ACS central science

Research Article

http://pubs.acs.org/journal/acscii

Metabologenomics: Correlation of Microbial Gene Clusters with Metabolites Drives Discovery of a Nonribosomal Peptide with an Unusual Amino Acid Monomer

Anthony W. Goering,[†] Ryan A. McClure,[†] James R. Doroghazi,[‡] Jessica C. Albright,[†] Nicole A. Haverland,[†] Yongbo Zhang,[§] Kou-San Ju,[‡] Regan J. Thomson,[†] William W. Metcalf,[*,‡] and Neil L. Kelleher[*,†]

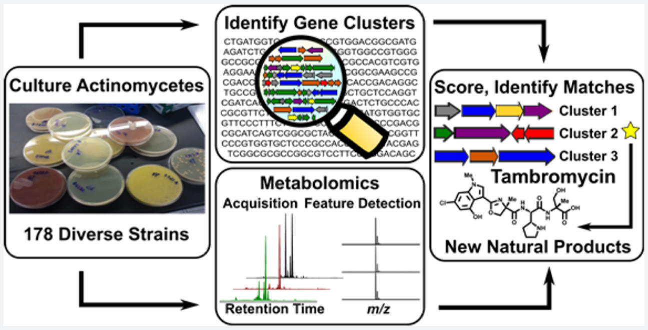
[†]Departments of Chemistry, Molecular Biosciences, and the Feinberg School of Medicine, Northwestern University, Evanston, Illinois 60208, United States

[‡]Department of Microbiology and the Carl R. Woese Institute of Genomic Biology, University of Illinois at Urbana−Champaign, Urbana, Illinois 61801, United States

[§]Integrated Molecular Structure Education and Research Center, Weinberg College of Arts and Sciences, Northwestern University, Evanston, Illinois 60208, United States

S *Supporting Information*

ABSTRACT: For more than half a century the pharmaceutical industry has sifted through natural products produced by microbes, uncovering new scaffolds and fashioning them into a broad range of vital drugs. We sought a strategy to reinvigorate the discovery of natural products with distinctive structures using bacterial genome sequencing combined with metabolomics. By correlating genetic content from 178 actinomycete genomes with mass spectrometry-enabled analyses of their exported metabolomes, we paired new secondary metabolites with their biosynthetic gene clusters. We report the use of this new approach to isolate and characterize tambromycin, a new chlorinated natural product, composed of several nonstandard amino acid monomeric units, including a unique pyrrolidine-containing amino acid we name tambroline. Tambromycin shows antiproliferative activity against cancerous human B- and T-cell lines. The discovery of tambromycin via large-scale correlation of gene clusters with metabolites (a.k.a. metabologenomics) illuminates a path for structure-based discovery of natural products at a sharply increased rate.



INTRODUCTION

Small molecule natural products produced by microbes have demonstrated high value through their utility as medicines.[1] Many of the most important microbe-derived medicines, such as streptomycin (the first effective treatment for tuberculosis) and doxorubicin (a potent cancer therapeutic), are produced by actinomycetes, a diverse group of bacteria present in soil and marine environments worldwide.[2,3]

Historical efforts to discover bioactive natural products from actinomycetes and other microbes on an industrial scale have largely involved screening spent media for biological activity. It has been estimated that more than 10 million bacterial extracts have been screened for antibiotic activity in academic and industrial laboratories over the last half of the 20th century.[4] Despite early successes, the productivity of the classical discovery approach has diminished in recent times—an assertion reaffirmed by the closing of natural product discovery efforts at most major pharmaceutical companies.

The advent of low-cost genome sequencing has led to the revelation that microbes possess a far greater biosynthetic potential than previously realized, and this has stood in contrast with the declining interest in natural products as a source of new pharmacophores.[5] Biosynthetic genes that function in concert to produce a single metabolite are typically found close together in bacterial genomes as a biosynthetic gene cluster (BGC). The disparity between the number of discrete BGCs and the current rate of discovery for the new natural products made by those BGCs suggests a need for novel approaches.[6] The genetic diversity that exists in even a single genus of the microbial world is far too large to characterize the metabolism it encodes for using traditional methods.[7]

To meet this challenge, we combined genome sequencing and automated gene cluster prediction with mass spectrometry-based metabolomics to detect new natural products and

Received: October 5, 2015
Published: January 20, 2016

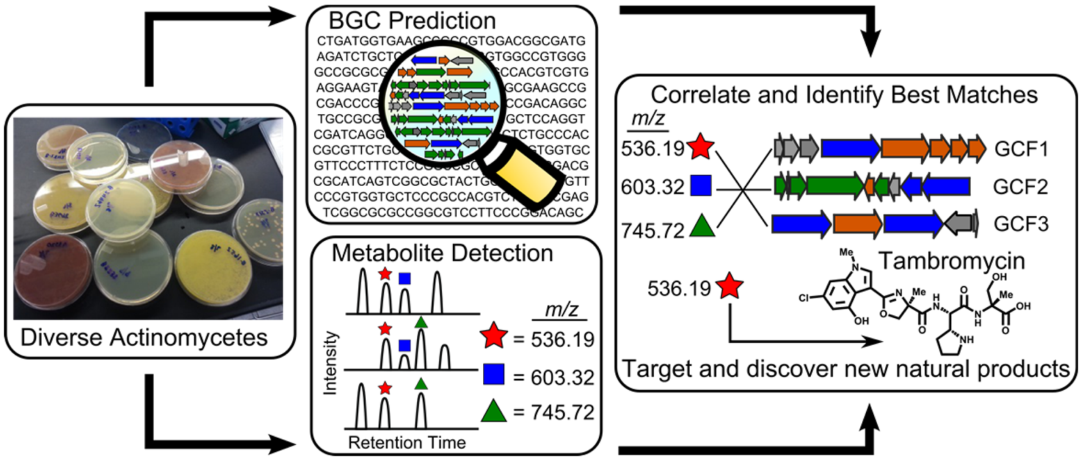


Figure 1. Work and data flows for a "metabologenomics" approach to natural products discovery from bacteria. Using information obtained from interpreting 178 sequenced genomes into gene clusters and gene cluster families (top center panel, described in ref 8) and from MS-based metabolomics of the same 178 strains with accurate mass (bottom center panel), pairwise correlation yields scores that associate metabolites with their gene cluster families (panel at right).

correlate them with their biosynthetic gene clusters on a large-scale basis: a discovery approach we call "metabologenomics" (Figure 1).[8] Metabologenomics enables discovery at a larger scale by using data from multiple organisms to identify the BGCs responsible for the biosynthesis of expressed metabolites. It is also possible to identify small molecules with related biosynthesis, or to focus efforts toward natural products that do not have well studied relatives. Metabologenomics works by grouping similar BGCs from diverse bacteria into gene cluster families (GCFs).[9] GCFs are then scored against metabolomics LC-MS data from the same set of strains to obtain a score indicating the likelihood that a particular observed metabolite and GCF are associated, based on the similarity of their occurrence across different strains.

Previously, we reported the use of this method to successfully correlate 27 known secondary metabolites with their published biosynthetic gene clusters.[8] Here, we report for the first time the use of metabologenomics to uncover a new bioactive natural product, tambromycin, and its biosynthetic gene clusters in 11 different actinomycetes. Tambromycin's distinguishing features are a hydroxylated and chlorinated indole, a methyl-oxazoline, and 2-methyl-serine, substructures also found in JBIR-34/35 (Figure 2A).[10] However, an unusual pyrrolidine-containing amino acid was found upon complete structural elucidation of tambromycin by NMR, also recently reported by Izumikawa et al. under the name JBIR-126.[11] Here we report the absolute stereochemistry of the tambroline monomer, the structures of tambromycin and two of its analogues, and the biosynthetic gene cluster for the compound, shown to be expressed in nine different strains. Further, we explore the phylogenetic distribution of the biosynthetic cassette for the new amino acid, formed by cyclization of lysine as determined by extensive metabolic labeling experiments with labeled precursors.

RESULTS

Binary Correlation To Identify a Novel Biosynthetic Gene Cluster and Metabolite Pair. To achieve targeted discovery of new small molecules and pathways, we analyzed the genetic content and exported metabolites of 178 actinomycete strains from the NRRL collection. The genome of each strain was sequenced using Illumina technology, and a routine for identification and categorization of gene clusters into families (GCFs) was used as previously described.[8] In this initial report, 11 422 natural product biosynthetic gene clusters were predicted across five categories (NPRS, PKS Type I, PKS Type II, lanthipeptides, and thiazole/oxazole modified microcins) (see www.igb.illinois.edu/labs/metcalf/gcf for an interactive display of all gene clusters). Next, the profile of exported metabolites from these 178 strains was accessed in four

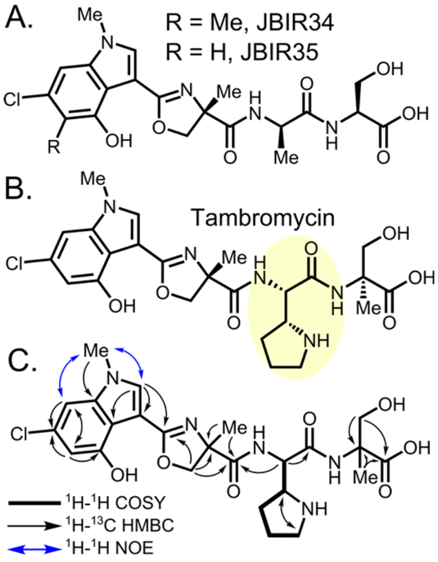


Figure 2. Structures of JBIR34 and 35 (A), tambromycin (B), and key NMR correlations (C). Structure of tambromycin is shown with the new amino acid residue tambroline highlighted in panel (B). The related structures JBIR 34 and 35 are shown in panel (A). Panel (C) highlights key correlations derived from ^{1}H-^{1}H COSY, ^{1}H-^{1}H NOESY, and ^{1}H-^{13}C HMBC experiments. ^{1}H-^{13}C HMBC correlations from the indole 2-position proton and 2-methyl-serine methylene protons to the same carbonyl carbon were used to determine the connectivity of these substructures as a methyl-oxazoline. Another important ^{1}H-^{13}C HMBC correlation from the alpha proton of tambroline to the carbonyl carbon of 2-methyl-serine associated these two substructures. These correlations and the overall sequence of the peptide were confirmed by tandem mass spectrometry (Figure S5). ^{1}H-^{1}H COSY correlations across the continuous spin system present in the pyrrolidine substructure are shown as widened bonds.

Table 1. Data Enabling the Metabologenomic Identification of Tambromycin and Its Biosynthetic Gene Cluster[a]

Strain ID	Strain Name	*m/z* 536 (Tambromycin)	Selected Ion Intensity	NRPS GCFs									
NRRL F-4474	*S. species*	yes	3.7 x 10^8	95	311	263	**519**		221				
NRRL S-515	*S. species*	yes	1.0 x 10^7	95			**519**				490	382	259
NRRL WC-3542	*S. lavendulae* subsp. *lavendulae*	yes	1.3 x 10^8	95	311	263	**519**	446	221	594			
ISP-5094	*S. virginiae* subsp. *virginae*	yes	1.7 x 10^8	95	311	263	**519**	446	221	594			
NRRL S-98	*S. species*	yes	1.3 x 10^8	95	311	263	**519**	446	221	594			
NRRL S-237	*S. species*	yes	2.1 x 10^7	95			**519**				490		
NRRL S-241	*S. species*	yes	2.5 x 10^6	95	311	263	**519**				490	382	
NRRL B-2375	*S. species*	no	0		311	263	**519**	446	221	594			
NRRL S-104	*S. katrae*	no	0	95		263	**519**	446	221	594			
NRRL S-118	*S. species*	yes	5.3 x 10^6				**519**						259
ISP-5269	*S. sclerotialus*	yes	7.1 x 10^7				**519**						

[a] This table shows the co-occurrence of tambromycin with members of 10 gene cluster families (GCFs) for 11 *Streptomycete* strains probed in this study. Strains are listed at the far left, tambromycin detection by MS is shown in columns 3 and 4, and columns 5−14 show all NRPS GCFs that are present in two or more of this set of 11 strains. Note that data for the strains listed in the top nine rows were obtained in the first pass application of LC-MS and automated data reduction. The strains listed in the bottom two rows were interrogated in a targeted fashion because their genomes harbored the tambroline biosynthetic gene cluster. Note that the highest level of co-occurrence is between tambroline and NRPS Gene Cluster Family #519 in the first-pass metabologenomics data set (i.e., top nine strains). The background color in column 4 highlights whether the observed MS intensity was above (green) or below (red) the threshold intensity of 5 × 10^6 set for automated metabolite detection; the selected ion intensity derived from the LC-MS data reflects the number of ion counts (i.e., the NL value). Numbers designating each gene cluster family appear as archived on the website at www.igb.illinois.edu/labs/metcalf/gcf.

different media conditions by quantitative LC-MS/MS, detecting 2520 individual metabolites with predicted accurate masses ranging from 250.074 to 4538.881 Da in the refined data set.

Metabolite and gene cluster data were used as inputs for a simple binary correlation algorithm, and scores ranged from 0 to ∼300, with scores >200 considered confident based on the successful correlation of knowns in the data set. Among all the GCF/NP pairs, tambromycin was identified as an ion with *m/z* 536.190 and was selected for further analysis based on its high raw correlation score of 229 with NRPS GCF 519. This metabolite was expressed by six *different* strains in the data set that each encoded the *same* nonribosomal peptide synthetase (NRPS) biosynthetic gene cluster (see Table 1). There were three additional strains that contained the same cluster, but in whose extracts the compound was not detected above the threshold levels used in the automated feature detection step. However, in one of these three strains (NRRL S-241, see Row 7 of Table 1) tambromycin production was in fact detected at lower levels upon manual inspection. Inclusion of this hit would raise the correlation score to 239 using the reported framework. In summary, *m/z* 536.190 was detected in 7 of 9 strains containing a gene cluster from this GCF (Table 1); thus an overall expression rate of 78% was observed among strains in the data set that possess a member of this GCF. Direct analysis of tandem mass spectra confirmed that the observed ion species represented the same secondary metabolite in all the strains in which it was detected.

To further verify the correlation between tambromycin and NRPS GCF 519, we asked whether the presence of the BGC could be used to predict whether new strains not included in our original screen of 178 could produce tambromycin. We tested two additional strains that contained the tambromycin BGC, NRRL B-5269 and S-118, the most distantly related from the main group of strains in the *S. virginiae* clade that contain NRPS GCF 519 but still contain a member of this GCF (Figure 3). Both of these strains produced the same *m/z* 519.190 ion species that was previously observed (Table 1, bottom two rows). Of all the strains analyzed by LC-MS, the highest intensity of the tambromycin ion was observed from *Streptomyces* strain F-4474. Therefore, this strain was selected for fermentation and purification of the metabolite.

Isolation and Structural Characterization of Tambromycin. Tambromycin (Figure 2B) was isolated from the fermentation broth of *Streptomyces* strain F-4474 using a solid phase resin and then purified by semipreparative reverse-phase HPLC. We found that significantly higher quantities of tambromycin were obtained when strain F-4474 was grown on solid agar media. The tambromycin ion displayed an isotopic distribution characteristic of the incorporation of a chlorine atom into the structure. With this constraint, the

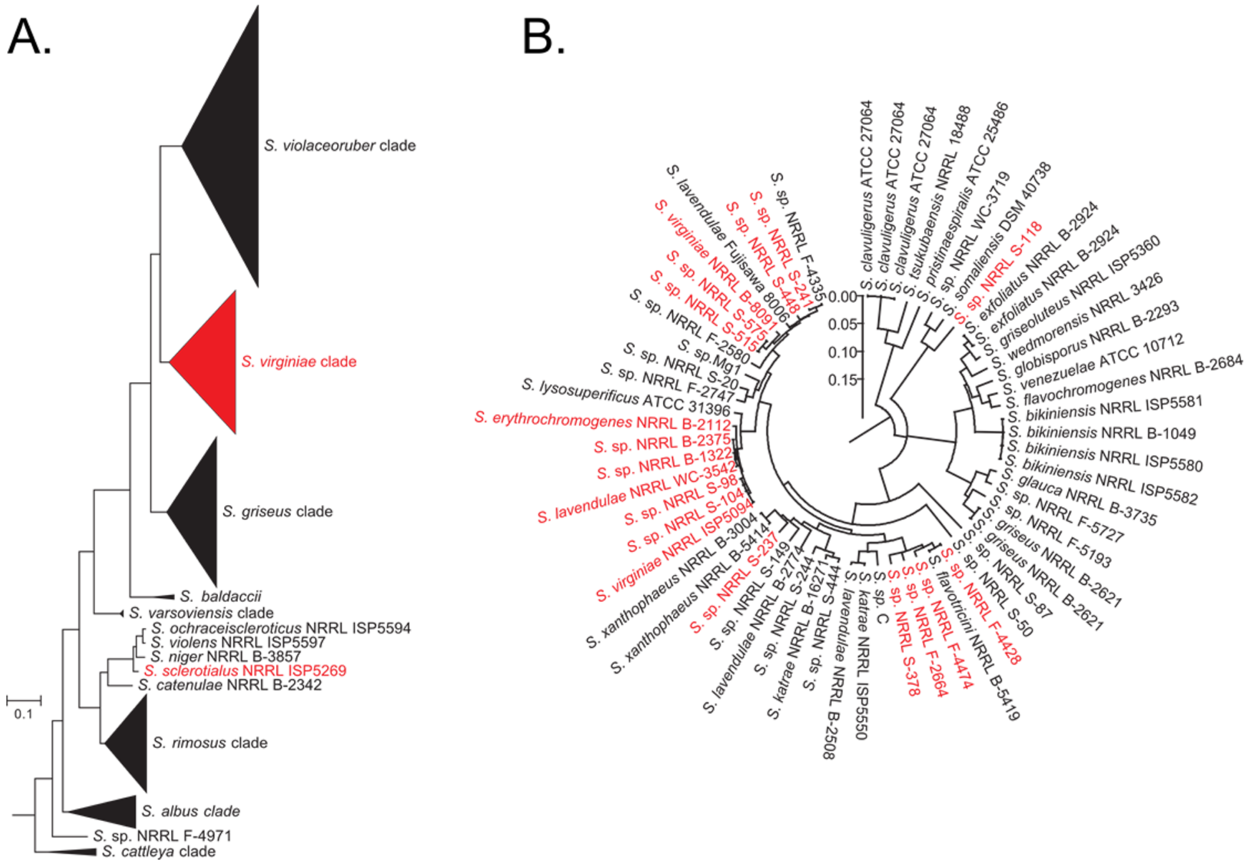


Figure 3. Distribution of the tambromycin gene cluster across diverse Streptomycetes. (A) The biosynthetic gene cluster of tambromycin is—with one exception—distributed throughout the *Streptomyces virginiae* clade (www.igb.illinois.edu/labs/metcalf/gcf/gcfDisplay.php?gcf=NRPS_GCF.519). (B) Eighteen strains within the *virginiae* clade were identified as having a member of this GCF. The plurality of strains containing a member of this gene cluster family made it possible to draw precise gene cluster boundaries from the ORFs that are conserved across all members of this group.

molecular formula of $C_{24}H_{30}N_5O_7Cl$ was determined for tambromycin by high mass-accuracy mass spectrometry.

^{1}H NMR spectroscopic analysis accounted for 24 of the protons, and the remaining six protons were absent because of exchange with the deuterated methanol solvent (Figure S3A). The presence of three proton signals in the aromatic region (δ_H 7.61, 6.57, and 6.89) was congruent with the presence of a substituted indole (Table S2). The methine with shifts δ_c/δ_H 133.4/7.61 was readily assigned to the indole 2-position. The other two aromatic protons were not strongly coupled, indicating that indole substitution must occur either at the 7- and 5-positions or at the 6- and 4-positions. An interesting singlet methyl (δ_c/δ_H 33.9/3.82) was identified as an indole *N*-methyl substituent based on ^{1}H-^{13}C HMBC correlations with aromatic carbons and ^{1}H-^{1}H NOESY cross peaks with the indole 2-position (δ_H 7.61). An additional NOESY cross peak between this *N*-methyl and one of the aromatic protons (δ_H 6.81) localized this proton to the indole 7-position, thus eliminating the possibility of substitutions at the 7- and 5-positions (Figure 2C). Therefore, the indole substitutions were assigned as follows: hydroxyl to the 4-position (δ_c 152.3), chlorine to the 6-position (δ_c 131.3), and the other aromatic proton assigned to the 5 position (δ_c/δ_H 108.2/6.57). Of the remaining indole quaternary carbons, position 7a (δ_c 140.1) was localized based on HMBC correlations from the indole *N*-methyl and 7-position, position 3a (δ_c 114.1) by HMBC correlations from the 2-, 5-, and 7-positions, and position 3 (δ_c 102.5) by HMBC correlations from the 2-position.

Amino acid α-proton resonances were notably absent, with only one methine signal observed at δ_H 4.93. The two 2-methyl-serines were each identified as a grouping of methyl, methylene, quaternary, and carbonyl carbons associated by HMBC signals. From TOCSY experiments, only a single spin system was observed, spanning a five carbon aliphatic moiety [δ_H 4.93 (d, 1H), 4.09 (ddd, 1H), 1.74/2.10 (m, m, 2H), 2.00/1.93 (m, m, 2H), 3.27 (m, 2H)]. COSY and NOESY spectra were used to determine direct connectivity and precise order of the members in this spin system. On the basis of splitting patterns, long-range ^{1}H-^{13}C correlations, and chemical shifts, the five-membered spin system was determined to be a pyrrolidine unit adjacent to an amino acid α-carbon. Upon the basis of its composition, we refer to this new amino acid as tambroline (two-amino-beta-homoproline).

The key NMR spectroscopic correlations leading to the determination of the sequence of the peptide included multiple-bond (^{1}H-^{13}C HMBC) correlations from one of the 2-methyl-serine methylenes to a predicted carbonyl carbon (δ_c 165.2) that also correlated with the proton at the 2 position on the indole (δ_H 7.61). An HMBC signal from the α-proton of tambroline to the carbonyl of 2-methyl-serine adjacent to the indole acid determined that this new amino acid appears next in the sequence. No long-range correlations were observed to link the tambroline to the final 2-methyl-serine, an expected result because there are no protons in either substructure that are positioned fewer than four bonds away from any carbon.

Table 2. Functional Annotation of Genes in the Biosynthetic Gene Cluster of Tambromycin (from Strain F-4474)

ORF	no. of amino acids	proposed function	homologue function	homologue accession (Uniprot)	fractional amino acid identity (%)
tbrA	207	flavin reductase	flavin reductase RbmH	Q8KI76	75/155 (48)
tbrB	534	tryptophan 6-halogenase	flavin-dependent tryptophan halogenase RebH	Q8KHZ8	354/527 (67)
tbrC	333	SARP family regulator	SARP family transcriptional regulator	R1FZM2	126/284 (44)
tbrD	528	aldehyde dehydrogenase	betaine-aldehyde dehydrogenase	R4SX69	290/484 (60)
tbrE	611	NRPS (A-T) (Trp)	uncharacterized protein	A4FJG4	301/547 (55)
tbrF	249	NRPS (TE)	thioesterase	A0A077JC92	109/237 (46)
tbrG	60	mbth	putative MbtH-like protein	W7IRY5	33/54 (61)
tbrH	405	cytochrome p450 (tryptophan oxygenase)	cytochrome P450 NovI	Q9L9F9	140/400 (35)
tbrI	1313	NRPS (A-T-Cyc-Mt) (Cl-indole-acid)	nonribosomal peptide synthetase	A0A077JBM3	736/1319 (56)
tbrJ	215	phosphopantetheinyl transferase	4′-phosphopantetheinyl transferase	A0A0C5FYL8	92/198 (46)
tbrK	344	regulatory element	uncharacterized protein	A0A0C1VA91	179/353 (51)
tbrL	426	transport	major facilitator superfamily protein	R4TAR4	191/396 (48)
tbrM	1364	NRPS (C-A-T-Te) (2-me-ser)	uncharacterized protein	N0CW84	567/1402 (40)
tbrN	1119	NRPS (C-A-T) (tambroline)	nonribosomal peptide synthetase	K0K8E4	474/1126 (42)
tbrO	469	NRPS (C)	long-chain-fatty-acid-CoA ligase	J2JV27	171/463 (37)
tbrP	397	acyl-CoA dehydrogenase	acyl-CoA dehydrogenase domain protein	C7QK66	162/341 (48)
tbrQ	382	acyl-CoA dehydrogenase	uncharacterized protein	V6K9A0	105/320 (33)
tbrR	194	*N*-acetyltransferase	GCN5 family acetyltransferase	A0A094M2Y4	94/191 (49)
tbrS	357	tryptophan aldolase	threonine aldolase	C7QGM4	178/331 (54)
tbrT	230	hypothetical protein	uncharacterized protein fmoF	A0A077JC94	111/222 (50)
tbrU	467	alanine hydroxymethyltransferase	serine hydroxymethyltransferase fmoH	A0A077JCX7	272/402 (68)
tbrV	1097	NRPS (Cyc-A-T) 2-me-serine	nonribosomal peptide synthetase fmoA3	A0A077JG85	605/1091 (55)
tbrW	625	regulatory element	regulatory protein	D9W2C9	331/607 (55)
tbrX	69	hypothetical protein	uncharacterized protein	A0A0F0GHR6	27/73 (37)
tbrY	1015	regulatory element	putative AfsR-like transcriptional regulator	B1VL80	522/989 (53)

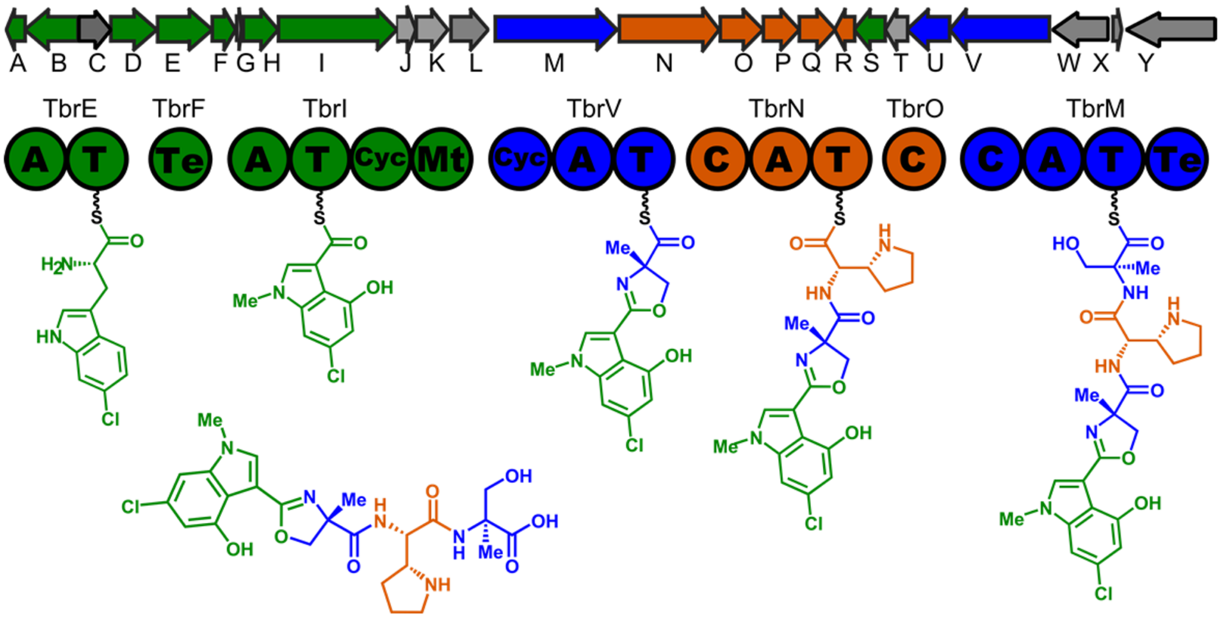


Figure 4. Overview of the tambromycin biosynthetic gene cluster. Organization of the tambromycin biosynthetic gene cluster is shown. ORFs and protein representations are color coordinated with the structure of tambromycin to indicate their proposed association with biosynthesis of a particular amino acid substructure. Proteins colored gray are not proposed to be involved directly in biosynthesis, but rather in cluster regulation, transport, or other supporting roles (e.g., phosphopantetheinylation of NRPS proteins).

Study of the tandem mass spectra of tambromycin determined that the final 2-methyl-serine was located at the terminus of the peptide opposite the indole, and the data were also consistent with the arrangement of the other monomers determined by NMR spectroscopy (from end-to-end: indole-acid, 2-methyl-serine, tambroline, and 2-methyl-serine). A recurring neutral loss of dihydropyrrole was also observed, that originates from tambroline (Figure S5). The mass of the sum of these predicted monomers was greater than the observed mass of tambromycin by the exact mass of a water molecule; therefore, a dehydration must be present within the structure. This dehydration was localized to the formation of an oxazoline at the position of the peptide bond between the indole-acid and its adjacent 2-methyl-serine. Chemical shifts of the groups constituting the 2-methyl-serine located at the C-terminus were also further upfield compared to their equivalents in the indole-acid adjacent monomer. The presence of an acid group was further supported by 1-D ^{1}H NMR spectroscopy of tambromycin in DMSO-d_6 (Figure S3B). We synthesized all four possible isomers of tambroline as standards for determining its absolute stereochemistry in tambromycin using Marfey's reagent (Figure S6). These results appear convergent with those of Izumikawa et al. reporting JBIR-126[11] and are augmented by the detection and elucidation of two additional analogues, including tambromycin C, featuring a portion of the phosphopantetheine cofactor stably appended to the C-terminus (Figures S8−S10 and Tables S3−S4).

Determination of Biosynthetic Gene Cluster Boundaries. The highest scoring biosynthetic GCF for tambromycin from the correlation results was NRPS GCF 519 (www.igb.illinois.edu/labs/metcalf/gcf/gcfDisplay.php?gcf=NRPS_GCF.519), with a score of 229. Continued sequencing of actinomycetes as part of a larger project identified additional members of this GCF, now numbering 24 in total.[12] This GCF is consistent with the expected biosynthesis of this molecule, because NRPS systems often incorporate noncanonical amino acids such as 2-methyl-serine into their products.[13] NRPS GCF 519 is concentrated primarily within the *S. virginiae* clade (Figure 3). Gene cluster boundaries were determined from ORFs that are conserved across all clusters in the tambromycin GCF. The gene cluster of strain NRRL F-4474, from which tambromycin was isolated and characterized fully, is reported here as the canonical biosynthetic gene cluster for tambromycin in Table 2.

Genes Encoding for the Biosynthesis of the Indole and Methyl-oxazoline. A schematic overview of the gene cluster ORFs is presented at the top of Figure 4 with gene annotations listed in Table 2. Structurally, tambromycin is related to the secondary metabolites JBIR-34 and JBIR-35, compounds previously isolated from a marine streptomycete, and which have a published biosynthetic scheme.[14] JBIR-34/35 also contain a 4-methyl-oxazoline resulting from the heterocyclization of 6-chloro-4-hydroxy-1-methylindole-3-carboxylic acid and 2-methly serine. Because these structural features are shared between tambromycin and JBIR-34/35, we expected the gene cluster producing tambromycin to also encode for highly similar enzymes. Indeed, tryptophan halogenase, tryptophan aldolase, aldehyde dehydrogenase, and methylation domain containing NRPS encoding genes were found in the cluster with reasonably high sequence similarity to their counterparts in the JBIR-34/35 cluster. Also, an alanine-hydroxymethyltransferase necessary to produce 2-methyl-serine from alanine was identified in the BGC with a high degree of homology to its JBIR-34/35 cluster counterpart (see Table S1 for a thorough comparison of the BGCs).[14a,15] Beyond the genes shared with the biosynthetic cluster of JBIR-34/35, tambromycin biosynthesis involves two additional modular NRPSs, each containing an A-domain for which a substrate could not be confidently assigned using NRPSpredictor2.[16] We predicted that one of these NRPS proteins is responsible for the loading of tambroline, either directly or as a lysine-derived precursor, possibly acetyl-lysine based on the presence of a predicted *N*-acetyltransferase, TbrR.

Genes Encoding NRPS Proteins. Figure 4 highlights the domain substructures tied to main NRPS carrier proteins. There are five A-domain containing NRPS proteins encoded by the tambromycin BGC: TbrE, TbrI, TbrM, TbrN, and TbrV. TbrE and TbrI have moderate sequence similarity with their counterparts FmoA1 and FmoA2 in the biosynthetic gene cluster of JBIR-34/35 (59 and 68%), their predicted domain structures are identical, and their 10 amino acid Stachelhaus specificity-conferring signatures match with 9/10 and 8/10 residues, respectively.[17] As in the biosynthesis of JBIR-34/35, TbrE activates and loads tryptophan, then facilitates the modification of the indole by TbrB and TbrH. Once modified, tryptophan is released from TbrE and then converted to the indole acid by the action of TbrS and TbrD. The indole acid is loaded onto TbrI via the action of a second A domain, which catalyzes the indole *N*-methylation and initiates the elongation of the tetrapeptide. We predict two of the remaining NRPS proteins, TbrM and TbrV, are selective for 2-methyl-serine, and both have the prototypical domain structure: a condensation domain, an adenylation domain, and a thiolation or peptide carrier protein domain, and TbrM also has a thiosterase domain. TbrV is orthologous to FmoA3 from the BGC of JBIR-34/35 (57% identity), its 10 amino acid specificity code is identical to that of FmoA3, and it is predicted to have the same cyclization functionality as FmoA3, which forms methyl-oxazoline from 2-methyl-serine (Figure S1). Interestingly, while TbrM and TbrV are predicted to select the same substrate, their NRPS codes are not similar, matching only 3 out of 10 residues. The A-domain signature of TbrM matches instead (9/10) to the NRPS protein AmiT in the biosynthetic gene cluster of amicetin. This protein activates 2-methyl-serine for incorporation into amicetin as the final step in the biosynthesis.[18] By the process of elimination, we propose that TbrN is involved in the incorporation of tambroline, although it remains to be shown whether this protein activates tambroline itself, or if the cyclization occurs while the uncyclized precursor is tethered to the NRPS.

Genes Encoding for Tambroline Biogenesis. To the best of our knowledge, tambroline has not been identified previously as a substructure of any other natural product, making the characterization of the enzymes responsible for its biosynthesis of particular interest. When comparing the gene clusters for JBIR-34/35 to that for tambromycin, large organizational differences between these clusters are apparent. Other than the pair *tbrE* and *tbrF* and the pair *tbrT* and *tbrU*, there are no two genes from the JBIR-34/35 BGC that remain proximal in strain F-4474. This lack of conservation affords insight into the proposed biosynthetic steps that result in these two compound families, including the candidate genes responsible for the pyrrolidine-containing amino acid (tambroline) which distinguishes tambromycin from these other compounds.

Among the biosynthetic genes in the cluster that had not been functionally assigned, the most likely candidates for tambroline biosynthesis are a pair of acyl-CoA dehydrogenases (ACADs) and an adjacent N-acetyl transferase, *tbrP*, *tbrQ*, and *tbrR*, respectively. The tambroline cassette is embedded within numerous and diverse gene-clusters for putative NRPS derived molecules (Figure S2). In all cases, two acyl-CoA dehydrogen-

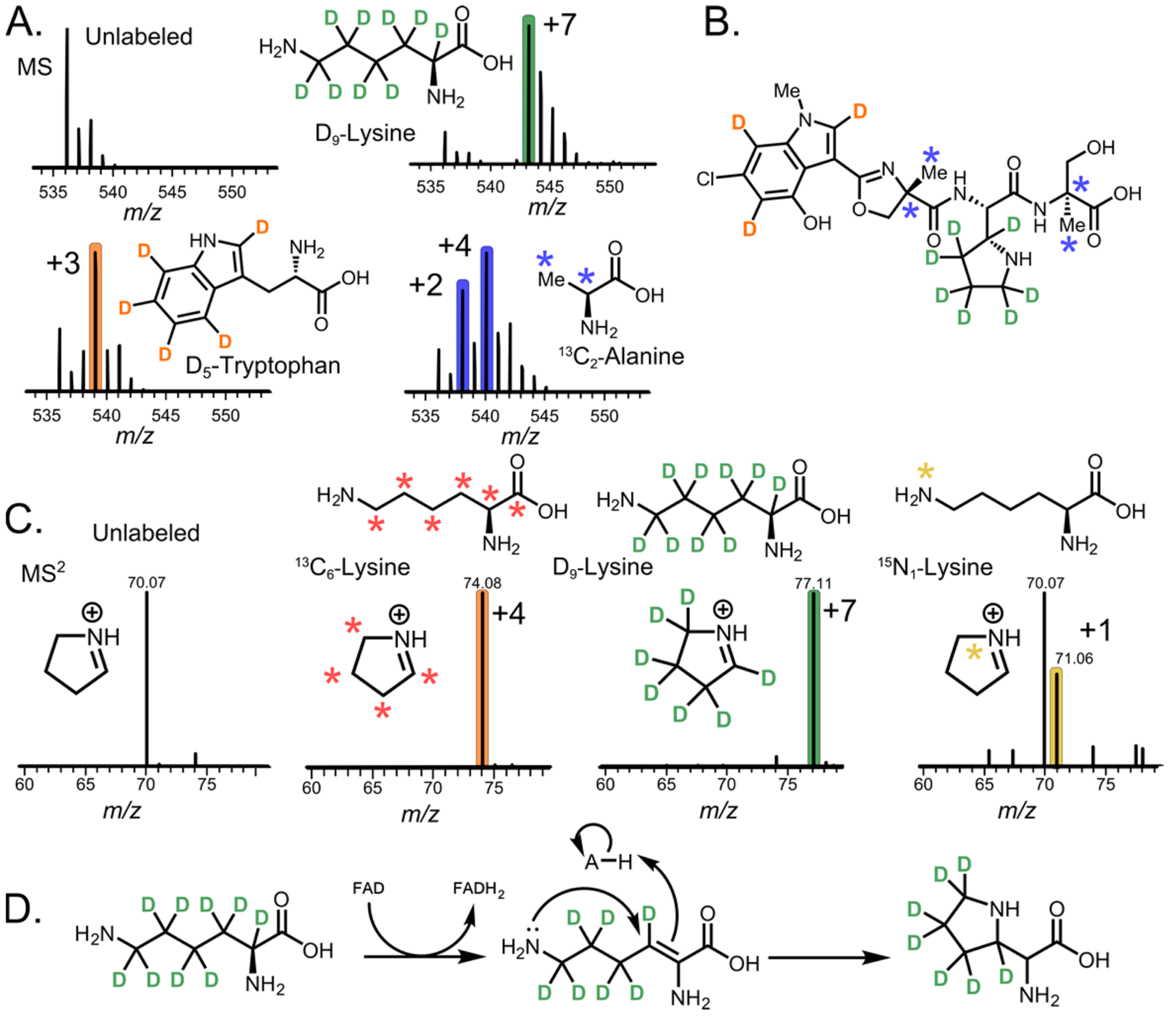


Figure 5. Summary of stable isotope feeding experiments and biosynthetic insights. (A) Stable isotope experiments were carried out by feeding labeled lysine, tryptophan, and alanine. When strain F-4474 was cultured with tryptophan-d_3, tambromycin was observed to incorporate three deuterons from the indole portion of the tryptophan label into its structure, consistent with two substitutions occurring on the indole ring. Tambromycin also appeared to incorporate two ^{13}C alanine monomers, which are modified via a hydroxymethyltransferase to form 2-methyl-serine. (B) Summary of heavy isotopes observed to be incorporated into the structure of tambromycin. (C) Lysine labeled with no heavy isotopes (panel at far left) deuterium, ^{13}C and ^{15}N (at the indicated positions in each panel) was used to illuminate the biosynthetic path leading to the formation of tambroline by comparing the change in mass of a tambroline-derived internal fragment ion observed in the MS2 spectrum of tambromycin. (D) Proposal for a biosynthesis of tambroline from lysine through an acyl-CoA dehydrogenase-catalyzed mechanism consistent with the labeling results shown in C).

ase genes are preceded by a highly conserved NRPS gene. Many of these are homologues of TbrN and are predicted to encode the same domains (C-A-T) and to load the same amino acid substrate. Collectively, these sequences inform the biosynthesis of tambroline and act as genetic markers to enable genome mining for natural products containing this unusual amino acid. We found that a similar biosynthetic cassette is present in the BGC responsible for the biosynthesis of azinomycin B, a genotoxic molecule that contains a pyrrolidine substructure resembling tambroline, although with extensive modifications. The biosynthesis of this substructure has been proposed to involve an ACAD protein (AziC8). Interestingly, the BGC for azinomycin B also contains an adjacent pair of ACAD proteins in the same family as TbrP and TbrQ (AziD3 and AziD2) and the putative *N*-acetyl transferase (AziC2, which has a functional counterpart in the tambromycin cluster, namely, TbrR). A possible difference is that the proposed substrate of the azinomycin biosynthesis enzymes is an extended acetyl-glutamate rather than lysine.[19] This leads us to propose that the pyrrolidine moiety is produced through the catalyzed cyclization of the side-chain of lysine following α,β-dehydrogenation (vide infra). The lysine precursor may also be acetylated prior to dehydrogenation, which we expect would promote its cyclization to form tambroline.

Investigation of Biosynthesis by Feeding with Stable Isotope Labeled Amino Acids. To investigate the biosynthesis of tambromycin, feeding experiments were performed with a panel of stable-isotope labeled amino acids. We found that labeled lysine, alanine, and tryptophan were efficiently incorporated into the metabolite (Figure 5A). Tambromycin was found to incorporate three deuterium atoms from tryptophan labeled with five deuterons in the indole region, suggesting the incorporation of an indole with two substitutions. When strain F-4474 was fed with $^{13}C_2$-labeled alanine, shifted $[M + H]^+$ signals for tambromycin were consistent with the incorporation of one and two units of the labeled substrate (Figure 5A, bottom right of panel). In further support of our hypothesis for tambroline biosynthesis, we observed that seven of nine deuterons were incorporated from lysine-d_9. By examining the tandem MS fragmentation pathway of unlabeled vs labeled tambromycin (Figure 5C), we localized the remaining seven deuterons precisely to the pyrrolidine ring. However, when labeling experiments with $^{13}C_6$-lysine were conducted, all six carbons were found to incorporate into the structure of tambromycin. These experiments indicate that one proton from each of the α and β positions of lysine must be removed during the biosynthesis of tambromycin, ruling out any biosynthesis involving the complete oxidation of the β position. Removing protons at the α and β positions is consistent with our hypothesized biosynthesis by ACAD enzymes (Figure 5D).

Initial Bioactivity Screening of Tambromycin. Partially purified extracts containing tambromycin were used to test growth inhibition of *Micrococcus luteus, Escherichia coli, Bacillus subtilis, Klebsiella pneumonia, Pseudomonas aeruginosa,* and *Staphylococcus epidermidis* using disk-diffusion assays. We observed no biological activity against any of these strains at concentrations up to 28 μM. Next, the activity of the purified natural product was also tested against five different cancerous B- and T-cell lines using a concentration of 37 μM (20 μg/mL). These five cell lines represent T-cell acute lymphocytic leukemia (Jurkat), Burkitt's lymphoma (Ramos), chronic lymphocytic leukemia (Hg-3), mantle cell lymphoma (Maver-1), and B-cell acute lymphocytic leukemia (RCH-ACV). Three days following tambroline exposure, the cells were stained with trypan blue, and the total counts of live and dead cells were determined. Significantly fewer total and live cells were present under treatment conditions (Figure S7). Antiproliferative activity testing was expanded by treating the following cell lines at concentrations of 50 μM: FaDu, HeLa, OVCar3, PC3. These four cell lines did not show a significant response to tambromycin exposure.

■ DISCUSSION

In a 2014 report, we first established that the frequency of co-occurrence of metabolites and gene clusters successfully assigns known secondary metabolites to their published biosynthetic gene clusters. The method was used to identify GCFs for the desertomycins and oasamycins, natural products for which the corresponding biosynthetic gene clusters were unknown.[20] Tambromycin highlights the benefits of this correlative approach for identifying compounds with new chemical scaffolds because it contains an amino acid monomer that has not been reported in the structure of previous natural products and displays promising bioactivity (Figure S7). While tambromycin shares its indole and methyl-oxazoline moieties with JBIR-34/35 and the enzymes proposed to be required for the production of this structural component are present in the cluster, the order and orientation of genes differs markedly between the two systems. The biosynthesis of JBIR-34/35 involves extensive preassembly line processing.[14a] These compounds were the first reported to possess the indole and methyl-oxazoline seen also in tambromycin and are among only a few known to incorporate 2-methyl-serine.

Tambromycin also shares some characteristic features with chlorocarcin A, a compound reported by Mikami et al. in 1976.[21] The reported mass observed for chlorocarcin A is within 1 Da of that which we have observed for tambromycin, well within the margin of error for the instrumentation used in that earlier report. Both compounds also contain chlorine and have been observed to be produced in subspecies of *Streptomyces lavendulae*. Chlorocarcin A was reported to have broad antibiotic activity, as well the ability to reduce the size and weight of transplanted animal tumors.[21] We have observed inhibition in the growth of selected human cell lines in the presence of tambromycin, which is consistent with these previous results. Despite these similarities, we report a different chemical formula for tambromycin than chlorocarcin A. To our knowledge, no structure has ever been proposed for chlorocarcin A.

We expect the set of remaining GCF/MS correlations contains many additional novel compounds beyond tambromycin. The correlation approach yields several direct outputs: (1) the strains are known to produce the compounds (and their approximate relative titers), (2) the gene clusters and small molecules are paired and associated (with a score), so the BGC responsible for production is linked, and (3) the compound produced is also known to be a new NP by virtue of the accurate mass and MS/MS data. Overall, the approach guides the discovery of new natural products such that confidently assigned biological sequence data can be used to aid nearly every level of the discovery process: enhancing the capabilities of natural products researchers to focus their pursuits toward identifying novel scaffolds at the dereplication stage, and assisting in determining monomer composition and connectivity during structure elucidation. The correlation approach, which we refer to as "metabologenomics", scales favorably, and correlations will become more reliable as the scale upon which it is implemented increases to hundreds more or even to several thousands of strains.

The driving force in natural products research has for a long time been the commercial pursuit of new bioactive scaffolds, with biosynthetic characterization too often sidelined as a largely academic effort. With technological advances of genome acquisition and analysis, and a growing body of knowledge regarding the logic of these often-modular systems, biosynthetic analysis at the genome level is no longer a secondary effort, but rather has the potential to focus discovery toward the unexplored corners of natural product chemical space. Taking these advances into consideration, it now appears feasible that individual laboratories could apply these technologies to navigate genetic and chemical space and elucidate dozens of new scaffolds per year.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acscentsci.5b00331. Biosynthetic gene cluster data are available at the following Web site: www.igb.illinois.edu/labs/metcalf/gcf.

Additional experimental methods and figures, including NMR and tandem mass spectra of tambromycin (PDF)

Accession Codes

The assembled genome of *Streptomyces* strain NRRL F-4474 is available from the National Center for Biotechnology Information under the accession number ASM71885v1.

Genes linked to the biosynthesis of tambromycin can be accessed from the NCBI under the accession numbers WP_030845393−WP_030845442. All sequences used for metabologenomics can be accessed through NCBI BioProject PRJNA238534.

AUTHOR INFORMATION

Corresponding Authors

*(W.W.M.) E-mail: metcalf@life.illinois.edu.
*(N.L.K.) E-mail: n-kelleher@northwestern.edu.

Author Contributions

A.W.G. and J.C.A. collected and analyzed metabolomics data, and A.W.G. performed the isolation and structure elucidation of tambromycin. R.A.M. performed the Marfey's test experiments including the synthesis of tambroline stereoisomers. Y.Z. aided the design and interpretation of NMR experiments. N.A.H. and R.A.M. performed bioactivity experiments against human cell lines. A.W.G., J.R.D., and K.S.J. performed *in silico* analysis of the biosynthetic gene cluster. W.W.M., J.R.D., and N.L.K. designed the correlative methodology. A.W.G., R.A.M., R.J.T., W.W.M., and N.L.K. performed manuscript preparation and revision.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We thank the staff at Northwestern University's Integrated Molecular Structure Education and Research Center (IMSERC), especially Dr. Derek Nelson, Dr. Andrew Ott, and Dr. Yuyang Wu. We also thank the Agricultural Research Service of the United States Department of Agriculture for providing the bacterial strains used in this publication. NMR spectroscopy experiments were supported by the National Institute of Health (NIH) Award NIH 1S10OD012016-01/1S10RR019071-01A1. This work was also supported by the Departments of Chemistry and Molecular Biosciences at Northwestern University and Grants GM067725 (to N.L.K.) and GM077596 (to W.W.M.) from the NIH. J.R.D. was supported by the Institute for Genomic Biology IGB Fellows Program at UIUC and K.S.J. by NIH Ruth L. Kirchstein National Service Research Award F32GM100658. Additional funding was provided by an Institute for Genomic Biology Proof of Concept grant.

REFERENCES

(1) (a) Cragg, G. M.; Newman, D. J. Natural products: a continuing source of novel drug leads. *Biochim. Biophys. Acta, Gen. Subj.* **2013**, *1830* (6), 3670−3695. (b) Newman, D. J.; Cragg, G. M. Natural products as sources of new drugs over the last 25 years. *J. Nat. Prod.* **2007**, *70* (3), 461−477.

(2) Jones, D.; Metzger, H. J.; Schatz, A.; Waksman, S. A. Control of Gram-Negative Bacteria in Experimental Animals by Streptomycin. *Science* **1944**, *100* (2588), 103−105.

(3) Arcamone, F.; Cassinelli, G.; Fantini, G.; Grein, A.; Orezzi, P.; Pol, C.; Spalla, C. Adriamycin, 14-hydroxydaunomycin, a new antitumor antibiotic from S. peucetius var. caesius. *Biotechnol. Bioeng.* **1969**, *11* (6), 1101−1110.

(4) Bachmann, B. O.; Van Lanen, S. G.; Baltz, R. H. Microbial genome mining for accelerated natural products discovery: is a renaissance in the making? *J. Ind. Microbiol. Biotechnol.* **2014**, *41* (2), 175−184.

(5) (a) 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.; 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* **2002**, *417* (6885), 141−147. (b) Oliynyk, M.; Samborskyy, M.; Lester, J. B.; Mironenko, T.; Scott, N.; Dickens, S.; Haydock, S. F.; Leadlay, P. F. Complete genome sequence of the erythromycin-producing bacterium Saccharopolyspora erythraea NRRL23338. *Nat. Biotechnol.* **2007**, *25* (4), 447−453. (c) 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.* **2007**, *104* (25), 10376−10381.

(6) (a) Harvey, A. L.; Edrada-Ebel, R.; Quinn, R. J. The re-emergence of natural products for drug discovery in the genomics era. *Nat. Rev. Drug Discovery* **2015**, *14* (2), 111−129. (b) Medema, M. H.; Fischbach, M. A. Computational approaches to natural product discovery. *Nat. Chem. Biol.* **2015**, *11* (9), 639−648.

(7) (a) Ziemert, N.; Lechner, A.; Wietz, M.; Millan-Aguinaga, N.; Chavarria, K. L.; Jensen, P. R. Diversity and evolution of secondary metabolism in the marine actinomycete genus Salinispora. *Proc. Natl. Acad. Sci. U. S. A.* **2014**, *111* (12), E1130−1139. (b) Labeda, D. P.; Goodfellow, M.; Brown, R.; Ward, A. C.; Lanoot, B.; Vanncanneyt, M.; Swings, J.; Kim, S. B.; Liu, Z.; Chun, J.; Tamura, T.; Oguchi, A.; Kikuchi, T.; Kikuchi, H.; Nishii, T.; Tsuji, K.; Yamaguchi, Y.; Tase, A.; Takahashi, M.; Sakane, T.; Suzuki, K. I.; Hatano, K. Phylogenetic study of the species within the family Streptomycetaceae. *Antonie van Leeuwenhoek* **2012**, *101* (1), 73−104.

(8) Doroghazi, J. R.; Albright, J. C.; Goering, A. W.; Ju, K. S.; Haines, R. R.; Tchalukov, K. A.; Labeda, D. P.; Kelleher, N. L.; Metcalf, W. W. A roadmap for natural product discovery based on large-scale genomics and metabolomics. *Nat. Chem. Biol.* **2014**, *10* (11), 963−968.

(9) Doroghazi, J. R.; Metcalf, W. W. Comparative genomics of actinomycetes with a focus on natural product biosynthetic genes. *BMC Genomics* **2013**, *14*, 611.

(10) Muliandi, A.; Katsuyama, Y.; Sone, K.; Izumikawa, M.; Moriya, T.; Hashimoto, J.; Kozone, I.; Takagi, M.; Shin-ya, K.; Ohnishi, Y. Biosynthesis of the 4-Methyloxazoline-Containing Nonribosomal Peptides, JBIR-34 and −35, in Streptomyces sp. Sp080513GE-23. *Chem. Biol.* **2014**, *21* (8), 923−934.

(11) Izumikawa, M.; Kawahara, T.; Kagaya, N.; Yamamura, H.; Hayakawa, M.; Takagi, M.; Yoshida, M.; Doi, T.; Shin-ya, K. Pyrrolidine-containing peptides, JBIR-126, −148, and −149, from Streptomyces sp. NBRC 111228. *Tetrahedron Lett.* **2015**, *56* (39), 5333−5336.

(12) Ju, K. S.; Gao, J.; Doroghazi, J. R.; Wang, K. K.; Thibodeaux, C. J.; Li, S.; Metzger, E.; Fudala, J.; Su, J.; Zhang, J. K.; Lee, J.; Cioni, J. P.; Evans, B. S.; Hirota, R.; Labeda, D. P.; van der Donk, W. A.; Metcalf, W. W. Discovery of phosphonic acid natural products by mining the genomes of 10,000 actinomycetes. *Proc. Natl. Acad. Sci. U. S. A.* **2015**, *112* (39), 12175−12180.

(13) (a) Miller, E. D.; Kauffman, C. A.; Jensen, P. R.; Fenical, W. Piperazimycins: cytotoxic hexadepsipeptides from a marine-derived bacterium of the genus Streptomyces. *J. Org. Chem.* **2007**, *72* (2), 323−330. (b) Stevens, C. L.; Nagarajan, K.; Haskell, T. H. The Structure of Amicetin. *J. Org. Chem.* **1962**, *27* (9), 2991−3005.

(14) (a) Muliandi, A.; Katsuyama, Y.; Sone, K.; Izumikawa, M.; Moriya, T.; Hashimoto, J.; Kozone, I.; Takagi, M.; Shin-Ya, K.; Ohnishi, Y. Biosynthesis of the 4-Methyloxazoline-Containing Nonribosomal Peptides, JBIR-34 and −35, in Streptomyces sp. Sp080513GE-23. *Chem. Biol.* **2014**, *21* (8), 923−934. (b) Motohashi, K.; Takagi, M.; Shin-Ya, K. Tetrapeptides possessing a unique skeleton, JBIR-34 and JBIR-35, isolated from a sponge-derived actinomycete, Streptomyces sp. Sp080513GE-23. *J. Nat. Prod.* **2010**, *73* (2), 226−228.

(15) Nozaki, H.; Kuroda, S.; Watanabe, K.; Yokozeki, K. Cloning of the gene encoding alpha-methylserine hydroxymethyltransferase from Aminobacter sp. AJ110403 and Ensifer sp. AJ110404 and characterization of the recombinant enzyme. *Biosci., Biotechnol., Biochem.* **2008**, 72 (11), 3002−3005.

(16) Rottig, 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.* **2011**, *39* (Web Server issue), W362−W367.

(17) Stachelhaus, T.; Mootz, H. D.; Marahiel, M. A. The specificity-conferring code of adenylation domains in nonribosomal peptide synthetases. *Chem. Biol.* **1999**, *6* (8), 493−505.

(18) Zhang, G.; Zhang, H.; Li, S.; Xiao, J.; Zhang, G.; Zhu, Y.; Niu, S.; Ju, J.; Zhang, C. Characterization of the amicetin biosynthesis gene cluster from Streptomyces vinaceusdrappus NRRL 2363 implicates two alternative strategies for amide bond formation. *Appl. Environ. Microbiol.* **2012**, *78* (7), 2393−2401.

(19) Zhao, Q.; He, Q.; Ding, W.; Tang, M.; Kang, Q.; Yu, Y.; Deng, W.; Zhang, Q.; Fang, J.; Tang, G.; Liu, W. Characterization of the azinomycin B biosynthetic gene cluster revealing a different iterative type I polyketide synthase for naphthoate biosynthesis. *Chem. Biol.* **2008**, *15* (7), 693−705.

(20) (a) Grabley, S.; Kretzschmar, G.; Mayer, M.; Philipps, S.; Thiericke, R.; Wink, J.; Zeeck, A. Secondary Metabolites by Chemical Screening, 24. Oasomycins, New Macrolactones of the Desertomycin Family. *Liebigs Annalen der Chemie* **1993**, *1993* (5), 573−579. (b) Uri, J. V. Desertomycin: a potentially interesting antibiotic. *Acta Microbiol. Hung.* **1986**, *33* (4), 271−283. (c) Uri, J.; Bognar, R.; Bekesi, I.; Varga, B. Desertomycin, a new crystalline antibiotic with antibacterial and cytostatic action. *Nature* **1958**, *182* (4632), 401.

(21) Mikami, Y.; Yokoyama, K.; Omi, A.; Arai, T. Identification of producer and biological activities of new antibiotics, mimosamycin and chlorocarcins. *J. Antibiot.* **1976**, *29* (4), 408−414.