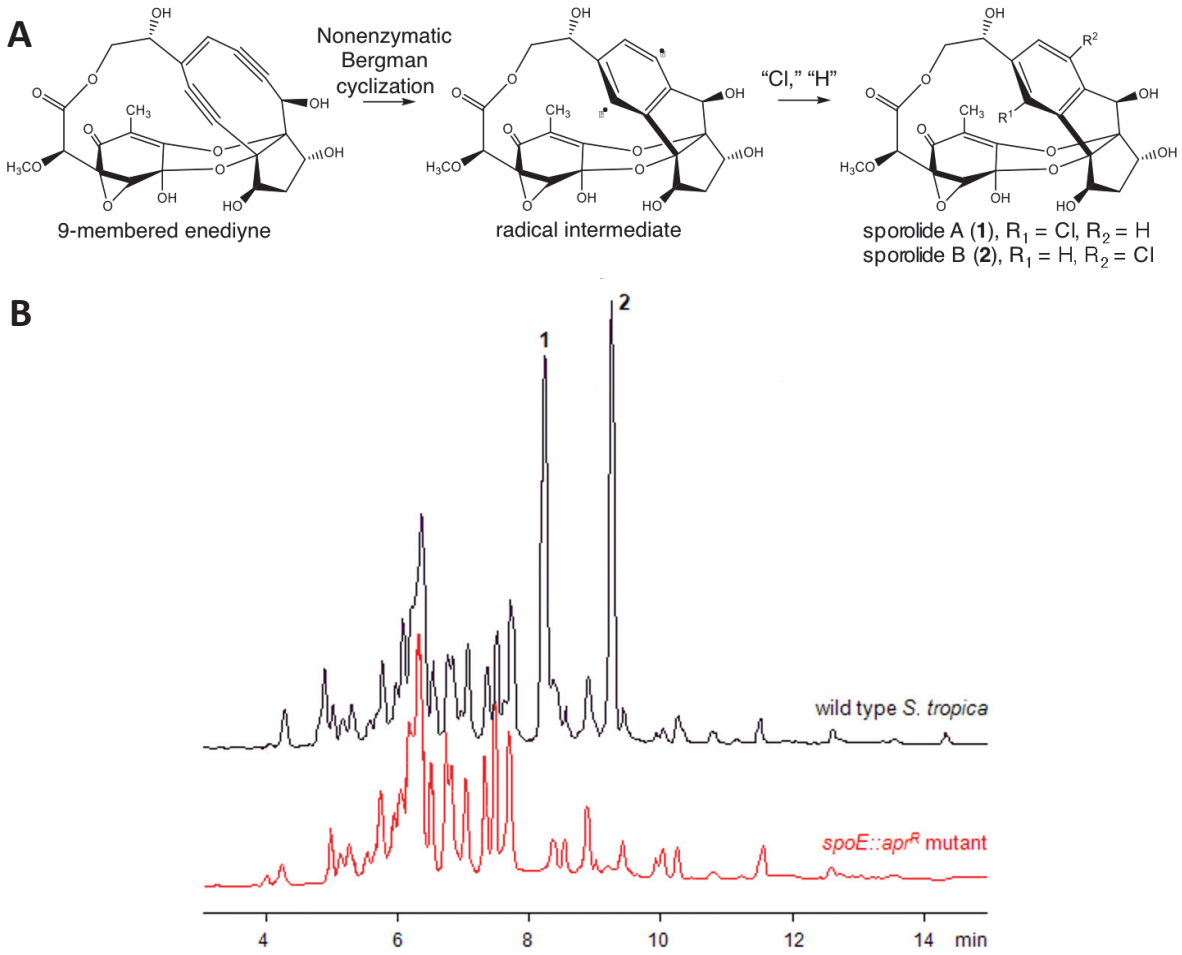
[39] S. J. Gould, S. T. Hong, J. R. Carney, *J. Antibiot.* **1998**, *51*, 50-57.

[40] T. Taguchi, T. Ebihara, A. Furukawa, Y. Hidaka, R. Ariga, S. Okamoto, K. Ichinose, *Bioorg. Med. Chem. Lett.* **2012**, 22, 5041-5045.

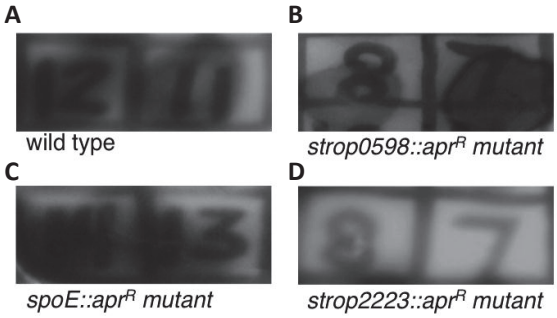
Table 1. Organization of the lomaiviticin biosynthetic gene cluster (*lom*) from *Salinispora tropica* CNB-440.

Gene	Size [aa]	Predicted function	Closest homolog [similarity/identity, (%/%)]	Reference	Kinamycin homolog
strop2172	524	Drug resistance transporter	Drug resistance transporter [Salinispora arenicola CNS-205] (84/90)	ABV98201	
strop2173	161	Hypothetical protein	Hypothetical protein [Salinispora arenicola CNS-205] (75/86)	ABV98202	
strop2174	276	ABC transporter	ABC transporter [Catenulispora acidiphila DSM 44928] (56/39)	ACU70166	
strop2175	316	ABC transporter	ABC transporter [Streptomyces sviceus ATCC 29083] (60/74)	EDY54089	
strop2176	171	Bleomycin resistance protein	Bleomycin resistance protein [Streptomyces pristinaespiralis ATCC 25486] (63/79)	EDY62258	
strop2177	258	Hypothetical protein	Alpha/beta hydrolase [Streptomyces sp. Mg1] (49/62)	EDX26142	
strop2178	278	Transcriptional activator	Transcriptional activator [Streptomyces coelicoflavus ZG0656] (68/52)	EHN80233	
strop2179	113	Hypothetical protein	Hypothetical protein [Lodderomyces elongisporus NRRL YB-4239] (63/40)	XP_001524904	
strop2180	281	AraC-family transcriptional regulator	AraC-family transcriptional regulator [Micromonospora sp. ATCC 39149] (61/45)	ZP_04603952	
strop2181	477	NDP-hexose 2,3-dehydratase	3-dehydratase [Streptomyces coelicoflavus ZG0656] (81/72)	EHN77727	
strop2182	221	Hypothetical protein	Hypothetical protein [Streptomyces coelicoflavus ZG0656] (79/63)	EHN77694	
strop2183	148	Carboxymuconolactone decarboxylase	Carboxymuconolactone decarboxylase [Streptomyces coelicoflavus ZG0656] (73/58)	EHN77723	
strop2184	275	Glutamine amidotransferases	Peptidase C26 [Streptomyces coelicoflavus ZG0656] (77/67)	EHN77724	
strop2185	590	Aromatic ring hydroxylase	XiaK [Streptomyces sp. SCSIO 02999] (61/47)	AFK78077	
strop2186	488	Protoporphyrinogen oxidase	Protoporphyrinogen oxidase [Streptomyces coelicoflavus ZG0656] (73/62)	EHN77710	
strop2187	368	ParB domain protein nuclease	Nuclease, partial [Streptomyces coelicoflavus ZG0656] (79/67)	HN77721	
strop2188	503	Anthrone hydroxylase	FAD-binding monooxygenase [Streptomyces coelicoflavus ZG0656] (78/71)	EHN77709	KinO2 (60/47)
strop2189	490	Anthrone hydroxylase	FAD-binding monooxygenase [Streptomyces coelicoflavus ZG0656] (75/66)	EHN77708	
strop2190	263	3-oxoacyl-ACP reductase	Short-chain dehydrogenase/reductase [Streptomyces coelicoflavus ZG0656] (82/71)	EHN77698	
strop2191	290	Putative dimerase	NmrA family protein [Streptomyces coelicoflavus ZG0656] (76/67)	EHN77697	
strop2192	205	DSBA oxidoreductase	DSBA oxidoreductase [Streptomyces hygroscopicus ATCC 53653] (71/57)	ZP_07297742	
strop2193	109	Polyketide synthesis cyclase	Cyclase [Streptomyces antibioticus] (85/75)	CAG14964	KinI (82/70)
strop2194	261	Polyketide ketoreductase	Short-chain dehydrogenase/reductase [Streptomyces bingchenggensis BCW-1] (85/77)	YP_004965103	KinE (80/71)
strop2195	345	Polyketide O-methyltransferase	O-methyltransferase [Streptomyces coelicoflavus] (78/63)	EHN77696	
strop2196	503	Anthrone hydroxylase	SaqE [Micromonospora sp. Tu 6368] (70/59)	ACP19351	KinOR (60/47)
strop2197	319	Polyketide aromatase	SaqL [Micromonospora sp. Tu 6368] (71/61)	ACP19357	KinD (67/55)
strop2198	479	FAD-dependent monooxygenase	Putative FAD-dependent monooxygenase [Streptomyces albaduncus] (61/52), JagF	CBH32087	
strop2199	492	Anthrone hydroxylase	Oxygenase-like protein [Streptomyces murayamaensis] (73/63)	AAO65343	KinO1 (73/63)
strop2200	247	Anthrone oxidase	JadG [Streptomyces venezuelae] (62/46)	AAV52247	KinG (68/52)
strop2201	627	Hypothetical protein	Hypothetical protein [Streptomyces ambofaciens] (71/61)	CAK51011	
strop2202	118	4Fe-4S ferredoxin	Putative ferredoxin [Streptomyces ambofaciens ATCC 23877] (83/75)	CAI78079	
strop2203	264	Hypothetical protein	Hypothetical protein [Salinispora tropica CNB-440] (78/65)	YP_001159312	
strop2204	500	Glutamine synthetase	Glutamine synthetase [uncultured bacterium BAC AB649/1850] (75/65)	AEE65491	
strop2205	484	Amidase	Putative amidase [Streptomyces coelicoflavus ZG0656] (73/63)	EHN77687	
strop2206	428	Adenylosuccinate lyase	Putative adenylosuccinate lyase [Streptomyces coelicoflavus ZG0656] (80/70)	EHN77655	
strop2207	137	N-acetyltransferase GCN5	Putative acetyltransferase [Streptomyces ambofaciens ATCC 23877] (79/72)	CAI78074	

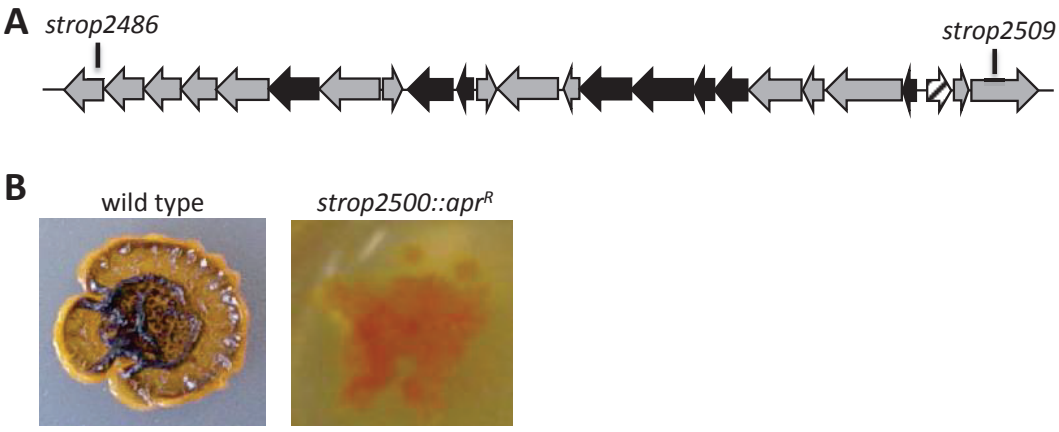
strop2208	523	ABC transporter-like protein	NovA [Streptomyces coelicoflavus ZG0656] (82/69)	EHN77703	
strop2209	106	Dioxygenase	Dioxygenase [Thermobifida fusca YX] (76/73)	YP_290041	
strop2210	136	Bleomycin resistance protein	Bleomycin resistance protein [Salinispora arenicola CNS-205] (97/92)	YP_001537194	
strop2211	263	Bleomycin resistance protein	Bleomycin resistance protein [Nesterenkonia sp. F] (61/42)	ZP_09541316	
strop2212	520	Secreted peptidase	TAP domain-containing protein [Streptomyces coelicoflavus ZG0656] (72/61)	EHN77722	
strop2213	380	Glycosyltransferase	Glycosyl transferase [uncultured bacterium] (57/41)	AEM44235	
strop2214	371	Polyketide O-methyltransferase	O-methyltransferase [Ktedonobacter racemifer DSM 44963] (66/56)	ZP_06965007	
strop2215	313	O-methyltransferase	NanM [Streptomyces nanchangensis] (77/60)	AAP42862	
strop2216	328	NDP-hexose 4-ketoreductase	4-ketoreductase [Streptomyces sp. TP-A0274] (68/56)	BAC55215	
strop2217	199	NDP-hexose 3,5-epimerase	3,5-epimerase [Streptomyces sp. TP-A0274] (77/67)	BAC55217	
strop2218	335	NDP-hexose 3-ketoreductase	Putative 3-ketoreductase [Streptomyces galilaeus] (68/55)	AAF73453	
strop2219	244	Glycosyltransferase (auxiliary)	Hypothetical protein [Salinispora arenicola CNS-205] (88/80)	YP_001537195	
strop2220	344	Glycosyltransferase	Glycosyl transferase [Streptomyces cyanogenus] (55/42)	AAD13553	
strop2221	401	NDP-hexose 3,4-dehydratase/ aminotransferase	3,4-dehydratase-like protein [Streptomyces sp. KCTC 0041BP] (74/61)	ABB52533	
strop2222	333	dTDP-glucose-4,6-dehydratase	Putative dTDP-glucose-4,6-dehydratase [Actinoplanes missouriensis 431] (79/69)	YP_005462170	
strop2223	423	Minimal type II PKS, KS	Beta-ketoacyl synthase [Streptomyces acidiscabies 84-104] (82/71)	ZP_10452176	KinA (79/66)
strop2224	423	Minimal type II PKS, CLF	Putative chain length factor (CLF) [Streptomyces ravidus] (76/62)	CBH32808	KinB (73/59)
strop2225	84	Minimal type II PKS, ACP	ACP [Thermomonospora curvata DSM 43183] (73/48)	YP_003300377	KinC (62/47)
strop2226	327	Propionate starter unit, AT	AknF [Streptomyces galilaeus] (64/53)	BAB72049	
strop2227	322	Propionate starter unit, KS	Putative modular polyketide synthase [Kitasatospora setae KM-6054] (45/30)	BAJ32815	
strop2228	107	Propionate starter unit, ACP	Actinorhodin polyketide dimerase, ACP [Thermobifida fusca YX] (71/51)	YP_289281	
strop2229	160	AraC family transcriptional regulator	AraC family transcriptional regulator [Salinispora arenicola CNS-205] (91/87)	YP_001537196	
strop2230	290	Glucose-1-phosphate thymidylyl-transferase	G1P thymidylyltransferase [Thermomonospora curvata DSM 43183] (87/72)	YP_003300347	



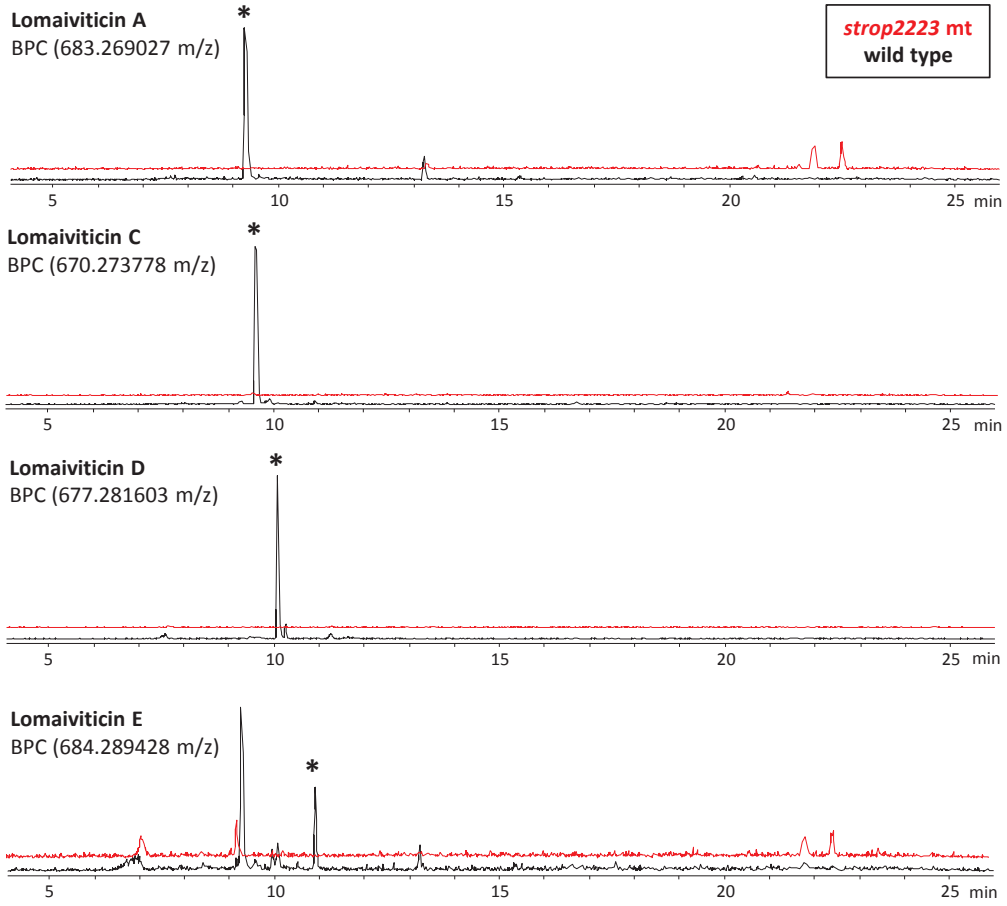
Supplementary Figure 1| (**A**) Proposed Bergman-type cyclization of a highly labile 9-membered enediyne precursor to yield observed sporolide A (**1**) and B (**2**). (**B**) LC profile comparison (monitored at 254 nm) of chemical extracts from wild type *S. tropica* and the *spoE* gene elimination mutant, illustrating eradication of sporolides A/B biosynthesis in the mutant and supporting the role of proposed enediyne PKS SpoE in sporolide biosynthesis.



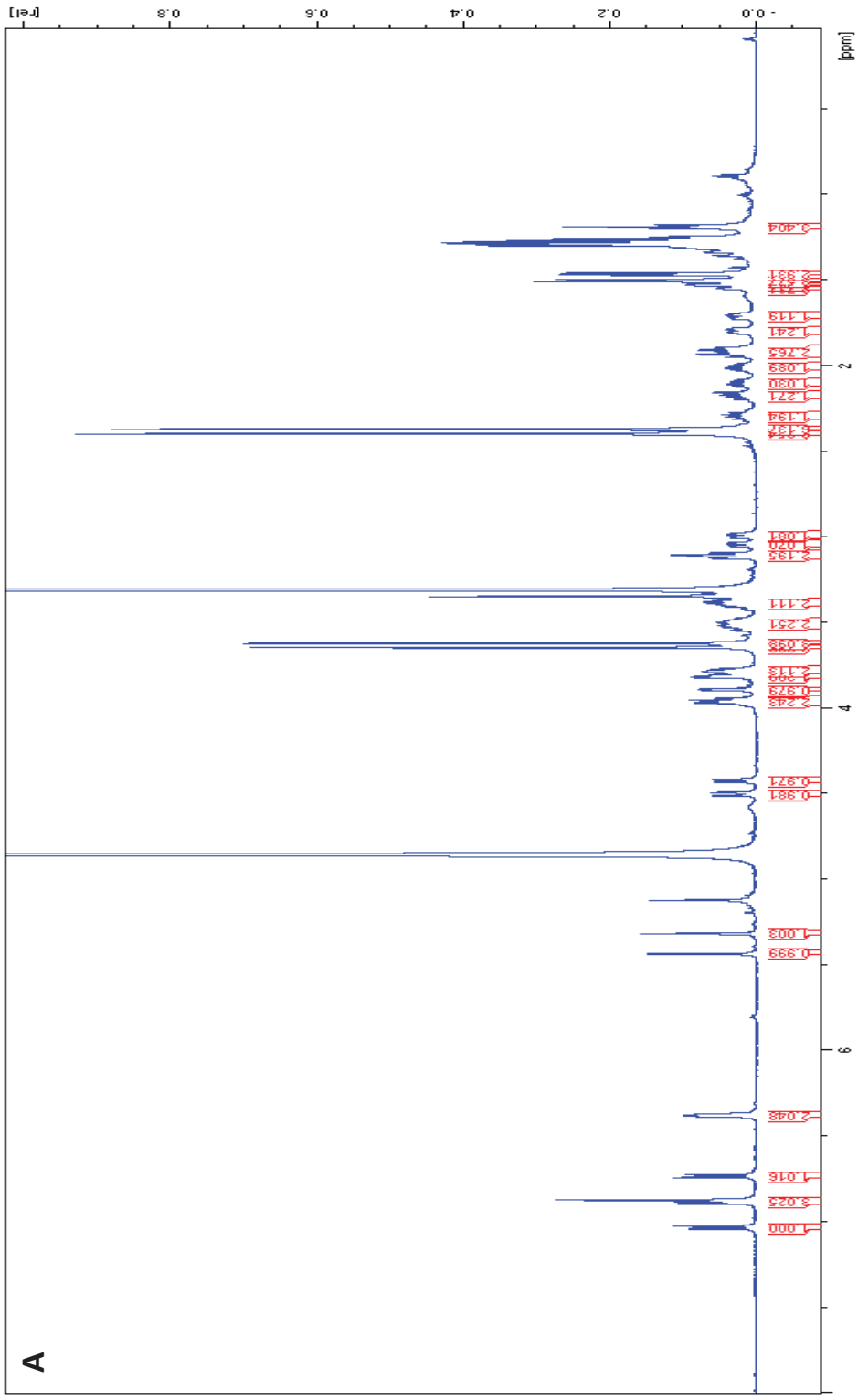
Supplementary Figure 2| BIA for methanol extracts from 7-day cultures of wild type *S. tropica* and gene elimination mutants. Dark spots on the assay plate indicate DNA interference activity, with a complete loss of activity observed for the *strop2223* ketosynthase gene elimination mutant (n = 2, 10 µL of a 1mg/mL stock). (**A**) Wild type *S. tropica*. (**B**) *ST_pks1* mutant. (**C**) *spo* mutant. (**D**) *ST_pks2 (lom)* mutant.



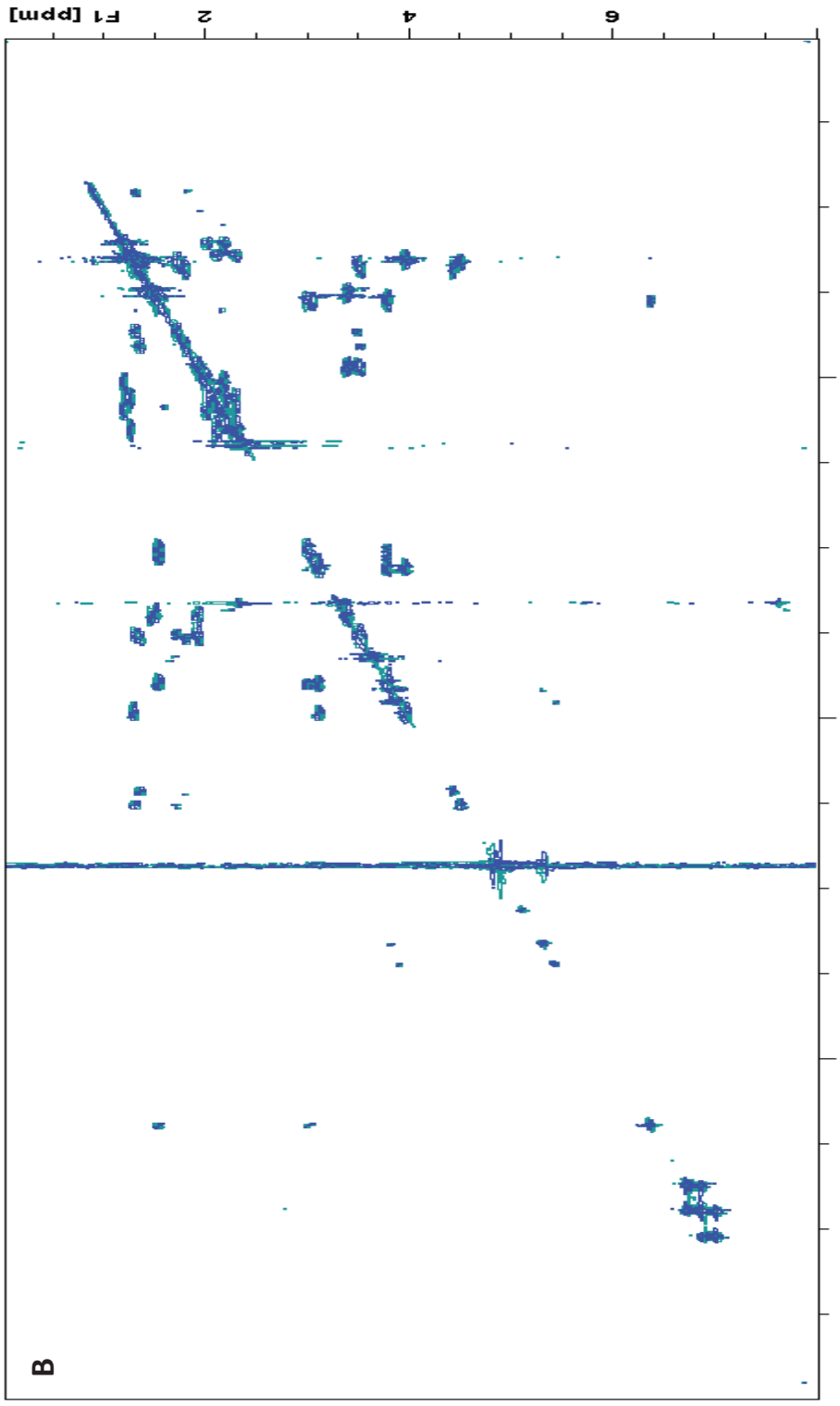
Supplementary Figure 3 | Characterization of *ST_pks3* as a spore pigment pathway. (**A**) The predicted *ST_pks3* gene cluster in *Salinispora tropica* CNB-440. See Figure 1 for color legend. (**B**) Sporulation phenotyping of wild type *S. tropica* and the *S. tropica ST_pks3* mutant on A1 agar after 14d growth at 28°C.



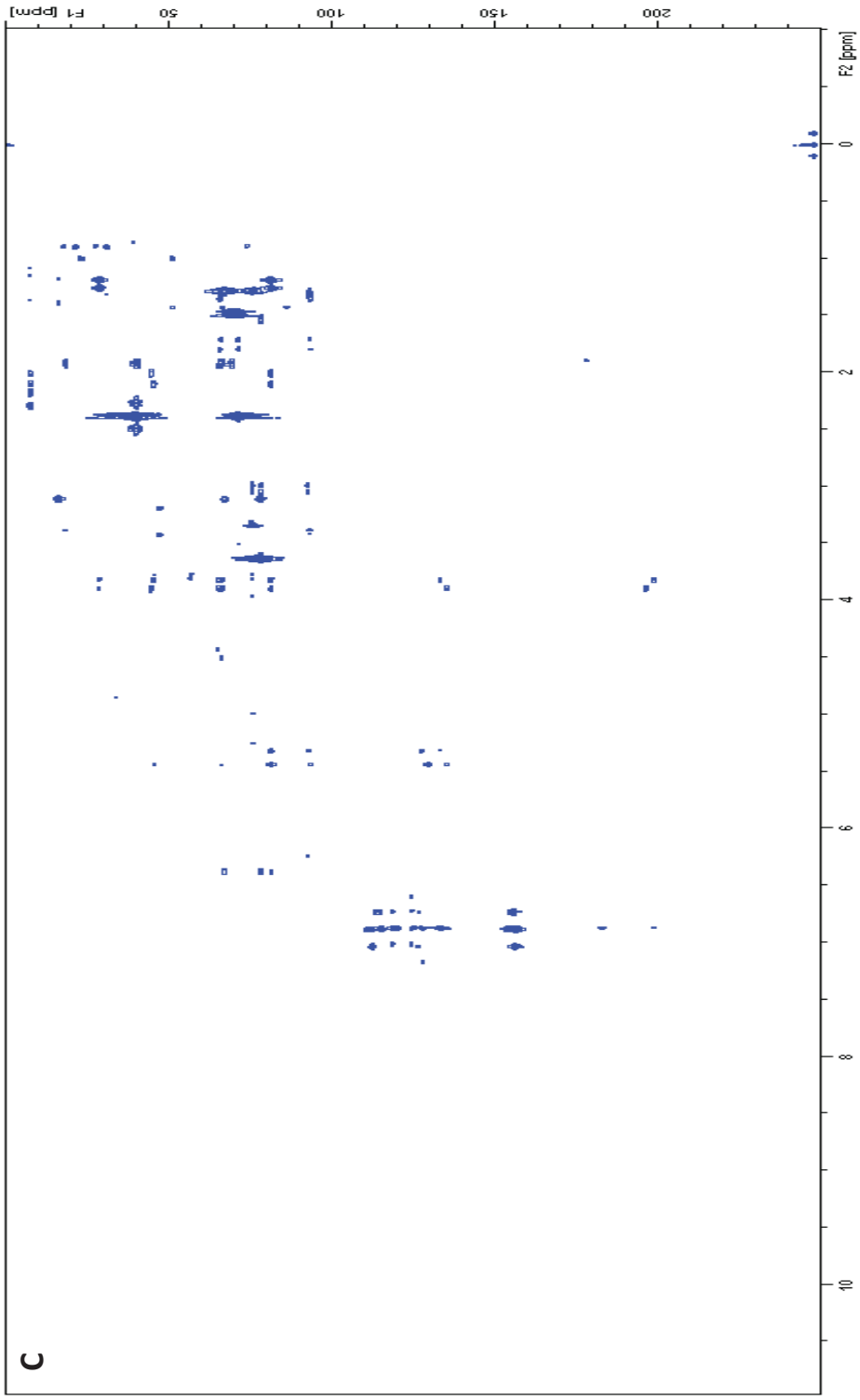
Supplementary Figure 4 | LC-MS comparative metabolic profiling of wild type (black) and *strop2223* mutant (red) *Salinispora tropica* CNB-440 revealing several abolished masses by *ST_pks2* inactivation. Target mass peaks in the wild type profiles are marked by asteriks.



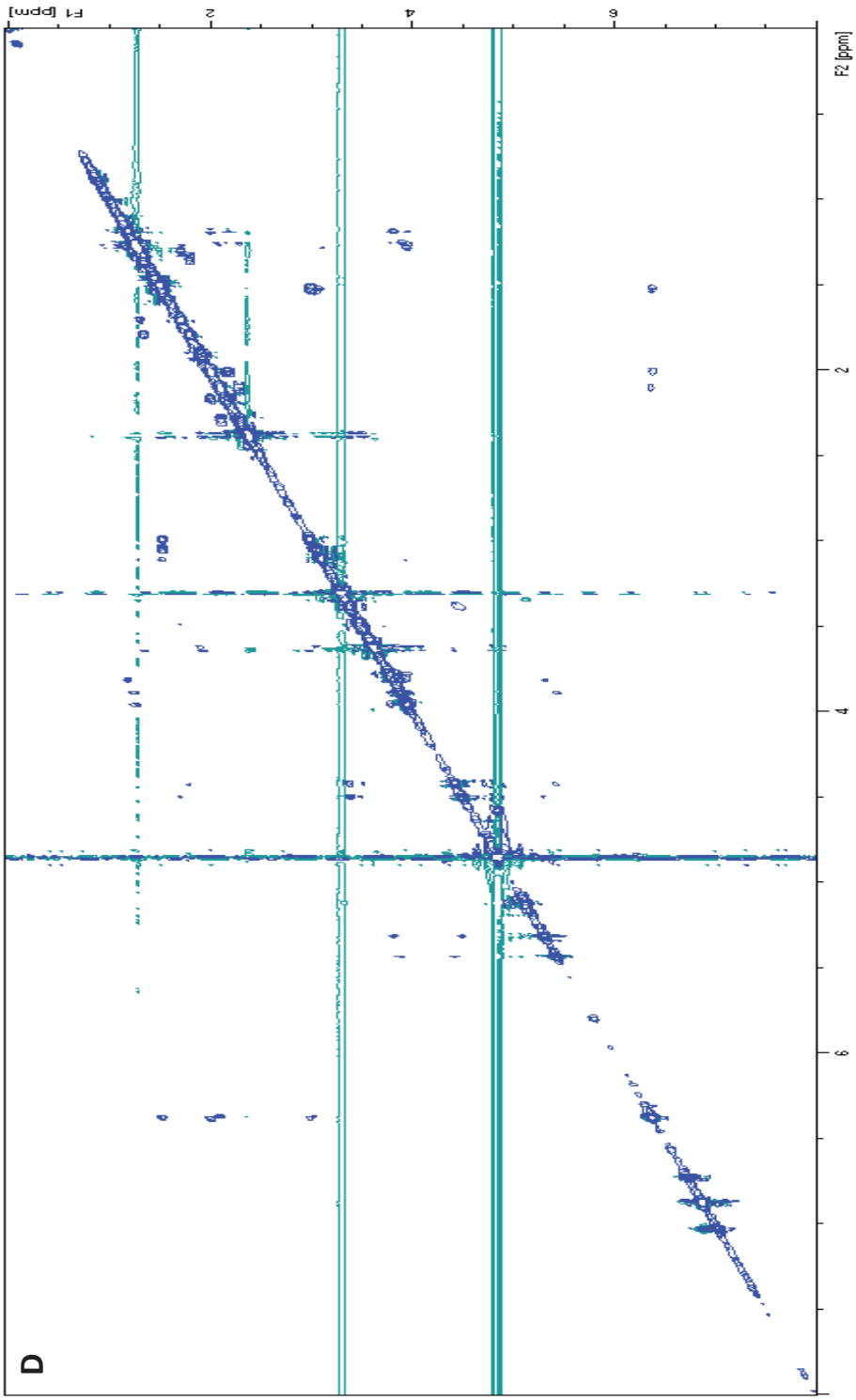
Supplementary Figure 5| NMR spectra of *ST_pks2 (lom)* product, lomaiviticin C. All spectra were observed in MeOD-d4 on a 600MHz instrument. For annotations, see Supplementary Table 3. (**A**) ^{1}H NMR spectrum.



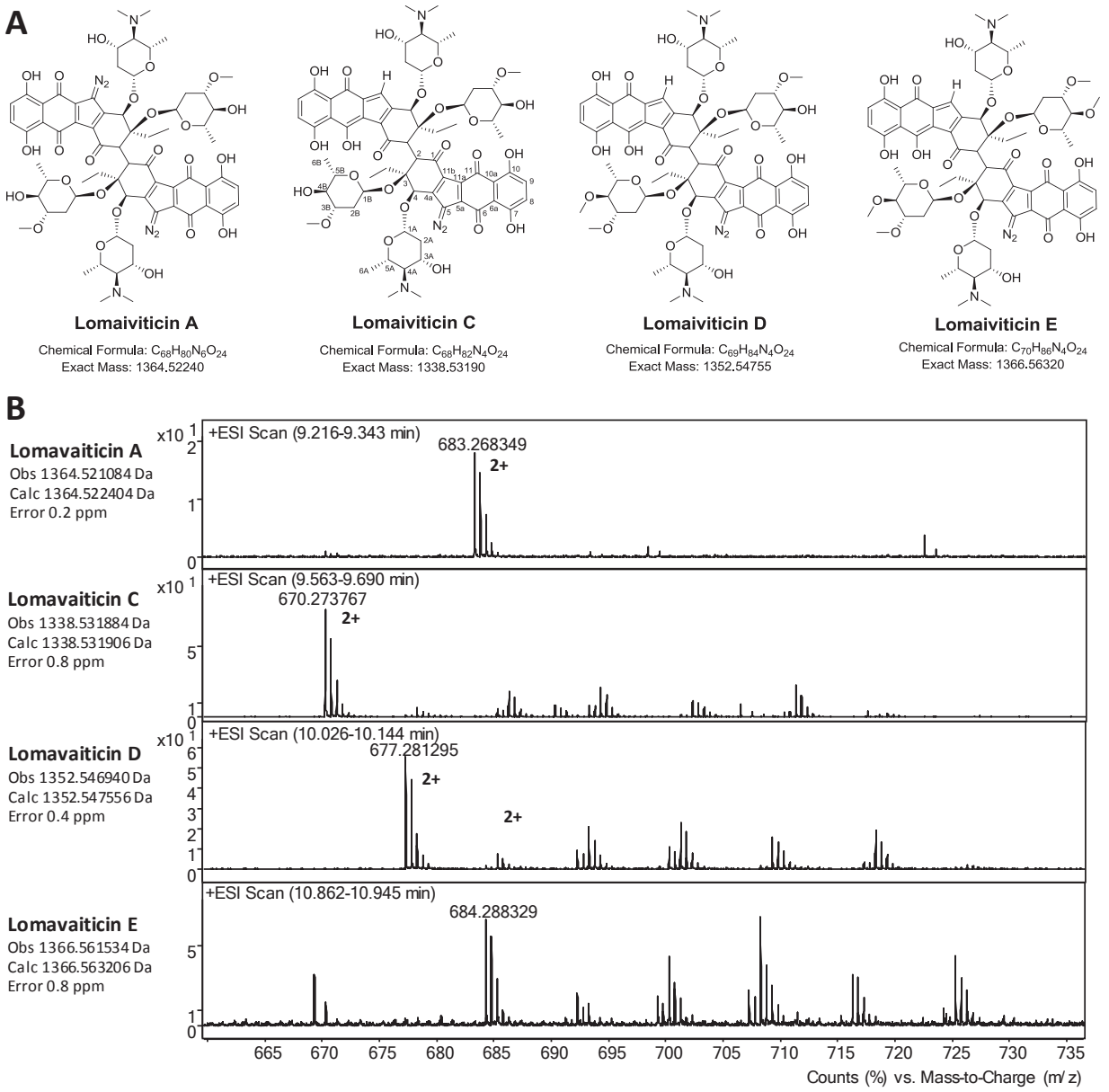
Supplementary Figure 5| NMR spectra of *ST_pks2 (lom)* product, lomaiviticin C. All spectra were observed in MeOD-d4 on a 600MHz instrument. For annotations, see Supplementary Table 3. (**B**) ^{1}H-^{1}H DFQ-COSY spectrum.



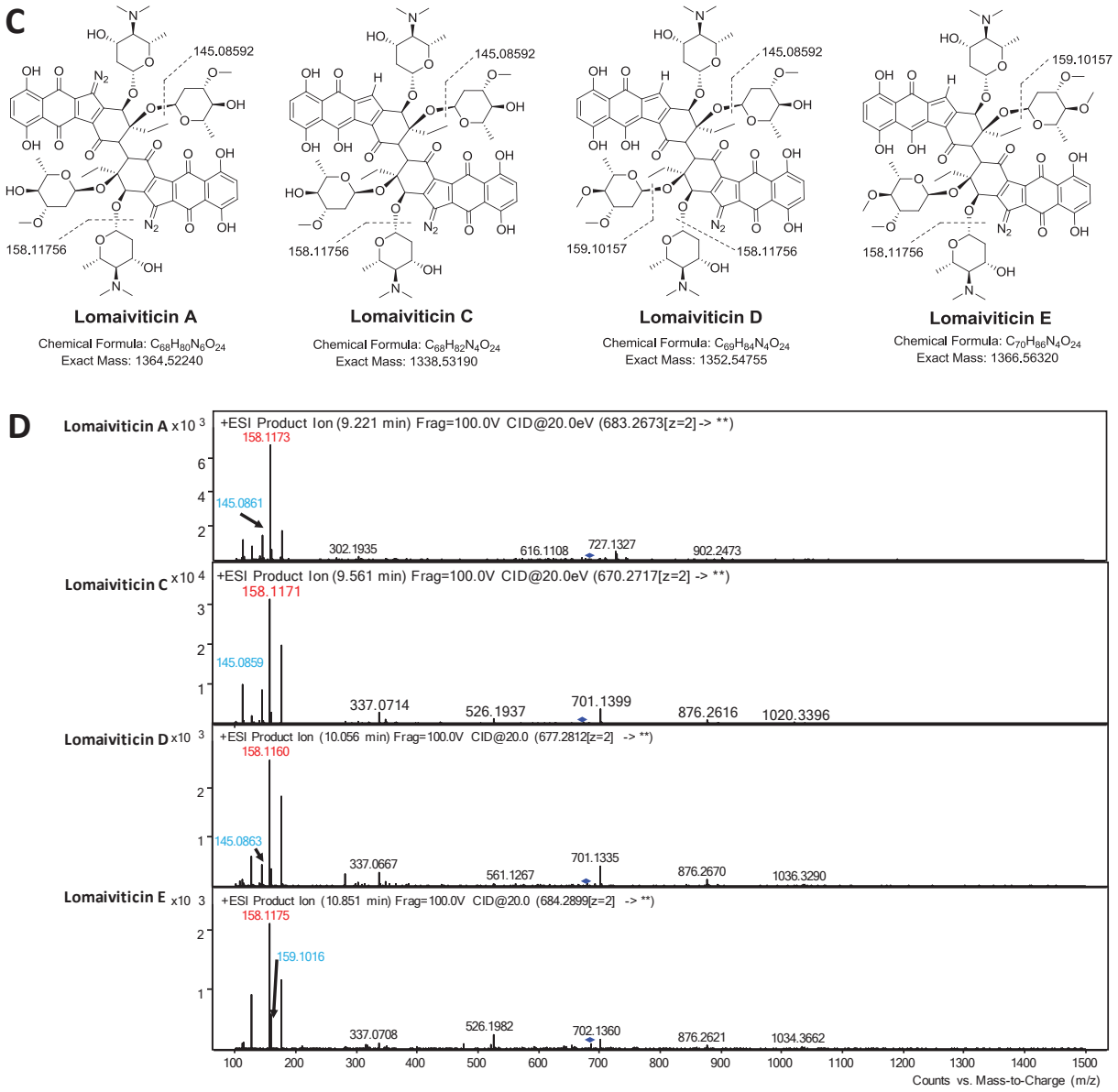
Supplementary Figure 5| NMR spectra of *ST_pks2 (lom)* product, lomaiviticin C. All spectra were observed in MeOD-d4 on a 600MHz instrument. For annotations, see Supplementary Table 3. (**C**) ^{1}H-^{13}C HMBC spectrum.



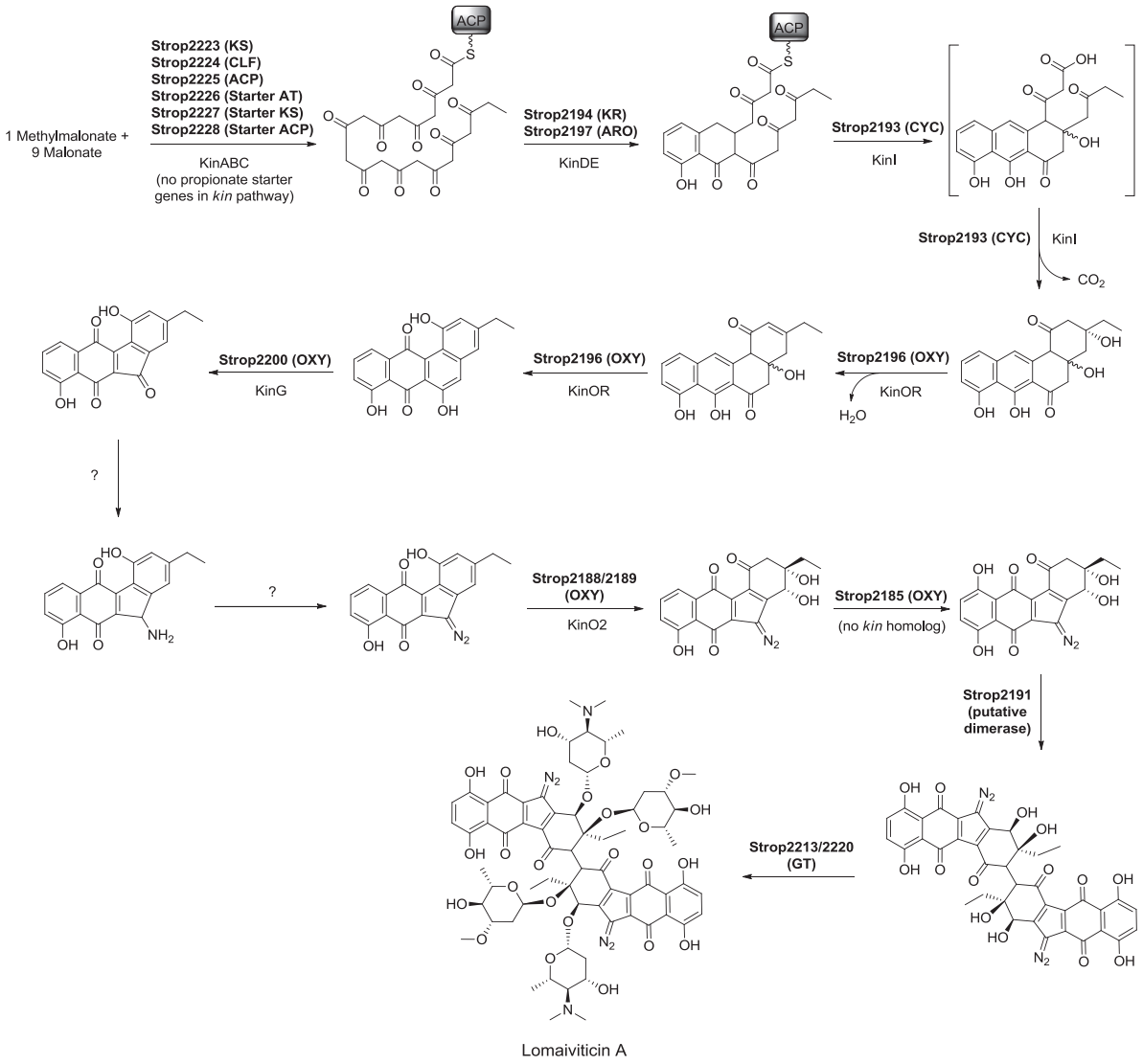
Supplementary Figure 5| NMR spectra of *ST_pks2 (lom)* product, lomaiviticin C. All spectra were observed in MeOD-d4 on a 600MHz instrument. For annotations, see Supplementary Table 3. (**D**) ^{1}H-^{1}H NOESY spectrum.



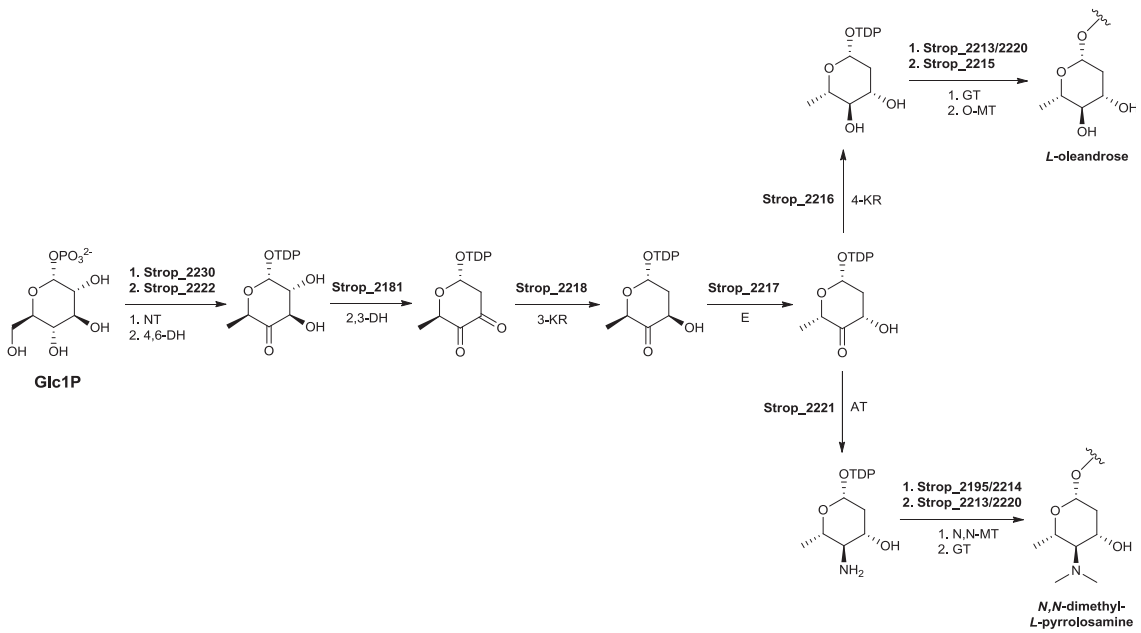
Supplementary Figure 6 | MS characterization of lomaiviticins A, C, D & E from a wild type *Salinispora tropica* CNB-440 extract. (**A**) Lomaiviticin A, C, D & E as reported by H. He, *et al. J. Am. Chem. Soc.* (**2001**) and C. M. Woo, N. E. Beizer, J. E. Janso, S. B. Herzon, *J. Am. Chem. Soc.* (**2012**). (**B**) HR-MS spectra of lomaiviticins A, C, D & E from LCMS/MS analysis.



Supplementary Figure 6 | MS characterization of lomaiviticins A, C, D & E from a wild type *Salinispora tropica* CNB-440 extract. (**C**) Predicted deoxysugar fragmentation of lomaiviticin A, C, D & E (**2012**). (**D**) MS/MS spectra of lomaiviticins A, B, C & D from LCMS/MS analysis. Observed B-ions of candidate MS/MS *N,N*-dimethylaminodeoxysugars are highlighted in red, candidate MS/MS *O*-methyldideoxysugars are highlighted in blue.



Supplementary Figure 7 | Proposed biosynthetic scheme of lomaiviticin A based on *lom* gene cluster and biosynthetic proposals by Gould [S. J. Gould, S. T. Hong, J. R. Carney, *J. Antibiot.* **1998**, *51*, 50-57] and Herzon [S. B. Herzon, C. M. Woo, *Nat. Prod. Rep.***2012**, *29*, 87-118]. *Kin* homologs are listed under pathway arrows.



Supplementary Figure 8 | Predicted biosynthesis of lomaiviticin deoxysugars, *L*-oleandrose and *N,N*-dimethyl-*L*-pyrrolosamine, by glycosylation genes in *ST_pks2 (lom)* gene cluster. Abbreviations: NT- nucleotidylyltransferase, 4,6-DH – 4,6-dehydratase, 2,3-DH – 2,3-dehydratase, 3-KR – 3-ketoreductase, AT – amino-Transferase, 4-KR – 4-ketoreductase, N,N-MT – N,N-dimethyltransferase, GT – glycosyltransferase, O-MT – O-methyltransferase

Supplementary Table 1| Bioinformatic analysis of *ST_pks1* gene cluster from *Salinispora tropica* CNB-440

Gene	Size [aa]	Predicted function	Closest homolog [similarity/identity, (%/%)]	Reference
strop0568	652	DNA-cytosine methyltransferase	DNA-cytosine methyltransferase [Saccharomonospora paurometabolica YIM 90007] (75/62)	ZP_09032022
strop0569	60	transposase	transposase IS4 family protein [Streptomyces violaceusniger Tu 4113] (75/64)	YP_004816282
strop0570	385	transposase	transposase IS4 family protein [Streptomyces violaceusniger Tu 4113] (79/66)	YP_004816282
strop0571	661	transposase	AIPR protein [Streptomyces cattleya NRRL 8057 = DSM 46488] (61/46)	YP_004912436
strop0572	504	ATPase	ATPase [Saccharothrix espanaensis DSM 44229] (70/55)	YP_007041495
strop0573	329	hypothetical protein	hypothetical protein BN6_73910 [Saccharothrix espanaensis DSM 44229] (59/44)	YP_007041494
strop0574	967	endonuclease	hypothetical protein BN6_73900 [Saccharothrix espanaensis DSM 44229] (61/49)	YP_007041493
strop0575	188	transcriptional regulator	hypothetical protein AMIS_7920 [Actinoplanes missouriensis 431] (71/57)	YP_005460528
strop0576	51	hydrolase	NUDIX hydrolase [Micromonospora lupini str. Lupac 08] (74/64)	ZP_21030025
strop0577	206	hypothetical protein	hypothetical protein AMIS_8300 [Actinoplanes missouriensis 431] (61/48)	YP_005460566
strop0578	143	hypothetical protein	hypothetical protein RHA1_ro10346 [Rhodococcus jostii RHA1] (62/49)	YP_708693
strop0579	158	hypothetical protein	hypothetical protein VAB18032_25335 [Verrucosispora maris AB-18-032] (71/61)	YP_004406765
strop0580	N/A	pseudogene	N/A	N/A
strop0581	174	hypothetical protein	hypothetical protein Sare_1943 [Salinispora arenicola CNS-205] (76/61)	YP_001536819
strop0582	87	DNA-binding protein	DNA-binding protein [Salinispora arenicola CNS-205] (81/62)	YP_001536821
strop0583	337	selenide, water dikinase	selenide, water dikinase [Catenulispora acidiphila DSM 44928] (85/74)	YP_003116389
strop0584	431	selenocysteine synthase	L-seryl-tRNA(Sec) selenium transferase [Streptosporangium roseum DSM 43021] (74/66)	YP_003339404
strop0585	604	selenocysteine-specific translation elongation factor	selenocysteine-specific translation elongation factor [Catenulispora acidiphila DSM 44928] (69/57)	YP_003116387
strop0586	478	transposase, mutator type	transposase mutator type [Salinispora arenicola CNS-205] (95/92)	YP_001536532
strop0587	111	transposase, mutator type	transposase mutator type [Micromonospora aurantiaca ATCC 27029] (63/54)	YP_003834873
strop0588	127	glyoxalase	glyoxalase [Streptomyces sviceus ATCC 29083] (73/66)	ZP_06921536
strop0589	192	nitroreductase	nitroreductase [Micromonospora sp. L5] (80/75)	YP_004085470
strop0590	590	peptidase S9B dipeptidylpeptidase IV subunit	peptidase S9B dipeptidylpeptidase IV domain-containing protein [Micromonospora aurantiaca ATCC 27029] (91/87)	YP_003835617
strop0591	287	type 11 methyltransferase	methyltransferase [Micromonospora aurantiaca ATCC 27029] (84/76)	YP_003835618
strop0592	315	hypothetical protein	hypothetical protein Micau_2505 [Micromonospora aurantiaca ATCC 27029] (81/75)	YP_003835620
strop0593	620	FG-GAP repeat-containing protein	fg-gap repeat protein [Micromonospora sp. L5] (89/84)	YP_004085465
strop0594	349	L-seryl-tRNA(Sec) selenium transferase	L-seryl-tRNA(Sec) selenium transferase [Micromonospora aurantiaca ATCC 27029] (79/77)	YP_003835622
strop0595	92	Beta-ketoacyl synthase	hypothetical protein Micau_2508 [Micromonospora aurantiaca ATCC 27029] (61/52)	YP_003835623
strop0596	231	hypothetical protein	hypothetical protein Micau_2510 [Micromonospora aurantiaca ATCC 27029] (92/81)	YP_003835625
strop0597	409	Unknown conserved protein	DynT3 [Micromonospora chersina] (87/73)	ACB47047
strop0598	1951	Enediyne core forming iterative polyketide synthase	DynE8 [Micromonospora chersina]	ACB47048
strop0599	138	4-hydroxybenzoyl-CoA thioesterase	DynE7 [Micromonospora chersina] (70/55)	ACB47049
strop0600	297	transcriptional regulator	DynR7 [Micromonospora chersina] (51/37)	ACB47062
strop0601	402	PBS lyase HEAT-like repeat protein	PBS lyase HEAT-like repeat protein [Micromonospora chersina] (71/54)	ACB47059
strop0602	436	ABC-1 domain-containing protein	hypothetical protein [Micromonospora chersina] (60/44)	ACB47063
strop0603	397	Cytochrome P450	PlaO3 [Streptomyces sp. Tu6071] (58/43)	ZP_08452004
strop0604	276	methyltransferase	type 11 methyltransferase [Roseiflexus sp. RS-1] (59/43)	YP_001277818
strop0605	127	bleomycin resistance protein	Glyoxalase/bleomycin resistance protein/dioxygenase [Saccharomonospora paurometabolica YIM 90007] (88/77)	ZP_09030707
strop0606	233	activator	activator of Hsp90 ATPase 1 family protein [Nakamurella multipartita DSM 44233] (58/44)	YP_003200395
strop0607	205	transcriptional regulator	ArsR family transcriptional regulator [Kribbella flavida DSM 17836] (75/65)	YP_003379641
strop0608	152	transposase	transposase [Frankia sp. EAN1pec] (66/56)	YP_001505769
strop0609	227	transposase	transposase [Frankia symbiont of Datisca glomerata] (62/51)	YP_004585588
strop0610	305	transcriptional regulator	DynR7 [Micromonospora chersina] (50/36)	ACB47062

Supplementary Table 2| Bioinformatic analysis of *ST_pks3* gene cluster from *Salinispora tropica* CNB-440

Gene	Size [aa]	Predicted function		Reference
strop2486	321	transketolase	transketolase [Salinispora arenicola CNS-205] (95/93)	YP_001537497
strop2487	323	dehydrogenase, E1 component	dehydrogenase E1 component [Salinispora arenicola CNS-205] (96/95)	YP_001537498
strop2488	270	hypothetical protein	hypothetical protein Sare_2671 [Salinispora arenicola CNS-205] (99/98)	YP_001537499
strop2489	309	phytanoyl-CoA dioxygenase	phytanoyl-CoA dioxygenase [Salinispora arenicola CNS-205] (94/89)	YP_001537500
strop2490	389	4-hydroxyphenylpyruvate dioxygenase	4-hydroxyphenylpyruvate dioxygenase [Salinispora arenicola CNS-205] (90/86)	YP_001537501
strop2491	414	beta-ketoacyl synthase	beta-ketoacyl synthase [Salinispora arenicola CNS-205] (95/92)	YP_001537502
strop2492	538	AMP-dependent synthetase and ligase	AMP-dependent synthetase and ligase [Salinispora arenicola CNS-205] (94/91)	YP_001537503
strop2493	166	hypothetical protein	N/A	N/A
strop2494	374	beta-ketoacyl synthase	beta-ketoacyl synthase [Salinispora arenicola CNS-205] (92/89)	YP_001537505
strop2495	88	ACP	hypothetical protein Sare_2679 [Salinispora arenicola CNS-205] (92/84)	YP_001537507
strop2496	226	4'-phosphopantetheinyl transferase	4'-phosphopantetheinyl transferase [Salinispora arenicola CNS-205] (82/78)	YP_001537508
strop2497	466	AMP-dependent synthetase and ligase	AMP-dependent synthetase and ligase [Salinispora arenicola CNS-205] (92/89)	YP_001537509
strop2498	119	antibiotic biosynthesis monooxygenase	antibiotic biosynthesis monooxygenase [Salinispora arenicola CNS-205] (92/90)	YP_001537510
strop2499	415	beta-ketoacyl synthase I	beta-ketoacyl synthase [Salinispora arenicola CNS-205] (94/92)	YP_001537511
strop2500	423	beta-ketoacyl synthase II	beta-ketoacyl synthase [Salinispora arenicola CNS-205] (96/95)	YP_001537512
strop2501	135	putative polyketide synthase, whiE	cupin [Salinispora arenicola CNS-205] (93/91)	YP_001537513
strop2502	239	polyketide cyclase/dehydrase	cyclase/dehydrase [Salinispora arenicola CNS-205] (92/90)	YP_001537514
strop2503	458	carbamoyl-phosphate synthase	carbamoyl-phosphate synthase L chain ATP-binding [Salinispora arenicola CNS-205] (96/94)	YP_001537515
strop2504	192	acetyl-CoA carboxylase, biotin carboxyl carrier protein	acetyl-CoA carboxylase, biotin carboxyl carrier protein [Salinispora arenicola CNS-205] (77/72)	YP_001537516
strop2505	566	acetyl-CoA carboxylase, carboxyl transferase subunit beta	acetyl-CoA carboxylase, carboxyl transferase subunit beta [Salinispora arenicola CNS-205] (94/92)	YP_001537517
strop2506	110	polyketide synthesis cyclase	polyketide synthesis cyclase [Salinispora arenicola CNS-205] (96/92)	YP_001537518
strop2507	275	transcriptional activator	SARP family transcriptional regulator [Salinispora arenicola CNS-205] (90/87)	YP_001537519
strop2508	115	polyketide synthesis cyclase	polyketide synthesis cyclase [Salinispora arenicola CNS-205] (96/94)	YP_001537520
strop2509	450	FAD-binding monooxygenase	FAD-binding monooxygenase [Salinispora arenicola CNS-205] (93/90)	YP_001537521

Supplementary Table 3| NMR analysis of *ST_pks2 (lom)* product, lomaiviticin C, at 600 MHz in MeOD-d4. Abbreviations: br- broad, d- doublet, *J* – coupling constant in Hertz [Hz], m – multiplet, s – singlet, t – triplet). Positions marked with ‘ in the table below correspond to hydroxyfulvene side of molecule. See Supplementary Figure 6A for positions.

Position	δH	δH - Integral, signal, *J*-value	δC	HMBC (H→C)	DFQ-COSY	NOESY
1	-	-	196.7	-	-	-
2	3.89	1H, d, *J* = 3.1 Hz	45.8	C-1, -3, -4, -11a, -11b, -2’	H-4, -2’	H-4, -12, -13, -2’
3	-	-	81.5	-	-	-
4	5.43	1H, s	65.8	C-2, -3, -4a, -11b, -1A	H-2	H-2, -1A
4a	-	-	129.8	-	-	-
5	-	-	N/A	-	-	-
5a/11a	-	-	N/A/126.8	-	-	-
6/11	-	-	183.0/180.8	-	-	-
6a/10a	-	-	112.7/111.6	-	-	-
7/10	-	-	156.5/157.1	-	-	-
8/9	7.03/6.88	1H, d, *J* = 9.2 Hz 1H, m	128.7/127.1	C-7, -6a, -6/C-10, -10a, -11	H-9/H-8	H-9/H-8
11b	-	-	135.3	-	-	-
12	2.08-2.12/ 2.27-2.31	1H, m 1H, m	28.7	C-2, -3, -13	H-12, -13	H-12, -13, -1B
13	1.26	3H, m	7.8	C-3, -12	H-12	H-2, -12, -3B
1A	4.42	1H, d, *J* = 9.2 Hz	93.9	C-4, -2A	H-2A	H-4, -2A, -3A, -5A
2A	1.36/1.79	1H, m/1H, dd, *J* = 10.5/4.8	39.5	C-1A, -3A, -4A	H-1A, -2A, -3A	H-1A, -2A, -3A
3A	3.47-3.54	1H, m	65.8	C-2A, -4A	H-2A, -4A	H-1A, -2A
4A	1.93	1H, m	71.7	C-3A, -4A ($N(CH_3)_2$), -5A, -6A	H-3A, -5A	H-2A, -3A, -4A ($N(CH_3)_2$), -6A, -3B (OCH_3)
4A ($N(CH_3)_2$)	2.39	6H, s	40.4	-	-	H-3A, -5A
5A	3.37-3.41	1H, m	69.3	C-1A, -3A, -4A, -6A	H-4A, -6A	H-1A, -3A, -4A, -4A ($N(CH_3)_2$), -6A
6A	1.50	3H, d, *J* = 6.1 Hz	18.9	C-5A	H-5A	H-4A, -5A
1B	6.38	1H, br s	92.7	C-3, -3B, -5B	H-2B	H-12, -2B
2B	1.53/2.99	1H, m/1H, dd, *J* = 11.4/4.8 Hz	35.9	C-1B, -3B, -4B	H-1B’, -2B’, -3B’	H-1B, -2B, -3B
3B	3.76-3.81	1H, m	78.5	C-3B (OCH_3), -4B	H-2B, -4B	H-2B, -3B (OCH_3), -4B, -5B
3B (OCH_3)	3.62/3.64	3H, s	57.0	C-3B		H-2B, -3B, -4B, -5B, -6B
4B	3.11	1H, t, *J* = 9.0 Hz	76.0	C-3B, -5B, -6B	H-3B, -5B	H-2B, -3B, -5B, -6B
5B	3.93-3.99	1H, m	67.5	C-3B, -4B, -6B	H-4B, -6B	H-13, -3B, -4B, -6B
6B	1.30	3H, m	16.5	C-1B, -4B, -5B	H-5B	H-4B, -5B
1’	-	-	198.9	-	-	-
2’	3.81	1H, m	44.8	C-1, -3, -4, -11a, -11b, -2’	H-4’, -2	H-4’, -12’, -13’, -2
3’	-	-	81.5	-	-	-
4’	5.32	1H, s	66.3	C-2’, -3’, -4a’, -11a’, -11b’, -1A’	H-2’	H-2’, -1A’
4a’	-	-	127.8	-	-	-
5’	6.87	1H, m	119.5	C-4a’, -5a’, -6’, -11a’, -11b’	-	-
5a’/11a’	-	-	125.4/120.1	-	-	-
6’/11’	-	-	183.5/181.2	-	-	-
6a’/10a’	-	-	114.5/115.5	-	-	-
7’/10’	-	-	155.7/154.7	-	-	-
8’/9’	6.73/6.88	1H, d, *J* = 9.2 Hz 1H, m	125.1/125.4	C-7, -6a, -6/C-10, -10a, -11	H-9/H-8	H-9’/H-8’
11b’	-	-	133.2	-	-	-
12’	1.99-2.02/ 2.14-2.19	1H, m/1H, m	28.7	C-2’, -3’, -13’	H-12’, -13’	H-12’, -13’, -1B’
13’	1.19	3H, t, *J* = 7.0 Hz	7.8	C-3’, -12’	H-12’	H-12’, -3B’, -5B’

Position	δH	δH - Integral, signal, *J*-value	δC	HMBC (H→C)	DFQ-COSY	NOESY
1A′	4.50	1H, d, *J* = 9.2 Hz	93.7	C-4′, -2A′	H-2A′	H-4′, -2A′, -3A′, -5A′
2A′	1.31/1.71	1H, m/1H, dd, *J* = 11.4/4.8	39.4	C-1A′, -3A′, -4A′	H-1A′, -2A′, -3A′	H-1A′, -2A′, -3A′
3A′	3-47-3.54	1H, m	66.3	C-2A′, -4A′	H-2A′, -4A′	H-1A′, -2A′
4A′	1.92	1H, m	71.7	C-3A′, -4A ($N(CH_3)_2$)′, -5A′, -6A′	H-3A′, -5A′	H-2A′, -3A′, -4A ($N(CH_3)_2$)′, -5A′, -6A′, -3B (OCH_3)′
4A ($N(CH_3)_2$)′	2.37	6H, s	40.4	-	-	H-3A′, -5A′
5A′	3.36-3.41	1H, m	69.7	C-1A′, -6A′	H-4A′, -6A′	H-1A′, -3A′, -4A ($N(CH_3)_2$)′, -6A′
6A′	1.46	3H, d, *J* = 6.1 Hz	18.5	C-5A′	H-5A′	H-4A′, -5A′
1B′	6.37	1H, br s	93.1	C-3′, -3B′, -5B′	H-2B′	H-12′, -2B′
2B′	1.54/3.04	1H, m/1H, dd, *J* = 11.9/4.9 Hz	35.7	C-1B′, -3B′, -4B′	H-1B′, -2B′, -3B′	H-1B′, -2B′, -3B′
3B′	3.76-3.81	1H, m	78.5	C-3B (OCH_3)′, -4B′	H-2B′, -4B′	H-2B′, -3B (OCH_3)′, -4B′, -5B′
3B (OCH_3)′	3.62/3.64	3H, s	57.0	C-3B′	-	H-2B′, -3B′, -4B′, -5B′, -6B′
4B′	3.11	1H, t, *J* = 9.0 Hz	76.0	C-3B′, -5B′, -6B′	H-3B′, -5B′	H-2B′, -3B′, -5B′, -6B′
5B′	3.93-3.99	1H, m	67.5	C-3B′, -4B′, -6B′	H-4B′, -6B′	H-13′, -3B′, -4B′, -6B′
6B′	1.30	3H, m	16.5	C-1B′, -4B′, -5B′	H-5B′	H-4B, -5B

Supplementary Table 4| Primers for *S. tropica* gene knockouts

Target pathway	Primers for *S. tropica* gene knockout	Primer sequences (5' --> 3')
ST_pks1	*strop0598* forward	CCGGGCCCGGCGGATTCGCCTAAGGAAGGCAGCCGCATGATTCCGGGGATCCGTCGACC
	strop0598 reverse	CATGTCGCCCCGATTTGGTCGCCACCGCCACCGCCACCTGTAGGCTGGAGCTGCTTC
lom	*strop2223* forward	CGGGAGAACGTCGGATCGACCGTGGCGGTGCCCCGGTGAATTCCGGGGATCCGTCGACC
	strop2223 reverse	GATCACCGCTGTCGTCATGCCGGACCCGTGCCCAACCGGTGTAGGCTGGAGCTGCTTC
spo	*spoE* forward	CGTCCCCGACTCAGGCAAGGAGGTTGTGCCGGCATGAGCATTCCGGGGATCCGTCGACC
	spoE reverse	AGTCCTCCATCTGCGTTACCCGTCCTTCCCAGTCAGCACTGTAGGCTGGAGCTGCTTC
ST_pks3	*strop2500* forward	TGTGCGAGCCCCGCGGTCGCGAAACGGAGTGGCACCGTGATTCCGGGGATCCGTCGACC
	strop2500 reverse	CGCCGCCCCAAGCAGGACCAGCGCACTGTTGAACCCGTCTGTAGGCTGGAGCTGCTTC

Chapter 5 Amendment – Characterization of two lomaiviticin genotypes in *Salinispora* genomes

Two candidate lomaiviticin genotypes have been characterized in *Salinispora* genomes: genotype 1 (*lom1*) is the *ST pks2*-type from *Salinispora tropica* CNB-440, genotype 2 (*lom2*) has seven additional genes and is only found in *Salinispora pacifica* genomes (Figure 31). LCMS-based metabolic profiling of a *lom* genotype 2-*Salinispora* strain, *Salinispora pacifica* CNR-114, revealed that lomaiviticin production occurred after 4 weeks (data not shown). In addition, a putative metabolite of mass 939 Da was detected in a *S. pacifica* CNR-114 extract and in *S. tropica lom*-knockout extracts after 10 days (Figure 32). This metabolite might compensate the slower or diminished production of lomaiviticins in *lom2*-genotypes or *lom1*-knockouts, respectively.

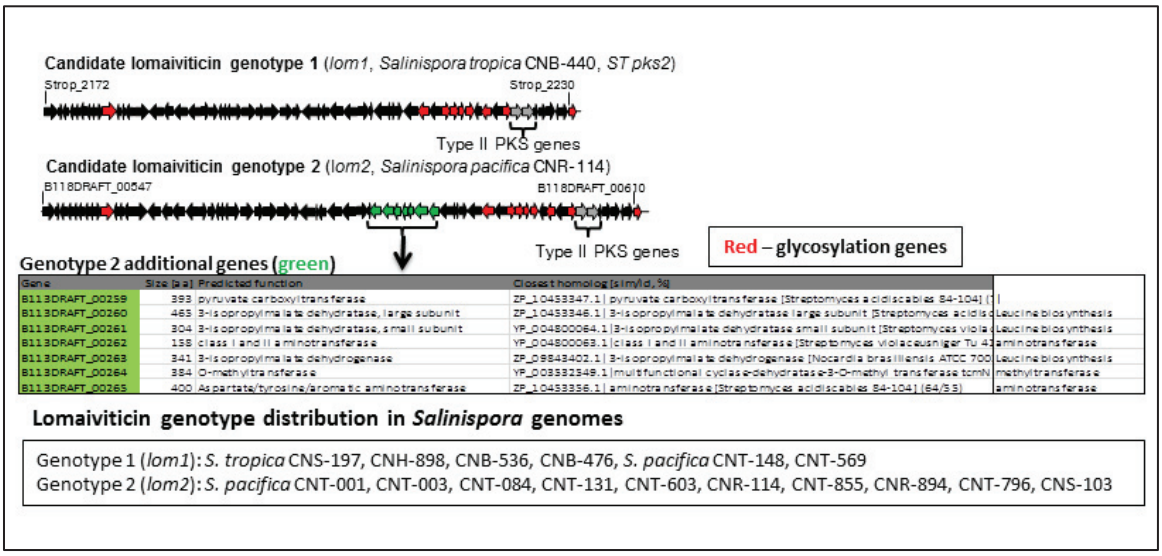


Figure 31: Characterization of two lomaiviticin genotypes (*lom1* and *lom2*) in *Salinispora* genomes.

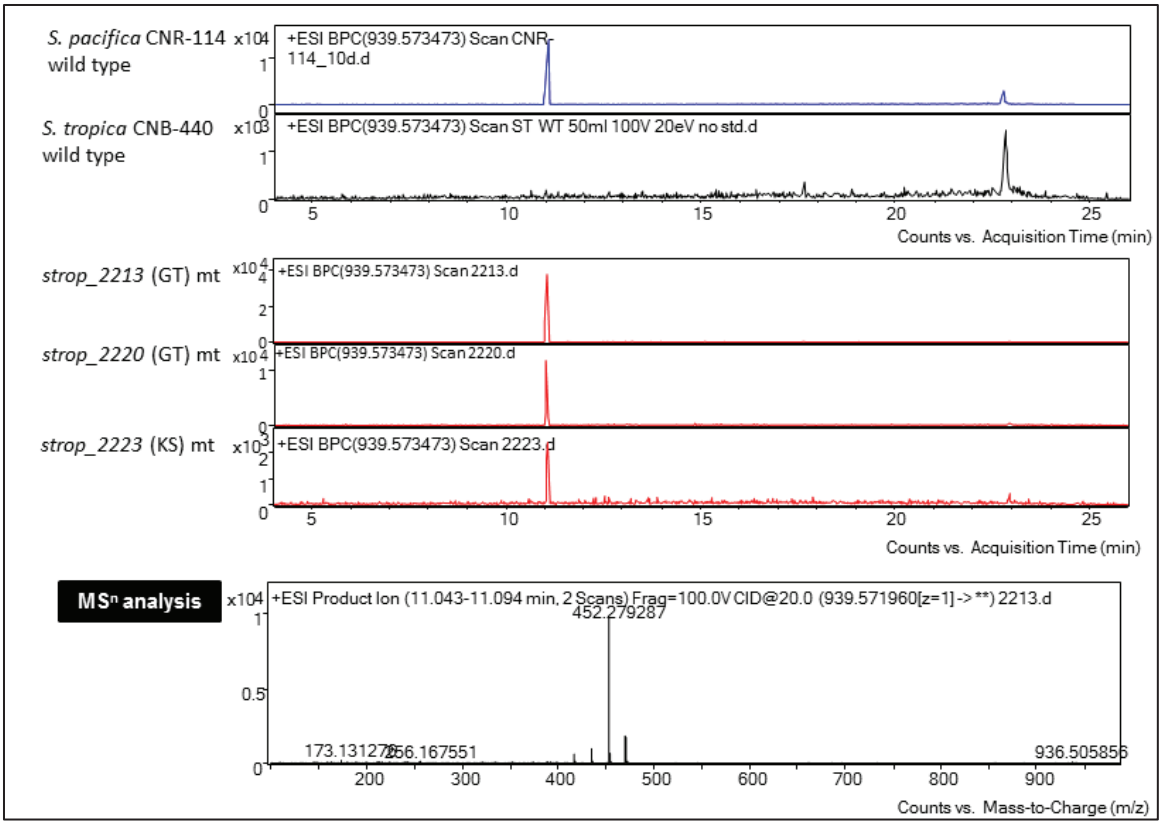


Figure 32: LCMS-based identification of putative upregulated *Salinispora* metabolite metabolic extracts of a *lom2*-strain and of *lom1*-mutant strains.

Chapter 5, in full, has been accepted for publication. It is shown as it may appear in ChemBioChem, Kersten, R.D., Lane, A.L., Nett, M., Richter, T.S.K., Duggan, B.M., Dorrestein, P.C., Moore, B.S. *Chembiochem*, 2013, DOI: 10.1002/cbic.2001300147. The dissertation author was the primary investigator and author of this paper.

R.D.K. designed and carried out cultivation, MS, chemical isolation and NMR experiments, analyzed data and wrote the paper, A.L.L. designed and carried out genetic knockout experiments, analyzed data and wrote the paper, M.N. and T.S.K.R. carried out genetic knockout experiments, B.M.D. carried out NMR experiments, P.C.D. designed experiments and analyzed data, B.S.M. designed experiments, analyzed data and wrote the paper.

Conclusions

In this dissertation, the concept of mass spectrometry-guided genome mining was introduced in two approaches, peptidogenomics and glycogenomics. Sequencing of microbial genomes has become cheaper and faster, thereby making it currently the most accessible 'omic'-dataset of a microbe. Consequently, genome databases are increasing exponentially [1]. Based on the realization that ~90% of secondary metabolic gene clusters in microbial genomes are uncharacterized [2], the increase of genomic information results in an increase of orphan secondary metabolic gene clusters. Traditional *in silico*-guided genome mining approaches do not match the pace with which genomes are sequenced nowadays as they only characterize one pathway chemically per experiment [3]. The motivation of this dissertation was to introduce a genome mining concept that could potentially link multiple pathways to their natural product chemotypes in one experiment and, thus establish a blueprint for automated natural product discovery by genome mining. The new approaches should then be tested by discovery of new chemo- and genotypes of microbial natural products.

MS-guided genome mining rapidly connects natural products (chemotypes) with their biosynthetic genes (genotypes) by matching *de novo* tandem MS structures to genomics-derived structures of predicted secondary metabolites. In peptidogenomics (Chapter 2) [4], amino acid mass shifts in peptide tandem MS spectra are connected to a ribosomal peptide precursor gene or to a NRPS gene cluster in a microbial genome. Thus, the peptidogenomic approach can readily differentiate between ribosomal and nonribosomal peptides. The candidate peptide biosynthetic gene cluster can enable verification, classification, partial structure elucidation and dereplication of the putative peptide natural product. In glycogenomics (Chapter 4), sugar mass signals in tandem MS spectra of *O*-/*N*-glycosylated natural products are connected to corresponding glycosylation genes in a secondary metabolic gene cluster via a MS-glycogenetic code. The biosynthetic genes of the aglycone can enable the classification, structure prediction and dereplication of the predicted GNP, and the glycosylation genes can enable further

verification of a GNP chemotype-genotype connection. MS-guided genome mining is an iterative approach in that a match of a tandem MS spectrum with a biosynthetic gene cluster is confirmed by reanalysis of tandem MS and genetic data. The presented genome mining concept could be automated as multiple chemotypes can be connected to their genotypes in one experiment. An example is the characterization of three different ribosomal peptide chemotypes and corresponding genotypes by tandem MS analysis of a *Streptomyces roseosporus* culture extract [4].

MS-guided genome mining can enable discovery of new peptide and GNP chemotypes. In this work, 18 new chemotypes were characterized by peptidogenomics and glycogenomics. Furthermore, MS-guided genome mining can enable discovery of new genotypes of known peptide and GNP chemotypes. Tandem MS analysis in combination with IMS and genomics revealed the didemnin biosynthetic pathway in the genome of the marine α-proteobacterium *Tistrella mobilis* (Chapter 3) [5]. In addition, LC-MS^n analysis of extracts of the marine actinobacterium *Salinispora tropica* [6] enabled the characterization of the lomaiviticins and their biosynthetic gene cluster in the genome of *S. tropica* (Chapter 5) [7]. The characterization of the didemnin and lomaiviticin biosynthetic gene clusters is an important discovery as both compounds are potent anti-cancer agents [8-10]. A biosynthetic gene cluster of a target compound produced in low yields from fermentation can open up an increased production by heterologous pathway expression or genetic engineering of pathway regulators [11,12]. For example, lomaiviticins have not entered clinical trials despite their high anticancer activities because yields from fermentation and organic synthesis are too low to supply required amounts of the most active derivative, lomaiviticin A [9]. The described lomaiviticin gene cluster could lead to a solution to this supply problem by heterologous expression or genetic upregulation. A biosynthetic gene cluster of a target compound can also enable production of structural derivatives by genetic engineering for structure-activity relationship studies and clinical trials [13]. Peptidogenomics has been already applied by the Dorrestein and Moore groups and by other

groups to characterize several chemo- and genotypes of new bioactive ribosomal and nonribosomal peptides from diverse bacteria [14-17].

MS-guided genome mining can also uncover new biosynthetic mechanisms by connecting peptidic and GNP chemotypes to their pathways. A combination of imaging mass spectrometry, tandem MS and genome analysis has characterized a new activation mechanism in NRP biosynthesis (Chapter 3) [5], in specific a putative didemnin activation mechanism. IMS and tandem MS showed the secretion of didemnin B precursors with N-terminal acyl-glutamine modifications prior to didemnin B occurrence. Genome analysis identified that the candidate didemnin gene cluster had a starter NRPS for N-terminal incorporation of acyl-glutamine chains into didemnin molecules. This led to the hypothesis that didemnins are biosynthesized as inactive didemnin precursors X/Y with N-terminal acyl-glutamine modifications, secreted and then cleaved by a putative extracellular esterase to form active didemnin B [5].

Ultimately, the introduced MS-guided genome mining approaches, peptidogenomics and glycogenomics, can expedite the discovery of PNP and GNP chemistry, biochemistry and bioactivity by rapid connections of PNP and GNP chemo- and genotypes from genome sequenced microbes. Integration of these approaches with new metabolomic and genomic tools, such as molecular networking [14] and AntiSMASH [18], respectively, and genetic and metabolomic databases [19,20] can lead to automation of genome mining for natural product discovery and integrate genome mining into natural product identification strategies.

References (Conclusion)

1. Pagani, I., Liolios, K., Jansson, J., Chen, I. M. A., Smirnova, T., Nosrat, B., Kyrpides, N. C. The Genomes OnLine Database (GOLD) v. 4: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res*. **40**, D571-D579 (2012).

2. Bentley, S.D., Chater, K.F., Cerdeño-Tárraga, A.M., Challis, G.L., Thomson, N.R., James, K.D., Harris, D.E., Quail, M.A., Kieser, H., Harper, D., Bateman, A., Brown, S., Chandra, G., Chen, C.W., Collins, M., Cronin, A., Fraser, A., Goble, A., Hidalgo, J., Hornsby, T., Howarth, S., Huang, C.H., Kieser, T., Larke, L., Murphy, L., Oliver, K., O'Neil, S., Rabbinowitsch, E., Rajandream, M.A., Rutherford, K., Rutter, S., Seeger, K., Saunders, D., Sharp, S., Squares, R., Squares, S., Taylor, K., Warren, T., Wietzorrek, A., Woodward, J., Barrell, B.G., Parkhill, J., Hopwood, D.A. Complete genome sequence of the model actinomycete Streptomyces coelicolor A3(2). *Nature* **417**, 141-147 (2002).

3. Zerikly, M., Challis, G.L. Strategies for the discovery of new natural products by genome mining. *Chembiochem*. **4**, 625-633 (2009).

4. Kersten, R.D., Yang, Y.L., Xu, Y., Cimermancic, P., Nam, S.J., Fenical, W., Fischbach, M.A., Moore, B.S., Dorrestein, P.C. A mass spectrometry-guided genome mining approach for natural product peptidogenomics. *Nat. Chem. Biol*. **7**, 794-802 (2011).

5. Xu, Y., Kersten, R.D., Nam, S.J., Lu, L., Al-Suwailem, A.M., Zheng, H., Fenical, W., Dorrestein, P.C., Moore, B.S., Qian, P.Y. Bacterial biosynthesis and maturation of the didemnin anti-cancer agents*. J. Am. Chem. Soc*. **134**, 8625-8632 (2012).

6. Udwary, D.W., Zeigler, L., Asolkar, R.N., Singan, V., Lapidus, A., Fenical, W., Jensen, P.R., Moore, B.S. Genome sequencing reveals complex secondary metabolome in the marine actinomycete *Salinispora tropica. Proc. Natl. Acad. Sci. U. S. A.* **104**, 10376-10381 (2007).

7. Kersten, R.D., Lane, A.L., Nett, M., Richter, T.K.S., Duggan, B.M., Dorrestein, P.C., Moore, B.S. Bioactivity-guided genome mining identifies the lomaiviticin biosynthetic gene cluster in *Salinispora tropica*. *Chembiochem*. DOI: 10.1002/cbic.2001300147 (2013).

8. Lee, J., Currano, J. N., Carroll, P. J., Joullié, M.M. Didemnins, tamandarins and related natural products. *Nat. Prod. Rep.* **29**, 404-424 (2012).

9. Herzon, S.B., Woo, C.M. The diazofluorene antitumor antibiotics: structural elucidation, biosynthetic, synthetic, and chemical biological studies. *Nat. Prod. Rep.* **29**, 87-118 (2012).

10. Woo, C.M., Beizer, N.E., Janso, J.E., Herzon, S.B. Isolation of lomaiviticins C-E, transformation of lomaiviticin C to lomaiviticin A, complete structure elucidation of lomaiviticin A, and structure-activity analyses. *J. Am. Chem. Soc*. **134**, 15285-15288 (2012).

11. Wenzel, S.C., Müller, R. Recent developments towards the heterologous expression of complex bacterial natural product biosynthetic pathways. *Curr. Opin. Biotechnol*. **16**, 594-606 (2005).

12. Laureti, L., Song, L., Huang, S., Corre, C., Leblond, P., Challis, G.L., Aigle, B. Identification of a bioactive 51-membered macrolide complex by activation of a silent polyketide synthase in Streptomyces ambofaciens. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 6258-6263 (2011).

13. Nett, M., Gulder, T.A., Kale, A.J., Hughes, C.C., Moore, B.S. Function-oriented biosynthesis of beta-lactone proteasome inhibitors in *Salinispora tropica*. *J. Med. Chem*. **52**, 6163-6167 (2009).

14. Watrous, J., Roach, P., Alexandrov, T., Heath, B.S., Yang, J.Y., Kersten, R.D., van der Voort, M., Pogliano, K., Gross, H., Raaijmakers, J.M., Moore, B.S., Laskin, J., Bandeira, N., Dorrestein, P.C. Mass spectral molecular networking of living microbial colonies. *Proc. Natl. Acad. Sci. U. S. A.* **109**, E1743-1752 (2012).

15. Liu, W.T., Kersten, R.D., Yang, Y.L., Moore, B.S., Dorrestein, P.C. Imaging mass spectrometry and genome mining via short sequence tagging identified the anti-infective

agent arylomycin in Streptomyces roseosporus. *J. Am. Chem. Soc.* **133**, 18010-18013 (2011).

16. Gonzalez, D.J., Okumura, C.Y., Hollands, A., Kersten, R., Akong-Moore, K., Pence, M.A., Malone, C.L., Derieux, J., Moore, B.S., Horswill, A.R., Dixon, J.E., Dorrestein, P.C., Nizet, V. Novel phenol-soluble modulin derivatives in community-associated methicillin-resistant Staphylococcus aureus identified through imaging mass spectrometry. *J. Biol. Chem*. **287**, 13889-13898 (2012).

17. Graupner, K., Scherlach, K., Bretschneider, T., Lackner, G., Roth, M., Gross, H., Hertweck, C. Imaging mass spectrometry and genome mining reveal highly antifungal virulence factor of mushroom soft rot pathogen. *Angew. Chem. Int. Ed. Engl*. **51**, 13173-13177 (2012).

18. Medema M.H., Blin, K., Cimermancic, P., de Jager, V., Zakrzewski, P., Fischbach, M.A., Weber, T., Takano, E., Breitling, R. antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences *Nucleic Acids Res*. **39**, W339-346 (2011).

19. Benson, D. A., Karsch-Mizrachi, I., Clark, K., Lipman, D. J., Ostell, J., Sayers, E. W. GenBank. *Nucleic Acids Res.* **40**, D48-D53 (2012).

20. Smith, C.A., O'Maille, G., Want, E.J., Qin, C., Trauger, S.A., Brandon, T.R., Custodio, D.E., Abagyan, R., Siuzdak, G. METLIN: a metabolite mass spectral database. *Ther. Drug Monit.* **27**, 747-751 (2005).