Title

Molecular Networking and Pattern-Based Genome Mining Improves Discovery of Biosynthetic Gene Clusters and their Products from Salinispora Species

Journal

Cell Chemical Biology, 22(4)

ISSN

2451-9456

Authors

Duncan, Katherine R
Crüsemann, Max
Lechner, Anna
et al.

Publication Date

2015-04-01

DOI

10.1016/j.chembiol.2015.03.010

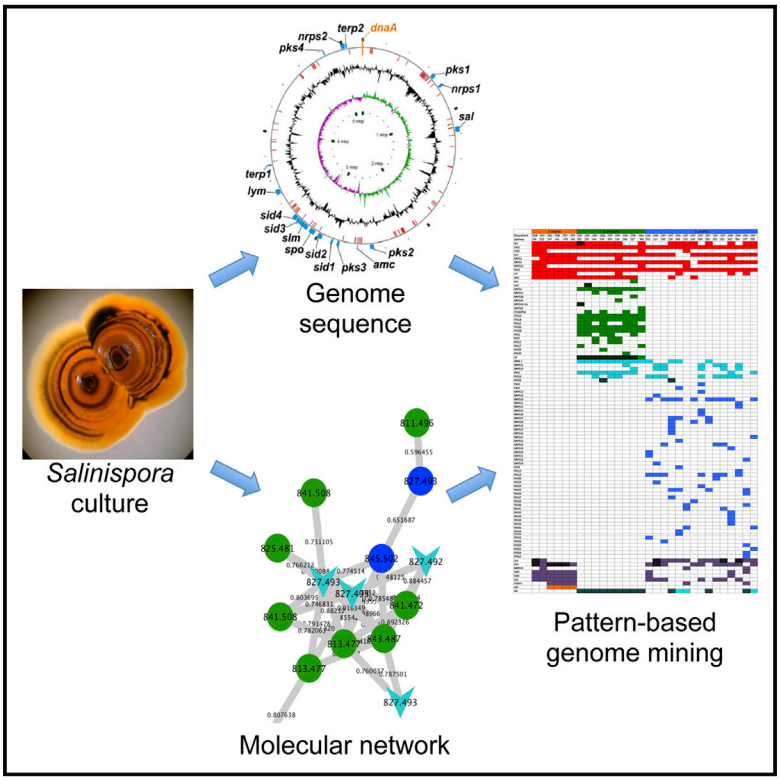
Peer reviewed

Article

Chemistry & Biology

Molecular Networking and Pattern-Based Genome Mining Improves Discovery of Biosynthetic Gene Clusters and their Products from *Salinispora* Species

Graphical Abstract



Authors

Katherine R. Duncan,
Max Crüsemann, ...,
Pieter C. Dorrestein, Paul R. Jensen

Correspondence

pjensen@ucsd.edu (P.R.J.),
bsmoore@ucsd.edu (B.S.M.),
pdorrestein@ucsd.edu (P.C.D.)

In Brief

Duncan et al. use pattern-based genome mining to help bridge the gap between the detection of biosynthetic gene clusters and their products. Coupled with molecular networking, these approaches facilitated the de-replication of known compounds, the detection of new analogs, and the prioritization of compounds for structure elucidation.

Highlights

- Pattern-based genome mining was applied to 35 *Salinispora* strains
- Molecular networking facilitated new compound discovery
- The quinomycin-type depsipeptide retimycin A was characterized

Duncan et al., 2015, Chemistry & Biology *22*, 460–471
April 23, 2015
http://dx.doi.org/10.1016/j.chembiol.2015.03.010

CellPress

Chemistry & Biology
Article

Molecular Networking and Pattern-Based Genome Mining Improves Discovery of Biosynthetic Gene Clusters and their Products from *Salinispora* Species

Katherine R. Duncan,[1,4] Max Crüsemann,[1,4] Anna Lechner,[1,4] Anindita Sarkar,[1] Jie Li,[1] Nadine Ziemert,[1] Mingxun Wang,[2] Nuno Bandeira,[2] Bradley S. Moore,[1,3,*] Pieter C. Dorrestein,[3,*] and Paul R. Jensen[1,*]

[1]Center for Marine Biotechnology & Biomedicine, Scripps Institution of Oceanography, University of California San Diego, La Jolla, CA 92093, USA

[2]Department of Computer Science and Engineering, University of California San Diego, La Jolla, CA 92093, USA

[3]Skaggs School of Pharmacy and Pharmaceutical Sciences, Departments of Pharmacology, Chemistry and Biochemistry, University of California San Diego, La Jolla, CA 92093, USA

[4]Co-first author

*Correspondence: pjensen@ucsd.edu (P.R.J.), bsmoore@ucsd.edu (B.S.M.), pdorrestein@ucsd.edu (P.C.D.)

http://dx.doi.org/10.1016/j.chembiol.2015.03.010

SUMMARY

Genome sequencing has revealed that bacteria contain many more biosynthetic gene clusters than predicted based on the number of secondary metabolites discovered to date. While this biosynthetic reservoir has fostered interest in new tools for natural product discovery, there remains a gap between gene cluster detection and compound discovery. Here we apply molecular networking and the new concept of pattern-based genome mining to 35 *Salinispora* strains, including 30 for which draft genome sequences were either available or obtained for this study. The results provide a method to simultaneously compare large numbers of complex microbial extracts, which facilitated the identification of media components, known compounds and their derivatives, and new compounds that could be prioritized for structure elucidation. These efforts revealed considerable metabolite diversity and led to several molecular family-gene cluster pairings, of which the quinomycin-type depsipeptide retimycin A was characterized and linked to gene cluster NRPS40 using pattern-based bioinformatic approaches.

INTRODUCTION

The analysis of genome sequence data has revealed that even well-studied bacteria can maintain the genetic potential to produce many more secondary or specialized metabolites than discoveries to date would suggest (Bentleyet al., 2002; Cimermancic et al., 2014; Doroghazi et al., 2014; Nett et al., 2009). While this revelation has generated renewed interest in the field of natural product discovery, there remain inefficiencies in the processes by which known compounds are detected and new compounds prioritized for isolation and structure elucidation. Molecular networking is a tandem mass spectrometry (MS/MS)-based computational approach that represents an important advance for the field of natural product research (Nguyen et al., 2013; Winnikoff et al., 2013; Yang et al., 2013). This technique allows for high-throughput multi-strain comparisons and, with the integration of authentic standards, a rapid method to de-replicate (Koehn, 2008) and identify new compounds with known structural scaffolds (Krug and Müller, 2014; Yang et al., 2013). Molecular networking also shows considerable potential as an aid to novel compound discovery, especially when complemented with genome sequence data and recently developed peptidogenomic and glycogenomic methods (Kersten et al., 2011, 2013). Together, these approaches provide a rapid method to create bioinformatic links between parent ions and the pathways responsible for their biosynthesis. Furthermore, by analyzing large numbers of related strains, it becomes possible to address relationships between taxonomy and metabolome content, which can help resolve the ecological significance and evolutionary history of specific functional traits.

Molecular networks are used to organize MS/MS spectra into groups based on similarities in their fragmentation patterns and the expectation that structurally related molecules will yield similar MS/MS spectra. In these networks, MS/MS spectra are represented as nodes, and the similarity between two spectra computed using a modified cosine score (Watrous et al., 2012), which defines the edges connecting two nodes (Bandeira, 2007; Watrous et al., 2012). A series of connected nodes generally indicates structurally related molecules or molecular families (Nguyen et al., 2013). Molecular networking provides a rapid and highly sensitive approach to compare metabolic profiles among strains without arduous data mining (Yang et al., 2013). It has been used for the global visualization of the molecules produced by one (Liu et al., 2014) or a large number of organisms (Nguyen et al., 2013), to discover a new suite of natural products from *Streptomyces coelicolor* (Sidebottom et al., 2013), characterize small molecules dependent on the colibactin pathway in *Escherichia coli* (Vizcaino et al., 2014), define the metabolomic potential of a new environmental taxon (Wilson et al., 2014), and study the chemical basis of microbial interactions on agar surfaces (Watrous et al., 2012). However, molecular networking has not been used to simultaneously assess the pan-metabolome of a large number of closely related bacterial strains.

The marine actinomycete genus *Salinispora* consists of three closely related species (Freel et al., 2013; Maldonado et al., 2005) that share 99% 16S rRNA gene sequence identity (Jensen and Mafnas, 2006). They are the source of a wide range of structurally novel secondary metabolites (Jensen et al., 2015) including salinosporamide A, which has undergone phase I clinical trials for the treatment of cancer (Feling et al., 2003). Genome sequence data support previous observations that some compounds are consistently produced by members of the same *Salinispora* species (Jensen et al., 2007) while also revealing extraordinary levels of pathway diversity and considerable potential for new compound discovery (Penn et al., 2009; Ziemert et al., 2014). Whereas genome mining has been used to target the products of individual *Salinispora* pathways (Udwary et al., 2007), molecular networking provides the opportunity to simultaneously assess metabolite production in large numbers of strains. When coupled with genome sequence data, metabolite and biosynthetic gene cluster (BGC) distributions can be compared in a process we describe here as "pattern-based genome mining," thus expanding on recent efforts to create bioinformatics links between BGCs and their small-molecule products (Doroghazi et al., 2014). Here we applied molecular networking to the analysis of 35 closely related *Salinispora* strains including 30 for which genome sequence data were available. This metabolomic approach complements a previous bioinformatic study (Ziemert et al., 2014) and further resolves the relationships among *Salinispora* species designations and secondary metabolite production (Jensen et al., 2007). A de-replication strategy (Yang et al., 2013) was used to populate the network with previously described *Salinispora* secondary metabolites, and aid in the identification of known compounds and new derivatives. We then searched for "patterns" in an effort to create bioinformatic links between the presence of uncharacterized BGCs and the production of specific compounds. This combined approach, which has the potential to be automated, was then used to select compounds of interest for structure elucidation.

RESULTS

Thirty-five *Salinispora* strains isolated from ten global collection sites were analyzed (Table S1). These strains include a broad representation of the diversity within the three currently recognized species (Freel et al., 2012) and 30 for which draft genome sequences are available, seven of which are new to this study. The fermentation conditions were standardized using the indicator phenol red such that all cultures were extracted upon entry into stationary phase (Figure S1), which occurred on days 9 to 30 depending on strain (Table S1) and is linked to a shift from primary to secondary metabolism (Nieselt et al., 2010). This transition was associated with a pH shift from acidic (yellow) to basic (red), which corresponds to a change from rapid growth and net acetate excretion in the presence of abundant nutrients to slower growth and acetate assimilation following nutrient depletion (Wolfe, 2005). The extracts were analyzed by high-resolution tandem mass spectrometry (HR-MS/MS), the results from biological replicates combined for each strain, and *m/z* values <300 excluded, which resulted in the generation of more than 200,000 MS1 HR-MS spectra over a mass range of 304.175–2485.4 *m/z*.

Molecular Network

Analysis of the MS/MS data led to the identification of 1,137 parent ions, which were visualized as nodes in a molecular network (Figure 1). The node size reflects the number of strains producing each parent ion and varies from strain specific (one strain), which represents the vast majority of nodes, to the most ubiquitous metabolites, which were observed in a maximum of 24 of the 35 strains (Figure 1). When media components are excluded, the most common parent ion was observed in 15 strains (Figure S2). The network was screened against a spectral database generated from authentic standards, which led to the identification of seven compound classes previously described from *Salinispora* spp. (Table 1). The large number of nodes that networked with many of the standards suggests the presence of additional analogs in these compound classes. Examples include the cyclomarin (Renner et al., 1999) and arenicolide (Williams et al., 2007) molecular families, where networking reveals the production of both known compounds and what appear to be new analogs. For example, in addition to cyclomarin A and D, inspection of the parent mass and the MS/MS data suggests the molecular family contains putative demethylated, methylated, and hydrated cyclomarin A analogs (*m/z* 1051.59, 1079.62, and 1083.61, respectively). Similarly, in the arenicolide cluster, putative dehydrogenated, methylated, and hydroxylated arenicolide A congeners were detected (*m/z* 825.48, 841.47, and 843.49, respectively) (Figure 1).

Patterns visualized in the network include a large number of nodes and molecular families that are specific to *Salinispora arenicola*. It is also clear that many of the ions observed from all three species networked with media components (shown as black nodes), suggesting that these molecular families are of low interest in terms of secondary metabolite discovery. The exclusion of metabolites with *m/z* values <300 prevented the detection of some previously identified *Salinispora* metabolites such as salinosporamide K (274.1 *m/z* $[M + Na]^+$) (Eustáquio et al., 2011) and salinipyrone (293.17 *m/z* $[M + H]^+$) (Oh et al., 2008). However, salinosporamide A (314.116 *m/z* $[M + H]^+$) (Feling et al., 2003), which has consistently been reported from *Salinispora tropica* (Jensen et al., 2007), was not observed in the molecular network or the raw experimental data. This may be due to feedback regulation of the biosynthetic pathway (Lechner et al., 2011), the absence of adsorbent resins in the fermentation medium (Tsueng et al., 2008), or the aqueous instability of salinosporamide A (Fenical et al., 2009).

Parent Ion Distributions

Only 22 of the 1,137 parent ions (1.9%) were observed in the medium blank. The majority of ions (87.1%) were not produced across species boundaries, with only 5.8% of the total shared by two and 4.1% shared by all three species (Figure 2). This apparent species specificity can be accounted for by the large number of ions that was observed in only one strain (Figure S2), thus indicating a high level of variability in secondary metabolite production among strains. These ions include a broad range of masses, suggesting there is no correlation between mass and the frequency with which a parent ion was detected. The results

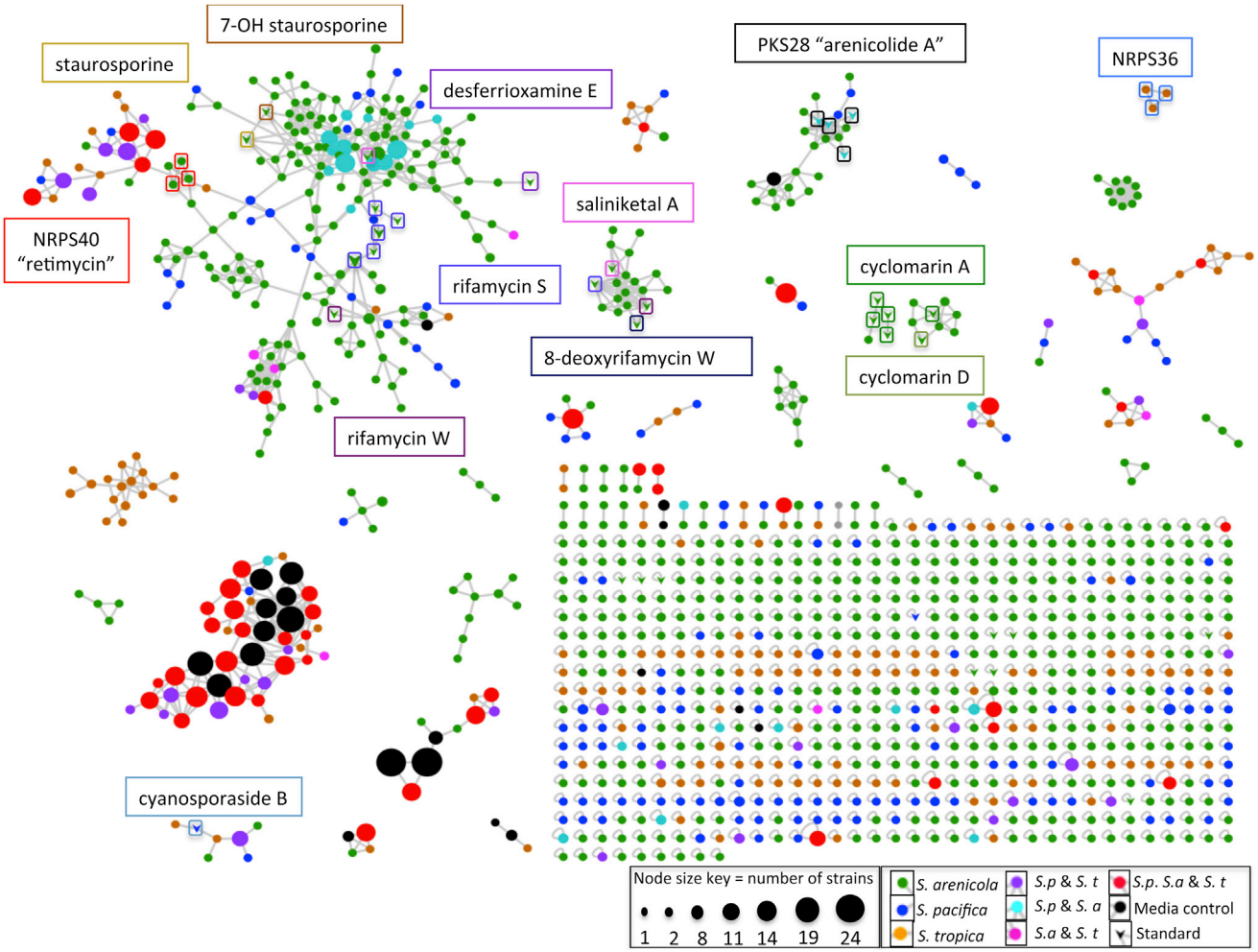


Figure 1. *Salinispora* Molecular Network

Related parent ions are networked based on similarities in MS/MS fragmentation patterns. Arrows indicate ions that matched known *Salinispora* secondary metabolites, with a representative of that compound molecular family named in a similarly colored box. Node size reflects the number of strains that produced each parent ion. Node color reflects the distribution of the parent ion among the three species.

are in agreement with a bioinformatics analysis of 75 *Salinispora* genome sequences, in which more than half of the polyketide synthase (PKS) and non-ribosomal peptide synthetase (NRPS) pathways were only observed in one or two strains (Ziemert et al., 2014). There were no cases where all 35 strains analyzed, or all strains from the same species, produced the same parent ion (Figure S2). The number of unique parent ions was far greater for *S. arenicola*, which averaged 57.0 per strain, relative to *S. tropica* and *Salinispora pacifica*, which averaged 28.0 and 8.7 per strain, respectively. Nodes not associated with edges (ions that are not part of a molecular family and potentially represent unique chemistry) were distributed among the species as follows: *S. arenicola* 371 (62.1%), *S. pacifica* 98 (16.4%), and *S. tropica* 98 (16.4%).

The phylogenetic relationships among the three *Salinispora* spp. are well resolved and reveal the basal position of *S. arenicola* relative to the more recently diverged sister taxa *S. tropica* and *S. pacifica* (Freel et al., 2013). Based on the assumption that horizontal gene transfer has not ameliorated all evidence of vertical inheritance among the genes responsible for secondary metabolite production, it is expected that parent ion similarities would be greater between *S. tropica* and *S. pacifica*, based on their relatively close evolutionary relationship, than between either of these species and *S. arenicola*. A species-level pairwise comparison reveals that *S. tropica* and *S. pacifica* shared 5.6% and 7.9% of the parent ions they produced, respectively, with *S. arenicola*, while *S. tropica* and *S. pacifica* shared 18.1%. Thus, despite the high level of among strain variability in parent ion production, species relationships are reflected in the levels of shared metabolite production.

Linking Compounds to Gene Clusters through Molecular Networking

The BGCs associated with the biosynthesis of ten of the secondary metabolites reported from *Salinispora* spp. have been experimentally linked to their secondary metabolic products. These encode the production of the cyclomarins (*cym*) (Schultz et al., 2008), cyanosporasides (*cya*) (Lane et al., 2013), rifamycins (*rif*)

Table 1. Linking Pathway Distributions to Parent Ion Detection

Biosynthetic Gene Cluster (BGC)			Product							
Name	No. of Strains Detected	Species Detected	Name	No. of Strains Detected (% of Total)	Observed Parent Ion *m/z*	MF	MW	GnPS Score	Expected Parent Ion *m/z*	Reference
PKS28	2	Sa, Sp	Arenicolide A[a]	2 (100)	827.49 $[M + Na]^+$	$C_{45}H_{72}O_{12}$	804.50	0.62–0.76 $[M + Na]^+$	827.492 $[M + Na]^+$	Williams et al. (2007)
cya	17	Sp, St	Cyanosporaside B	5 (29.4)	440.26 $[M + Na]^+$	$C_{21}H_{20}ClNO_6$	417.00	0.81 $[M + Na]^+$	440.088 $[M + Na]^+$	Oh et al. (2006)
cym	1	Sa	Cyclomarin A	1 (100)	1065.60 $[M + Na]^+$	$C_{56}H_{82}N_8O_{11}$	1042.61	0.97 $[M + Na]^+$	1065.60 $[M + Na]^+$	Schultz et al. (2008)
des	21	Sp	Desferrioxamine E	1 (4.7)	601.21 $[M + H]^+$	$C_{27}H_{48}N_6O_9$	600.35	0.54 $[M + H]^+$	601.36 $[M + H]^+$	Roberts et al. (2012)
rif	9	Sa	Rifamycin S	8 (88.9)	718.29 $[M + Na]^+$	$C_{37}H_{45}NO_{12}$	695.30	0.67 $[M + Na]^+$	718.29 $[M + Na]^+$	Kim et al. (2006)
sta	15	Sa, Sp, St	Staurosporine	11 (73.3)	467.22 $[M + H]^+$	$C_{28}H_{26}N_4O_3$	466.20	0.68 $[M + H]^+$	467.208 $[M + H]^+$	Freel et al. (2011)
lom	16	Sp, St	Lomaiviticin C	5 (31.3)	670.27 $[M + 2H]^{2+}$	$C_{68}H_{82}N_4O_{24}$	1339.39	NA	670.277 $[M + 2H]^{2+}$	He et al. (2001)
slm	13	Sp, St	Salinilactam[b]	0 (0)	ND	NA	NA	NA	NA	Udwary et al. (2007)
sal	13	Sa, Sp, St	Salinosporamides	0 (0)	ND	NA	NA	NA	NA	Feling et al. (2003)
spo	4	St	Sporolides[b]	0 (0)	ND	NA	NA	NA	NA	Buchanan et al. (2005)
lym	28	Sa, Sp, St	Lymphostins	0 (0)	ND	NA	NA	NA	NA	Aotani et al. (1997)
rtm	1	Sa	Retimycin A[a]	1 (100)	1185.44 $[M + H]^+$	$C_{54}H_{68}N_{10}O_{15}S_2$	1184.33	NA	1185.439 $[M + H]^+$	This study

Biosynthetic pathways associated with 13 structurally characterized secondary metabolites were identified in the 30 *Salinispora* genome sequences analyzed. Products from eight of these pathways (including *rtm*, which is new to this study) were identified using HR-MS/MS data in comparison with authentic standards. All MS data, including raw data that did not appear in the network, were screened for parent ions associated with each compound class. Data are provided for only one representative of each structure class. GnPS, Global Natural Products Social Molecular Networking (http://gnps.ucsd.edu/ProteoSAFe/static/gnps-splash.jsp); MF, molecular formula; MW, molecular weight; NA, not applicable; ND, not detected; Sa, *Salinispora arenicola*; Sp, *Salinispora pacifica*; St, *Salinispora tropica*.

[a]Links between BGC and product are bioinformatic-based.

[b]Standards not available.

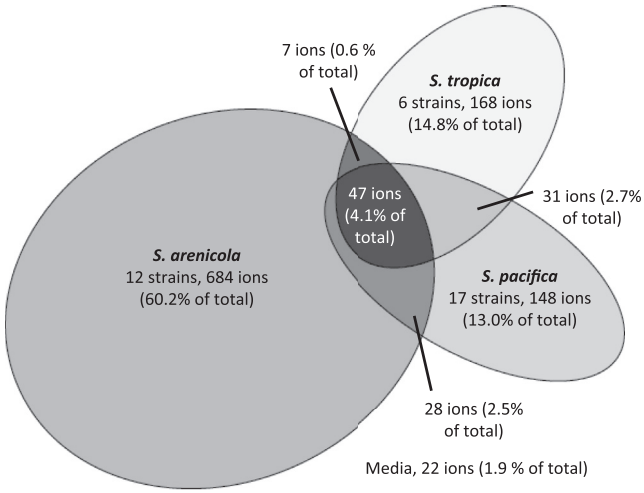


Figure 2. Parent Ion Distribution

A proportional Euler diagram reveals that most ions were species specific. Parent ions were considered shared when produced by at least one strain from two different species. Twenty-two ions (1.9% of total) assigned to the medium were not included in the analysis.

(including saliniketal) (Wilson et al., 2010; Ziemert et al., 2014), lomaiviticins (*lom*) (Kersten et al., 2013), desferrioxamines (*des*) (Roberts et al., 2012), staurosporines (*sta*) (Onaka et al., 2002), salinilactam (*slm*) (Udwary et al., 2007), salinosporamides (*sal*) (Eustáquio et al., 2009, 2011), sporolides (*spo*) (McGlinchey et al., 2008), and lymphostin (*lym*) (Miyanaga et al., 2011). Furthermore, the *arn* cluster has been bioinformatically linked to arenimycin biosynthesis (Asolkar et al., 2010) via glycogenomics (Kersten et al., 2013) while the *rtm* cluster (Table S2) has been bioinformatically linked to retimycin A biosynthesis (this study) using peptidogenomics (Kersten et al., 2011). Overall, products associated with eight *Salinispora* BGCs were detected among the extracts analyzed (Table 1).

Molecular networking coupled with genome sequence data provides a rapid method to assess the relationships between the presence of a BCG and the detection of its secondary metabolic products. We refer to these correlative analyses as pattern-based genome mining, whereby the detection of a parent ion within a molecular family is used as a proxy for the expression of the associated BCG (Figure 3). In some cases, pattern-based genome mining revealed a perfect correlation between BCGs and products. This was observed for the arenicolides and cyclomarins, which were detected from all three of the strains that possessed the respective clusters (Table 1). In most cases, however, the correlations were less perfect. For example, rifamycins were detected in eight of the nine strains that possessed the BCG, whereas the desferrioxamines were only observed in 1 of 21 strains. In the latter case it was surprising that even one strain produced these iron-chelating compounds, as the medium was not specifically designed to be iron limited, which is known to support production (Roberts et al., 2012). While it remains unclear how much of the variability observed can be linked to the nuances associated with fermentation conditions, extraction protocols, and MS, pattern-based genome mining holds promise as a method to generate links between the presence of BCGs and the small molecules they ultimately produce. In total, products were detected in 34 of 140 cases (24%) in which the associated BGCs were detected. This suggests that many of the BGCs were not expressed under the culture conditions used or that, if expressed, the products were either not extracted or went undetected in the MS analyses.

Linking Compounds to Uncharacterized Gene Clusters

It was possible to link an ion that matched the previously described metabolite arenicolide A (Williams et al., 2007) to an uncharacterized BGC (PKS28) based on a number of lines of evidence. First, pattern-based analysis revealed that the two strains producing the arenicolide A ion (827.492 *m/z* [M + Na]) were the only strains in which PKS28 was observed (*S. arenicola* strain CNQ-748 and *S. pacifica* strain CNT-138) (Figure 3). While incomplete genome assembly prevented a precise interpretation of this gene cluster, the KS sequences in PKS28 show a high level of sequence identity (92%) to KS sequences previously linked to arenicolide A production (Edlund et al., 2011). The molecular network places arenicolide A within a larger molecular family of structurally related analogs (Figure 4), suggesting that additional diversity remains to be discovered in this class of compounds. This is the first evidence of arenicolide A production by *S. pacifica*. In a second example of pattern-based genome mining, the pathway NRPS40 was identified as unique to strain CNT-005 (Ziemert et al., 2014) (Figure 3). In an effort to search for the products of this pathway, a series of nodes that were similarly unique to strain CNT-005 was explored in more detail (Figure 5). MS/MS analysis of the 1171.42 and 1185.43 *m/z* parent ions revealed peptide mass shifts while the characteristic UV profiles suggested that the compounds may be related to the quinomycins, a group of highly cytotoxic and antibiotic dimeric depsipeptides (Dawson et al., 2007; Zolova et al., 2010). Analysis of the MS/MS spectra revealed alanine, several dehydrated threonine residues, and unassigned mass shifts, suggesting modified amino acid residues (Figure S3). An analysis of the NRPS gene clusters in strain CNT-005 led to NRPS40 as a candidate for the biogenesis of the 1171.42 and 1185.43 *m/z* parent ions.

Isolation and Characterization of Retimycin A

A detailed analysis of NRPS40 revealed considerable homology to the BGC responsible for the production of the quinomycin-like compound SW-163 (Watanabe et al., 2009), including genes for the production of hydroxyquinaldic acid (HQA) and the cyclopropane-containing norcoronamic acid (NCA). However, the adenylation domain specificity of the second NRPS module differs from SW-163 and was not predictable by bioinformatic tools. In addition, NRPS40 includes a CYP450 oxygenase, which collectively suggests the product differs from that of the previously characterized compound SW-163. Large-scale fermentation followed by the isolation of the 1171.42 *m/z* metabolite and subsequent 2D nuclear magnetic resonance (NMR) characterization (Figures S4 and S5; Table S3) confirmed the presence of HQA and NCA units. Assignment of the unknown amino acid revealed a threonine moiety, a previously unprecedented residue in the quinomycin family. The characterized members of this family contain D-serine (echinomycins), D-cysteine (thiocoralines), or D-diaminobutyric acid (quinaldopeptin) at

this position (Fernández et al., 2014). Marfey analysis of the hydrolyzed compound showed the stereochemistry of the threonine residue to be D-*allo*-Thr, which corresponds to the presence of an epimerase domain in the first module of RtmO. Assignment of the β protons of the cysteine residues revealed a thioacetal moiety, a common motif in the quinomycin family. MS/MS analysis showed a central fragment loss of *m/z* 63.99, which corresponds to a molecular formula of CH_4SO (Figure S3). Taken together with the observation of a methyl singlet that shows correlation to the tertiary carbon of the thioacetal, we concluded that the central moiety is a methylated thioacetal that is oxidized at the rearranged sulfur, which is a novel feature in this family. Interestingly, in a recently described quinomycin, the non-rearranged sulfur was shown to be oxidized (Lim et al., 2014). The 1171.42 *m/z* compound was named "retimycin A," after the Latin word "reticulum" meaning network. It is the first fully characterized novel natural product discovered by molecular networking. For the analog with *m/z* 1185.43, preliminary NMR data suggest that this molecule is not simply a methylated version of retimycin A, but lacks symmetry in the peptide backbone. A lack of material prevented the elucidation of this structure, which warrants further investigation. Bioactivity results revealed comparable cytotoxicity (half-maximal inhibitory concentration <0.076 μg/ml) against an HCT-116 cell line for retimycin A and the structurally related compound echinomycin.

DISCUSSION

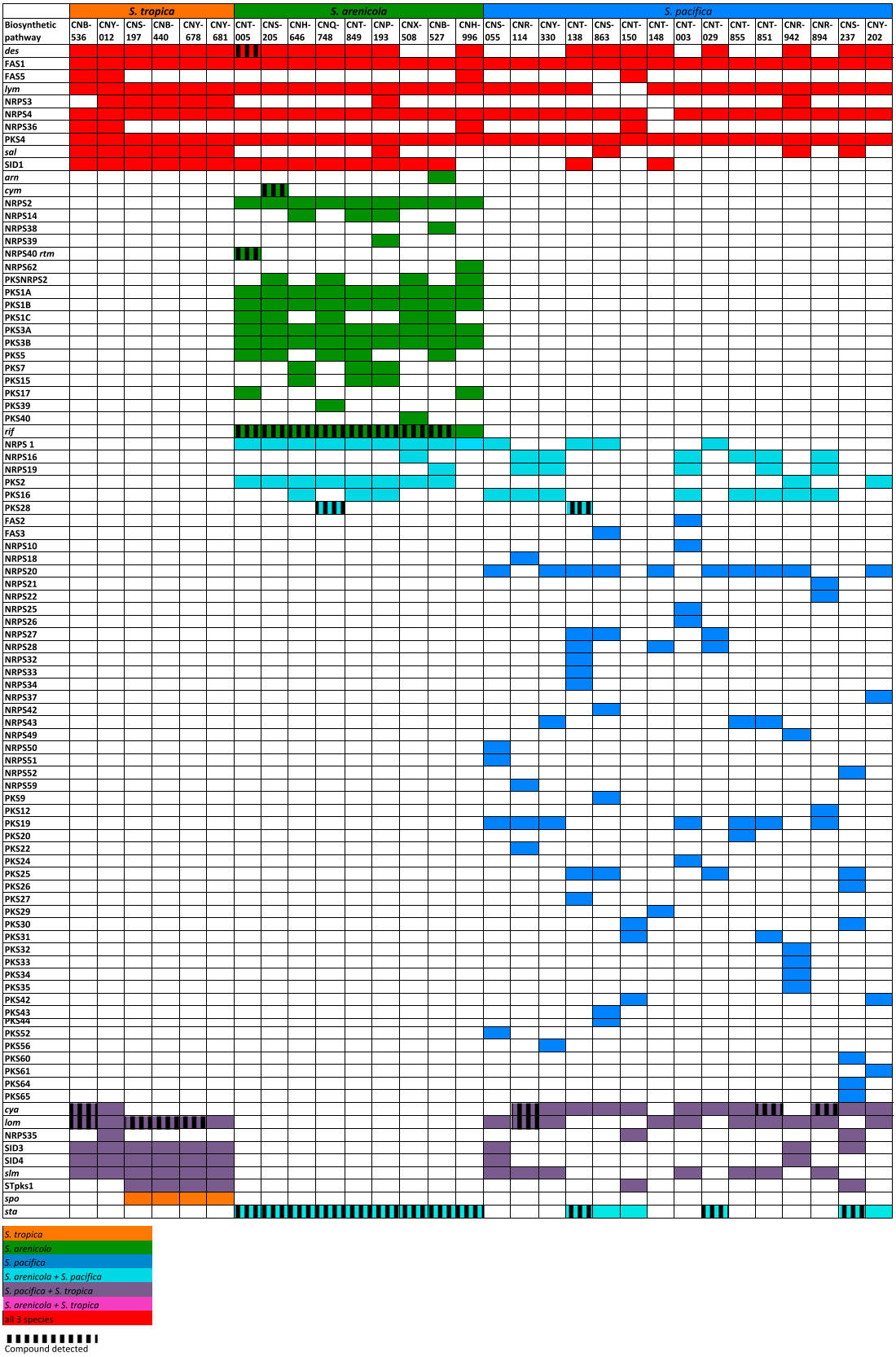
The field of natural products chemistry was invigorated by the observation that even well-studied bacteria can maintain the genetic potential to produce many new secondary or specialized metabolites (Bentley et al., 2002). Spearheading this renewed interest is the application of genome mining techniques (Bachmann et al., 2014) and focused efforts to develop new discovery platforms including heterologous expression (Bonet et al., 2014; Yamanaka et al., 2014), the activation of silent gene clusters (Seyedsayamdost, 2014), peptidogenomics (Kersten et al., 2011), glycogenomics (Kersten et al., 2013), and proteomics (Chen et al., 2013), all of which have helped to realize this potential. While there have been efforts to automate the process of genome mining (Medema et al., 2014; Mohimani et al., 2014; Zhang et al., 2014), there remain major gaps between the detection of BGCs in genome sequence data and the identification of the metabolites they produce (Jensen et al., 2014). Bridging this gap presents considerable challenges because of our poor understanding of the relationships between BGC distributions, expression, and the successful biogenesis and detection of a secondary metabolite. Here we present pattern-based genome mining as a highly sensitive and scalable approach to link molecules detected by MS to the BGCs responsible for their biogenesis. When applied to the marine actinomycete genus *Salinispora*, the results provided a rapid method to recognize previously described secondary metabolites, generate bioinformatic links between unidentified parent ions and their putative BGCs, and prioritize compounds for isolation and structure elucidation.

The global *Salinispora* molecular network revealed high levels of species- and strain-specific production, with only 15% of the parent ions crossing species boundaries. This variability is not entirely due to differences in genome content, as the products of only 34 of 140 (24%) of the characterized BGCs were detected. Variables such as culture conditions likely contribute to this discrepancy, which was similarly reported in the analysis of 830 actinobacterial metabolomes (Doroghazi et al., 2014). Considering the *des* pathway, it is not surprising that the iron-chelating desferrioxamines were only detected in 1 of 21 cases, given that the iron-replete growth medium was not designed to elicit production. The variables that affect the expression of other BGCs are less clear. Future transcriptome analyses will help define these variables and distinguish between "silent" BGCs and those for which the products are produced but remain undetected. An extended molecular network including different growth conditions and extraction methods is expected to provide better consistency between predicted compounds and detected ions. It is also anticipated that subjecting larger numbers of organisms to pattern-based genome mining will improve the correlative power of the approach.

It was interesting that *S. arenicola* produced the largest number of unique parent ions per strain. This was surprising, considering that in a prior study *S. pacifica* was observed to maintain greater PKS and NRPS diversity (Ziemert et al., 2014). Nonetheless, this observation is supported by an independent liquid chromatography (LC)-MS analysis in which twice the number of compounds was detected from *S. arenicola* (Bose et al., 2014). A more comprehensive comparison of the BGCs and their levels of expression in these two species will provide a better understanding of the relationships between BGC distributions and specialized metabolite production.

As has been shown previously with other bacteria (Yang et al., 2013), seeding the *Salinispora* molecular network with previously identified secondary metabolites made it possible to rapidly identify known compounds and related molecular families among the 1,137 nodes present in the network. It was also possible to identify common media components and what appear to be new derivatives of previously described molecules such as the cyclomarins and arenicolides, which in many cases (e.g. methylation, hydroxylation) could be readily identified by MS. This study benefited from the analysis of a large number of closely related strains and data from previously described *Salinispora* secondary metabolites. These methods will become increasingly useful for the analysis of diverse collections of bacteria as shared knowledge databases, such as that available through the Global Natural Products Social Molecular Networking site (http://gnps.ucsd.edu/ProteoSAFe/static/gnps-splash.jsp), provide infrastructure to easily share MS/MS data with the larger community.

Using pattern-based genome mining in combination with peptidogenomics, NRPS40 was linked to a 1171.423 *m/z* parent ion, which was targeted for isolation and subsequently identified as retimycin A, a new quinomycin-like depsipeptide in the thiocoraline family. The detection of the retimycin gene cluster (*rtm*) in only 1 of the 30 genome sequences indicates the value of studying large numbers of closely related strains. Interestingly, the related metabolite thiocoraline has been reported from two *Micromonospora* strains isolated from marine invertebrates (Lombó et al., 2006), suggesting that this pathway has been exchanged horizontally between these



(legend on next page)

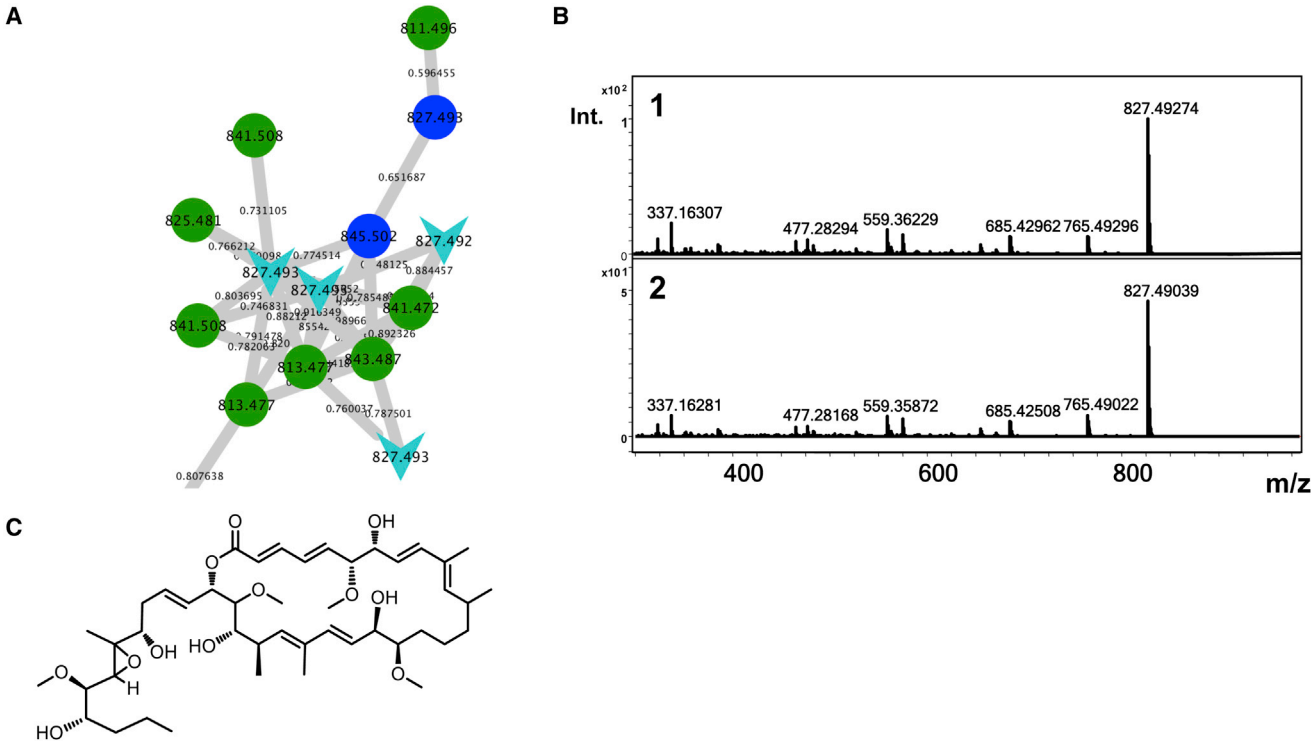


Figure 4. Arenicolide A Production

(A) An arenicolide A standard (expected 827.492 *m/z*) networked to 827.492 and 827.493 *m/z* ions (light blue arrowheads) produced by strains *S. pacifica* CNT-138 and *S. arenicola* CNQ-748. Structurally related analogs are produced by CNT-138 (dark blue nodes) and CNQ-748 (green nodes).

(B) MS/MS fragmentation pattern of the 827.492 *m/z* parent ion (1) in comparison with the arenicolide A standard (2).

(C) Structure of arenicolide A.

two closely related actinomycete genera. Retimycin A has now been added to the *Salinispora* MS/MS spectral library, thus providing a useful reference to better assess the production of this and related compounds among other strains. This growing spectral library provides unique opportunities to obtain a more global view of the *Salinispora* secondary metabolome.

SIGNIFICANCE

This study represents the first application of molecular networking to a large collection of environmental bacteria. It introduces the concept of pattern-based genome mining, a scalable MS method with the potential to be automated. The molecular network made it possible to simultaneously visualize the molecular composition of organic extracts generated from 35 closely related strains belonging to the marine actinomycete genus *Salinispora*. Populating the network with standards facilitated the identification of known compounds and their derivatives and compounds of high priority for isolation and structure elucidation. A majority of parent ions were species or strain specific, resulting in a high degree of secondary metabolite diversity. When complemented with genome sequence data, pattern-based genome mining revealed non-perfect correlations between gene cluster distributions and the detection of the associated products, suggesting that gene expression may be highly variable among strains. Pattern-based genome mining was used to identify NRPS40 as a candidate for natural product discovery while peptidogenomics was used to provide a bioinformatics link between this BGC and a parent ion observed in the molecular network. This compound was structurally characterized and named retimycin A, a new member of the quinomycin family of depsipeptide antitumor antibiotics. The growing MS/MS database of *Salinispora* natural products will continue to improve the effectiveness by which molecular networking can be used to rationalize chemical space and improve the effectiveness by which natural products are discovered from this genus.

Figure 3. Pattern-Based Genome Mining in 30 *Salinispora* Strains

Gene clusters, listed on the left, follow previous nomenclature (Ziemert et al., 2014), and include the addition of the staurosporine (*sta*) and desferrioxamine (*des*) pathways. Colored boxes indicate the presence of a biosynthetic gene cluster; vertical lines (shading) indicate the HR-MS detection of the products of that gene cluster. Colors indicate the distribution of the gene cluster across the three species: red (all three species), light blue (*S. arenicola* and *S. pacifica*), purple (*S. tropica* and *S. pacifica*), blue (*S. pacifica* only), orange (*S. tropica* only), and green (*S. arenicola* only). Three *S. arenicola* and two *S. pacifica* strains were excluded from the figure because genome sequences were not available.

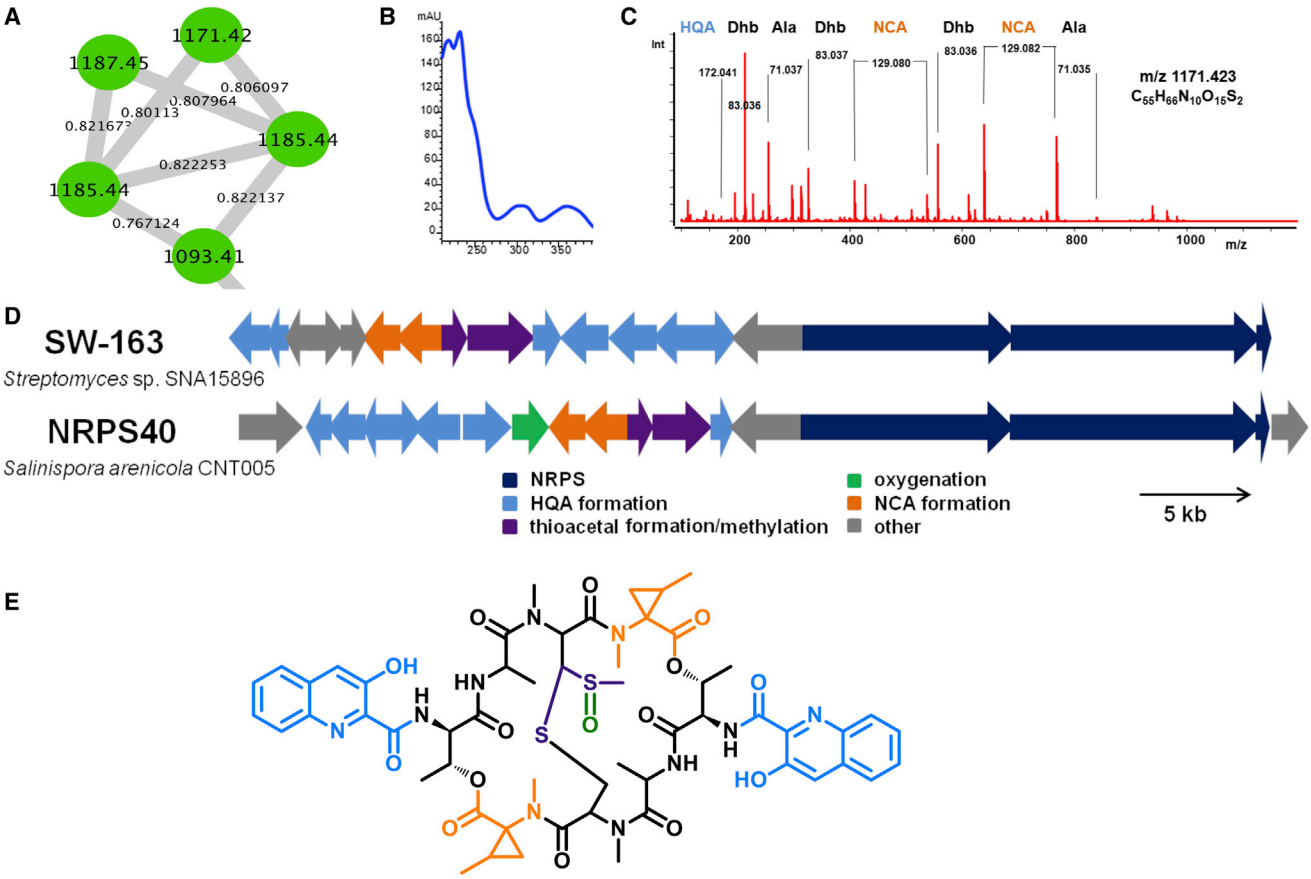


Figure 5. Identification of the Novel Non-ribosomal Peptide Retimycin A

(A) Analysis of the molecular network revealed a cluster of four parent ions produced only by *S. arenicola* strain CNT-005.

(B) UV spectra associated with parent ion 1171.423 *m/z*.

(C) MS/MS analysis of the 1171.423 *m/z* parent ion revealed amino acid shifts that corresponded to alanine, dehydrated threonine, and two unknown residues later assigned as HQA and NCA.

(D) Bioinformatic analysis of NRPS40 (*rtm*) in comparison with the related pathway responsible for the production of SW-163, a quinomycin-like depsipeptide.

(E) Chemical structure of retimycin A. The colors of the building blocks correspond to the colors of their respective biosynthetic genes in Figure 5D. See also Figures S3–S5 and Table S2.

EXPERIMENTAL PROCEDURES

Strain Culture Conditions and Extraction

Salinispora strains were selected based on 16S rRNA phylotype, isolation location, and the availability of genome sequence data (Table S1). Pre-cultures were grown in medium A1M1 (5 g/l soluble starch [Affymetrix], 2 g/l peptone [Fischer Scientific], 2 g/l yeast extract [Affymetrix], 22 g/l instant ocean [Marineland], 100 ml deionized water, adjusted to pH 6.5 before autoclaving at 121°C for 45 min) for 7–9 days, after which 5 ml was inoculated in duplicate into 100 ml A1M1 with 100 µl of 10 mg/ml filter-sterilized phenol red solution (Sigma). Fermentations were performed in 500-ml Erlenmeyer flasks at 28°C and shaking at 160 rpm. Stainless steel springs were added to reduce cell clumping. The cultures were extracted with an equal volume of ethyl acetate (Fischer Scientific) when the indicator changed from yellow to red (corresponding to a pH of 8.0), and the ethyl acetate layers were collected, dried in vacuo, and stored at −20°C.

Mass Spectral Data Acquisition

Extracts were dissolved in MeOH at the final concentration of 0.1 mg/ml and injected onto a Phenomenex Kinetex C18 reversed-phase high-performance LC (HPLC) column (2.6 mm, 100 × 4.6 mm). Samples were analyzed using an Agilent 6530 Accurate-Mass Q-TOF spectrometer coupled to an Agilent 1260 LC system under the following LC conditions with 0.1% trifluoroacetic acid (TFA): 1–5 min (10% acetonitrile [MeCN] in H_2O), 5–26 min (10–100% MeCN), 26–30 min (100% MeCN). The divert valve was set to waste for the first 5 min. Q-TOF MS settings during the LC gradient were as follows: positive ion mode mass range 300–2,500 *m/z*, MS scan rate 1/s, MS/MS scan rate 5/s, fixed collision energy 20 eV; source gas temperature 300°C, gas flow 11 l/min, nebulizer 45 psig; scan source parameters: VCap 3000, fragmentor 100, skimmer1 65, octopoleRFPeak 750. The MS was auto-tuned using 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). The spectral data generated from the biological replicates were combined for each strain prior to the network analyses.

Crude Extract and Standard Compound Analysis

HR-MS/MS fragmentation data were generated for the known *Salinispora* compounds listed in Table 1 under the acquisition conditions described above. Ions were selected via peak extraction using MassHunter software and the data converted to mzXML format using the Trans-Proteomic pipeline (Institute for Systems Biology) (Keller et al., 2005). The data were uploaded to the molecular networking server as a *Salinispora* standards library. All HR-MS/MS raw data files were searched for high-resolution masses that matched

the 12 classes of compounds previously reported from *Salinispora* species (Table 1) including those for which standards were not available. Methods include (1) the Mass Hunter ion extraction function, (2) searching for HR masses in raw molecular networking files, and (3) searching for HR masses in the Cytoscape file. Searches were made for expected parent ions corresponding to multiple ions or analogs within each compound class.

Molecular Networking

The MS/MS data of 35 *Salinispora* strains were converted from MassHunter data files (.d) to mzXML file format using the Trans-Proteomic pipeline (Deutsch et al., 2010) and clustered using structure-independent spectral alignment (MS-Cluster) (Frank et al., 2007; Guthals et al., 2012; Keller et al., 2005). For network visualization, these were imported into Cytoscape 3.0 (Guthals et al., 2012). Each node corresponds to a consensus spectrum and each edge represents a significant pairwise alignment. Computationally, spectra were converted into unit vectors in *n*-dimensional space; pairs of vectors were compared with a dot product calculation, which includes the cosine of the angle between the two vectors, referred to as the cosine similarity score. Identical spectra were combined into consensus spectra that have a minimum of six ions that match. Cosine similarity scores range from 0 to 1, where identical spectra have a cosine score of 1. Two nodes are required to be in the top ten cosine scores (K parameter) in both directions for an edge to connect them in Cytoscape. The pairs with a cosine score higher than 0.95 were combined into consensus spectra. The algorithm parameters include mass tolerance for fragment peaks (0.3 Da), parent mass tolerance (2.0 Da), a minimum number of matched peaks per spectral alignment (6), a maximum component size of 1, and a minimum cosine score of 0.5. This latter value was selected empirically to eliminate the clustering of different compound classes into the same molecular family (i.e., false positives). Cytoscape was used to visually display the data as a network of nodes and edges (Cline et al., 2007) and organized with the edge-weighted force-directed layout plugin.

Chemical Isolation and Structural Characterization

For the isolation of retimycin A, the fermentation of strain CNT-005 was scaled up to a total volume of 7 × 1 l in medium A1M1 in 2.8-l Fernbach flasks. After 7 days of cultivation, a 1:1 mixture of XAD7HP:XAD16 resin (Amberlite) was added and, after 2 hr, collected by filtration and eluted 2× with acetone. The extract was concentrated under vacuum and separated via preparative HPLC (Agilent Prostar, Synergi-10u Hydro-RP, 250 × 21.2 mm [Phenomenex]) using a gradient from 30% to 95% acetonitrile containing 0.1% TFA over 30 min (15 ml/min). A fraction eluting from 23 to 26 min was collected, dried, and subjected to semi-preparative HPLC purification (Agilent 1200, Luna 5u C18, 100A, 250 × 10 mm [Phenomenex]) with isocratic 55% acetonitrile (0.1% TFA, 2.5 ml/min) to yield pure retimycin (0.5 mg, t_R = 36 min). Subsequent NMR characterization (^{1}H, ^{1}H-^{1}H COSY, HSQC, HMBC, TOCSY) was carried out on a Bruker Avance spectrometer (600 MHz).

Marfey Analysis of Retimycin

Retimycin (200 μg) was hydrolyzed (6 M HCl, 160°C, 5 min) and dried under nitrogen. The resulting solid was re-dissolved in 1 M sodium bicarbonate (200 μl) followed by addition of 1 ml of 1-fluoro-2,4-dinitrophenyl-D-alanine amide (D-FDAA) in acetone (1.5 mg/ml). The reaction was stirred at 50°C for 1 hr, quenched with 1 M HCl (200 μl), and dried under nitrogen. The resulting solid was re-dissolved in 1:1 H_2O/CH_3CN (200 μl) and filtered. The resulting solution was analyzed (20 μl) by LC-MS (positive mode) on a Luna C_{18} column, 250 × 4.6 mm, 5u (Phenomenex) with a gradient from 0% CH_3CN to 40% CH_3CN in H_2O over 95 min (0.4 ml/min). The mass for Thr-FDAA (*m/z* 372.1) was extracted and determined to elute at 53.28 min. This was compared with the retention times for all four threonine isomers, derivatized with D-FDAA (D-Thr: 52.90 min; D-*allo*-Thr: 53.20 min; L-*allo*-Thr: 56.70 min; L-Thr: 61.50 min). The optical rotation was determined on a Jasco P200 polarimeter (c = 0.2, $CHCl_3$).

SUPPLEMENTAL INFORMATION

Supplemental Information includes five figures and three tables and can be found with this article online at http://dx.doi.org/10.1016/j.chembiol.2015.03.010.

AUTHOR CONTRIBUTIONS

K.R.D., A.L., and P.R.J. designed the study. K.R.D. and A.L. performed and analyzed the molecular networking experiments. K.R.D. and P.R.J. wrote the manuscript. M.C., B.S.M., and J.L. isolated and solved the structure of retimycin. A.S. performed the fermentation studies. N.Z. and A.L. analyzed the gene clusters. M.W. and P.C.D. assisted with the networking analyses.

ACKNOWLEDGMENTS

This work was funded by the NIH under grants RO1GM085770 (to P.R.J. and B.S.M.), U19TW007401 (to P.R.J.), and R01GM097509 (to B.S.M., N.B., and P.C.D.). M.C. was funded by a DFG postdoctoral fellowship. Genome sequencing was conducted by the U.S. Department of Energy Joint Genome Institute and supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. The authors thank W. Fenical for supplying previously identified *Salinispora* secondary metabolites as standards used to seed the molecular network. N. Millán-Aguiñaga is acknowledged for assistance with strain identification. We thank B.M. Duggan for assistance in NMR measurements and J.C. Busch for HCT cytotoxicity testing. D.A. Phillips is acknowledged for graphic design work on Figure 2.

Received: December 11, 2014
Revised: February 26, 2015
Accepted: March 10, 2015
Published: April 9, 2015

REFERENCES

Aotani, Y., Nagata, H., and Yoshida, M. (1997). Lymphostin (LK6-A), a novel immunosuppressant from *Streptomyces* sp. KY11783: structural elucidation. J. Antibiot. *50*, 543–545.

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

Bachmann, B.O., Van Lanen, S.G., and Baltz, R.H. (2014). Microbial genome mining for accelerated natural products discovery: is a renaissance in the making? J. Ind. Microbiol. Biotechnol. *41*, 175–184.

Bandeira, N. (2007). Spectral networks: a new approach to de novo discovery of protein sequences and posttranslational modifications. BioTechniques *42*, 687–695.

Bentley, S.D., Chater, K.F., Cerdeno-Tarraga, A.M., Challis, G.L., Thomson, N.R., James, K.D., Harris, D.E., Quail, M.A., Kieser, H., Harper, D., et al. (2002). Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2). Nature *417*, 141–147.

Bonet, B., Teufel, R., Crüsemann, M., Ziemert, N., and Moore, B.S. (2014). Direct capture and heterologous expression of *Salinispora* natural product genes for the biosynthesis of enterocin. J. Nat. Prod. http://dx.doi.org/10.1021/np500664q.

Bose, U., Hodson, M.P., Shaw, P.N., Fuerst, J.A., and Hewavitharana, A.K. (2014). Two peptides, cycloaspeptide A and nazumamide A from a sponge associated marine actinobacterium *Salinispora* sp. Nat. Prod. Comm. *9*, 545–546.

Buchanan, G.O., Williams, P.G., Feling, R.H., Kauffman, C.A., Jensen, P.R., and Fenical, W. (2005). Sporolides A and B: structurally unprecedented halogenated macrolides from the marine actinomycete *Salinispora tropica*. Org. Lett. *7*, 2731–2734.

Chen, Y., Unger, M., Ntai, I., McClure, R.A., Albright, J.C., Thomson, R.J., and Kelleher, N.L. (2013). Gobichelin A and B: mixed-ligand siderophores discovered using proteomics. MedChemComm *4*, 233–238.

Cimermancic, P., Medema, M.H., Claesen, J., Kurita, K., Brown, L.C.W., Mavrommatis, K., Pati, A., Godfrey, P.A., Koehrsen, M., and Clardy, J. (2014). Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters. Cell *158*, 412–421.

Cline, M.S., Smoot, M., Cerami, E., Kuchinsky, A., Landys, N., Workman, C., Christmas, R., Avila-Campilo, I., Creech, M., and Gross, B. (2007). Integration of biological networks and gene expression data using Cytoscape. Nat. Protoc. *2*, 2366–2382.

Dawson, S., Malkinson, J.P., Paumier, D., and Searcey, M. (2007). Bisintercalator natural products with potential therapeutic applications: isolation, structure determination, synthetic and biological studies. Nat. Prod. Rep. *24*, 109–126.

Deutsch, E.W., Mendoza, L., Shteynberg, D., Farrah, T., Lam, H., Tasman, N., Sun, Z., Nilsson, E., Pratt, B., and Prazen, B. (2010). A guided tour of the Trans-Proteomic Pipeline. Proteomics *10*, 1150–1159.

Doroghazi, J.R., Albright, J.C., Goering, A.W., Ju, K.-S., Haines, R.R., Tchalukov, K.A., Labeda, D.P., Kelleher, N.L., and Metcalf, W.W. (2014). A roadmap for natural product discovery based on large-scale genomics and metabolomics. Nat. Chem. Biol. *10*, 963–968.

Edlund, A., Loesgen, S., Fenical, W., and Jensen, P.R. (2011). Geographic distribution of secondary metabolite genes in the marine actinomycete *Salinispora arenicola*. Appl. Environ. Microbiol. *77*, 5916–5925.

Eustáquio, A.S., McGlinchey, R.P., Liu, Y., Hazzard, C., Beer, L.L., Florova, G., Alhamadsheh, M.M., Lechner, A., Kale, A.J., Kobayashi, Y., et al. (2009). Biosynthesis of the salinosporamide A polyketide synthase substrate chloroethylmalonyl-coenzyme A from S-adenosyl-L-methionine. Proc. Natl. Acad. Sci. USA *106*, 12295–12300.

Eustáquio, A.S., Nam, S.-J., Penn, K., Lechner, A., Wilson, M.C., Fenical, W., Jensen, P.R., and Moore, B.S. (2011). The discovery of salinosporamide K from the marine bacterium "*Salinispora pacifica*" by genome mining gives insight into pathway evolution. ChemBioChem *12*, 61–64.

Feling, R.H., Buchanan, G.O., Mincer, T.J., Kauffman, C.A., Jensen, P.R., and Fenical, W. (2003). Salinosporamide A: a highly cytotoxic proteasome inhibitor from a novel microbial source, a marine bacterium of the new genus *Salinispora*. Angew. Chem. Int. Ed. Engl. *42*, 355–357.

Fenical, W., Jensen, P.R., Palladino, M.A., Lam, K.S., Lloyd, G.K., and Potts, B.C. (2009). Discovery and development of the anticancer agent salinosporamide A (NPI-0052). Bioorg. Med. Chem. *17*, 2175–2180.

Fernández, J., Marín, L., Álvarez-Alonso, R., Redondo, S., Carvajal, J., Villamizar, G., Villar, C.J., and Lombó, F. (2014). Biosynthetic modularity rules in the bisintercalator family of antitumor compounds. Mar. Drugs *12*, 2668–2699.

Frank, A.M., Bandeira, N., Shen, Z., Tanner, S., Briggs, S.P., Smith, R.D., and Pevzner, P.A. (2007). Clustering millions of tandem mass spectra. J. Proteome Res. *7*, 113–122.

Freel, K.C., Nam, S.-J., Fenical, W., and Jensen, P.R. (2011). Secondary metabolite gene evolution in three closely related marine actinomycete species. Appl. Environ. Microbiol. *77*, 7261–7270.

Freel, K.C., Edlund, A., and Jensen, P.R. (2012). Microdiversity and evidence for high dispersal rates in the marine actinomycete '*Salinispora pacifica*'. Environ. Microbiol. Rep. *14*, 480–493.

Freel, K.C., Millan-Aguinaga, N., and Jensen, P.R. (2013). Multilocus sequence typing reveals evidence of homologous recombination linked to antibiotic resistance in the genus *Salinispora*. Appl. Environ. Microbiol. *79*, 5997–6005.

Guthals, A., Watrous, J.D., Dorrestein, P.C., and Bandeira, N. (2012). The spectral networks paradigm in high throughput mass spectrometry. Mol. Biosystems *8*, 2535–2544.

He, H., Ding, W.-D., Bernan, V.S., Richardson, A.D., Ireland, C.M., Greenstein, M., Ellestad, G.A., and Carter, G.T. (2001). Lomaiviticins A and B, potent antitumor antibiotics from *Micromonospora lomaivitiensis*. J. Am. Chem. Soc. *123*, 5362–5363.

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

Jensen, P.R., Williams, P.G., Oh, D.C., Zeigler, L., and Fenical, W. (2007). Species-specific secondary metabolite production in marine actinomycetes of the genus *Salinispora*. Appl. Environ. Microbiol. *73*, 1146–1152.

Jensen, P.R., Chavarria, K.L., Fenical, W., Moore, B.S., and Ziemert, N. (2014). Challenges and triumphs to genomics-based natural product discovery. J. Ind. Microbiol. Biotechnol. *41*, 203–209.

Jensen, P.R., Moore, B.S., and Fenical, W. (2015). The marine actinomycete genus *Salinispora*: a model organism for secondary metabolite discovery. Nat. Prod. Rep. http://dx.doi.org/10.1039/c4np00167b.

Keller, A., Eng, J., Zhang, N., Li, X.j., and Aebersold, R. (2005). A uniform proteomics MS/MS analysis platform utilizing open XML file formats. Mol. Sys. Biol. *1*, 2005.0017.

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

Kersten, R.D., Ziemert, N., Gonzalez, D.J., Duggan, B.M., Nizet, V., Dorrestein, P.C., and Moore, B.S. (2013). Glycogenomics as a mass spectrometry-guided genome-mining method for microbial glycosylated molecules. Proc. Natl. Acad. Sci. USA *110*, 4407–4416.

Kim, T.K., Hewavitharana, A.K., Shaw, P.N., and Fuerst, J.A. (2006). Discovery of a new source of rifamycin antibiotics in marine sponge actinobacteria by phylogenetic prediction. Appl. Environ. Microbiol. *72*, 2118–2125.

Koehn, F.E. (2008). High impact technologies for natural products screening. In Natural Compounds as Drugs Volume I, F. Petersen and R. Amstutz, eds. (Springer), pp. 175–210.

Krug, D., and Müller, R. (2014). Secondary metabolomics: the impact of mass spectrometry-based approaches on the discovery and characterization of microbial natural products. Nat. Prod. Rep. *31*, 768–783.

Lane, A.L., Nam, S.-J., Fukuda, T., Yamanaka, K., Kauffman, C.A., Jensen, P.R., Fenical, W., and Moore, B.S. (2013). Structures and comparative characterization of biosynthetic gene clusters for cyanosporasides, enediyne-derived natural products from marine actinomycetes. J. Am. Chem. Soc. *135*, 4171–4174.

Lechner, A., Eustáquio, A., Gulder, T.A.M., Hafner, M., and Moore, B.S. (2011). Selective overproduction of the proteasome inhibitor salinosporamide A via precursor pathway regulation. Chem. Biol. *18*, 1527–1536.

Lim, C.L., Nogawa, T., Uramoto, M., Okano, A., Hongo, Y., Nakamura, T., Koshino, H., Takahashi, S., Ibrahim, D., and Osada, H. (2014). RK-1355A and B, novel quinomycin derivatives isolated from a microbial metabolites fraction library based on NPPlot screening. J. Antibiot. *67*, 323–329.

Liu, W.-T., Lamsa, A., Wong, W.R., Boudreau, P.D., Kersten, R., Peng, Y., Moree, W.J., Duggan, B.M., Moore, B.S., and Gerwick, W.H. (2014). MS/MS-based networking and peptidogenomics guided genome mining revealed the stenothricin gene cluster in *Streptomyces roseosporus*. J. Antibiot. *67*, 99–104.

Lombó, F., Velasco, A., Castro, A., De la Calle, F., Braña, A.F., Sánchez-Puelles, J.M., Méndez, C., and Salas, J.A. (2006). Deciphering the biosynthesis pathway of the antitumor thiocoraline from a marine actinomycete and its expression in two *Streptomyces* species. ChemBioChem *7*, 366–376.

Maldonado, L.A., Fenical, W., Jensen, P.R., Kauffman, C.A., Mincer, T.J., Ward, A.C., Bull, A.T., and Goodfellow, M. (2005). *Salinispora arenicola* gen. nov., sp. nov. and *Salinispora tropica* sp. nov., obligate marine actinomycetes belonging to the family Micromonosporaceae. Int. J. Syst. Evol. Microbiol. *55*, 1759–1766.

McGlinchey, R.P., Nett, M., and Moore, B.S. (2008). Unraveling the biosynthesis of the sporolide cyclohexenone building block. J. Am. Chem. Soc. *130*, 2406–2407.

Medema, M.H., Paalvast, Y., Nguyen, D.D., Melnik, A., Dorrestein, P.C., Takano, E., and Breitling, R. (2014). Pep2Path: automated mass spectrometry-guided genome mining of peptidic natural products. PLoS Comp. Biol. *10*, e1003822.

Miyanaga, A., Janso, J.E., McDonald, L., He, M., Liu, H., Barbieri, L., Eustáquio, A.S., Fielding, E.N., Carter, G.T., and Jensen, P.R. (2011). Discovery and assembly-line biosynthesis of the lymphostin pyrroloquinoline alkaloid family of mTOR inhibitors in *Salinispora* bacteria. J. Am. Chem. Soc. *133*, 13311–13313.

Mohimani, H., Kersten, R.D., Liu, W.-T., Wang, M., Purvine, S.O., Wu, S., Brewer, H.M., Pasa-Tolic, L., Bandeira, N., Moore, B.S., et al. (2014). Automated genome mining of ribosomal peptide natural products. ACS Chem. Biol. *9*, 1545–1551.

Nett, M., Ikeda, H., and Moore, B.S. (2009). Genomic basis for natural product biosynthetic diversity in the actinomycetes. Nat. Prod. Rep. *26*, 1362–1384.

Nguyen, D.D., Wu, C.-H., Moree, W.J., Lamsa, A., Medema, M.H., Zhao, X., Gavilan, R.G., Aparicio, M., Atencio, L., Jackson, C., et al. (2013). MS/MS networking guided analysis of molecule and gene cluster families. Proc. Natl. Acad. Sci. USA *110*, E2611–E2620.

Nieselt, K., Battke, F., Herbig, A., Bruheim, P., Wentzel, A., Jakobsen, Ø.M., Sletta, H., Alam, M.T., Merlo, M.E., and Moore, J. (2010). The dynamic architecture of the metabolic switch in *Streptomyces coelicolor*. BMC Genomics *11*, 10.

Oh, D.-C., Williams, P.G., Kauffman, C.A., Jensen, P.R., and Fenical, W. (2006). Cyanosporasides A and B, cyano- and chloro-cyclopenta [*a*]indene glycosides from the deep-sea marine actinomycete "*Salinispora pacifica*". Org. Lett. *8*, 1021–1024.

Oh, D.C., Gontang, E.A., Kauffman, C.A., Jensen, P.R., and Fenical, W. (2008). Salinipyrones and pacificanones, mixed-precursor polyketides from the marine actinomycete *Salinispora pacifica*. J. Nat. Prod. *71*, 570–575.

Onaka, H., Taniguchi, S., Igarashi, Y., and Furumai, T. (2002). Cloning of the staurosporine biosynthetic gene cluster from *Streptomyces* sp. TP-A0274 and its heterologous expression in *Streptomyces lividans*. J. Antibiot. *55*, 1063–1071.

Penn, K., Jenkins, C., Nett, M., Udwary, D.W., Gontang, E.A., McGlinchey, R.P., Foster, B., Lapidus, A., Podell, S., Allen, E.E., et al. (2009). Genomic islands link secondary metabolism to functional adaptation in marine Actinobacteria. ISME J. *3*, 1193–1203.

Renner, M.K., Shen, Y.-C., Cheng, X.-C., Jensen, P.R., Frankmoelle, W., Kauffman, C.A., Fenical, W., Lobkovsky, E., and Clardy, J. (1999). Cyclomarins A-C, new antiinflammatory cyclic peptides produced by a marine bacterium (*Streptomyces* sp.). J. Am. Chem. Soc. *121*, 11273–11276.

Roberts, A.A., Schultz, A.W., Kersten, R.D., Dorrestein, P.C., and Moore, B.S. (2012). Iron acquisition in the marine actinomycete genus *Salinispora* is controlled by the desferrioxamine family of siderophores. FEMS Microbiol. Lett. *335*, 95–103.

Schultz, A.W., Oh, D.C., Carney, J.R., Williamson, R.T., Udwary, D.W., Jensen, P.R., Gould, S.J., Fenical, W., and Moore, B.S. (2008). Biosynthesis and structures of cyclomarins and cyclomarazines, prenylated cyclic peptides of marine actinobacterial origin. J. Am. Chem. Soc. *130*, 4507–4516.

Seyedsayamdost, M.R. (2014). High-throughput platform for the discovery of elicitors of silent bacterial gene clusters. Proc. Natl. Acad. Sci. USA *111*, 7266–7271.

Sidebottom, A.M., Johnson, A.R., Karty, J.A., Trader, D.J., and Carlson, E.E. (2013). Integrated metabolomics approach facilitates discovery of an unpredicted natural product suite from *Streptomyces coelicolor* M145. ACS Chem. Biol. *8*, 2009–2016.

Tsueng, G., Teisan, S., and Lam, K.S. (2008). Defined salt formulations for the growth of *Salinispora tropica* strain NPS21184 and the production of salinosporamide A (NPI-0052) and related analogs. Appl. Microbiol. Biotechnol. *78*, 827–832.

Udwary, D.W., Zeigler, L., Asolkar, R.N., Singan, V., Lapidus, A., Fenical, W., Jensen, P.R., and Moore, B.S. (2007). Genome sequencing reveals complex secondary metabolome in the marine actinomycete *Salinispora tropica*. Proc. Natl. Acad. Sci. USA *104*, 10376–10381.

Vizcaino, M.I., Engel, P., Trautman, E., and Crawford, J.M. (2014). Comparative metabolomics and structural characterizations illuminate colibactin pathway-dependent small molecules. J. Am. Chem. Soc. *136*, 9244–9247.

Watanabe, K., Hotta, K., Nakaya, M., Praseuth, A.P., Wang, C.C., Inada, D., Takahashi, K., Fukushi, E., Oguri, H., and Oikawa, H. (2009). *Escherichia coli* allows efficient modular incorporation of newly isolated quinomycin biosynthetic enzyme into echinomycin biosynthetic pathway for rational design and synthesis of potent antibiotic unnatural natural product. J. Am. Chem. Soc. *131*, 9347–9353.

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., et al. (2012). Mass spectral molecular networking of living microbial colonies. Proc. Natl. Acad. Sci. USA *109*, 1743–1752.

Williams, P.G., Miller, E.D., Asolkar, R.N., Jensen, P.R., and Fenical, W. (2007). Arenicolides A-C, 26-membered ring macrolides from the marine actinomycete *Salinispora arenicola*. J. Org. Chem. *72*, 5025–5034.

Wilson, M.C., Gulder, T.A.M., Mahmud, T., and Moore, B.S. (2010). Shared biosynthesis of the saliniketals and rifamycins in *Salinispora arenicola* is controlled by the sare1259-encoded cytochrome P450. J. Am. Chem. Soc. *132*, 12757–12765.

Wilson, M.C., Mori, T., Ruckert, C., Uria, A.R., Helf, M.J., Takada, K., Gernert, C., Steffens, U.A.E., Heycke, N., Schmitt, S., et al. (2014). An environmental bacterial taxon with a large and distinct metabolic repertoire. Nature *506*, 58–62.

Winnikoff, J.R., Glukhov, E., Watrous, J., Dorrestein, P.C., and Gerwick, W.H. (2013). Quantitative molecular networking to profile marine cyanobacterial metabolomes. J. Antibiot. *67*, 105–112.

Wolfe, A.J. (2005). The acetate switch. Mol. Biol. Rev. *69*, 12–50.

Yamanaka, K., Reynolds, K.A., Kersten, R.D., Ryan, K.S., Gonzalez, D.J., Nizet, V., Dorrestein, P.C., and Moore, B.S. (2014). Direct cloning and refactoring of a silent lipopeptide biosynthetic gene cluster yields the antibiotic taromycin A. Proc. Natl. Acad. Sci. USA *111*, 1957–1962.

Yang, J.Y., Sanchez, L.M., Rath, C.M., Liu, X., Boudreau, P.D., Bruns, N., Glukhov, E., Wodtke, A., de Felicio, R., Fenner, A., et al. (2013). Molecular networking as a dereplication strategy. J. Nat. Prod. *76*, 1686–1699.

Zhang, Q., Ortega, M., Shi, Y., Wang, H., Melby, J.O., Tang, W., Mitchell, D.A., and van der Donk, W.A. (2014). Structural investigation of ribosomally synthesized natural products by hypothetical structure enumeration and evaluation using tandem MS. Proc. Natl. Acad. Sci. USA *111*, 12031–12036.

Ziemert, N., Lechner, A., Wietz, M., Millan-Aguinaga, N., Chavarria, K.L., and Jensen, P.R. (2014). Diversity and evolution of secondary metabolism in the marine actinomycete genus *Salinispora*. Proc. Natl. Acad. Sci. USA *111*, E1130–E1139.

Zolova, O.E., Mady, A.S., and Garneau-Tsodikova, S. (2010). Recent developments in bisintercalator natural products. Biopolymers *93*, 777–790.

Chemistry & Biology, Volume *22*

Supplemental Information

Molecular Networking and Pattern-Based Genome Mining Improves Discovery of Biosynthetic Gene Clusters and their Products from *Salinispora* Species

Katherine R. Duncan, Max Crüsemann, Anna Lechner, Anindita Sarkar, Jie Li, Nadine Ziemert, Mingxun Wang, Nuno Bandeira, Bradley S. Moore, Pieter C. Dorrestein, and Paul R. Jensen

Supplemental Figures

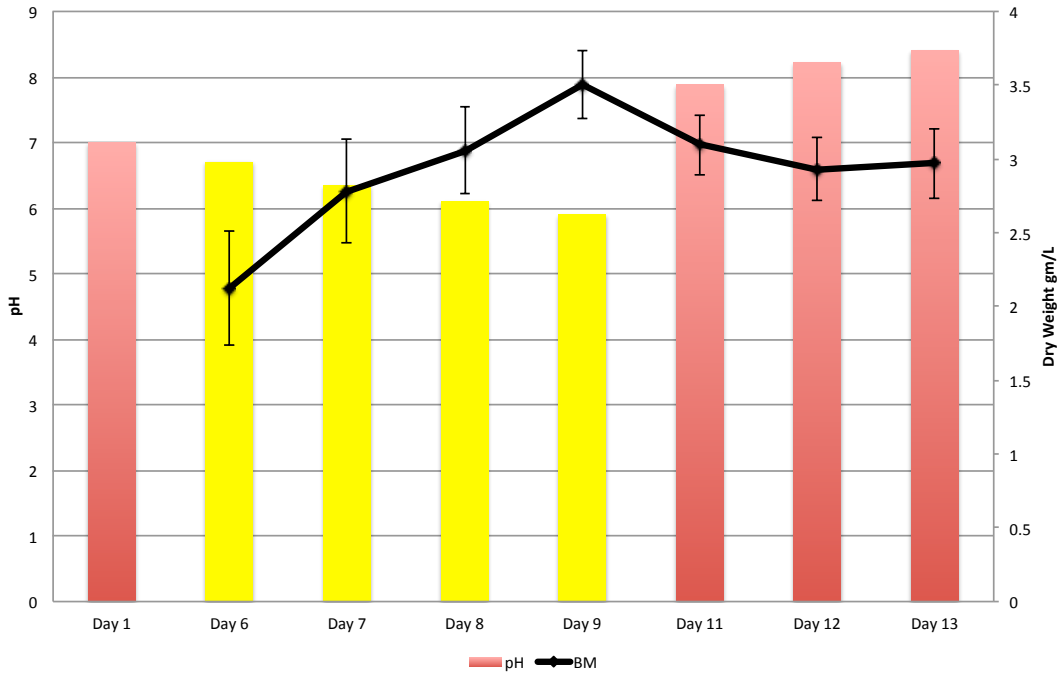


Figure S1, Related to Figure 1. Representative phenol red growth curve (*S. tropica* strain CNY-681). The culture pH dropped below 7 at day 6, which corresponded to a phenol red color change from red to yellow. Upon entering stationary phase after day 9, the pH increased to above 7, which resulted in a color change from yellow back to red. All cultures were extracted when this yellow to red color change was observed. Standard deviations around the mean were calculated for the dry weight biomasses (BM) from three independent cultures.

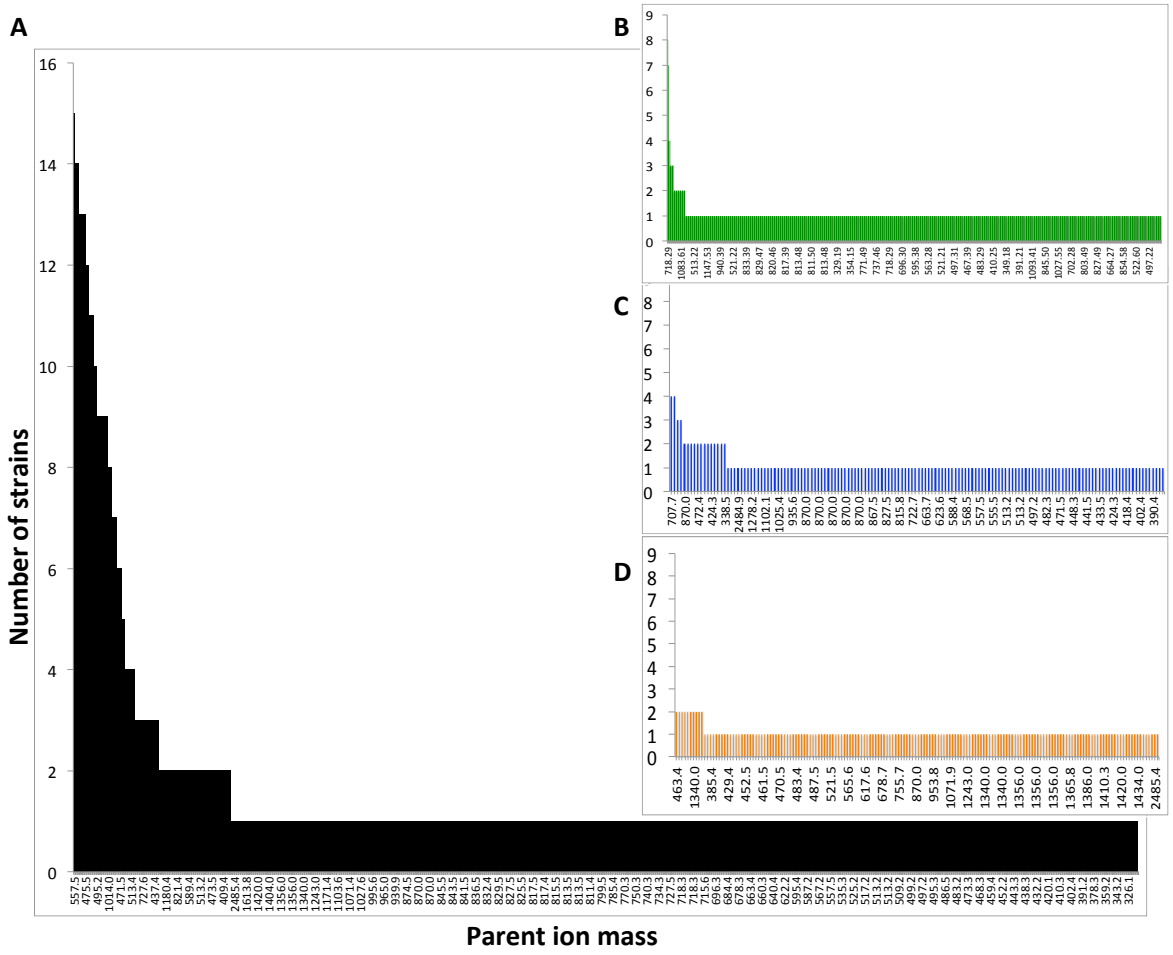


Figure S2, Related to Figure 2. Rank-abundance curve of parent ion distributions. A) Excluding media components, a total of 1114 parent ions were observed of which 85.3 % were detected in only one strain. B) *S. arenicola* parent ion distribution among 12 strains. C) *S. pacifica* parent ion distribution among 17 strains. D) *S. tropica* parent ion distribution among 7 strains. Note: Y-axis = number of strains for figures B-C; scales vary among figures.

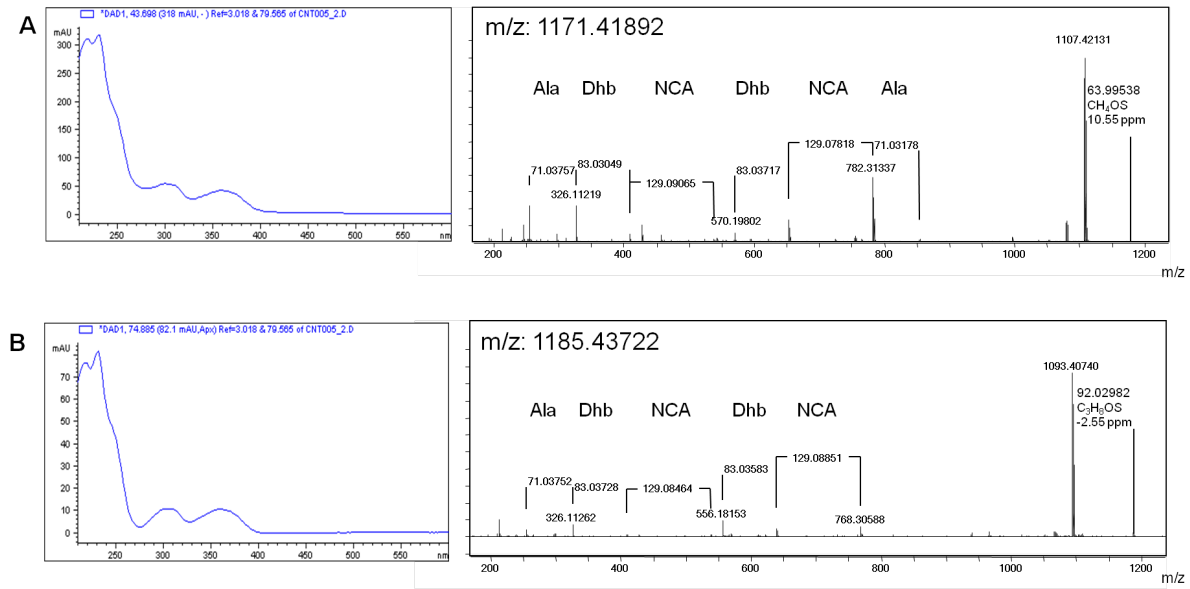


Figure S3. Related to Figure 5. : UV-Vis spectra and MS2 fragmentation paterns for (A) retimycin A and (B) the retimycin *m/z* 1185 analogue.

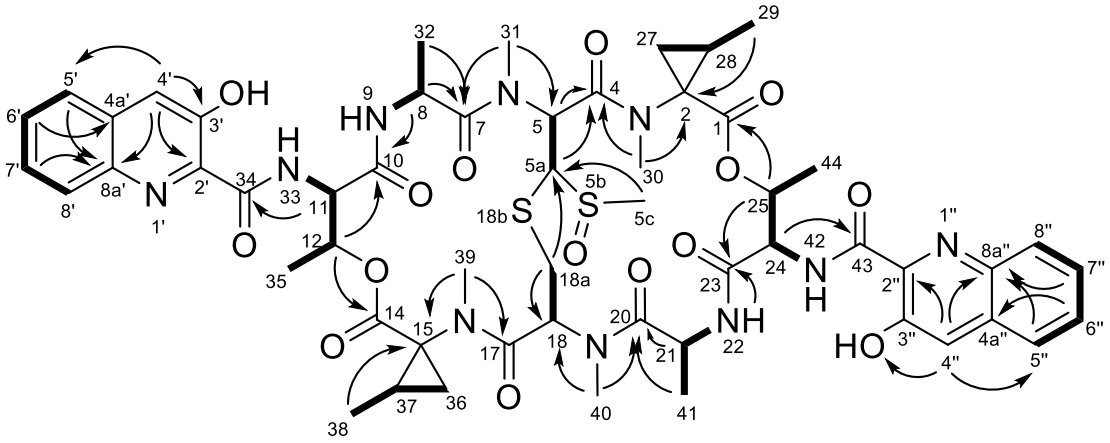
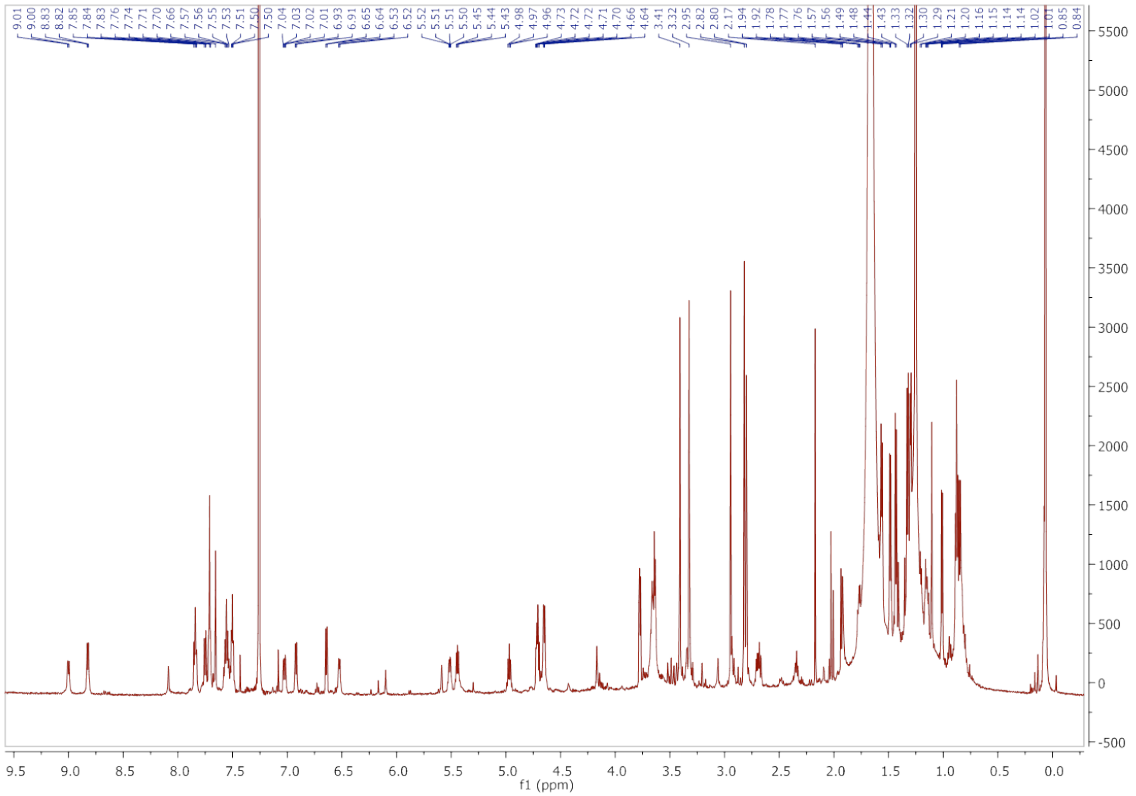
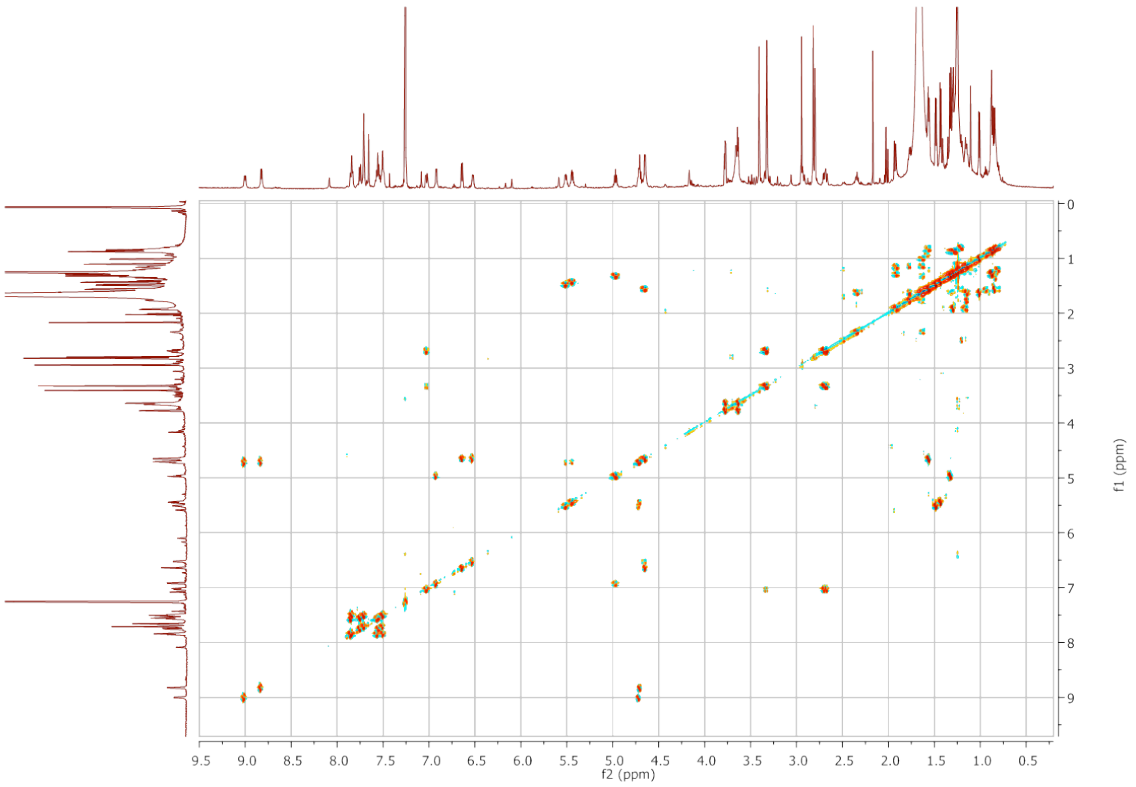


Figure S4, Related to Figure 5. ^{1}H–^{1}H COSY (bold lines) and key HMBC (arrows) correlations of retimycin A

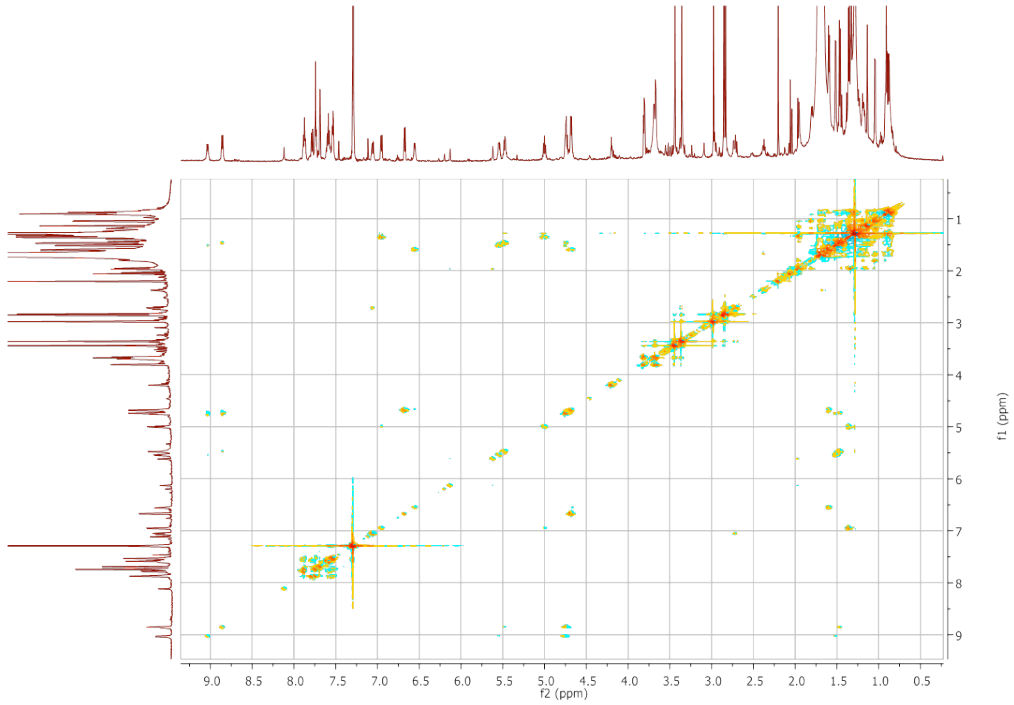
A)



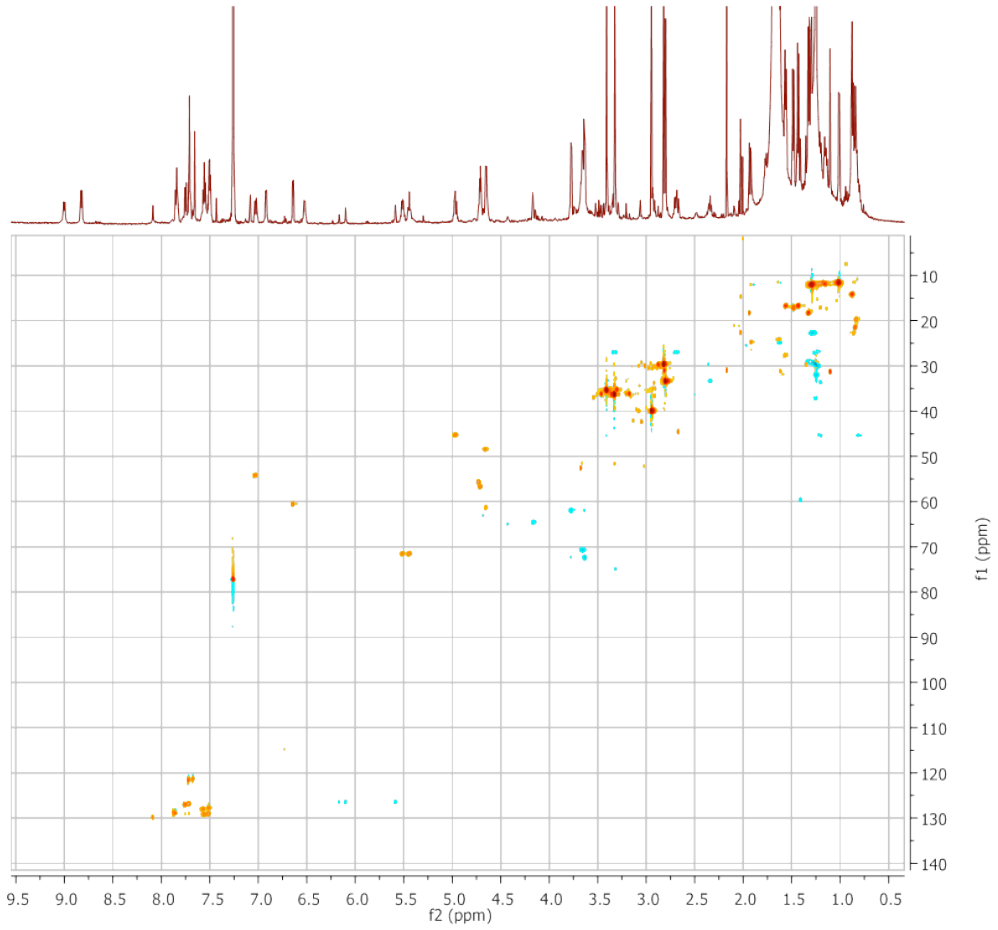
B)



C)



D)



E)

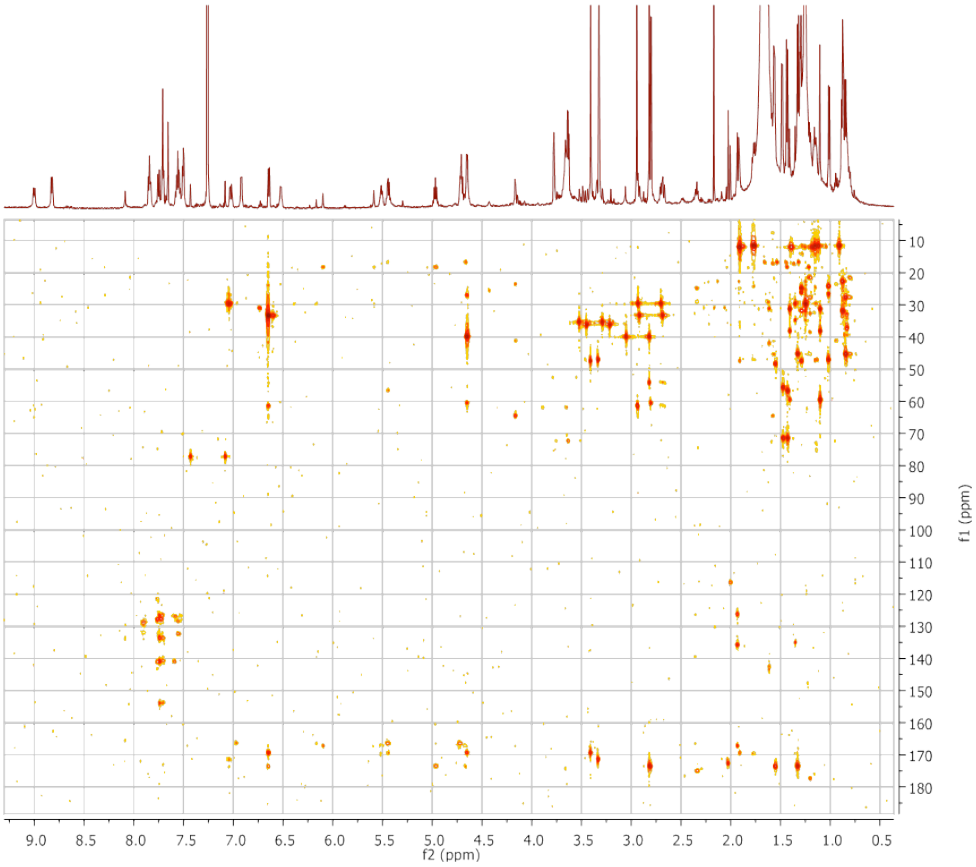


Figure S5, Related to Figure 5. A) ^{1}H NMR spectrum of retemycin A ($CDCl_3$, 600 MHz). B) DQF-COSY spectrum of retemycin A ($CDCl_3$, 600 MHz). C) TOCSY spectrum of retemycin A ($CDCl_3$, 600 MHz). D) HSQC spectrum of retemycin A ($CDCl_3$, 600 MHz). E) HMBC spectrum of retemycin A ($CDCl_3$, 600 MHz).

Supplemental Tables

Table S1, Related to Table 1. *Salinispora* strains used in this study. NA = not available, IMG = Integrated Microbial Genomes (http://img.jgi.doe.gov/).

Number of strains	Species	Strain	Location	16S Sequence type[a] (accession number)	Extraction time (days)	Genome sequence	New to this study	IMG Genome ID
1	*S. arenicola*	CNB-527	Bahamas	ST (JN999716)	19	Yes	No	2515154093
2	*S. arenicola*	CNH-646	Bahamas	ST (AY040620)	13	Yes	No	2515154181
3	*S. arenicola*	CNH-725	Red Sea	ST (AY040621)	19	No	NA	NA
4	*S. arenicola*	CNH-996	Sea of Cortez	A (JN999727)	30	Yes	Yes	2561511104
5	*S. arenicola*	CNP-188	US Virgin Islands	ST (KM205629)	17	No	NA	NA
6	*S. arenicola*	CNP-193	Sea of Cortez	B (NA)	18	Yes	No	2518285552
7	*S. arenicola*	CNQ-748	Guam	ST (NA)	18	Yes	No	2515154180
8	*S. arenicola*	CNR-647	Bahamas	ST (FJ887039)	19	No	NA	NA
9	*S. arenicola*	CNS-205	Palau	ST (NR_074612)	11	Yes	No	641228504
10	*S. arenicola*	CNT-005	Fiji	ST (JN161827)	19	Yes	No	2517572137
11	*S. arenicola*	CNT-849	Hawaii	ST (NA)	15	Yes	No	2518285550
12	*S. arenicola*	CNX-508	Palmyra	ST (NA)	15	Yes	No	2515154188
13	*S. pacifica*	CNR-894	Palau	ST (NA)	14	Yes	No	2515154194
14	*S. pacifica*	CNR-114	Guam	ST (DQ224161)	13	Yes	No	2515154178
15	*S. pacifica*	CNR-551	Guam	A (HQ642881)	13	No	NA	NA
16	*S. pacifica*	CNR-942	Palau	E (HQ642877)	11	Yes	No	2518285561
17	*S. pacifica*	CNS-055	Palau	A (DQ224159)	13	Yes	No	2518285562
18	*S. pacifica*	CNS-237	Palau	B (HQ642850)	13	Yes	Yes	2524614807
19	*S. pacifica*	CNS-863	Fiji	C (HQ642851)	25	Yes	No	2517572194
20	*S. pacifica*	CNT-003	Fiji	ST (NA)	30	Yes	No	2515154126
21	*S. pacifica*	CNT-029	Fiji	F (HQ642852)	13	Yes	No	2515154177
22	*S. pacifica*	CNT-138	Fiji	C (HQ642853)	16	Yes	No	2516493032
23	*S. pacifica*	CNT148	Fiji	A (HQ642899)	29	Yes	No	2517287023
24	*S. pacifica*	CNT-150	Fiji	B (HQ642900)	24	Yes	No	2517434008
25	*S. pacifica*	CNT-851	Hawaii	D (NA)	14	Yes	No	2517572162

26	*S. pacifica*	CNT-855	Hawaii	A (NA)	14	Yes	No	2515154128
27	*S. pacifica*	CNY-202	Sea of Cortez	K (HQ873948)	21	Yes	Yes	2528311034
28	*S. pacifica*	CNY-330	Sea of Cortez	ST (NA)	19	Yes	No	2518645626
29	*S. pacifica*	CNY-368	Sea of Cortez	C (HQ873947)	19	No	NA	NA
30	*S. tropica*	CNB-440	Bahamas	ST (AY040617)	11	Yes	No	640427140
31	*S. tropica*	CNB-536	Bahamas	ST (AY040618)	9	Yes	No	2517572212
32	*S. tropica*	CNS-197	Bahamas	ST (JN999709)	15	Yes	No	2515154163
33	*S. tropica*	CNY-012	Bahamas	ST (NA)	16	Yes	Yes	2540341192
34	*S. tropica*	CNY-678	Yucatan	ST (NA)	24	Yes	Yes	2561511109
35	*S. tropica*	CNY-681	Yucatan	ST (NA)	14	Yes	Yes	2561511108

a) Sequence types defined as per Freel et al., 2012.

Table S2, Related to Figure 5. The retimycin gene cluster (NRPS40) from *S. arenicola* CNT-005. Genes are listed as presented from left to right in Figure 5D for NRPS40.

Gene	Size (aa)	Annotation	Closest homolog [source] (% similarity/identity)	Predicted function
rtmA	488	Amino acid adenylation domain	ribosomal peptide synthetase [*Streptomyces davawensis*] (61/49)	Unknown
rtmB	236	Tryptophan 2,3-dioxygenase	tryptophan 2,3-dioxygenase [uncultured bacterium esnapd9] (90/83)	HQA biosynthesis
rtmC	421	Aspartate/tyrosine/aromatic aminotransferase	kynurenine aminotransferase [*Micromonospora* sp. ML1] (83/76)	HQA biosynthesis
rtmD	240	Dehydrogenase	NAD or NADP oxidoreductase [*Micromonospora* sp. ML1] (83/75)	HQA biosynthesis
rtmE	423	Cytochrome P450	quinaldate 3-hydroxylase [*Micromonospora* sp. ML1] (85/77)	HQA biosynthesis
rtmF	528	(2,3-dihydroxybenzoyl) adenylate synthase	putative 3-hydroxy-quinaldate-AMP-Ligase [*Micromonospora* sp. ML1] (89/76)	HQA biosynthesis
rtmG	599	Amino acid adenylation domain	NRPS protein [*Micromonospora* sp. ML1] (78/70)	HQA biosynthesis
rtmH	391	Cytochrome P450	cytochrome P450 [*Saccharopolyspora spinosa*] (80/68)	Core oxygenation
rtmI	385	Kynurenine aminotransferase	putative aminotransferase [*Streptomyces* sp. SNA15896] (83/77)	NCA biosynthesis
rtmJ	634	Fe-S oxidoreductase	hypothetical protein [*Streptomyces* sp. SNA15896] (88/81)	NCA biosynthesis
rtmK	237	Methyltransferase	putative SAM-dependent methyltransferase [*Streptomyces* sp. SNA15896] (78/64)	Methylation
rtmL	661	Fe-S oxidoreductase	radical SAM protein [*Saccharomonospora* sp. CNQ490] (79/66)	Thioacetal formation
rtmM	244	Thioesterase	putative thioesterase [*Streptomyces* sp. SNA15896] (83/74)	HQA biosynthesis
rtmN	820	Excinuclease ATPase subunit	UvrA-like protein [*Streptomyces* sp. SNA15896] (89/80)	Unknown
rtmO	2565	Non-ribosomal peptide synthetase	long-chain-fatty-acid--CoA ligase [uncultured bacterium esnapd9] (77/67)	Peptide backbone formation
rtmP	3126	Non-ribosomal peptide synthetase	putative non-ribosomal peptide synthetase [*Streptomyces* sp. SNA15896] (85/77)	Peptide backbone formation
rtmQ	70	MbtH-like protein	MbtH-like protein [*Streptomyces* sp. SNA15896] (90/71)	Peptide backbone formation
rtmR	167	Signal transduction response regulator, C-terminal	OmpR family regulator [*Micromonospora* sp. ML1] (85/74)	Unknown

Table S3, related to Figure 5. ^{1}H and ^{13}C-NMR spectroscopic data of retimycin A[a] ([α] -12.2° (c.0.2, $CHCl_3$)

no.	δ_H [b]	δ_C [c]	no.	δ_H [b]	δ_C [c]
1		169.4	33-NH	9.00 (1H, d, 7.5)	
2		47.4	34		167.8
4		169.3	35	1.48 (3H, d, 6.5)	17.1
5	6.64 (1H, d, 6.9)	60.5	36	1.80 (1H, m)	26.6
5a	4.66 (1H, d, 6.9)	61.4		1.18 (1H, m)	
5c	2.95 (3H, s)	39.9	37	1.64 (1H, m)	24.0
7		173.6	38	1.04 (3H, d, 6.2)	11.3
8	4.67 (1H, m)	48.3	39	3.32 (3H, s)	36.2
9-NH	6.53 (1H, br s)		40	2.82 (3H, s)	29.5
10		167.1	41	1.33 (1H, d, 6.8)	18.2
11	4.72 (1H, m)	55.7	42-NH	8.82 (1H, d, 7.5)	
12	5.51 (1H, m)	71.5	43		167.5
14		169.5	44	1.43 (1H, d, 6.5)	16.7
15		47.1	2′		133.6
17		171.5	3′		153.9
18	7.03 (1H, dd, 10.2, 3.9)	54.2	4′	7.66 (1H, s)	121.1
18a	3.35 (1H, dd, 15.3, 3.9)	26.9	4a′		132.4
	2.72 (1H, dd, 15.3, 10.2)		5′	7.71 (1H, m)	126.9
20		173.5	6′	7.50 (1H, m)	129.0
21	4.97 (1H, m)	45.2	7′	7.55 (1H, m)	127.6
22-NH	6.92 (1H, d, 7.4)		8′	7.84 (1H, m)	128.9
23		166.3	8a′		140.8
24	4.71 (1H, m)	56.7	2″		133.6
25	5.44 (1H, m)	71.5	3″		153.9
27	1.92 (1H, m)	25.9	4″	7.71 (1H, m)	121.4
	1.18 (1H, m)		4a″		132.4
28	1.90 (1H, m)	24.7	5″	7.75 (1H, m)	127.0
29	1.30 (3H, d, 6.2)	11.9	6″	7.51 (1H, m)	129.2
30	3.41 (3H, s)	35.2	7″	7.57 (1H, m)	127.9
31	2.80 (3H, s)	33.3	8″	7.85 (1H, m)	129.1
32	1.56 (3H, d, 6.9)	16.7	8a″		141.0

[a] NMR data obtained in $CDCl_3$. Assignments are based on ^{1}H–^{1}H COSY, HSQC, and HMBC spectroscopic data. The assignments for the two 3-hydroxyquinaldic acid rings are interchangeable due to the absence of HMBC correlations from 33-NH and 42-NH protons [b] Measured at 600 MHz for ^{1}H NMR; δ in ppm, (mult. J in Hz). [c] Deduced from HSQC and HMBC spectroscopic data; δ in ppm.