

STRATEGIES FOR THIAZOLE/OXAZOLE-MODIFIED MICROCIN DISCOVERY

BY

COURTNEY LYNNE COX

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Microbiology
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2015

Urbana, Illinois

Doctoral Committee:

Assistant Professor Douglas A. Mitchell, Chair
Associate Professor Rachel J. Whitaker
Professor Gary J. Olsen
Professor Willem A. van der Donk

ABSTRACT

Natural products continue to be an important source of therapeutically-relevant compounds. With the advent of inexpensive genome sequencing it has become apparent that bacteria produce a larger array of natural products than was previously believed. This new wealth in sequence data has potential to be helpful for the discovery of novel compounds by using genome mining. Although strategies of genome mining have become more efficient and capable of identifying novel biosynthetic gene clusters, it remains difficult to correlate gene clusters with natural products. In this dissertation I discuss limitations with the current methods of genome mining and correlating individual natural products with gene clusters. Furthermore, I characterize a rapidly growing family of natural products, the thiazole/oxazole-modified microcins (TOMMs), and discuss novel methods used to correlate the gene clusters to natural products from this family of metabolites. In chapter 2, I establish the sequence diversity and structural capability of bacteria and archaea to produce TOMM natural products. This genome mining characterization was used to identify nine novel classes of TOMMs, including one class from archaeal producers. In chapter 3, I discuss the utilization of genetic techniques to identify and isolate the TOMM natural product from the archaeal species *Sulfolobus acidocaldarius*. I demonstrate that although genetic manipulation has been previously used for the identification of natural products, comparative metabolomics is difficult to use for routine identification of low-abundance natural products such as the TOMM from *S. acidocaldarius*. Very few methods have been created to identify natural products from particular gene clusters. Therefore, in chapter 4, I discuss the creation of a novel method for the rapid identification of natural products following bioinformatics prioritization of antibiotic producing strains. This method utilizes the combination of genome mining and the chemical reactivity of natural products to discover new compounds. Dehydrated amino acids are modified residues commonly found in natural products such as TOMMs. I utilize the mild electrophilic chemical reactivity of dehydrated amino acids to label these natural products using a

soft nucleophilic probe. These labeled natural products were easily detected using comparative mass spectrometry. Bacterial strains were prioritized by the genome mining established in chapter 2 to reduce the screening time to find a novel natural product. This dissertation presents the addition of novel genome mining and natural product discovery techniques to increase the discovery and production of therapeutically-relevant compounds.

To Grandpooops for my determination and Whoopsie for my love of shoes and lobster

ACKNOWLEDGEMENT

I must first and foremost thank my advisor, Professor Doug Mitchell. Your tireless guidance in my research and your endless encouragement to break down the walls of science exceeded my expectations and will not be forgotten. I would like to thank my committee members, Professor Rachel Whitaker, Professor Gary Olsen, Professor Wilfred van der Donk, and Professor Bill Metcalf, for all of your kind support and thoughtful suggestions. I am grateful to all of my collaborators, particularly James Doroghazi for his unrivaled support in my bioinformatics endeavors. I must also thank members of the Whitaker lab for their patience and scientific assistance while I was working with *S. acidocaldarius*. I am grateful to all members of the Mining Microbial Genomes theme at the Institute for Genomic Biology. I could not imagine a more helpful or compassionate group of scientists to work with.

I am incredibly lucky to have worked with such wonderful colleagues. I would like to thank all current and former members of the Mitchell lab for making my time at UIUC unforgettable. Particularly, I must thank Joel Melby and Jonathan Tietz for their essential collaborations on the reactivity-based screening project. Katie Molohon and Kate Woodall, I would not have survived the advisory or prelim without you nor will I ever meet anyone that is as good at rolling their letters. Karol Sokolowski was the *best* undergraduate researcher and I am grateful for his never ending enthusiasm for the research projects we worked together on. I have had quite an adventure living in Champaign-Urbana and none of that would have been possible without all of the friends I have made here. I am thankful for Gwendolyn Humphreys, Chelsea Lloyd, Kate Woodall, Katie Molohon, Caitlin Deane, Patti Blair, Michelle Goettge, Joel Melby, Jason Bouvier, BFF, James Doroghazi (and family), and Brian San Francisco. I will never forget the girl's nights, game nights or rowdy nights. To the members of Six Innings, this is our year to bite the Bullet and win the finals!

Finally, I would like to thank my mommy for showing me you can be smart and beautiful, my dad for supporting both my scientific career and my love for shoes, my Kalon for his unending hours of lessons in math and computer science, my brother Chris for continually pushing me to be my best, my brother Kaden for teaching me that all dreams take hard work, my entire extended family, and Luke for his never ending support. You not only instilled in me a love for science, but have also shown unconditional love throughout all of my pursuits, and for that I am truly thankful.

TABLE OF CONTENTS

CHAPTER I: INTRODUCTION	1
1.1 Genome Mining for Natural Product Discovery	1
1.2 Ribosomally Synthesized and Post-Translationally Modified Peptide Natural Products	2
1.3 Thiazole/Oxazole-Modified Microcins (TOMMs)	3
1.4 TOMM Machinery for Genome Mining.....	4
1.4.1 Microcin B17	4
1.4.2 Streptolysin S	5
1.4.3 Cyanobactins.....	6
1.4.4 Thiopeptides.....	6
1.4.5 Plantazolicin.....	7
1.4.6 Bottromycin	7
1.5 Natural Product Discovery Methods Potentiated by Genome Mining.....	8
1.5.1 Gene deletions and heterologous production using comparative metabolic profiling	8
1.5.2 Chemoselective enrichment, genomisotopic labeling, and selective enzymatic derivatization.....	9
1.5.3 Mass spectrometry guided isolation.....	10
1.5.4 Reactivity-based screening	11
1.6 Summary and Outlook	12
1.7 Figures	15
1.8 References.....	18
CHAPTER II: THE GENOMIC LANDSCAPE OF RIBOSOMAL PEPTIDES CONTAINING THIAZOLE AND OXAZOLE HETEROCYCLES	25
2.1 Introduction.....	26
2.2 Genome Mining and Isofunctional Grouping	28
2.3 Isofunctional Groups with Explored TOMMs.....	32
2.3.1 Microcin B17	32
2.3.2 Cytolysin.....	32
2.3.3 Cyanobactin	34
2.3.4 Nitrile hydratase-related leader peptides and Nif11-related precursor peptides	35
2.3.5 Goadsporin.....	36

2.3.6 Thiopeptides.....	36
2.3.7 Plantazolicin.....	37
2.3.8 Hakacin	38
2.3.9 Heterocycloanthracin	38
2.3.10 Bottromycin and other TOMMs with a stand-alone D protein	40
2.4 Presumed Isofunctional Groups with no Characterized Members.....	41
2.4.1 Faecalisin	41
2.4.2 Propionisin	41
2.4.3 Helicobactin	42
2.4.4 Mobilisin	42
2.4.5 Haloazolisin	43
2.4.6 Thermoacidophsin	44
2.4.7 Gallolytisin.....	44
2.4.8 Anabaenasin.....	45
2.4.9 Coryneazolisin type 1 and type 2.....	45
2.5 Distribution of TOMM Gene Clusters.....	46
2.6 Summary and Outlook	47
2.7 Experimental.....	47
2.7.1 Biosynthetic gene cluster discovery and comparison	47
2.7.2 Sequence similarity networks	48
2.7.3 Precursor sequence discovery	48
2.7.4 Phylogenetic analysis.....	49
2.8 Figures	50
2.9 References.....	64
CHAPTER III: TOMM BIOSYNTHESIS IN ARCHAEA	72
3.1 Introduction.....	72
3.1.1 Thiazole/oxazole-modified microcins.....	74
3.1.2 Sulfolobus acidocaldarius TOMM cluster	75
3.1.3 Other related TOMM clusters	76
3.2 Multipronged Approach.....	77
3.3 Heterologous Expression and Purification of the <i>S. acidocaldarius</i> TOMM Proteins.....	78

3.4 Heterologous Expression of the <i>Bacillus cereus</i> Rock3-44 TOMM Proteins	78
3.5 MS of Unmodified Precursor Peptides	79
3.6 RT-PCR.....	79
3.7 Gene Deletion Strains.....	80
3.8 Fosmid Library Generation	81
3.9 Summary and Outlook	82
3.10 Experimental	82
3.10.1 Protein overexpression and purification	82
3.10.2 MALDI-TOF mass spectrometric analysis of the precursor peptides.....	83
3.10.3 <i>Sulfolobus</i> cultivation conditions	83
3.10.4 RT-PCR.....	83
3.10.5 SaciC gene deletion production	84
3.10.6 Fosmid library generation	84
3.10.7 Liquid chromatography mass spectrometric comparisons	85
3.11 Figures.....	86
3.12 References.....	95
CHAPTER IV: NUCLEOPHILIC 1,4-ADDITIONS FOR NATURAL PRODUCT DISCOVERY	98
4.1 Introduction.....	98
4.2 Rationale and Overview of a New Natural Product Discovery Method	100
4.3 Validation of the DTT-Labeling Strategy	101
4.4 Bioinformatics Guided Strain Prioritization	103
4.5 MS-Based Screening of Prioritized Strains.....	104
4.6 Verification of the Cyclothiazomycin C Structure.....	105
4.7 Conservation Analysis of the Cyclothiazomycin C Biosynthetic Gene Cluster	106
4.8 Assessment of Cyclothiazomycin Bioactivity	107
4.9 Summary and Outlook	108
4.10 Experimental	109
4.10.1 Preparation of cell extracts for screening.....	109
4.10.2 DTT-labeling.....	110
4.10.3 Bioinformatics based strain prioritization.....	110
4.10.4 MALDI-TOF mass spectrometric analysis	111

4.10.5 Isolation of cyclothiazomycin C	111
4.10.6 Isolation of cyclothiazomycin B	112
4.10.7 FT-MS/MS analysis of cyclothiazomycin B and C	112
4.10.8 NMR spectroscopy of cyclothiazomycin C	113
4.10.9 Analysis of NMR data.....	113
4.10.10 Evaluation of cyclothiazomycin B and C antibiotic activity.....	114
4.10.11 Evaluation of cyclothiazomycin B and C antifungal activity	114
4.11 Figures and Tables.....	116
4.12 References.....	139
APPENDIX A: FURTHER PUBLICATIONS WITH MINOR CONTRIBUTIONS	144
A.1 Discovery of a New ATP-Binding Motif Involved in Peptidic Azoline Biosynthesis	144
A.2 HIV Protease Inhibitors Block Streptolysin S Production.....	174
A.3 Undecaprenyl Diphosphate Synthase Inhibitors: Antibacterial Drug Leads	185
A.4 References.....	195

CHAPTER I: INTRODUCTION

I am grateful to Joel Melby, Brian San Francisco, and Jonathan Tietz for critically editing this chapter.

1.1 Genome Mining for Natural Product Discovery

Natural products have historically been the most prolific source of antibiotics (Newman, *et al.* 2012). Nearly 80% of all approved antibacterials are natural products or derivatives thereof (Newman, *et al.* 2012). Some notable examples include ciprofloxacin (DNA synthesis inhibitor) (Wise, *et al.* 1983), chloramphenicol (protein synthesis inhibitor) (Gottlieb, *et al.* 1954), and rifampicin (RNA polymerase inhibitor) (Sensi, *et al.* 1959). Traditionally, scientists have relied on bioactivity-based screening to identify fractions of bacterial crude extracts for compounds with antibiotic properties. Difficulties intrinsic to this discovery method, including frequent rediscovery and necessarily large, expensive screening efforts, have led most large pharmaceutical companies to reduce or eliminate their antibiotic discovery efforts (Lewis 2013, Payne, *et al.* 2007). This has coincided with the emergence of multidrug-resistant strains of pathogenic bacteria, accelerated by the overuse and misuse of antibiotics. It has been estimated that the U.S. alone spends \$20 billion annually on direct healthcare costs to treat patients with antibiotic-resistant bacterial infections (U.S. Department of Health and Human Services 2013). Therefore, it is imperative that novel methods for identifying antibacterial therapies be developed. With the maturation of genome sequencing, it has become apparent that the collection of characterized natural products represents only a fraction of the true genomic potential of microbes (Bentley, *et al.* 2002). Consequently, genome mining is emerging as a powerful method for the identification and isolation of novel compounds (Bachmann, *et al.* 2014, Challis 2008, Deane, *et al.* 2014, Doroghazi, *et al.* 2014, Doroghazi, *et al.* 2013, Velasquez, *et al.* 2011).

Genome sequencing has profoundly changed the discovery process of natural products by providing access to the biosynthetic potential of microbes prior to the intense traditional

screening process (Bachmann, *et al.* 2014, Challis 2008, Van Lanen, *et al.* 2006). The recognition that even the well-characterized bacterium *Streptomyces coelicolor* harbors many additional biosynthetic pathways beyond its characterized natural products (Bentley, *et al.* 2002) led to the wide acceptance that many microbes encode a plethora of biosynthetic gene clusters with unknown natural products (Baltz 2008, Fischbach, *et al.* 2009, Jensen, *et al.* 2014). This observation implies that many microbial natural products are either not produced in laboratory conditions or are not being detected using the traditional discovery techniques. Genome mining provides an approach wherein strains can be prioritized for their biosynthetic potential, while providing researchers with information used to avoid re-isolating known compounds, forecast the properties of the expected compounds, or even, in some cases, make *exact* predications of the final natural product structure, all of which enhance the discovery platform. Many bioinformatics tools have emerged to characterize novel biosynthetic pathways (Blin, *et al.* 2013, de Jong, *et al.* 2006, Li, Qu, *et al.* 2012, Mohimani, *et al.* 2014), the majority of which have focused on two common, predictable classes of natural products: type I polyketide synthases (PKS) and nonribosomal peptide synthetases (NRPS). Recently, genome characterization of ribosomally produced natural products (RiPPs) has shown that the biosynthetic and structural potential of this class is much more diverse than originally believed, meaning increased effort should be directed at bioinformatic evaluation of these types of molecules (Lee, *et al.* 2008, Letzel, *et al.* 2014, Velasquez, *et al.* 2011).

1.2 Ribosomally Synthesized and Post-Translationally Modified Peptide Natural Products

Bacterial natural product research over the past century has largely been focused on nonribosomal peptide (NRP) and polyketide (PK) natural products. There are many examples of NRP and PK natural products or derivatives that have been approved as antibiotics (Fischbach, *et al.* 2006). Additionally, genome sequencing combined with enzymatic characterization of the PKS and NRPS machinery has bolstered the genome mining approach for NRP and PK natural

product discovery (Fischbach, *et al.* 2006). More recently, genome sequencing has revealed another widely distributed subgroup of natural products, RiPPs (Haft, *et al.* 2010, Lee, *et al.* 2008, Letzel, *et al.* 2014, Maksimov, *et al.* 2014, Velasquez, *et al.* 2011). RiPPs comprise both a diverse chemical and genetic landscape, including but not limited to lanthipeptides, thiazole/oxazole-modified microcins (TOMMs), lasso peptides, and linaridins. In all cases a ribosomally produced precursor peptide undergoes modification by a set of tailoring enzymes (Figure 1.1) (Arnison, *et al.* 2013, Dunbar, *et al.* 2013). These modifications include cyclizations, dehydrations, methylations, and disulfide bond formations, among others (Arnison, *et al.* 2013). The precursor peptide sequence and posttranslational modifications govern the structure of the final product; this template-based biosynthetic strategy simplifies structural predictions from genomic information. Of the characterized RiPPs, all show diverse combinations of chemical modifications and structures leading to a plethora of bioactivities. This simple peptide-oriented strategy leads to highly evolvable biosynthetic pathways with the capability to produce structurally diverse compounds (variable templates) using minimal genetic space. Additionally, microbes can further evolve the natural product with the addition or deletion of modification enzymes (variable tailoring).

1.3 Thiazole/Oxazole-Modified Microcins (TOMMs)

Thiazole/oxazole-modified microcins (TOMMs) comprise a subclass of RiPPs characterized by the presence of nitrogenous five-membered heterocycles derived from cysteine, serine, and threonine residues (Li, *et al.* 1996, Mitchell, *et al.* 2009). The hallmark of a TOMM gene cluster is the presence of a cyclodehydratase (termed the C and D protein complex if separate or the CD fusion protein if merged) that executes an ATP-dependent cyclization of cysteine, serine, and threonine residues to form the thiazoline and (methyl)oxazoline (azoline) heterocycles (Figure 1.1) (Dunbar, *et al.* 2014, Dunbar, *et al.* 2012, Dunbar, *et al.* 2013). Some TOMM gene clusters also contain an FMN-dependent dehydrogenase (termed the B protein) that catalyzes the oxidation of azoline heterocycles to their respective azoles (Melby, *et al.* 2014).

These heterocycles restrict the conformational flexibility to provide the peptide with a rigid structure, which ultimately endows the final product with a specific activity. TOMM biosynthetic gene clusters often contain additional tailoring enzymes that increase structural diversity within the TOMM family. As with many bacterial natural product classes, the enzymes responsible for heterocycle formation are often clustered in biosynthetic pathways, making them easily discernable by genome mining (Figure 1.1) (Lee, *et al.* 2008, Velasquez, *et al.* 2011).

TOMM biosynthetic genes clusters have been identified in both bacterial and archaeal genomes (Lee, *et al.* 2008). There are many categories of TOMM clusters, defined by the type of natural product produced. Examples of studied TOMMs include microcin B17 (DNA gyrase inhibitor) (Belshaw, *et al.* 1998), streptolysin S (cytolysin) (Mitchell, *et al.* 2009, Molloy, *et al.* 2011), thiopeptides (ribosome inhibitor) (Just-Baringo, *et al.* 2014), and the cyanobactins (anticancer properties)(Schmidt, *et al.* 2009) (Figure 1.2). Although some TOMMs have been extensively characterized, the majority of TOMMs are unknown and could be an unexploited reservoir of therapeutically useful compounds (Melby, *et al.* 2011).

1.4 TOMM Machinery for Genome Mining

Genome mining for TOMMs requires an understanding of the prevalence and function of the genes essential for their biosynthesis. Progress in recent years has shed light on these biosynthetic pathways and the biosynthetic enzymes responsible for natural product maturation.

1.4.1 Microcin B17

Microcin B17 (MccB17), the first characterized TOMM, is a DNA gyrase inhibitor produced by select *E. coli* strains. In a pioneering study, the proteins responsible for MccB17 production were discovered by genome mining and reconstituted, allowing for the general biosynthetic route for the production of TOMMs to be determined (Li, *et al.* 1996). In this study it was shown that the trimeric complex composed of McbB (C protein), McbC (B protein), and McbD (D protein) installed thiazole and oxazole moieties onto the McbA precursor peptide. This

seminal publication as well as follow up studies introduced the minimal set of modification enzymes necessary for the production of TOMMs. The functions of these enzymes have been further elucidated in other TOMM systems. The D protein was demonstrated to catalyze the cyclodehydration reaction to convert a subset or all Cys, Ser, and Thr residues to thiazolines and (methyl)oxazolines (Dunbar, *et al.* 2014, Dunbar, *et al.* 2012, Dunbar, *et al.* 2013). The C protein is responsible for the recognition of the peptide substrate and the regulation of the D protein activity (Dunbar, *et al.* 2013). In MccB17 biosynthesis a third protein is necessary for the installation of heterocycles. The B-protein is a flavin mononucleotide (FMN)-dependent dehydrogenase that oxidizes the azoline heterocycles to azoles.

1.4.2 Streptolysin S

Although the β -hemolytic phenotype from group A *Streptococcus* was first identified over a century ago, the genes responsible for the biosynthesis of streptolysin S (SLS), a toxin that causes the phenotype, were not determined until 1998 (Betschel, *et al.* 1998, Nizet, *et al.* 2000). In a transposon mutational study, the entire *SLS-associated gene (sag)* operon was discovered. Similar to MccB17 production, the biosynthetic gene cluster is composed of SagBCD enzymes that are required for the posttranslational modification of the precursor peptide, SagA. The Sag enzymes are similar enough to the MccB17 enzymes that they are capable of modifying the MccA precursor peptide to produce an active gyrase inhibitor (Mitchell, *et al.* 2009). Characterization of these two clusters led to the discovery that it was precursor sequence and not modification enzymes that dictated final product activity (Lee, *et al.* 2008, Mitchell, *et al.* 2009). Furthermore, the discovery of two similar clusters in disparate organisms prompted researches to use genome mining to identify similar cluster with homologs of the B, C, and D proteins, which lead to the identification of the entire family of TOMM natural product gene clusters (Velasquez, *et al.* 2011). Although the MccB17 and Sag operons require only the BCD enzymes for

modification, characterization of other TOMM clusters revealed that many clusters have additional modification enzymes to enhance the chemical structures.

1.4.3 Cyanobactins

Cyanobactins are a family of small macrocyclized compounds produced by cyanobacteria. In a pioneering study, it was determined that cyanobactins were not produced by a sea squirt, but rather a bacterial symbiont, *Prochloron didemni* (Schmidt, *et al.* 2005).

Furthermore, it was identified that these cyanobacterial natural products were indeed ribosomally produced and heterocycles were installed in a similar manner to MccB17 production.

Cyanobactins have many activities including antitumor, cytotoxic, and multi-drug reversing effects (Sivonen, *et al.* 2010). A key difference between cyanobactins and other TOMMs is that after installation of the heterocycles, the cyanobactin peptides are cleaved from the precursor by not one but two proteases. In the patellamide biosynthetic cluster, PatA was shown to cleave the N-termini and PatG the C-termini. PatG also catalyzes the head-to-tail cyclization of the peptides. The permissive substrate tolerance of PatG was demonstrated through the characterization of macrocyclization of synthetic substrates with varying lengths and amino acid compositions (Agarwal, *et al.* 2012, Lee, *et al.* 2009).

1.4.4 Thiopeptides

Thiopeptides are a family of TOMMs with highly modified macrocyclic core peptides defined by a central pyridine ring. Although the first of this class was described in 1948 (Su 1948), the biosynthetic machinery responsible for producing thiopeptides was not determined until 2009, when four independent research groups reported the genes responsible for the production of multiple thiopeptides; at this time it became apparent that thiopeptides were manufactured in a similar fashion to linear TOMMs such as microcin B17 (Crone, *et al.* 2012, Gomez-Escribano, *et al.* 2012, Hou, *et al.* 2012, Huo, *et al.* 2012). Furthermore, nearly all identified thiopeptide gene clusters contain a lanthipeptide-like dehydratase that simply

dehydrates, rather than cyclodehydrates, select Ser and Thr residues to form dehydroalanine (Dha) and dehydrobutyrines (Dhb) (Li, *et al.* 2010, Ortega, *et al.* 2014). Another TOMM, goadsporin – a promoter of secondary metabolism and morphogenesis in actinomycetes – also contains enzymes to install Dha and Dhb residues, but is not macrocyclized like the thiopeptides.

It is widely believed that in thiopeptide biosynthesis two dehydroalanines undergo an enzyme-catalyzed formal [4+2] cycloaddition to form the central 6-membered nitrogen ring which can be found in various oxidation states. This was first supported by feeding experiments, and confirmed by the identification of the ‘hetero-Diels-Alderase’ enzyme responsible for the reaction (TclM in thiocillin biosynthesis) (Bowers, *et al.* 2010).

1.4.5 Plantazolicin

Plantazolicin (PZN) is a TOMM from the soil bacterium *Bacillus amyloliquefaciens* FZB42 recently discovered using gene deletions and mass spectrometry. It has highly discriminating antibacterial activity against *Bacillus anthracis*, the causative agent of anthrax (Scholz, *et al.* 2011). The PZN precursor peptide is modified by an extremely specific S-adenosylmethionine (SAM)-dependent methyltransferase present in the gene cluster (Lee, *et al.* 2013, Molohon, *et al.* 2011). The X-ray crystal structure of the methyltransferase revealed a deep and narrow cavity for substrate binding, supporting the specificity of the enzyme for modification only of PZN derivatives (Lee, *et al.* 2013).

1.4.6 Bottromycin

Bottromycin is a TOMM with potent antimicrobial activity against methicillin resistant *Staphylococcus aureus* (MRSA) and vancomycin resistant *Enterococcus* (VRE). Characterized bottromycin gene clusters contain two genes with a YcaO-like domain similar to the D proteins, but they contain no distinguishable C protein, which is used in other TOMM biosynthesis to potentiate the D protein’s activity (Crone, *et al.* 2012, Gomez-Escribano, *et al.* 2012, Hou, *et al.*

2012, Huo, *et al.* 2012). This is the first example of a TOMM that does not contain both a C and D protein, complicating genome mining for TOMMs.

There are two groups of YcaO domain-containing proteins (homologs of D proteins not associated with a C protein): the non-TOMM YcaOs and the TfuA-associated non-TOMM YcaOs. The latter is found co-occurring in clusters with a gene encoding for the protein TfuA, which has been implicated in trifolitoxin production (Breil, *et al.* 1996). Although all of these YcaO proteins contain the canonical ATP-binding pocket, the substrate of the non-TOMM YcaOs are unknown (Dunbar, *et al.* 2014). With the elucidation of bottromycin biosynthesis, it became apparent that YcaO domain-containing proteins have the potential to make natural products even without a canonical C protein, lending previously unanticipated potential to many uncharacterized ‘stand-alone’ YcaO proteins.

1.5 Natural Product Discovery Methods Potentiated by Genome Mining

While genome mining has become an extremely useful tool for natural product discovery, it has also become apparent that finding natural products produced by certain gene clusters is much more difficult than identifying the gene clusters of interest. This has created a roadblock in the discovery process, generating an overabundance of gene clusters with unknown natural products. It is imperative to develop novel strategies to quickly identify natural products from predicted gene clusters in order to realize the full potential of genome mining-based natural product discovery. Therefore many research groups have begun to introduce strategies to connect genome clusters to natural products.

1.5.1 Gene deletions and heterologous production using comparative metabolic profiling

Perhaps the most definitive method to establish the connection between genes and natural products is through the genetic deletion of production enzymes. Traditionally, comparison of the wild type strain with the genetic deletion mutant, using metabolomics (using mass spectrometry) or bioactivity, allows for the identification of natural products (Figure 1.3 A). Additionally, the

discovery of plantazolicin highlights a secondary use for genetic deletions, in which deletions of other natural product clusters were used to reduce the interfering mass spectrometry signal from abundant natural products, thus unmasking the less abundant natural product: plantazolicin. Furthermore, genetic deletions of the necessary modification machinery for the production of plantazolicin were made to confirm the gene cluster (Scholz, *et al.* 2011).

In the event that the original host is not genetically tractable, the genes from a biosynthetic cluster can instead be transferred to a genetically tractable host for the production and identification of natural products. Recently, genome mining identified the gene cluster responsible for the production of the lasso peptide natural products including, caulosegnins I-III (Hegemann, Zimmermann, Xie, *et al.* 2013, Maksimov, *et al.* 2014, Maksimov, *et al.* 2012). Although small amounts of caulosegnin I and II were detected after cultivation of the native hosts, the production levels were too low for isolation. Heterologous expression of these gene clusters was optimized in an *E. coli* host to produce large enough quantities of each natural product for isolation and structure elucidation (Hegemann, Zimmermann, Xie, *et al.* 2013). A follow-up study combined genome mining with this optimized strain of *E. coli* to identify a further 10 lasso peptides from sequenced Proteobacteria (Hegemann, Zimmermann, Zhu, *et al.* 2013). This method can be used for not only genetically intractable hosts, but also to distinguish between products in hosts that produce a variety.

1.5.2 Chemoselective enrichment, genomisotopic labeling, and selective enzymatic derivatization

Detection of lower-abundance metabolites is often hindered by the overwhelming signal of abundant natural products. Therefore, Erin Carlson and coworkers devised a strategy to enrich for a specific subset of natural products from a biological systems. Therein, derivatized-resin chemical probes, designed to target specific functional groups or post-translational modifications, were used for the chemoenrichment of specific natural products. This method was initially used

for the identification of biomarkers and therapeutic targets (Carlson, *et al.* 2007). Recently, the technology has been used to enrich for specific natural products within crude bacterial extract (Figure 1.3 B) (Odendaal, *et al.* 2011). When combined with genome-guided bacterial prioritization, this becomes a robust natural product discovery approach.

The genomisotopic approach combines genome mining of natural product gene clusters with isotope-guided fractionation to isolate novel natural products. This method was used for the isolation of the nonribosomal peptide (NRP) orfamide A. The researchers first identified a gene cluster expected to produce a novel NRP. An isotope-labeled amino acid that was predicted to exclusively be incorporated into the unknown compound was then selected, and incorporation was detected using NMR (an isotope-sensitive technique) to guide identification and isolation (Gross, *et al.* 2007). These labeling tools represent a good alternative to heterologous expression for clusters that are too large or difficult to express in a heterologous host.

To that same end, an enzymatic approach was created to rapidly enrich phosphonate natural products from the spent medium of bacteria (Gao, *et al.* 2014). Phosphonate natural products have high polarity and water solubility, making them relatively difficult to purify. The phosphonate O-methyltransferase DhpI, from the dihydrophos biosynthetic pathway, was found to non-specifically methylate other phosphonate natural products (Lee, *et al.* 2010). Gao *et al.* accordingly used DhpI to develop a method to label novel phosphonates in bacterial spent medium with either a CH₃ or CD₃ group. This altered stable isotope composition is detectable using mass spectrometry. This method was used to identify and purify novel phosphonate biosynthetic intermediates from the strain *Streptomyces* sp. WM6372 (Gao, *et al.* 2014).

1.5.3 Mass spectrometry guided isolation

Mass spectrometry (MS) has traditionally been an essential technique for the identification and structural elucidation of many natural products. However, recently, a MS method was created particularly to correlate biosynthetic gene clusters to natural products

(Kersten, *et al.* 2011). Initial MS data was obtained from a bacterial organism of interest. Putative natural products from the initial MS were then subjected to tandem MS and analyzed for a “sequence tag”. The sequence tag—generated from the mass shift sequence in the tandem MS—represents a subset of amino acids that are in the natural product and can subsequently be used as a genome-mining query for the identification of the gene cluster (Figure 1.3 C). This method can be used to identify both ribosomal and non-ribosomal natural products, provided they are peptidic. Furthermore, common post-translational modifications were taken into account for the generation of the sequence database, expanding the method’s utility for identifying natural products from uncharacterized gene clusters. The tag-based search strategy was used for the identification of nine novel natural products and their associated gene clusters (Kersten, *et al.* 2011).

This technique was expanded with the introduction of tandem MS networking (Nguyen, *et al.* 2013). MS/MS networking was used to relate all detectable metabolites according to their fragmentation spectra—which in turn depend on each metabolite’s molecular structure. The MS-based genome mining from the previous study was then used to correlate the MS network clusters with gene clusters (Figure 1.3 C). This method was shown to be capable of identifying natural products from unsequenced organisms by utilizing the MS network to identify similar natural products from a sequenced organism. This natural product production data can be extrapolated quickly to new samples solely based on MS signatures without the effort in genome sequencing (Nguyen, *et al.* 2013).

1.5.4 Reactivity-based screening

Reactivity-based screening (RBS) utilizes the intrinsic chemical reactivity of natural products to create probes that allow for the easy MS detection of natural products. RBS relies on the predictable presence of functional groups with known, selective reactivity, such as dehydrated amino acids (Cox, *et al.* 2014).

The dehydrated amino acids (DHAAs) dehydroalanine and dehydrobutyrine are frequently found in natural products, in particular the thiopeptide subclass of TOMMs. As mentioned previously, almost every identified thiopeptide gene cluster contains a lanthipeptide-like dehydratase that eliminates water from Ser and Thr residues to give α,β -unsaturated amino acids. These dehydratases have previously been used in conjunction with CD fusion proteins and the 'hetero-Diels-Alderase' to identify thiopeptide gene clusters (Li, Qu, *et al.* 2012).

It has been demonstrated that thiol nucleophiles participate in nucleophilic 1,4-addition to α,β -unsaturated carbonyl/imine DHAAs under mild conditions to yield covalent thioether adducts. Thus, it was envisioned this well-established, reliable chemistry could be used as a chemical handle for the discovery of DHAAs-containing natural products. First, bioinformatically identified bacteria containing a lanthipeptide-like dehydratase are grown, and the natural products are extracted with an organic solvent. A portion of this extract then undergoes treatment with dithiothreitol (DTT) in the presence of base. DTT was chosen as the thiol probe owing to its low cost and ubiquity in many chemical laboratories. If reactive DHAAs moieties are present in the cell-surface extract, the resulting DTT adducts increase the mass of the exported metabolite by multiples of 154.0 Da. Differential mass spectrometry between the unreacted control and the DTT-reacted extracts readily identifies the compounds containing DHAAs within a predetermined mass range. This reactivity-based screening method, when combined with bioinformatic analysis of biosynthetic gene clusters to prioritize strains, can be used to quickly identify, isolate, dereplicate and characterize natural products (Cox, *et al.* 2014, Li, Girard, *et al.* 2012).

1.6 Summary and Outlook

TOMMs are a large class of peptidic natural products that are post-translationally modified to install heterocycles. Some TOMMs have been extensively characterized, including SLS and MccB17, however the majority of TOMMs are unknown natural products which

represent an untapped reservoir of potentially therapeutically useful compounds. Given the remarkable structural and functional diversity of characterized TOMMs, a basic understanding of the synthetic capability of microbes to produce TOMMs is desirable. TOMMs are produced ribosomally, making them easily identified using bioinformatics by identification of the modification enzymes. In Chapter 2, we use genome mining to characterize the diversity of TOMM biosynthetic gene clusters. We identify 20 families of TOMMs, 9 of which had never been characterized. In Chapter 2, we also discuss the phylogenetic diversity of the identified TOMM clusters along with the precursor peptide and modification enzyme diversity.

Traditional natural product discovery often identifies compounds using activity based screening. Because of rediscovery, without a strain prioritization step this method becomes burdensome, forcing scientists to screen thousands (or millions) of strains to find novel compounds. Furthermore, many compounds are missed because they lack the single, specific activity that is used in the screen. With the advent of sequencing it became apparent that many compounds were not being identified using this activity screening technique. In Chapter 3, we identify a TOMM biosynthetic gene cluster from *Sulfolobus acidocaldarius* and use this gene cluster to attempt to characterize the natural product. This is one of the few TOMM gene clusters identified in an Archaeal species and interestingly, similar clusters were also identified in bacterial producers. Although the natural product remains elusive, we outline some preliminary characterizations of this cluster and similar bacterial clusters.

Genome mining has become an extremely useful tool to identify novel biosynthetic gene clusters, and here we identify nine novel families of TOMMs. However, as suggested in Chapter 3, identification of novel compounds produced from these gene clusters of interest is much more difficult than identifying the gene clusters themselves. Therefore, new, rapid methods to correlate compounds with biosynthetic gene clusters are desired. In Chapter 4, we discuss a novel method for the identification of compounds containing DHAAs. Our strategy uses a commercially

available thiol, DTT, for the covalent labeling of activated alkenes by nucleophilic 1,4-addition. Modification is easily discerned by comparing MS of reacted and unreacted cell surface extracts. We combined this reactivity-based screening method with the bioinformatics analysis of TOMMs to prioritize strains capable of producing natural products with DHAAs. We anticipate, with the rise in sequencing data, that this method, along with other chemoselective reactions, will provide a powerful tool to correlate natural products with biosynthetic gene clusters.

1.7 Figures

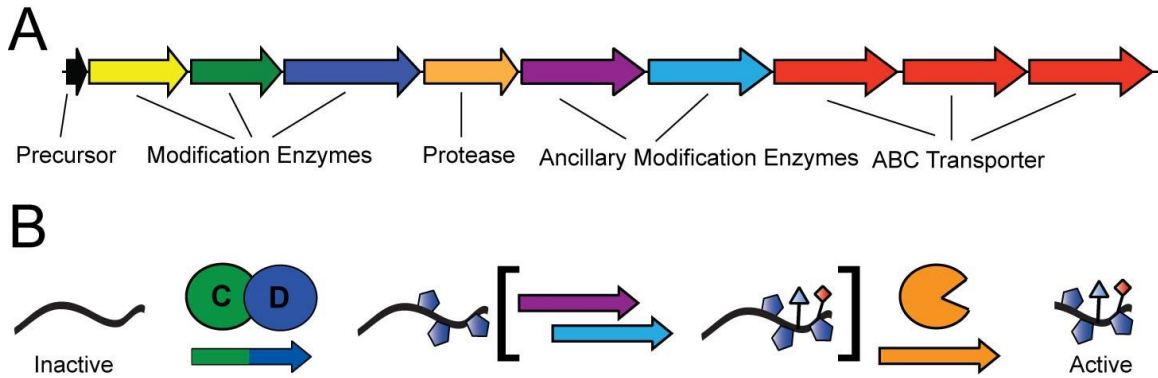


Figure 1.1 | Overview of a TOMM gene cluster and biosynthesis. (A) Example of a TOMM biosynthetic gene cluster. (B) The precursor peptide (black) is modified by the cyclodehydratase (C and D proteins, green and blue) to install heterocycles. Heterocycles can be further oxidized with the dehydrogenase (B protein, yellow). Additionally, further modifications can be installed onto the precursor peptide by ancillary modification enzymes. A protease (orange) then cleaves off the leader peptide (or follower). Further modifications can occur after leader peptide cleavage (not depicted).

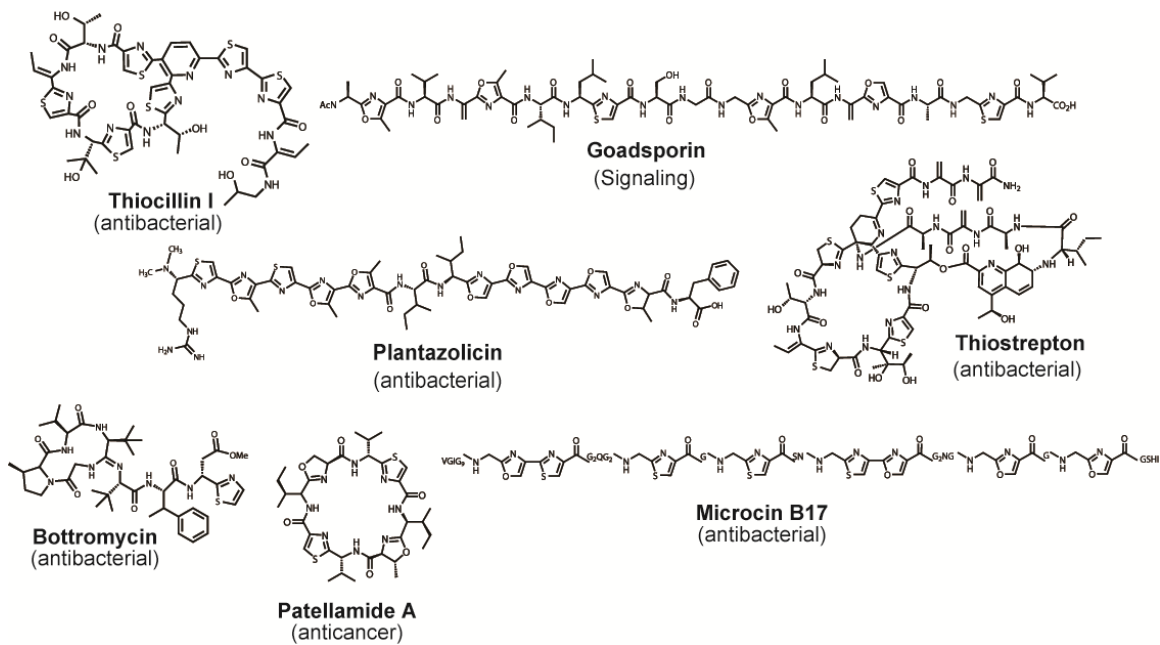


Figure 1.2 | Representative TOMM natural products. Examples of structures of TOMMs from different classes are depicted.

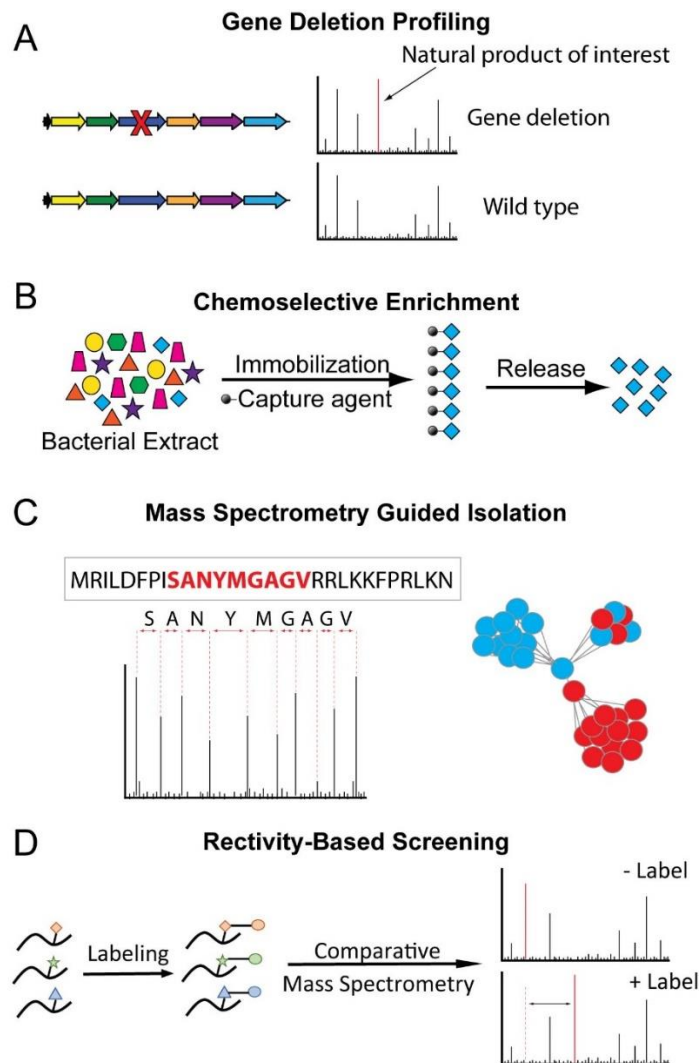


Figure 1.3 | Methods for correlating biosynthetic gene clusters to natural products. (A) Gene deletion profiling uses a gene deletion from either the natural host or a heterologous host (depicted as blue gene). Comparative metabolic profiling (using mass spectrometry or liquid chromatography) is used to identify missing peaks from the deletion strain (red peak). (B) chemoselective enrichment uses a capture agent to capture natural products from bacterial extracts. Once enriched, the natural products can be released and further characterization can be performed. (C) Mass spectrometry can be used to identify a sequence tag to relate back to natural product gene clusters (left) or natural product cluster can be used to identify similar natural products from similar species (right). (D) Reactivity-based screening utilizes a specific probe to label particular functional groups (diamond, star, triangle labeled with circles). Comparative mass spectrometry can then be utilized to identify natural products labeled with the probe.

1.8 References

1. Agarwal V, Pierce E, McIntosh J, Schmidt EW, and Nair SK. (2012) Structures of cyanobactin maturation enzymes define a family of transamidating proteases. *Chemistry & biology* 19: 1411-1422.
2. Arnison PG, Bibb MJ, Bierbaum G, Bowers AA, Bugni TS, Bulaj G, Camarero JA, Campopiano DJ, Challis GL, Clardy J, Cotter PD, Craik DJ, Dawson M, Dittmann E, Donadio S, Dorrestein PC, Entian KD, Fischbach MA, Garavelli JS, Goransson U, Gruber CW, Haft DH, Hemscheidt TK, Hertweck C, Hill C, Horswill AR, Jaspars M, Kelly WL, Klinman JP, Kuipers OP, Link AJ, Liu W, Marahiel MA, Mitchell DA, Moll GN, Moore BS, Muller R, Nair SK, Nes IF, Norris GE, Olivera BM, Onaka H, Patchett ML, Piel J, Reaney MJ, Rebuffat S, Ross RP, Sahl HG, Schmidt EW, Selsted ME, Severinov K, Shen B, Sivonen K, Smith L, Stein T, Sussmuth RD, Tagg JR, Tang GL, Truman AW, Vederas JC, Walsh CT, Walton JD, Wenzel SC, Willey JM, and van der Donk WA. (2013) Ribosomally synthesized and post-translationally modified peptide natural products: overview and recommendations for a universal nomenclature. *Natural product reports* 30: 108-160.
3. Bachmann BO, Van Lanen SG, and Baltz RH. (2014) Microbial genome mining for accelerated natural products discovery: is a renaissance in the making? *Journal of industrial microbiology & biotechnology* 41: 175-184.
4. Baltz RH. (2008) Renaissance in antibacterial discovery from actinomycetes. *Current opinion in pharmacology* 8: 557-563.
5. Belshaw PJ, Roy RS, Kelleher NL, and Walsh CT. (1998) Kinetics and regioselectivity of peptide-to-heterocycle conversions by microcin B17 synthetase. *Chemistry & biology* 5: 373-384.
6. Bentley SD, Chater KF, Cerdeno-Tarraga AM, Challis GL, Thomson NR, James KD, Harris DE, Quail MA, Kieser H, Harper D, Bateman A, Brown S, Chandra G, Chen CW, Collins M, Cronin A, Fraser A, Goble A, Hidalgo J, Hornsby T, Howarth S, Huang CH, Kieser T, Larke L, Murphy L, Oliver K, O'Neil S, Rabbinowitsch E, Rajandream MA, Rutherford K, Rutter S, Seeger K, Saunders D, Sharp S, Squares R, Squares S, Taylor K, Warren T, Wietzorrek A, Woodward J, Barrell BG, Parkhill J, and Hopwood DA. (2002) Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2). *Nature* 417: 141-147.
7. Betschel SD, Borgia SM, Barg NL, Low DE, and De Azavedo JC. (1998) Reduced virulence of group A streptococcal Tn916 mutants that do not produce streptolysin S. *Infection and immunity* 66: 1671-1679.
8. Blin K, Medema MH, Kazempour D, Fischbach MA, Breitling R, Takano E, and Weber T. (2013) antiSMASH 2.0--a versatile platform for genome mining of secondary metabolite producers. *Nucleic acids research* 41: W204-212.

9. Bowers AA, Walsh CT, and Acker MG. (2010) Genetic interception and structural characterization of thiopeptide cyclization precursors from *Bacillus cereus*. *Journal of the American Chemical Society* 132: 12182-12184.
10. Breil B, Borneman J, and Triplett EW. (1996) A newly discovered gene, *tfuA*, involved in the production of the ribosomally synthesized peptide antibiotic trifolitoxin. *Journal of bacteriology* 178: 4150-4156.
11. Carlson EE, and Cravatt BF. (2007) Chemoselective probes for metabolite enrichment and profiling. *Nature Methods* 4: 429-435.
12. Challis GL. (2008) Genome mining for novel natural product discovery. *Journal of medicinal chemistry* 51: 2618-2628.
13. Challis GL. (2008) Mining microbial genomes for new natural products and biosynthetic pathways. *Microbiology* 154: 1555-1569.
14. Cox CL, Tietz JI, Sokolowski K, Melby JO, Doroghazi JR, and Mitchell DA. (2014) Nucleophilic 1,4-additions for natural product discovery. *ACS chemical biology* 9: 2014-2022.
15. Crone WJK, Leeper FJ, and Truman AW. (2012) Identification and characterisation of the gene cluster for the anti-MRSA antibiotic bottromycin: expanding the biosynthetic diversity of ribosomal peptides. *Chemical Science* 3: 3516-3521.
16. de Jong A, van Hijum SA, Bijlsma JJ, Kok J, and Kuipers OP. (2006) BAGEL: a web-based bacteriocin genome mining tool. *Nucleic acids research* 34: W273-279.
17. Deane CD, and Mitchell DA. (2014) Lessons learned from the transformation of natural product discovery to a genome-driven endeavor. *Journal of industrial microbiology & biotechnology* 41: 315-331.
18. Doroghazi JR, Albright JC, Goering AW, Ju KS, Haines RR, Tchalukov KA, Labeda DP, Kelleher NL, and Metcalf WW. (2014) A roadmap for natural product discovery based on large-scale genomics and metabolomics. *Nature chemical biology* 10: 963-968.
19. Doroghazi JR, and Metcalf WW. (2013) Comparative genomics of actinomycetes with a focus on natural product biosynthetic genes. *BMC genomics* 14: 611.
20. Dunbar KL, Chekan JR, Cox CL, Burkhart BJ, Nair SK, and Mitchell DA. (2014) Discovery of a new ATP-binding motif involved in peptidic azoline biosynthesis. *Nature chemical biology* 10: 823-829.

21. Dunbar KL, Melby JO, and Mitchell DA. (2012) YcaO domains use ATP to activate amide backbones during peptide cyclodehydrations. *Nature chemical biology* 8: 569-575.
22. Dunbar KL, and Mitchell DA. (2013) Insights into the mechanism of peptide cyclodehydrations achieved through the chemoenzymatic generation of amide derivatives. *Journal of the American Chemical Society* 135: 8692-8701.
23. Dunbar KL, and Mitchell DA. (2013) Revealing nature's synthetic potential through the study of ribosomal natural product biosynthesis. *ACS chemical biology* 8: 473-487.
24. Fischbach MA, and Walsh CT. (2006) Assembly-line enzymology for polyketide and nonribosomal Peptide antibiotics: logic, machinery, and mechanisms. *Chemical reviews* 106: 3468-3496.
25. Fischbach MA, and Walsh CT. (2009) Antibiotics for emerging pathogens. *Science* 325: 1089-1093.
26. Gao JT, Ju KS, Yu XM, Velasquez JE, Mukherjee S, Lee J, Zhao CM, Evans BS, Doroghazi JR, Metcalf WW, and van der Donk WA. (2014) Use of a Phosphonate Methyltransferase in the Identification of the Fosfazinomycin Biosynthetic Gene Cluster. *Angewandte Chemie International Edition* 53: 1334-1337.
27. Gomez-Escribano JP, Song LJ, Bibb MJ, and Challis GL. (2012) Posttranslational beta-methylation and macrolactamidation in the biosynthesis of the bottromycin complex of ribosomal peptide antibiotics. *Chemical Science* 3: 3522-3525.
28. Gottlieb D, Carter HE, Legator M, and Gallicchio V. (1954) The biosynthesis of chloramphenicol. I. Precursors stimulating the synthesis. *Journal of bacteriology* 68: 243-251.
29. Gross H, Stockwell VO, Henkels MD, Nowak-Thompson B, Loper JE, and Gerwick WH. (2007) The genomisotopic approach: A systematic method to isolate products of orphan biosynthetic gene clusters. *Chemistry & biology* 14: 53-63.
30. Haft DH, Basu MK, and Mitchell DA. (2010) Expansion of ribosomally produced natural products: a nitrile hydratase- and Nif11-related precursor family. *BMC biology* 8: 70.
31. Hegemann JD, Zimmermann M, Xie X, and Marahiel MA. (2013) Caulosegnins I-III: a highly diverse group of lasso peptides derived from a single biosynthetic gene cluster. *Journal of the American Chemical Society* 135: 210-222.
32. Hegemann JD, Zimmermann M, Zhu S, Klug D, and Marahiel MA. (2013) Lasso peptides from proteobacteria: Genome mining employing heterologous expression and mass spectrometry. *Biopolymers* 100: 527-542.

33. Hou Y, Tianero MD, Kwan JC, Wyche TP, Michel CR, Ellis GA, Vazquez-Rivera E, Braun DR, Rose WE, Schmidt EW, and Bugni TS. (2012) Structure and biosynthesis of the antibiotic bottromycin D. *Organic letters* 14: 5050-5053.
34. Huo L, Rachid S, Stadler M, Wenzel SC, and Muller R. (2012) Synthetic biotechnology to study and engineer ribosomal bottromycin biosynthesis. *Chemistry & biology* 19: 1278-1287.
35. Jensen PR, Chavarria KL, Fenical W, Moore BS, and Ziemert N. (2014) Challenges and triumphs to genomics-based natural product discovery. *Journal of industrial microbiology & biotechnology* 41: 203-209.
36. Just-Baringo X, Albericio F, and Alvarez M. (2014) Thiopeptide antibiotics: retrospective and recent advances. *Marine drugs* 12: 317-351.
37. Kersten RD, Yang YL, Xu YQ, Cimermancic P, Nam SJ, Fenical W, Fischbach MA, Moore BS, and Dorrestein PC. (2011) A mass spectrometry-guided genome mining approach for natural product peptidogenomics. *Nature chemical biology* 7: 794-802.
38. Lee J, Hao Y, Blair PM, Melby JO, Agarwal V, Burkhardt BJ, Nair SK, and Mitchell DA. (2013) Structural and functional insight into an unexpectedly selective N-methyltransferase involved in plantazolicin biosynthesis. *Proceedings of the National Academy of Sciences of the United States of America* 110: 12954-12959.
39. Lee J, McIntosh J, Hathaway BJ, and Schmidt EW. (2009) Using marine natural products to discover a protease that catalyzes peptide macrocyclization of diverse substrates. *Journal of the American Chemical Society* 131: 2122-2124.
40. Lee JH, Bae B, Kuemin M, Circello BT, Metcalf WW, Nair SK, and van der Donk WA. (2010) Characterization and structure of Dhpl, a phosphonate O-methyltransferase involved in dehydrophos biosynthesis. *Proceedings of the National Academy of Sciences of the United States of America* 107: 17557-17562.
41. Lee SW, Mitchell DA, Markley AL, Hensler ME, Gonzalez D, Wohlrab A, Dorrestein PC, Nizet V, and Dixon JE. (2008) Discovery of a widely distributed toxin biosynthetic gene cluster. *Proceedings of the National Academy of Sciences of the United States of America* 105: 5879-5884.
42. Letzel AC, Pidot SJ, and Hertweck C. (2014) Genome mining for ribosomally synthesized and post-translationally modified peptides (RiPPs) in anaerobic bacteria. *BMC genomics* 15: 983.
43. Lewis K. (2013) Platforms for antibiotic discovery. *Nature reviews Drug discovery* 12: 371-387.

44. Li C, and Kelly WL. (2010) Recent advances in thiopeptide antibiotic biosynthesis. *Natural product reports* 27: 153-164.
45. Li J, Girard G, Florea BI, Geurink PP, Li N, van der Marel GA, Overhand M, Overkleeft HS, and van Wezel GP. (2012) Identification and isolation of lantibiotics from culture: a bioorthogonal chemistry approach. *Organic & Biomolecular Chemistry* 10: 8677-8683.
46. Li J, Qu X, He X, Duan L, Wu G, Bi D, Deng Z, Liu W, and Ou HY. (2012) ThioFinder: a web-based tool for the identification of thiopeptide gene clusters in DNA sequences. *PLoS one* 7: e45878.
47. Li YM, Milne JC, Madison LL, Kolter R, and Walsh CT. (1996) From peptide precursors to oxazole and thiazole-containing peptide antibiotics: microcin B17 synthase. *Science* 274: 1188-1193.
48. Maksimov MO, and Link AJ. (2014) Prospecting genomes for lasso peptides. *Journal of industrial microbiology & biotechnology* 41: 333-344.
49. Maksimov MO, Pelczer I, and Link AJ. (2012) Precursor-centric genome-mining approach for lasso peptide discovery. *Proceedings of the National Academy of Sciences of the United States of America* 109: 15223-15228.
50. Melby JO, Li X, and Mitchell DA. (2014) Orchestration of enzymatic processing by thiazole/oxazole-modified microcin dehydrogenases. *Biochemistry* 53: 413-422.
51. Melby JO, Nard NJ, and Mitchell DA. (2011) Thiazole/oxazole-modified microcins: complex natural products from ribosomal templates. *Current opinion in chemical biology* 15: 369-378.
52. Mitchell DA, Lee SW, Pence MA, Markley AL, Limm JD, Nizet V, and Dixon JE. (2009) Structural and functional dissection of the heterocyclic peptide cytotoxin streptolysin S. *The Journal of biological chemistry* 284: 13004-13012.
53. Mohimani H, Kersten RD, Liu WT, Wang M, Purvine SO, Wu S, Brewer HM, Pasa-Tolic L, Bandeira N, Moore BS, Pevzner PA, and Dorrestein PC. (2014) Automated genome mining of ribosomal peptide natural products. *ACS chemical biology* 9: 1545-1551.
54. Molloy EM, Cotter PD, Hill C, Mitchell DA, and Ross RP. (2011) Streptolysin S-like virulence factors: the continuing saga. *Nature reviews Microbiology* 9: 670-681.
55. Molohon KJ, Melby JO, Lee J, Evans BS, Dunbar KL, Bumpus SB, Kelleher NL, and Mitchell DA. (2011) Structure determination and interception of biosynthetic intermediates for the plantazolicin class of highly discriminating antibiotics. *ACS chemical biology* 6: 1307-1313.

56. Newman DJ, and Cragg GM. (2012) Natural products as sources of new drugs over the 30 years from 1981 to 2010. *Journal of natural products* 75: 311-335.
57. Nguyen DD, Wu CH, Moree WJ, Lamsa A, Medema MH, Zhao XL, Gavilan RG, Aparicio M, Atencio L, Jackson C, Ballesteros J, Sanchez J, Watrous JD, Phelan VV, van de Wiel C, Kersten RD, Mehnaz S, De Mot R, Shank EA, Charusanti P, Nagarajan H, Duggan BM, Moore BS, Bandeira N, Palsson BO, Pogliano K, Gutierrez M, and Dorrestein PC. (2013) MS/MS networking guided analysis of molecule and gene cluster families. *Proceedings of the National Academy of Sciences of the United States of America* 110: E2611-E2620.
58. Nizet V, Beall B, Bast DJ, Datta V, Kilburn L, Low DE, and De Azavedo JC. (2000) Genetic locus for streptolysin S production by group A streptococcus. *Infection and immunity* 68: 4245-4254.
59. Odendaal AY, Trader DJ, and Carlson EE. (2011) Chemoselective enrichment for natural products discovery. *Chemical Science* 2: 760-764.
60. Ortega MA, Hao Y, Zhang Q, Walker MC, van der Donk WA, and Nair SK. (2014) Structure and mechanism of the tRNA-dependent lantibiotic dehydratase NisB. *Nature*.
61. Payne DJ, Gwynn MN, Holmes DJ, and Pompliano DL. (2007) Drugs for bad bugs: confronting the challenges of antibacterial discovery. *Nature reviews Drug discovery* 6: 29-40.
62. Schmidt EW, and Donia MS. (2009) Chapter 23. Cyanobactin ribosomally synthesized peptides--a case of deep metagenome mining. *Methods in enzymology* 458: 575-596.
63. Schmidt EW, Nelson JT, Rasko DA, Sudek S, Eisen JA, Haygood MG, and Ravel J. (2005) Patellamide A and C biosynthesis by a microcin-like pathway in *Prochloron didemni*, the cyanobacterial symbiont of *Lissoclinum patella*. *Proceedings of the National Academy of Sciences of the United States of America* 102: 7315-7320.
64. Scholz R, Molohon KJ, Nachtigall J, Vater J, Markley AL, Sussmuth RD, Mitchell DA, and Borriss R. (2011) Plantazolicin, a novel microcin B17/streptolysin S-like natural product from *Bacillus amyloliquefaciens* FZB42. *Journal of bacteriology* 193: 215-224.
65. Sensi P, Greco AM, and Ballotta R. (1959) Rifomycin. I. Isolation and properties of rifomycin B and rifomycin complex. *Antibiotics annual* 7: 262-270.
66. Sivonen K, Leikoski N, Fewer DP, and Jokela J. (2010) Cyanobactins-ribosomal cyclic peptides produced by cyanobacteria. *Applied microbiology and biotechnology* 86: 1213-1225.
67. Su TL. (1948) Micrococcin, an antibacterial substance formed by a strain of *Micrococcus*. *British journal of experimental pathology* 29: 473-481.

68. U.S. Department of Health and Human Services CfDcAP. **(2013)** Antibiotic Resistance Threats in the United States, 2013.
69. Van Lanen SG, and Shen B. **(2006)** Microbial genomics for the improvement of natural product discovery. *Current opinion in microbiology* 9: 252-260.
70. Velasquez JE, and van der Donk WA. **(2011)** Genome mining for ribosomally synthesized natural products. *Current opinion in chemical biology* 15: 11-21.
71. Wise R, Andrews JM, and Edwards LJ. **(1983)** In vitro activity of Bay 09867, a new quinoline derivative, compared with those of other antimicrobial agents. *Antimicrobial agents and chemotherapy* 23: 559-564.

CHAPTER II: THE GENOMIC LANDSCAPE OF RIBOSOMAL PEPTIDES CONTAINING THIAZOLE AND OXAZOLE HETEROCYCLES

I am grateful to K. Whalen for running the Enzyme Similarity Tool for the enzyme protein network and also J. Doroghazi for advice throughout the research. This chapter was critically edited by J. Melby, C. Deane and J. Doroghazi.

Abstract

Ribosomally synthesized and post-translationally modified peptides (RiPPs) are a burgeoning class of natural products with diverse activity that share a similar origin and common features in their biosynthetic pathways. The precursor peptides of these natural products are ribosomally produced, upon which a combination of modification enzymes installs diverse functional groups. This genetically encoded peptide-based strategy allows for rapid diversification of these natural products by mutation in the precursor genes merged with unique combinations of modification enzymes. Thiazole/oxazole-modified microcins (TOMMs) are a class of RiPPs defined by the presence of heterocycles derived from cysteine, serine, and threonine residues in the precursor peptide. TOMMs encompass a number of different families, including but not limited to the linear azol(in)e-containing peptides (*e.g.* streptolysin S, microcin B17, and plantazolicin), cyanobactins, thiopeptides, and bottromycins. Although many TOMMs have been explored, the increased availability of genome sequences has illuminated several unexplored TOMM producers. Given the remarkable structural and functional diversity displayed by known TOMMs, a comprehensive bioinformatic study to catalog and classify the entire RiPP class was undertaken. Here we report the characterization of nearly 1,500 TOMM gene clusters from genomes in the The European Molecular Biology Laboratory (EMBL) and The European Bioinformatics Institute (EBI) sequence repository. Genome mining suggests a complex diversification of modification enzymes and precursor peptides to create more than 20 distinct families of TOMMs, nine of which have not heretofore been described. Many of the identified TOMM families have an abundance of diverse precursor peptide sequences as well as unfamiliar

combinations of modification enzymes, signifying a potential wealth of novel natural products on known and unknown biosynthetic scaffolds. Phylogenetic analysis suggests a widespread distribution of TOMMs across multiple phyla; however, producers of similar TOMMs are generally found in the same phylum with few exceptions. The comprehensive genome mining study described herein has uncovered a myriad of unique TOMM biosynthetic clusters and provides an atlas to guide future discovery efforts. These biosynthetic gene clusters are predicted to produce diverse final products, and the identification of additional combinations of modification enzymes could expand the potential of combinatorial natural product biosynthesis.

2.1 Introduction

Recently, genome mining has revealed the tremendous sequence diversity of a pharmaceutically relevant family of natural products, the ribosomally synthesized and post-translationally modified peptides (RiPPs) (Haft, *et al.* 2010, Lee, *et al.* 2008, Leikoski, *et al.* 2013, Letzel, *et al.* 2014, Maksimov, *et al.* 2014, Mohimani, *et al.* 2014, Velasquez, *et al.* 2011). The gene clusters for these natural products have been discovered in all three domains of life, and their structural diversity continues to expand as more knowledge accumulates regarding these natural products and their biosynthesis. RiPPs populate a diverse chemical and genetic landscape, including, but not limited to, lanthipeptides, thiazole/oxazole-modified microcins (TOMMs), lasso peptides, and linaridins (Arnison, *et al.* 2013). The ribosomal origin of the starting material unites this otherwise disparate group of natural products. While the genes for most precursor peptides are located near to those for the modification enzymes within the genome, there are examples of precursors located elsewhere (e.g. heterocycloanthracins (Haft 2009) and prochlorosins (Li, Sher, *et al.* 2010, Zhang, *et al.* 2014)). With few exceptions, the C-terminal portion of the precursor peptide (often referred to as the core region) is post-translationally modified while the N-terminal portion (leader region) harbors binding motifs that recruit the modification enzymes. Common core modifications include heterocycles, dehydrated amino

acids, methylations, acetylations, backbone crosslinks, and many others (Arnison, *et al.* 2013, Dunbar, *et al.* 2013, McIntosh, *et al.* 2010). A number of these modifications restrict the conformational flexibility of the peptide, which plays a part in endowing the final product with a specific activity. Following the enzymatic processing of the core, the unmodified leader region is typically removed by a protease, resulting in either the fully mature product or a substrate for further modifications (Figure 2.1A) (Oman, *et al.* 2010). Certain RiPPs swap the functions of the N- and C-terminal regions (e.g. bottromycins (Crone, *et al.* 2012, Gomez-Escribano, *et al.* 2012, Hou, *et al.* 2012, Huo, *et al.* 2012)), while others have co-opted macrocyclization enzymes to excise the leader peptide (e.g. cyanobactins (Agarwal, *et al.* 2012, Koehnke, *et al.* 2012) and thiopeptides (Bowers, *et al.* 2010, Malcolmson, *et al.* 2013, Tocchetti, *et al.* 2013)). Regardless, the RiPP biosynthetic strategy is capable of producing structurally diverse compounds with minimal genetic space because the ribosome is utilized to synthesize the majority of the natural product scaffold. Furthermore, natural product variation can be expanded with the simple mutation of the core peptide, or addition and deletion of modification enzymes, leading to a variety of structures and bioactivities within the class. The particular combinations of precursor sequence and modification enzymes ultimately define the classes of RiPPs, and bioinformatics can readily identify and classify RiPP gene clusters using homology to these common enzymes (Velasquez, *et al.* 2011).

TOMMs are a large subclass of RiPPs encompassing a wide array of structures and bioactivities that are defined by the presence ofazole and azoline heterocycles derived from Cys, Ser, Thr residues in the precursor peptide (Lee 2008, Melby, *et al.* 2011). Examples of studied TOMMs include microcin B17 (DNA gyrase inhibitor) (Li, *et al.* 1996), streptolysin S (cytolysin) (Mitchell 2009), plantazolicin (antibacterial) (Molohon, *et al.* 2011, Scholz, *et al.* 2011), cyanobactins (anticancer, antimalarial, and others) (Schmidt, *et al.* 2005), and the thiopeptides (translation inhibitors) (Kelly, *et al.* 2009, Liao, *et al.* 2009, Morris, *et al.* 2009)

(Figure 2.2). The hallmark of a TOMM gene cluster is the presence of a cyclodehydratase that installs azoline heterocycles onto a precursor peptide (Dunbar, *et al.* 2012, Dunbar, *et al.* 2013, Li, *et al.* 1996, Melby, *et al.* 2012). In some cases, a locally-encoded dehydrogenase then oxidizes the azoline to the corresponding azole heterocycle (Belshaw, *et al.* 1998, Melby, *et al.* 2014, Milne, *et al.* 1999). TOMM biosynthetic clusters regularly encode ancillary modification enzymes that increase structural complexity.

Given the structural and functional diversity of previously explored TOMMs, a fundamental understanding of the synthetic capabilities of bacteria and archaea to produce these natural products is desirable. Here we have analyzed sequences from The European Molecular Biology Laboratory (EMBL) and The European Bioinformatics Institute (EBI) sequence database to view the distribution, evolution and structural potential of TOMMs. Nearly 1,500 biosynthetic gene clusters were identified, many of which appear to encode novel natural products.

Additionally, some gene clusters contain heretofore-undescribed combinations of ancillary modification enzymes, potentially expanding the chemical complexity of TOMMs. Furthermore, precursor peptides from both characterized and uncharacterized families were analyzed to identify common motifs. This study defines the genomic landscape of TOMM natural products.

2.2 Genome Mining and Isofunctional Grouping

TOMM biosynthetic gene clusters are defined by the presence of the aforementioned cyclodehydratase, which is composed of an E1 ubiquitin-activating enzyme homolog (C protein) and a member of the YcaO superfamily (D protein). In roughly half of all TOMM clusters, the genes encoding the C and D proteins are fused and expressed as a single polypeptide (CD fusion). This fusion underscores the important collaboration of the C and D proteins in cyclodehydratase function. Recently, it was demonstrated that the D protein formally catalyzes the cyclodehydration reaction, while the C protein engages the leader peptide and potentiates the cyclodehydration reaction by several orders of magnitude (Dunbar, *et al.* 2014, Dunbar, *et al.*

2012, Mitchell 2009). In at least two cases (e.g. bottromycin and trifolitoxin), the D protein is believed to act in the absence of a C protein (Breil, *et al.* 1996, Crone, *et al.* 2012, Gomez-Escribano, *et al.* 2012, Hou, *et al.* 2012, Huo, *et al.* 2012). In a sizeable percentage of TOMM gene clusters, a flavin mononucleotide (FMN)-dependent dehydrogenase (B protein) is encoded, which has been shown to oxidize select azoline rings to azoles (Li, *et al.* 1996, Melby, *et al.* 2014, Milne, *et al.* 1999).

In an attempt to catalog all TOMM biosynthetic gene clusters, the local genomic regions of YcaO homologs within UniProtKB were characterized (Figure 2.1B). YcaO homologs were chosen as the focus of this search primarily because it has been demonstrated that the B and C proteins can be omitted in TOMM production, whereas D proteins (YcaO homologs) are always present (*e.g.* bottromycin). Furthermore, the YcaO domain has considerably fewer non-TOMM related homologs than the B and C proteins (*i.e.* bona fide E1-family enzymes like ThiF and MoeB for the C protein and other FMN-dependent dehydrogenases), therefore producing fewer false positives. Notwithstanding, a subset of YcaO homologs are known to be present in non-TOMM related settings (previously referred to as “non-TOMM YcaO” and “TfuA-associated YcaO”) (Dunbar, *et al.* 2014), and therefore, multiple methods have been used to distinguish TOMM-producing gene clusters from non-producers. Using the genomic region surrounding *ycaO* genes (10 kb on either side), a combination of BLAST score and synteny was used to classify biosynthetic gene clusters into families (Figure 2.1B – Step 1 and 1C). Potential TOMM gene clusters were first analyzed for the presence of a C protein or CD fusion protein within the flanking genomic region (10 kb on either side of the *ycaO* gene). The gene cluster was also analyzed for the presence of a precursor peptide. Often, precursors evade automated gene finders due to their short lengths; therefore, intergenic regions were also analyzed for potentially unannotated precursor peptide genes. Precursor peptides were annotated under the assumption that they are short open reading frames (<150 amino acids) and typically contain an abundance of

Gly, Cys, Ser, and Thr residues in the core region. This approach does not locate precursor peptides that are not in close proximity to the D proteins (>10 kb away) or those with a low proportion of heterocyclizable residues. Of the potential TOMM clusters identified in the present study, 46% contained an identifiable precursor peptide gene within 10 kb of the D protein. As the bottromycins do not contain a C protein homolog (*i.e.* stand-alone D proteins) and do not have Gly-Cys-Ser-Thr rich precursor peptides, a manual analysis was performed to identify bottromycin-like gene clusters. If a TOMM cluster was identified using this criteria, all gene clusters in a family were annotated as TOMMs, regardless of whether the other clusters contained an identifiable C protein or precursor peptide. This cataloging procedure identified nearly 1,500 putative TOMM biosynthetic gene clusters in the prokaryotic genomes available from EMBL (Figure 2.1). This is likely an underestimate because (*i*) very little is known about stand-alone TOMM clusters (lacking C protein) (*ii*) it is unknown whether TfuA-associated YcaO proteins can adorn peptides with azoline andazole heterocycles and (*iii*) highly unusual precursor peptides would not be detected by the strategy employed. Furthermore, RefSeq (NCBI) records are not systematically included in UniProtKB, thus many TOMM proteins from RefSeq were not included in this identification process.

To visualize the relationship landscape of TOMM families, a sequence similarity network was produced using the D proteins from each gene cluster (Figure 2.3 and 2.4). Characterized gene cluster families, identified by similarity to previously explored TOMM clusters, were then mapped onto the network. D proteins from similar TOMM families were more similar to each other, irrespective of the phyla from which the gene clusters originated. This suggests, similar to other natural products like lanthipeptides (Yu, Zhang, *et al.* 2013, Zhang, *et al.* 2012) and phosphonates (Ju, *et al.* 2014, Yu, Doroghazi, *et al.* 2013), that the structure and function of a particular TOMM can be predicted not only by the sequence of the precursor peptide, but also by the similarity of the modification enzymes. Therefore, it is not necessary in all cases to identify

the putative precursor peptide to predict the family of the TOMM natural product. Examining isofunctional clusters in multiple genomic backgrounds also allows inference of gene cluster boundaries and the encoded enzymes that are involved in biosynthesis (Doroghazi, *et al.* 2013). Using a BLAST expectation value of E-54, there are 11 anticipated isofunctional groups that contain at least one previously explored TOMM. The groups have been designated as follows: cytolysin, cyanobactin, thiopeptide, microcin B17 (MccB17), NHLP/Nif11, goadsporin, heterocycloanthracin (HCA), hakacin, plantazolicin (PZN), YM-216319, and bottromycin (Figure 2.3).

As illustrated on the sequence similarity network, the families for nearly 60% of predicted TOMMs can be inferred from their similarity to a characterized D protein. However, a considerable number of presumed isofunctional groups contain no characterized TOMMs, leaving a vast area of the natural product space yet to be characterized (Figure 2.3 and 2.5). There are 10 presumed isofunctional groups with no explored TOMM clusters, which we have designated as the following: haloazolisin, faecalisin, helicobactin, mobilisin, propionisin, coryneazolisin type 1 and type 2, thermoacidophisin, anabaenasin, and gallolytisin (Figure 2.4 and 2.6). These TOMM biosynthetic gene clusters encode a variety of unique peptides rich in Gly, Ser, Thr, and Cys, suggesting that they are the TOMM precursor peptide. Although defined by the installation of azoline heterocycles, the majority of TOMM gene clusters contain additional post-translational modification enzymes (Figure 2.7) as well as a plethora of novel precursor peptides (Figure 2.8). To analyze enzymatic commonalities between TOMM families, the proteins encoded in the genomic region surrounding the D proteins were clustered by similarity (Figure 2.9, 2.10, 2.11). These family-specific modification enzymes are described further within each TOMM family discussed below.

2.3 Isofunctional Groups with Explored TOMMs

2.3.1 Microcin B17

Microcin B17 (MccB17) is a quintessential example of a TOMM cluster containing a discrete cyclodehydratase (*i.e.* separate C and D proteins). The enzymes encoded by this cluster extensively modify the MccB17 core peptide to yield a DNA gyrase inhibitor (Li, *et al.* 1996, Yorgey, *et al.* 1994). The current analysis identified 30 gene clusters from *Escherichia coli*, *Pseudomonas syringae*, *Pseudomonas putida*, and *Pseudomonas fluorescens*, all of which have been previously identified as MccB17 producers (Lee, *et al.* 2008, Metelev, *et al.* 2013). The gene clusters from *E. coli* and *Pseudomonas sp.* are similar to the previously characterized clusters, and all contain homologs to the C protein (Figure 2.9: group 41), D protein (Figure 2.9: group 54), and three ATP-binding cassette (ABC)-like transporters (Figure 2.9: groups 2, 66, 67). The 19 identified MccB17 precursor peptides in *E. coli* clusters are identical in the core region and bear only a single substitution in the leader peptide; however, these peptides vary in the length of the glycine linker region at the N-terminus of the core. The nine precursors from *Pseudomonas* are considerably more divergent, only sharing the glycine-rich cyclized region with the *E. coli* precursors (Figure 2.8).

2.3.2 Cytolysin

Streptolysin S (SLS) is a potent cytolysin responsible for the characteristic β -hemolytic phenotype exhibited by *Streptococcus pyogenes* (Molloy, *et al.* 2011). The cytolysin family continues to grow, with over 300 clusters identified since the pioneering identification of the SLS gene cluster (Betschel, *et al.* 1998, Nizet, *et al.* 2000). Homologous clusters have been identified in other pathogenic bacteria including *Listeria monocytogenes*, *Clostridium botulinum*, *Staphylococcus aureus*, and *Brachyspira murdochii* (Letzel, *et al.* 2014, Molloy, *et al.* 2011). Of particular interest are the clusters identified in *Spirochaetes* and *Lactobacillus crispatus* because the *Spirochaetes* are not known to produce any toxins and *Lactobacillus crispatus* is a commensal

bacteria and therefore it is unknown why they both have the genetic capacity to produce SLS-like compounds (Molloy *et al.*, *submitted*). Although the cytolysins form a single isofunctional group, the precursor peptides differ based on species. All of the identified clusters contain a discrete cyclodehydratase, a dehydrogenase, ABC transporters, and a CaaX-like protease (Maxson, *et al.* 2015, Pei, *et al.* 2011) (Figure 2.7). Of the 312 identified clusters, 294 had identifiable precursor peptides. Six cytolysin TOMM clusters encode two precursor peptides, in line with a previous finding (Tabata, *et al.* 2013). All of the identified cytolysin precursor peptide cores contain a Gly residue followed by 10 or more potentially heterocyclized residues, suggesting that contiguous heterocyclization may be important for activity. The C-terminal regions of the core peptides (following the conserved, contiguous, heterocyclizable region) vary by species or are missing entirely (*Spirochaetes*). The core regions, including the variable C-termini, and the leader peptide of the precursor peptides from *Streptococcus* and *Clostridium* are more similar to each other than they are to the peptides from *Staphylococcus* and *Listeria*, which themselves share similarity. This is consistent with previous studies that showed that the *Streptococcus* enzymes could modify the *Clostridium* precursor peptide, but not the native *Listeria* precursor (Lee, *et al.* 2008, Mitchell, *et al.* 2009). Furthermore, the core region of the precursor peptide from *Borrelia* is more similar to that from *Streptococcus* than it is to that from *Listeria*, solidifying the previous findings that these peptides can be modified by the *Streptococcus* enzymes (Molloy *et al.*, *submitted*). The C proteins involved in cytolysin biosynthesis are split by organism into two different enzyme groups (Figure 2.9), further corroborating the ability of only certain cyclodehydratases to modify precursor peptides in this family. The *Streptococcus*, *Borrelia*, *Brachyspira*, and *Clostridium* C proteins cluster together (Group 22), and the *Listeria* and *Staphylococcus* C proteins form a different cluster (Group 37).

2.3.3 Cyanobactin

The cyanobactins represent one of the largest families of TOMMs with fused cyclodehydratases TOMMs. Cyanobactins are cyclic peptides produced by *Cyanobacteria* and are best known for their anticancer, antiviral and antimalarial effects (Namikoshi, *et al.* 1996, Nunnery, *et al.* 2010). This exercise only included cyanobactin biosynthetic gene clusters that were bioinformatically identified as TOMM gene clusters from UniProtKB sequences (there are known cyanobactins which lackazole/azoline heterocycles and the requisite D protein is missing from the cluster) (Sivonen, *et al.* 2010). The 56 cyanobactin clusters identified here often encode precursor peptides with hypervariable core regions. These diverse natural product template sequences are flanked by highly conserved cleavage sites that ultimately direct the excision and macrocyclization of the mature cyanobactin from the precursor peptide (Agarwal, *et al.* 2012, Koehnke, *et al.* 2012, Lee, *et al.* 2009). In most clusters, a PatA-like protease recognizes and cleaves the N-terminal site. Then, a PatG-like protease recognizes the C-terminal site and catalyzes the N-to-C macrocyclization (Agarwal, *et al.* 2012, Houssen, *et al.* 2010, Koehnke, *et al.* 2012). In nearly one-third of the identified TOMM cyanobactins identified (18 total), PatG homologs are fused as a single polypeptide to FMN-dependent dehydrogenases for the oxidation of azoline heterocycles to the corresponding azoles (Figure 2.7). The identified PatA and PatG homologs form a group with other B proteins (lacking the protease) from similar clusters such as goadsporin and NHLP/Nif11 (Figure 2.9, group 9, also see section 2.3.4). This enzyme group also contains homologs of the prenyltransferases in the cyanobactin gene clusters because there are homologous methyltransferase domains that are fused to either a PatA or the prenyltransferase domain, thus combining the group by similarity. Of the 56 total TOMM cyanobactin gene clusters, prenyltransferases were identified in 18 and these enzymes are expected to prenylate Ser, Tyr, and Thr residues within the precursor peptide core regions (Bent, *et al.* 2013, McIntosh, *et al.* 2011). Although cyanobactin gene clusters often encode multiple precursor peptides, they are

relatively long (~100 amino acids) and have a reduced richness of Cys, Ser, Thr (~20-30% in predicted core peptides) compared to other TOMM precursor peptides. Therefore, few cyanobactin precursor peptides were identified with the parameters used in this study. However, identification of many cyanobactin precursor peptides has been previously reported (Donia, *et al.* 2008, Donia, *et al.* 2011, Schmidt, *et al.* 2009, Sivonen, *et al.* 2010).

2.3.4 Nitrile hydratase-related leader peptides and Nif11-related precursor peptides

Cyanobactin D proteins group with those for two other families of TOMMs, the nitrile hydratase-related leader peptides (NHLPs or NHLP-Burk, for clusters produced by *Burkholderia* species) and the Nif11-related precursor peptides (Figure 2.3 and 2.5) (Haft, *et al.* 2010). Unlike the cyanobactins, however, the NHLP and Nif11 families do not contain PatA/G-like proteases (Figure 2.7).

NHLP precursors share sequence similarity to the alpha subunit of nitrile hydratases but are missing the catalytic requisite CxxCSC motif (Haft, *et al.* 2010). Nif11-derived peptides are only found in bacteria capable of fixing nitrogen and have similarity to the Nif11 protein, whose function is unknown. In some clusters, NHLP and Nif11 peptides are found concurrently. Similar to cyanobactins, both of these families of precursors again have hypervariable core regions, and some NHLP-Burk peptides appear to have multiple cleavage sites, suggesting the production of two compounds from a single precursor peptide (Haft, *et al.* 2010). The NHLP-Burk clusters contain tandem precursor peptide genes. In some NHLP-burk gene clusters, these precursors are fused, suggesting they may form a two-peptide product. Similar to cyanobactin precursor peptides, the NHLP, NHLP-Burk and Nif11 precursor peptides are long, making the proportion of Cys, Thr, and Ser within the predicted core peptide low. Therefore, these peptides were not identified using the parameters from this bioinformatics study (Haft, *et al.* 2010).

2.3.5 Goadsporin

Only two biosynthetic gene clusters for goadsporin production were identified in this study. Goadsporin promotes secondary metabolism and morphogenesis in actinomycetes at low concentration, but inhibits bacterial growth at higher concentrations (Onaka, *et al.* 2005). In addition to a fused TOMM cyclodehydratase and B protein, the goadsporin biosynthetic gene cluster contains a dehydratase for the conversion of Ser and Thr to dehydroalanine (Dha) and dehydrobutyrine (Dhb), respectively. These lanthipeptide-like dehydratase proteins are split into separate proteins (glutamylation and elimination domains, respectively) rather than a single polypeptide with two-domains that is often found in lanthipeptide gene clusters (Li and Kelly 2010, Mavaro, *et al.* 2011, Ortega, *et al.* 2014, Yu, Zhang, *et al.* 2013, Zhang, *et al.* 2012). These two proteins form distinct enzyme groups containing the dehydratases from not only goadsporin, but also thiopeptide and coryneazolisin producers (discussed below, Groups 8 and 15).

2.3.6 Thiopeptides

Thiopeptides are highly modified macrocyclic TOMMs best known for their inhibition of protein synthesis by interacting with the 50S ribosomal subunit or elongation factor Tu (Bagley, *et al.* 2005). The D proteins involved in thiopeptide biosynthesis do not form a single isofunctional group at e-value E-54 like the D proteins from most other TOMM clusters. Instead, roughly half form a unique group while the other half clusters with the HCA producers (Figure 2.3) (Haft 2009). After further investigation, it became apparent that there exists two different families of thiopeptide biosynthetic gene clusters. One family has a single fused cyclodehydratase that clusters with the heterocycloanthracin (HCA) producers, while the other class encodes a discrete C and D cyclodehydratase; occasionally, this type also contains an additional fused cyclodehydratase.

Thiopeptide gene clusters that group with HCA gene clusters at E-54 include those responsible for production of thiostrepton, thiocillin, and other well-characterized thiopeptides.

Within these clusters, 85% contain a ThiF-like domain containing protein (TOMM F protein, Figures 2.7 and 2.9) that is presumably responsible for precursor peptide binding, as has been demonstrated during HCA biosynthesis (*vide infra*, Dunbar *et al.*, *submitted*). Only two natural products have been identified from the thiopeptide gene clusters that contain the discrete cyclodehydratase, TP-1161(Engelhardt, Degnes, Kemmler, *et al.* 2010, Engelhardt, Degnes and Zotchev 2010) and berninamycin(Malcolmson, *et al.* 2013). Only 25% of these gene clusters contain an F protein, suggesting that the C proteins from these gene clusters are capable of engaging the precursor peptide on their own.

The distinguishing feature of thiopeptides is a central nitrogen-containing six-membered ring formed from two dehydroalanines (Bowers, *et al.* 2010, Malcolmson, *et al.* 2013, Tocchetti, *et al.* 2013). The [4+2] cycloaddition enzyme responsible for the formation of the thiocillin macrocycle was recently heterologously produced and studied (Wever, *et al.* 2015). Homologs to this protein are ubiquitous in thiopeptide gene clusters (Bowers, *et al.* 2010).

2.3.7 Plantazolicin

Plantazolicin (PZN), a recently characterized natural product with highly discriminating antibiotic activity, forms a small isofunctional group in the sequence similarity network with 13 members (Figure 2.3) (Molohon, *et al.* 2011, Scholz, *et al.* 2011). The PZN gene cluster was initially identified in *Bacillus amyloliquefaciens*, but has since been identified in additional *Bacillus* species as well as from actinomycetes such as *Clavibacter*, *Brevibacterium*, and *Corynebacterium* (Molohon, *et al.* 2011). The current study identifies additional PZN clusters in the *Nesterenkonia* and *Sorangium* genera. In an early report on PZN (Scholz, *et al.* 2011), it was determined that dimethylation of the N-terminal arginine was required for activity. The PZN S-adenosyl methionine (SAM)-dependent methyltransferase responsible for this dimethylation was later reconstituted and found to be specific for PZN-like substrates, appearing to require an N-

terminal arginine followed by a thiazole (Hao, *et al.* 2015, Lee, *et al.* 2013, Piwowarska, *et al.* 2013). Due to this specificity, it is not surprising that the PZN methyltransferase forms a distinct enzyme group within the modification enzymes (Figure 2.9, group not shown). The precursor genes from these clusters are smaller (~130 bp) than most TOMM precursor peptides and consequently, all were identified manually. Of the identified clusters, 12 contain the PZN-specific methyltransferase (all but the *Nesterenkonia* cluster) and 10 have a core peptide region predicted to begin with Arg. The core regions of these 10 precursor peptides are very similar to the core of the initially-described PZN peptide from *B. amyloliquefaciens*, containing 5 heterocyclizable residues near the N-terminus, followed by two nonpolar amino acids, and 5 or 6 more heterocyclizable residues near the C-terminus (Figure 2.8).

2.3.8 Hakacin

The TOMMs of the hakacin group (Figure 2.3) have discrete cyclodehydratases, and although the C and D proteins have been extensively characterized *in vitro*, the final structure and function of any hakacin remains undetermined (Dunbar, *et al.* 2012, Dunbar, *et al.* 2013, Melby, *et al.* 2012, Melby, *et al.* 2014). The current analysis identified similar clusters from 16 *Bacillus cereus* and *Bacillus thuringiensis* strains. In addition to the cyclodehydratase, hakacin gene clusters encode a B protein, protease, ABC transporters, and a group-specific protein of unknown function (Figure 2.7). Interestingly, there are three groups of hakacin precursor peptides that vary in the core region; however, the leader regions are nearly identical (Figure 2.8).

2.3.9 Heterocycloanthracin

The heterocycloanthracin (HCA) producers comprise a large group of TOMMs with 254 being identified in this study. First bioinformatically identified in 2009,(Haft 2009) the genes responsible for the production of HCA were recently reconstituted *in vitro* (Dunbar *et al.*, *submitted*). These genes are widely distributed in the *Bacillus cereus* group, with the majority of the sequenced strains containing a HCA gene cluster. All HCA producers contain a fused

cyclodehydratase that is missing ~100 amino acids from the N-terminal C protein domain. This truncation means that the cyclodehydratase lacks the critical residues involved in peptide recognition. It was recently demonstrated that the ThiF-like protein (TOMM F protein, IPR022291) identified in all HCA clusters (and most thiopeptides) is responsible for leader peptide binding. The TOMM F protein also forms a complex with the truncated cyclodehydratase, which is now dependent on the F protein for activity (Dunbar *et al.*, *submitted*). F proteins have so far only been found in the gene clusters of HCA producers and the thiopeptides that group with them, and they form a single cohesive group within the modification enzymes (Figure 2.9, group 4).

The clusters of the *B. cereus* HCA clusters contain additional modification enzymes, including a B protein, a SAM-dependent methyltransferase, a succinyltransferase, and a 2-oxoglutarate dehydrogenase, suggesting additional modifications could decorate these natural products. However, the genomic regions of these clusters are almost identical between strains, making it difficult to assign gene cluster boundaries. After comparison of the entire HCA family, only the fused cyclodehydratase, F protein, and B protein are present within all clusters and are potentially the only necessary enzymes within this cluster (unless there are other essential enzymes elsewhere on the chromosome).

Until 2009, an HCA precursor peptide could not be identified because in a majority of the *B. cereus* HCA clusters, the gene encoding the precursor peptide is not located in the local genomic context of the cyclodehydratase (Haft 2009). Using the method outlined above, any precursor peptides further than 10 kb from the D protein were not annotated, and therefore the majority of the precursor peptides from these clusters were left unannotated. However, 14 HCA precursor peptides were located close to the D proteins and thus were successfully annotated. These precursor peptides were similar to the ones identified in previous studies, and most contained either Cys-Ser or Gly-Cys repeats (Haft 2009).

2.3.10 Bottromycin and other TOMMs with a stand-alone D protein

Bottromycins display potent antimicrobial activity against methicillin-resistant *Staphylococcus aureus* and vancomycin-resistant enterococci (Kobayashi, *et al.* 2010). Characterized bottromycin gene clusters each contain two genes with YcaO-like domains similar to the D protein component of the TOMM cyclodehydratase, but no recognizable C protein (Crone, *et al.* 2012, Gomez-Escribano, *et al.* 2012, Hou, *et al.* 2012, Huo, *et al.* 2012). One of the D proteins is suspected to convert Cys to thiazoline while the second is postulated to be responsible for the formation of the macroamidine. The absence of a C protein in these stand-alone D protein TOMM clusters makes genome mining for TOMMs inherently more difficult. Bottromycin gene clusters contain methyltransferases necessary for the *O*-methylation of the aspartic acid and the methylation at non-nucleophilic β -carbons in bottromycin biosynthesis, respectively. For this study, similarity of these proteins as well as similarity of the D proteins were used to identify bottromycin and other stand-alone D protein clusters.

There are two known groups of YcaO domain-containing proteins (homologs of D proteins, but not associated with a C protein), the “non-TOMM YcaOs” and the “TfuA-associated non-TOMM YcaOs”. The latter co-occurs in clusters with a gene encoding for the protein TfuA, which is implicated in trifolitoxin production (Breil, *et al.* 1996, Dunbar, *et al.* 2014). Although all of these YcaO proteins contain the canonical ATP-binding pocket, the substrate of the non-TOMM and TfuA-associated YcaOs are unknown. These proteins were not included in this study; however, with the discovery of bottromycin biosynthesis, it is apparent that YcaO domain-containing proteins have the potential to synthesize natural products without a canonical C protein. Many of these uncharacterized YcaO proteins have the potential to produce novel natural products. Further bioinformatic and biochemical analysis will be necessary to determine if the non-TOMM YcaO enzymes are indeed involved in natural product biosynthesis.

2.4 Presumed Isofunctional Groups with no Characterized Members

A significant number of TOMM natural product classes do not group with any characterized biosynthetic clusters, thus representing an untapped source of structure and functional novelty (Figures 2.3 and 2.4).

2.4.1 Faecalisin

The largest group of uncharacterized TOMMs, referred to here as faecalisins, is comprised of 124 gene clusters found predominantly in *Enterococcus faecalis*. These clusters have discrete cyclodehydratases, and the D protein from the cluster is most related to those of MccB17 and a few of the stand-alone clusters (Figures 2.5 and 2.6). However, the C protein, responsible for leader peptide binding, does not form a group with C proteins from other TOMM classes (Figure 2.9, group not shown), implying that these clusters differ significantly from the MccB17 clusters. The faecalisin gene clusters also contain ABC transporters along with two hypothetical proteins that could be responsible for further modifications, but have no similarity with other TOMM ancillary modification enzymes (Figure 2.7).

Precursor peptide genes were identified for 102 of the faecalisin producers in this study, 20 of which contained two precursor genes within its cluster. Each of the identified precursor peptides has a core region containing a Gly repeat linker followed by a Cys repeat region (Figure 2.6), and all but three differ only by the length of the Gly linker, plus contain identical leader peptides (Figure 2.6).

2.4.2 Propionisin

A group of 19 TOMM gene clusters from *Propionibacterium* contain a discrete cyclodehydratase with the D protein most related to the cytolysin family (Figures 2.5, 2.6 and 2.7) though the C protein does not form a group with the other C proteins (Figure 2.9, group not shown). These propionisin gene clusters contain ABC transporters as well as hypothetical proteins that do not share any similarity to other TOMM enzymes, but could potentially modify

the natural product (Figure 2.7). Unlike most TOMM clusters, the propionisin gene clusters also contain multiple CaaX-like proteases (Maxson, *et al.* 2015).

A precursor peptide gene was identified for all predicted propionisin gene clusters. The majority of the strains (14/19) contained two identified precursor peptide genes, and three strains contained three. The precursor peptides cluster by similarity into three groups. The first two groups differ dramatically in leader peptide sequence but contain nearly identical core regions. These core regions appear similar to those of the cytolysin precursor peptides because they contain contiguous heterocyclizable residues followed by C-terminus with no Cys, Ser, and Thr. The third group of propionisin precursor peptides, meanwhile, have almost no similarity to the other two. Further experimentation is necessary to establish if these are actual TOMM precursor peptides (Figure 2.8).

2.4.3 Helicobactin

Another putative type of TOMM uncovered, the helicobactins, are encoded by 10 *Helicobacter pylori* strains. These TOMM clusters contain a discrete cyclodehydratase with a D protein most closely related to those of the hakacins and thermoacidophilins (Figures 2.5 and 2.6), while the C protein groups by itself when compared to other homologs (Figure 2.9, group 83). These clusters also contain a B protein and a hypothetical protein that shares similarity only with other *H. pylori* enzymes. Some helicobactin clusters contain ABC transporters as well as a protease (Figure 2.7); however, this is not strictly conserved throughout the family. Precursor peptides were identified for eight of the helicobactin clusters. These precursor peptides are nearly identical, with only a single substitution in the predicted leader peptide.

2.4.4 Mobilisin

The mobilisins, a family of TOMMs produced mainly by strains of *Mobiluncus*, form a predicted isofunctional group with 52 D proteins (Figure 2.3). The D proteins from these clusters are most similar to those from the gallolytisin and haloazolisin clusters (Figure 2.5 and 2.6). The

mobilisin gene clusters appear to only have the B, C, and D proteins (Figure 2.7). Precursor peptides were not identified bioinformatically for these clusters, implying that these precursor peptides could either be extremely different from previously identified or be encoded elsewhere in the genome. Further manual analysis identified a short peptide near the fused cyclodehydratase, however the core region contains a low percentage of Cys, Ser, and Thr residues explaining the lack of automatic identification.

2.4.5 Haloazolisin

Halophilic archaea contain a family of nearly 100 TOMM gene clusters, which we term the haloazolisins. These gene clusters have very divergent fused cyclodehydratases with a barely recognizable C protein domain; however, some clusters do contain a recognizable precursor peptide, which allowed for their classification as TOMM gene clusters (Figure 2.8). This cyclodehydratase is most similar to those from other uncharacterized TOMM clusters, including the anabaenasin, mobilisin, and gallolytisin clusters (Figure 2.5 and 2.6). After further analysis, a precursor peptide was located near a F-like protein elsewhere on the chromosome of *Haloterrengina turkmenica*. Similar to the thiopeptide and HCA clusters, haloazolisin gene clusters encode a truncated, fused CD cyclodehydratase (missing roughly 200 amino acids from the N-terminus); however, the precursor peptide binding region is also missing from the F-like protein. Therefore, it is suspected that another uncharacterized protein within the cluster would be responsible for leader peptide binding, if these clusters do indeed generate a TOMM.

The haloazolisin precursor peptides are highly divergent, suggesting that this family may produce additional TOMMs. We identified 31 precursor peptides in these clusters with most having a Ser-rich core region (Figure 2.8). These clusters offer not only a wealth of potential novel TOMM structures and modification machinery, but also provide the opportunity to characterize natural product biosynthesis in archaea, which has been largely overlooked.

2.4.6 Thermoacidophisin

An additional archaeal family of TOMMs was identified in the genus *Sulfolobus*, specifically strains of *S. acidocaldarius* and *S. islandicus*. Four other related clusters were discovered in bacterial organisms, *Thermoanaerobacter mathranii* subsp. *mathranii* Str. A3, *Actinomyces odonolyticus* F0309, *Bacillus cereus* Rock3-44 and *Caldisericum exile* DSM 21853. All of these clusters have discrete cyclodehydratases, and their D proteins are most closely related to the helicobactin and PZN proteins (Figures 2.5 and 2.6), while the C proteins make up a single group of proteins unrelated to other C proteins (Figure 2.9, group not shown). The thermoacidophisin gene clusters also contain a B protein, ABC transporters, a regulator, and many hypothetical proteins (Figure 2.7).

Precursor peptides were identified for four of the thermoacidophisin clusters, all of which contain an abundance of Tyr and Gly residues. Characterization of these archaeal and bacterial TOMMs will potentially provide insight into the evolution of TOMM biosynthesis and horizontal transfer. The thermoacidophisin cluster has clearly disseminated over large phylogenetic distances through horizontal gene transfer, as it is present in four different phyla (Crenarchaeota, Firmicutes, Actinobacteria, and Caldiserica). Interestingly, three of the five strains that contain this particular cluster are known thermophiles despite residing in different phyla.

2.4.7 Gallolytisin

A few presumed isofunctional clusters have exceptionally unique precursor peptide sequences and gene composition. The gallolytisins are TOMMs encoded by a subset of only 20 strains, including *Streptococcus gallolyticus*. These clusters contain a discrete cyclodehydratase, and the D proteins are most similar to the D proteins from the PZN cluster (Figures 2.5 and 2.6). The C proteins from these clusters form a separate clade when compared to all other modification enzymes (Figure 2.9, group not shown). The gallolytisin clusters also contain ABC transporters

and a regulator (Figure 2.7). Seven gallolytisin precursor peptides were identified, all of which contain a highly conserved CCCCXCCCC motif, where X is Pro, Ala, or Asp (Figure 2.8).

2.4.8 Anabaenasin

Anabaenasins are encoded by 11 varied species. Their gene cluster contain a discrete cyclodehydratase; with a D protein most similar to the D proteins from the haloazolisin and mobilisin gene clusters and a unique C protein (Figure 2.9, group not shown). Surprisingly, the cluster from *Anabeana* sp. 90 contains a transposase gene directly between the C and D proteins, suggesting that these clusters could be either mobile or inactive. This cluster architecture is not conserved within all of the anabaenasin family members. Five precursor peptides were identified in these clusters, all of which are Gly and Cys rich.

2.4.9 Coryneazolisin type 1 and type 2

The strains of *Corynebacterium* associated with TOMM clusters are all disease-causing, including *C. diphtheriae*, *C. ulcerans*, and *C. pseudotuberculosis*. Although prominent AB toxins from these strains have been characterized, (Pappenheimer 1977) the TOMMs from these classes have not, and as such, it remains unknown whether these coryneazolisins play a role in pathogenesis akin to SLS (Molloy, *et al.* 2011). These gene clusters contain two D proteins which form distinct groups; one discrete (type 1) and one that is fused with a C protein (type 2) (Figure 2.3). The coryneazolisin clusters also contain lanthipeptide-like dehydratases, and similar to goadsporin, they lack the canonical [4+2] cycloaddition protein common to the thiopeptides, suggesting that these coryneazolisins are not macrocyclic (Figure 2.7).

Precursor peptides were identified in 24 coryneazolisin gene clusters. These precursor peptides are highly similar to each other, with only a single substitution in the leader region among them; however, they differ significantly from other TOMMs, making it difficult to predict the final product. The core region contains 10 Cys/Ser/Thr residues followed by an Ile, then 5-7 additional Cys/Ser/Thr residues (Figure 2.8). A subset of coryneazolisin gene clusters do not

contain identifiable precursor peptide or cyclodehydratase genes, suggesting that they may be inactive. Furthermore, these clusters are surrounded by transposable elements, and in some cases the D protein is fused to a transposable element, which can be indicative of horizontal gene transfer (Figure 2.12)

2.5 Distribution of TOMM Gene Clusters

Transfer of biosynthetic gene clusters has been previously discussed for many natural products. Although horizontal gene transfer of TOMMs has not been extensively studied, it is intriguing that many biosynthetic gene clusters contain or are flanked by transposase genes, remnants of transposable elements or tRNA genes. Although not a predominant group of genes identified in TOMMs, there are transposase genes found in the proximity of HCA, PZN, cyanobactin, hakacin, cytolysin, NHLP, faecalisin, microcin B17, thermoacidophisin, thiopeptide and coryneazolisin clusters (Figure 2.9, Groups 49, 51, 71, 77 and 78). This suggests a potential mechanism for gene cluster transfer between organisms.

To explore the distribution and transmission of TOMM clusters, a phylogenetic tree was created using the 16S sequences from each TOMM producing organism. The TOMM clusters produced by each organism were then mapped onto the tree (Figures 2.13 and 2.14). TOMM gene clusters are found in 6% of bacteria and 35% of archaea among the sequenced organisms in Ensembl. At first glance, the *Firmicutes* appear to be the major producers of TOMMs. While *Firmicutes* may encode the greatest number of TOMM gene clusters, many are extremely similar (e.g. the HCA and cytolysin clusters). Most TOMM diversity is presented by other phyla, such as the *Proteobacteria*, *Actinobacteria* and *Euryarchaea*. Although similar TOMM families are most often produced by related organisms, there are striking examples of possible horizontal transmission of a TOMM between distantly related organisms. For example, the cytolysins are primarily found in *Firmicutes* (*Streptococcus*, *Clostridium*, *Listeria*, etc.), but they are also present in *Spirochaetes* (*Brachyspira*, *Borrelia*, etc.). When assessed *in vitro*, the cytolysin from

Borrelia did possess a similar hemolytic phenotype as that of streptolysin S (Molloy *et al.*, *submitted*). In addition, thermoacidophilin-like clusters are found in *Crenarchaeota*, *Firmicutes* and *Actinobacteria*, suggesting these clusters may have been transferred between archaea and bacteria.

2.6 Summary and Outlook

This study characterized the database containing genomic complexity of TOMM natural product biosynthetic gene clusters. An in-depth analysis of TOMM clusters was used to identify nine novel TOMM families, as well as identify the predominant accessory enzymes that bestow additional structural diversity. Precursor peptides were also analyzed for sequence diversity within each class. This study revealed the diversity of TOMM clusters as well as the phylogenetic distribution of clusters in not only bacteria, but also archaea. With the geometric expansion in the rate of genome sequencing, it is expected that TOMM cluster diversity will increase as well, providing a large and growing source of new enzymes and natural products with potential medical or industrial implications.

2.7 Experimental

All YcaO domain-containing proteins (InterPro IPR003776, D protein) were obtained from InterPro on October 28th, 2014. An attempt was made to include all YcaO domain-containing proteins that have been sequenced, but many protein sequences from NCBI were not correlated with genomes or were not added to UniProt and therefore were not included in the characterization.

2.7.1 Biosynthetic gene cluster discovery and comparison

10-kb genomic regions on either side of the YcaO domain-containing proteins were obtained from NCBI, and predicted protein sequences were used as annotated. Genome regions were clustered using MultiGeneBlast (Blin, *et al.* 2013, Medema, *et al.* 2011). The database used was created from all of the genomic regions obtained from NCBI. 100 BLAST hits were mapped

with a synteny conservation hit weight of 0.5 and a BLAST hit weight of 0.5. The minimal BLAST sequence coverage was 25 and the minimal percent identity for BLAST hits was 30%. Genomic regions with a MultiGeneBlast score above 10 were grouped into families.

To identify TOMM biosynthetic gene clusters, profile Hidden Markov Models (pHMMs) and the program HMMER (Finn, *et al.* 2011) were used to identify C proteins from TOMM clusters. TIGR03603 and TIGR03882 were used to identify C proteins and CD fusion proteins, respectively. New pHMMs were created to identify short CD fusions similar to those in the haloazolisin clusters. Precursor peptides were identified as described below. Genomic regions were considered TOMMs if any members of the families identified with MultiGeneBlast contained a C or CD fusion protein identified with the pHMMs, the genomic region contained a precursor peptide (described below), or the genomic regions clustered with known bottromycin producers (Crone, *et al.* 2012, Gomez-Escribano, *et al.* 2012, Hou, *et al.* 2012, Huo, *et al.* 2012) (a TOMM with no identifiable C protein and a non-canonical precursor peptide).

2.7.2 Sequence similarity networks

The D proteins from all of the identified TOMM gene clusters were used to make the D-only sequence similarity networks. Similarity was evaluated using an all-vs-all BLAST with an e-value cutoff of E-54. To create the network with all of the TOMM proteins, proteins were predicted from NCBI gene annotations. All proteins within the genomic region were submitted to the Enzyme Function Initiative – Enzyme Similarity Tool (enzymefunction.org) for analysis. The similarity was calculated at an e-value of E-30 with a representative node cluster of 100%. Both networks were visualized with Cytoscape (cytoscape.org) using the organic layout.

2.7.3 Precursor sequence discovery

Precursor peptides were identified using two methods. In one, the NCBI-annotated genes from all of the genomic regions surrounding a YcaO domain-containing protein were analyzed, and any genes smaller than 450 bp were considered precursor peptides if the residues in the C-

terminal half of the encoded product were at least 45% Cys, Ser, or Thr. Because gene annotation programs often have difficulty annotating small open reading frames, the second method first determined all possible open reading frames in each genomic region. Any potential protein under 150 amino acids with at least 65% of the residues in the C-terminal half being Cys, Ser, or Thr were considered precursor peptides. Duplicates were removed. Precursor peptides vary in both sequence and length, and therefore, it is likely that many precursor peptides remained unidentified using this stringent method. Furthermore, any precursor peptides encoded elsewhere in the genome would be left unannotated with this analysis, as is the case with many HCA precursor peptides.

2.7.4 Phylogenetic analysis

D protein sequences were obtained from UniProt, and 16S rRNA sequences were obtained from SILVA by searching for the organism name from UniProt. All phylogenetic analysis was done using Molecular Evolutionary Genetics Analysis (MEGA) (Tamura, *et al.* 2013). Sequences were aligned using MUSCLE (Edgar 2004, Edgar 2004, Goujon, *et al.* 2010, McWilliam, *et al.* 2013) with all standard parameters. Maximum likelihood phylogenetic trees were created in MEGA using the standard parameters.

2.8 Figures

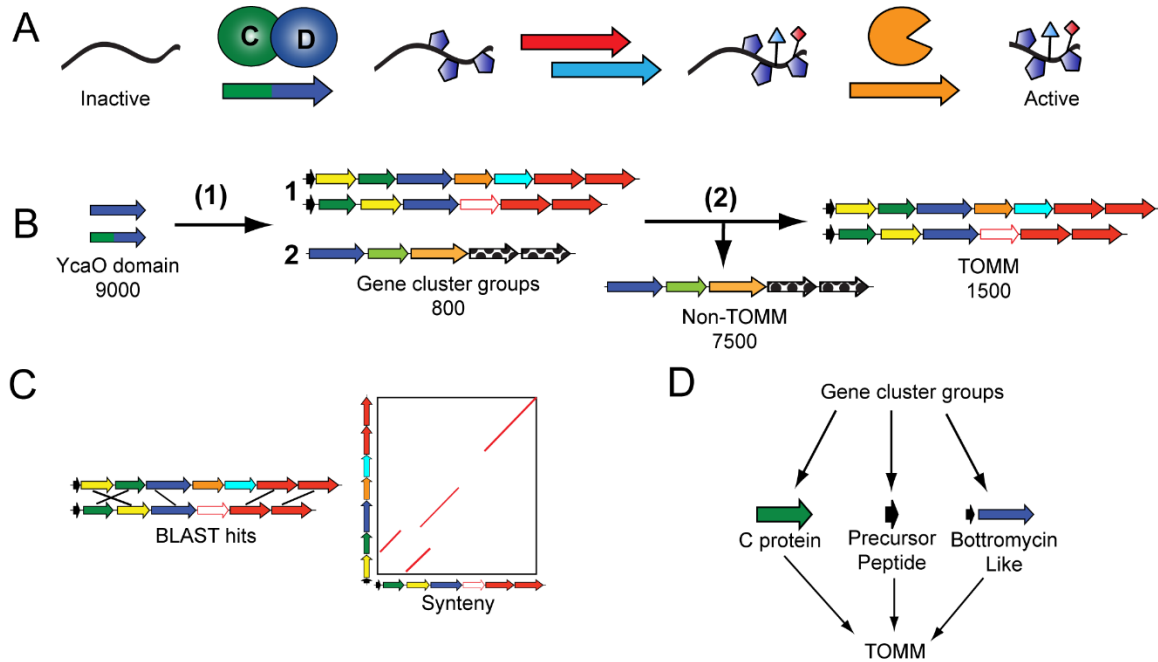


Figure 2.1 | Schematic of bioinformatics analysis. (A) TOMM biosynthesis begins with the ribosomal synthesis of a precursor peptide. The characteristic thiazoline/oxazoline heterocycles of a TOMM are installed by the C and D protein complex colored green and blue, respectively. Other tailoring enzymes (red and teal) often install additional modifications on the maturing product before the proteolytic cleavage (orange) of the leader peptide. (B) To identify TOMMs, all proteins containing a YcaO domain were identified using InterPro (IPR003776). The genomic regions surrounding the YcaO domains were retrieved, analyzed, and grouped by their cumulative BLAST bit score and synteny (Step 1). TOMM clusters were then separated from non-TOMM gene clusters determined by the inclusion of a C protein, precursor peptide, or bottromycin-like D protein (Step 2). (C) BLAST and synteny values from MultiGeneBlast were used to group TOMM clusters (Step 1). (D) A gene cluster was classified as a TOMM if it contained a C protein, precursor peptide, or was similar to bottromycin (Step 2).

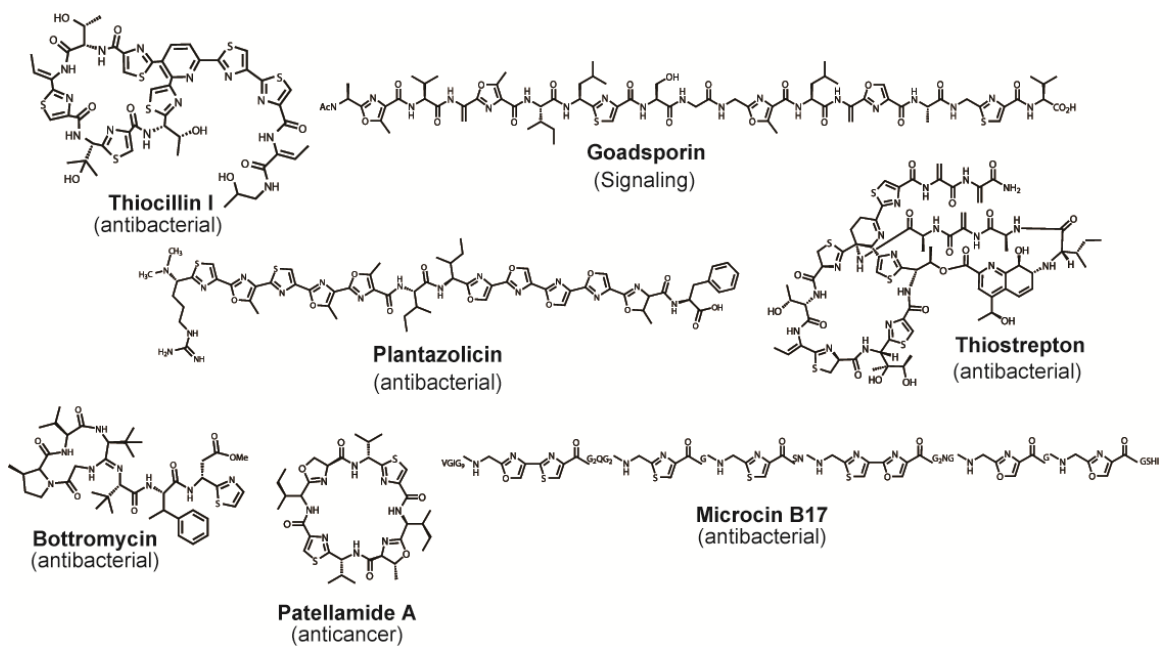


Figure 2.2 | Structures of a representative group explored TOMM compounds. Chemical structures from a few of the major classes of known TOMMs. Compound names and activities are listed below each structure.

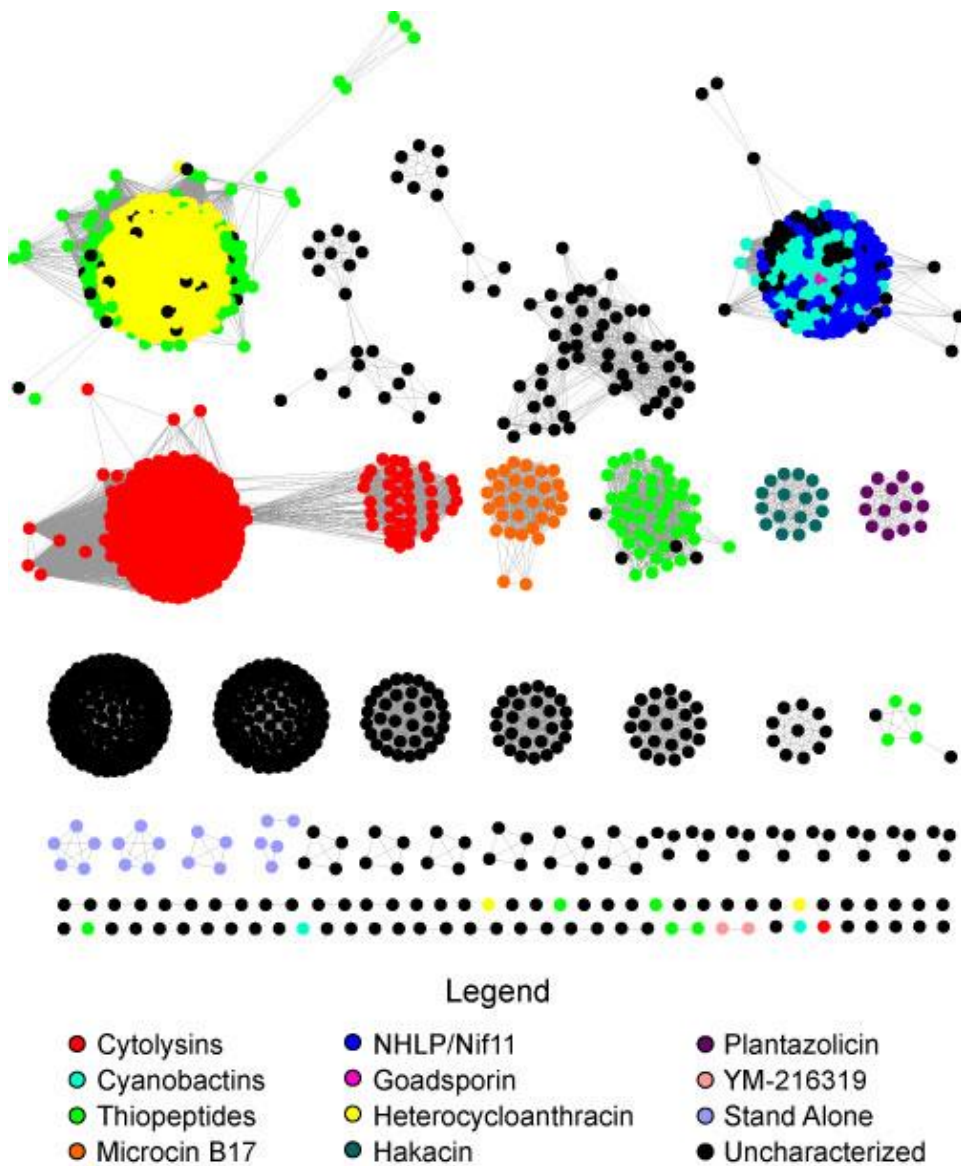


Figure 2.3 | Sequence similarity network of TOMM D-proteins. Each node represents a unique D protein (YcaO, from InterPro family IPR003776), while an edge indicates that two proteins have a BLAST expectation value $< 1E-54$. All nodes belonging to TOMM families with at least one characterized gene cluster (structure of final product not necessary) are colored as noted in the legend. Black isofunctional groups indicate that no member of the group has been characterized.

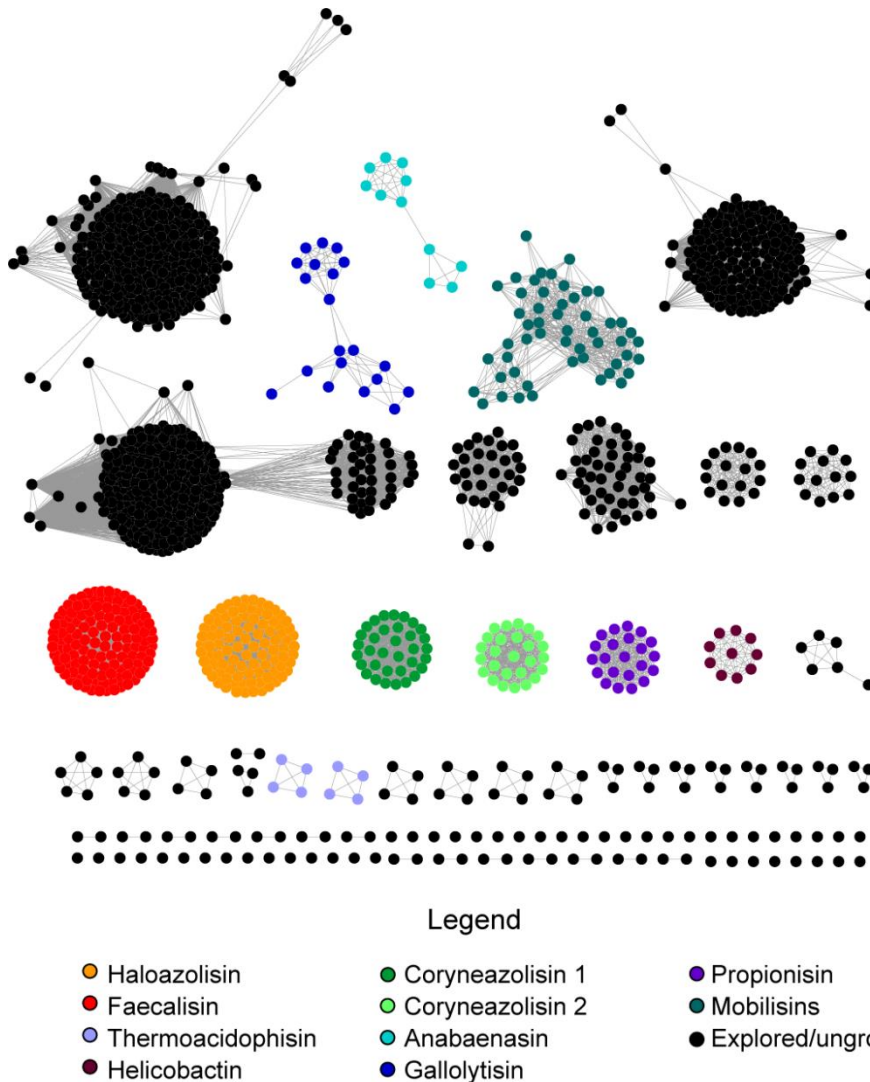


Figure 2.4 | Sequence similarity network of TOMM D proteins. Each node represents a unique D-protein, while an edge indicates that two proteins have a BLAST expectation value $< 1E-54$. All nodes from uncharacterized TOMM families are colored as noted in the legend. All nodes in TOMM families with at least one characterized gene cluster (structure of final product not necessary) are colored black.

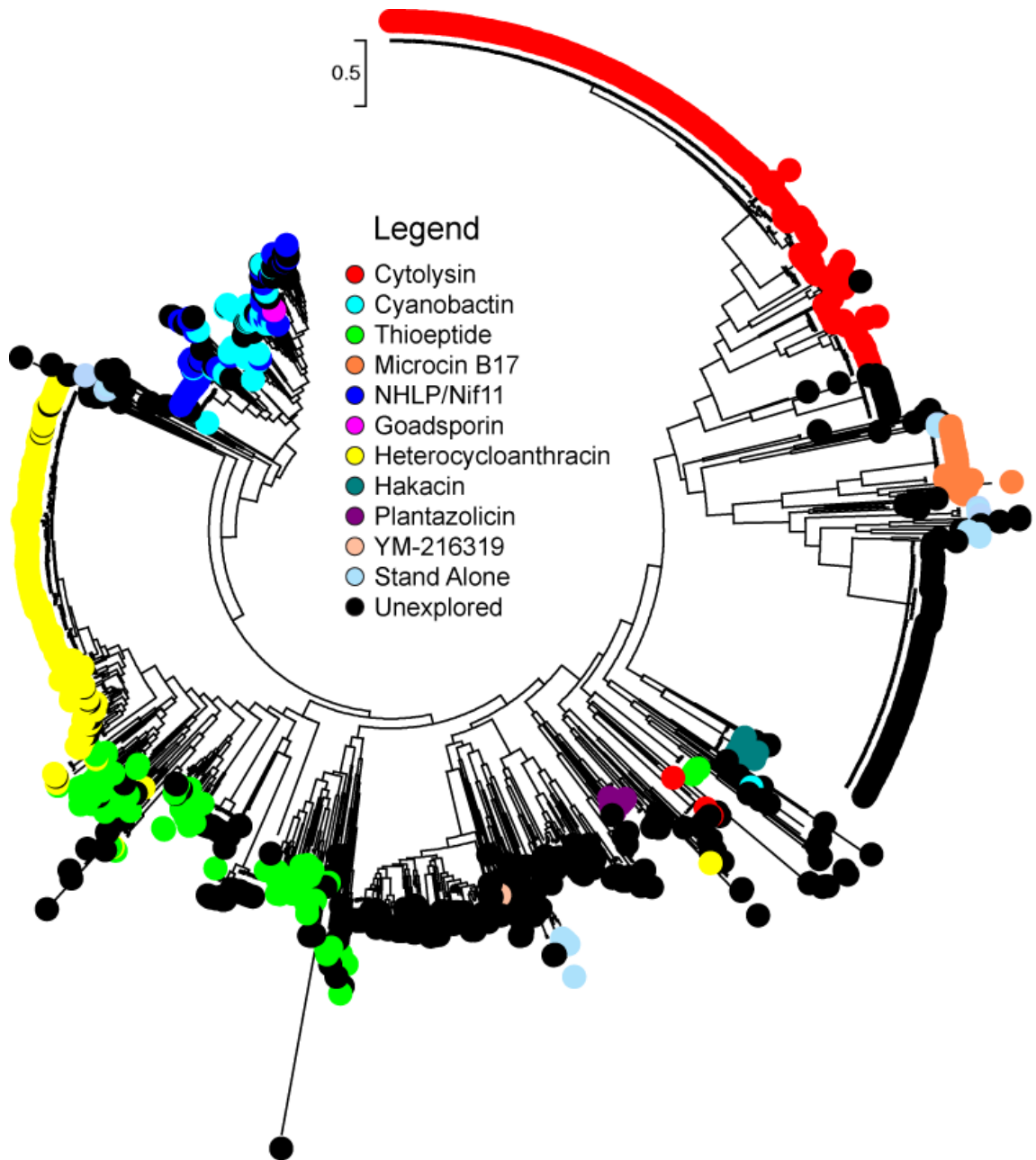


Figure 2.5 | Phylogenetic analysis of TOMM D proteins. A maximum likelihood tree was constructed using the D protein sequence from all TOMM producers. The class of characterized TOMM was then mapped on with colored circles as represented in the legend. Similar TOMM clusters seen in the sequence similarity network (Figure 2.3) are seen grouping here.

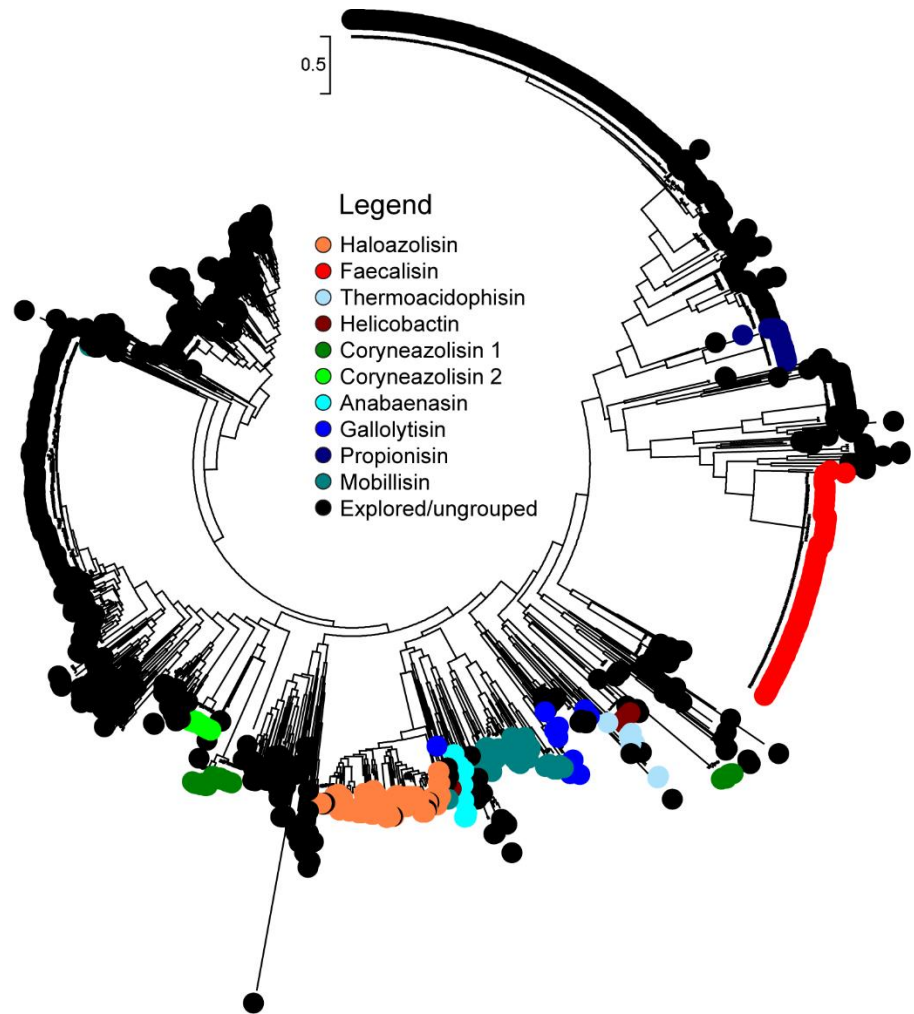


Figure 2.6 | Phylogenetic analysis of TOMM D proteins. A maximum likelihood tree was constructed using the D protein sequence from all TOMM producers. The class of uncharacterized TOMM was then mapped on with colored circles as represented in the legend. Similar TOMM clusters seen in the sequence similarity network (Figure 2.3) are seen grouping here. This tree is identical to the tree from Additional File 3: Figure S3, but with different colors mapped onto the tree for identification of the uncharacterized TOMM classes.

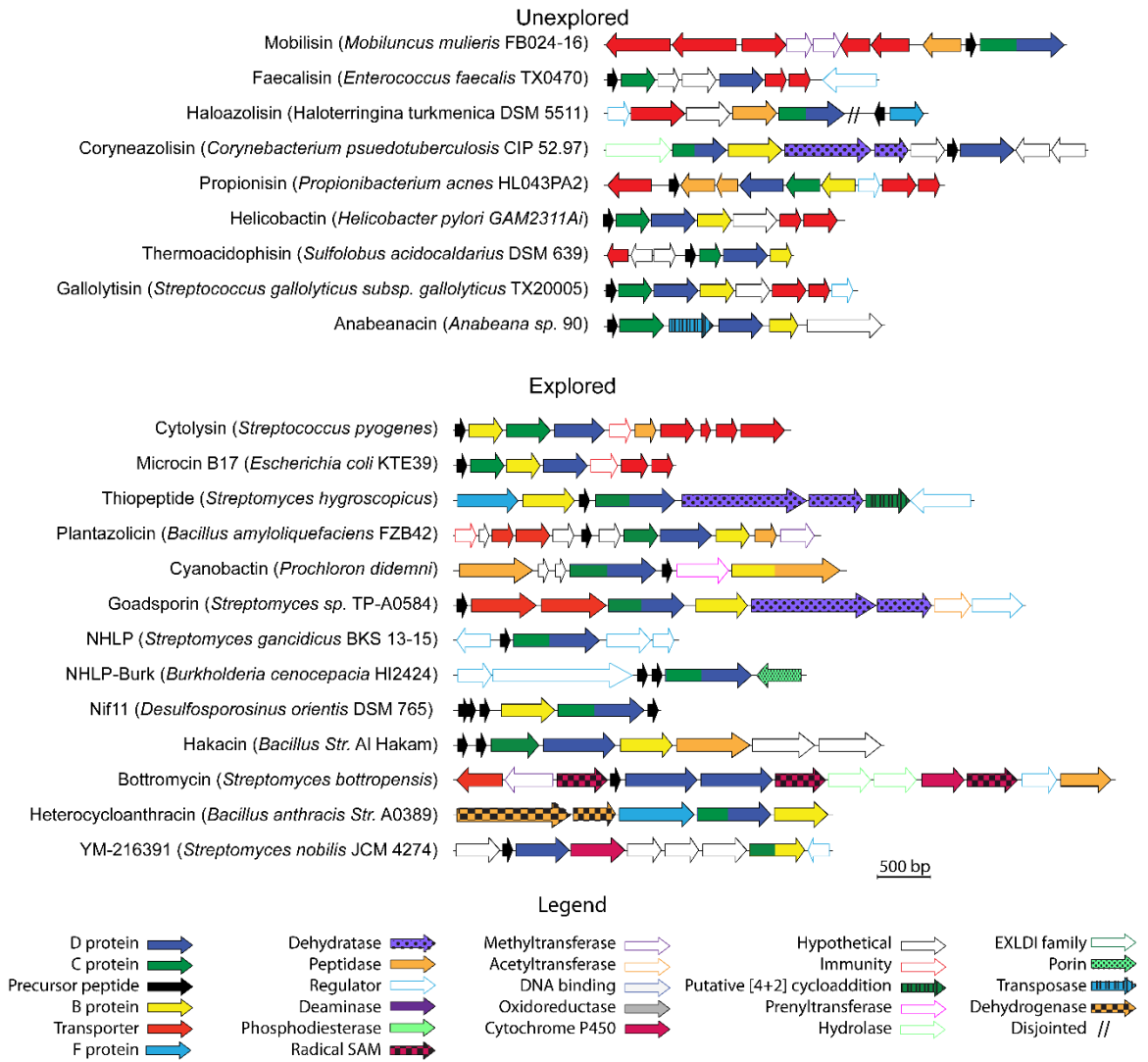


Figure 2.7 | Representative gene clusters from each TOMM subclass. Open reading frame diagrams are shown for a representative organism of each TOMM family. Uncharacterized gene clusters represent subclasses of TOMMs from which no gene clusters that have explored. Characterized clusters represent subclasses from which at least one gene cluster has been explored.

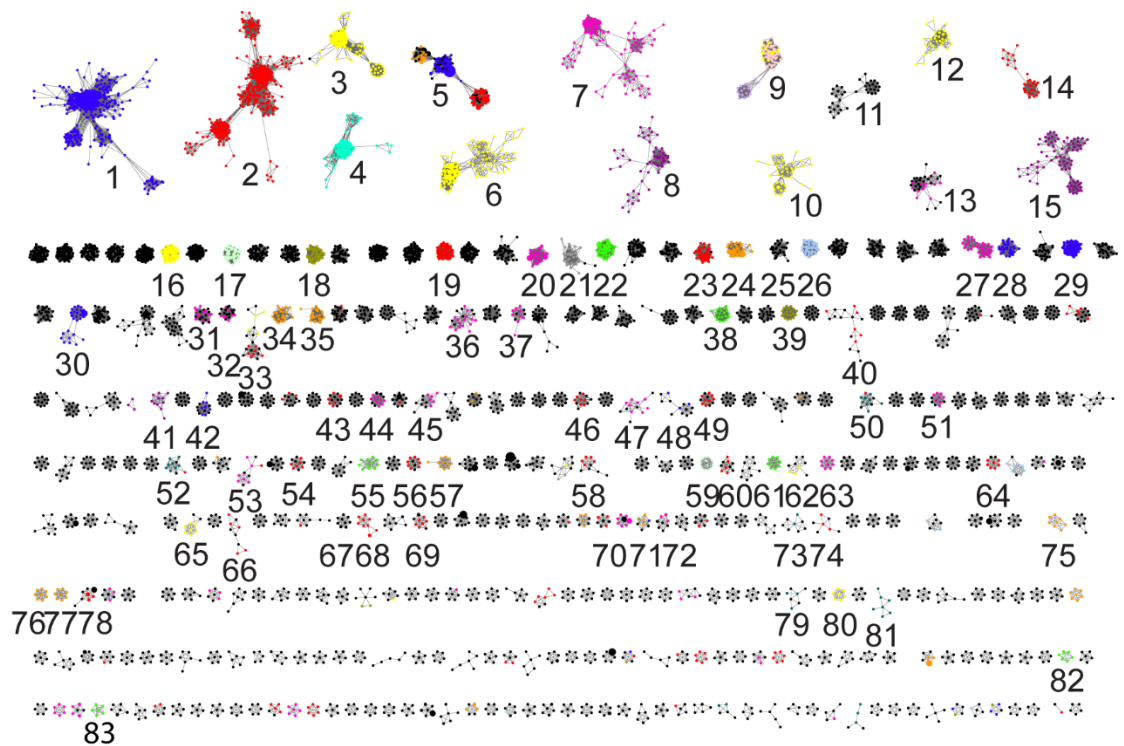


Figure 2.9 | The prevalence and distribution of enzymes involved in TOMM biosynthesis. A sequence similarity network was constructed with all proteins in the TOMM biosynthetic gene clusters visualized at a BLAST expectation value of 1E-30. All proteins with 100% identity were removed and are represented as larger nodes on the network (size is dependent on the number of redundant proteins). Groups are number for reference within the manuscript.

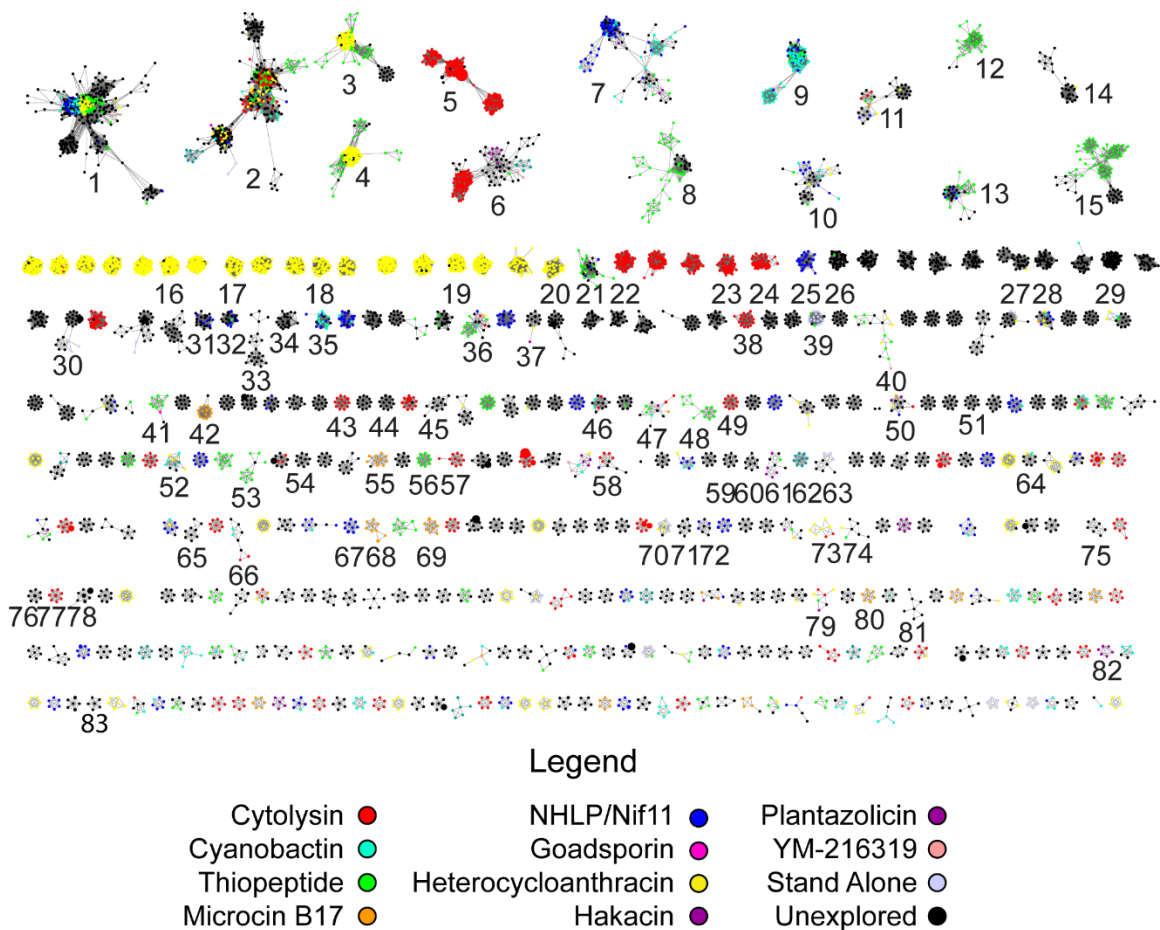


Figure 2.10 | The prevalence and phylogenetic distribution of enzymes involved in TOMM biosynthesis. A sequence similarity network with all proteins in the TOMM biosynthetic gene clusters visualized at an e-value of 1E-30. All proteins with 100% identity were removed and are represented as larger nodes on the network (size is dependent on the number of removed proteins).

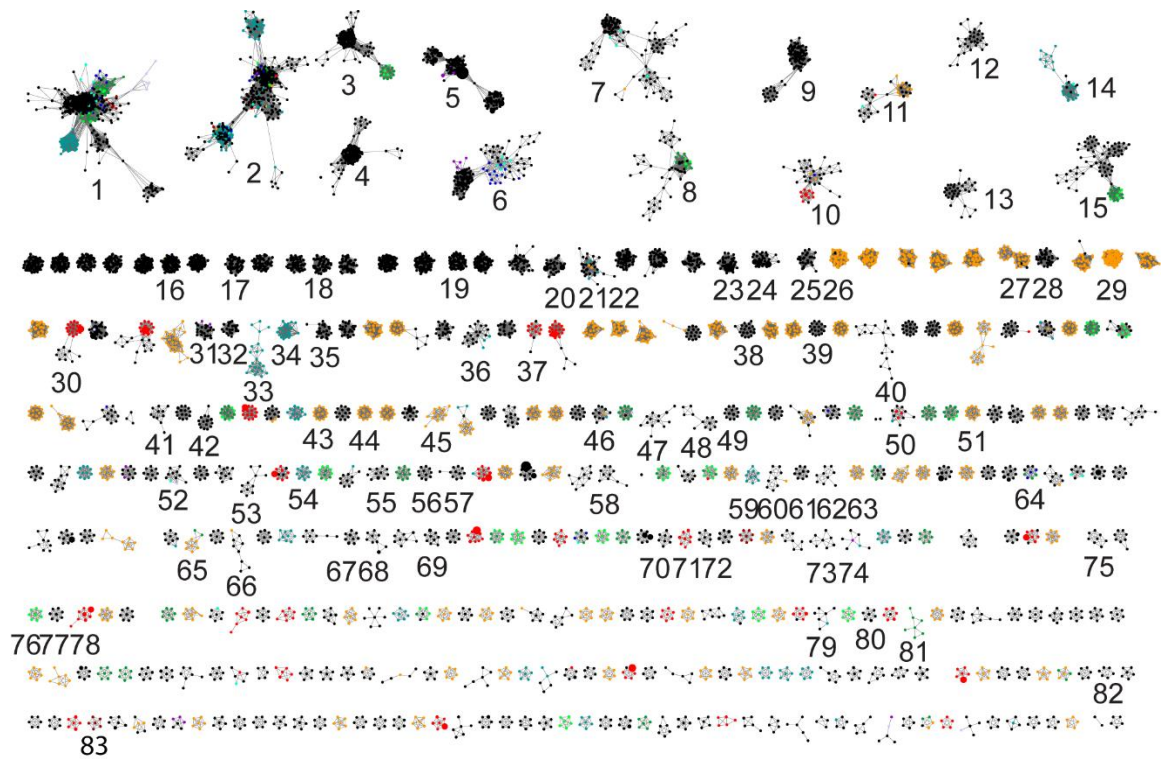


Figure 2.11 | The prevalence and phylogenetic distribution of enzymes involved in TOMM biosynthesis. A sequence similarity network with all proteins in the TOMM biosynthetic gene clusters visualized at an e-value of $1E-30$. All proteins with 100% identity were removed and are represented as larger nodes on the network (size is dependent on the number of removed proteins).

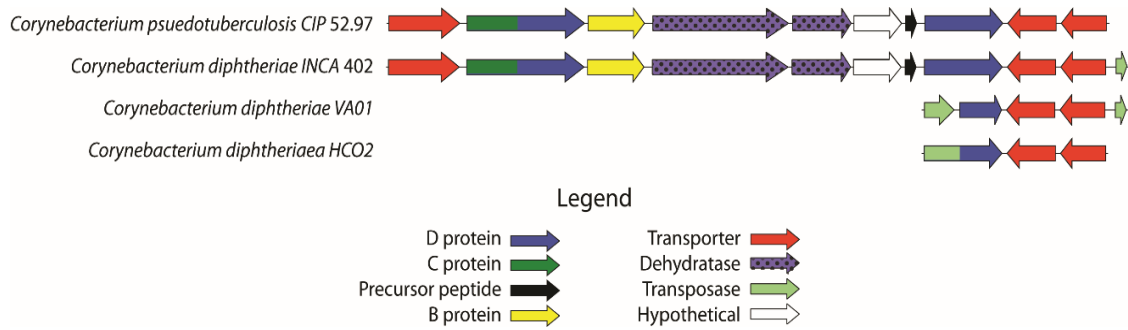


Figure 2.12 | Inactivated coryneazolisin cluster comparisons. Gene clusters from four potential coryneazolisin clusters are depicted. The two topmost clusters contain all the predicted enzymes required for coryneazolisin production. The second cluster from the top contains an additional transposase gene on the end. The third cluster is truncated and surrounded by transposable elements, and the fourth cluster contains a D protein that has been fused to a transposable element. It is likely that the two bottommost clusters have been inactivated.

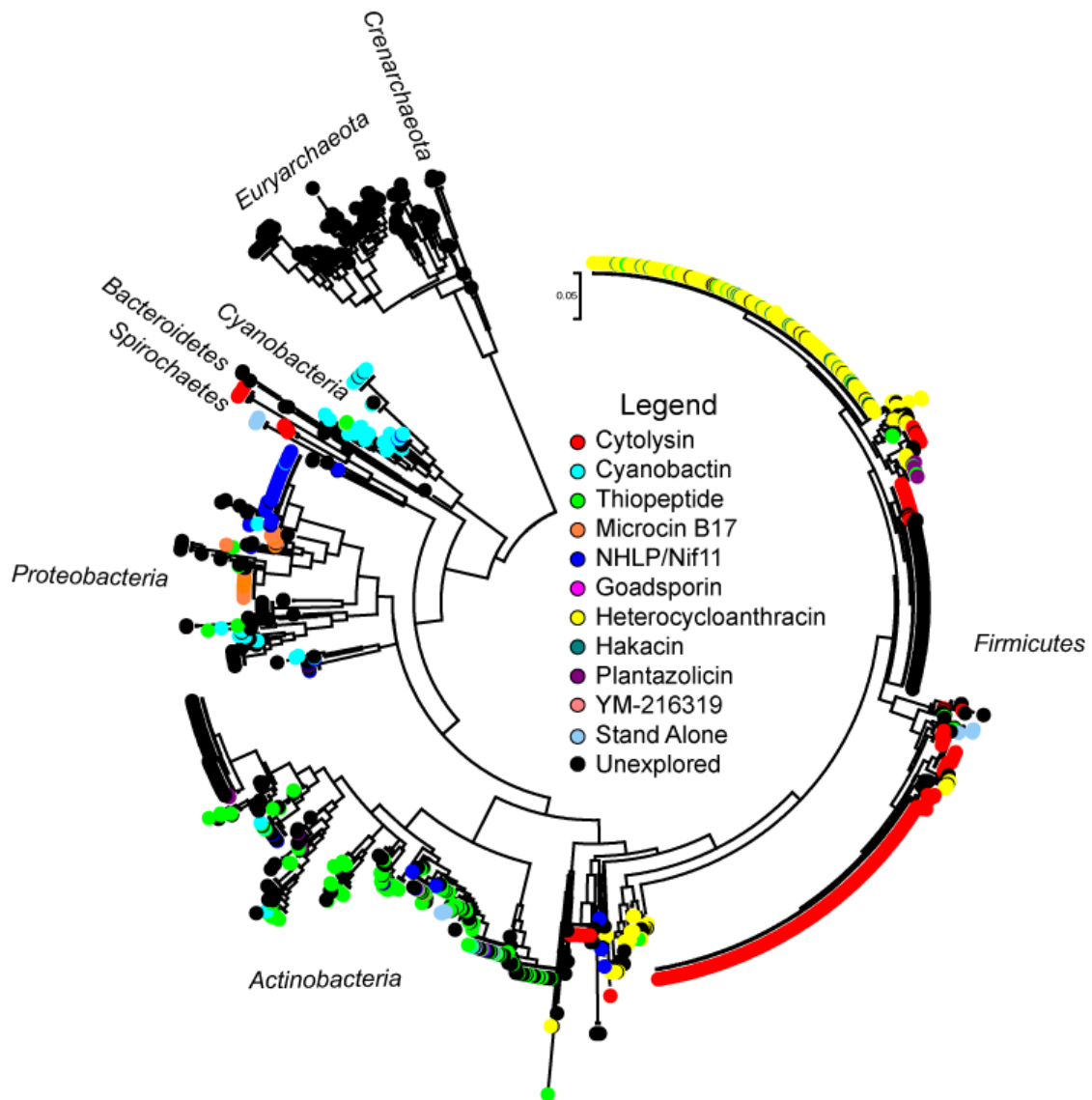


Figure 2.13 | Phylogenetic analysis of TOMM producers. A maximum likelihood tree was constructed using 16S sequences from all organisms that contain a TOMM gene cluster. Coloring indicates which class of TOMM that particular organism contains, per the legend. The phyla of the producing organisms are labeled around the tree. Most classes of TOMMs appear to be produced within the same phylum; however, some classes are found in multiple phyla.

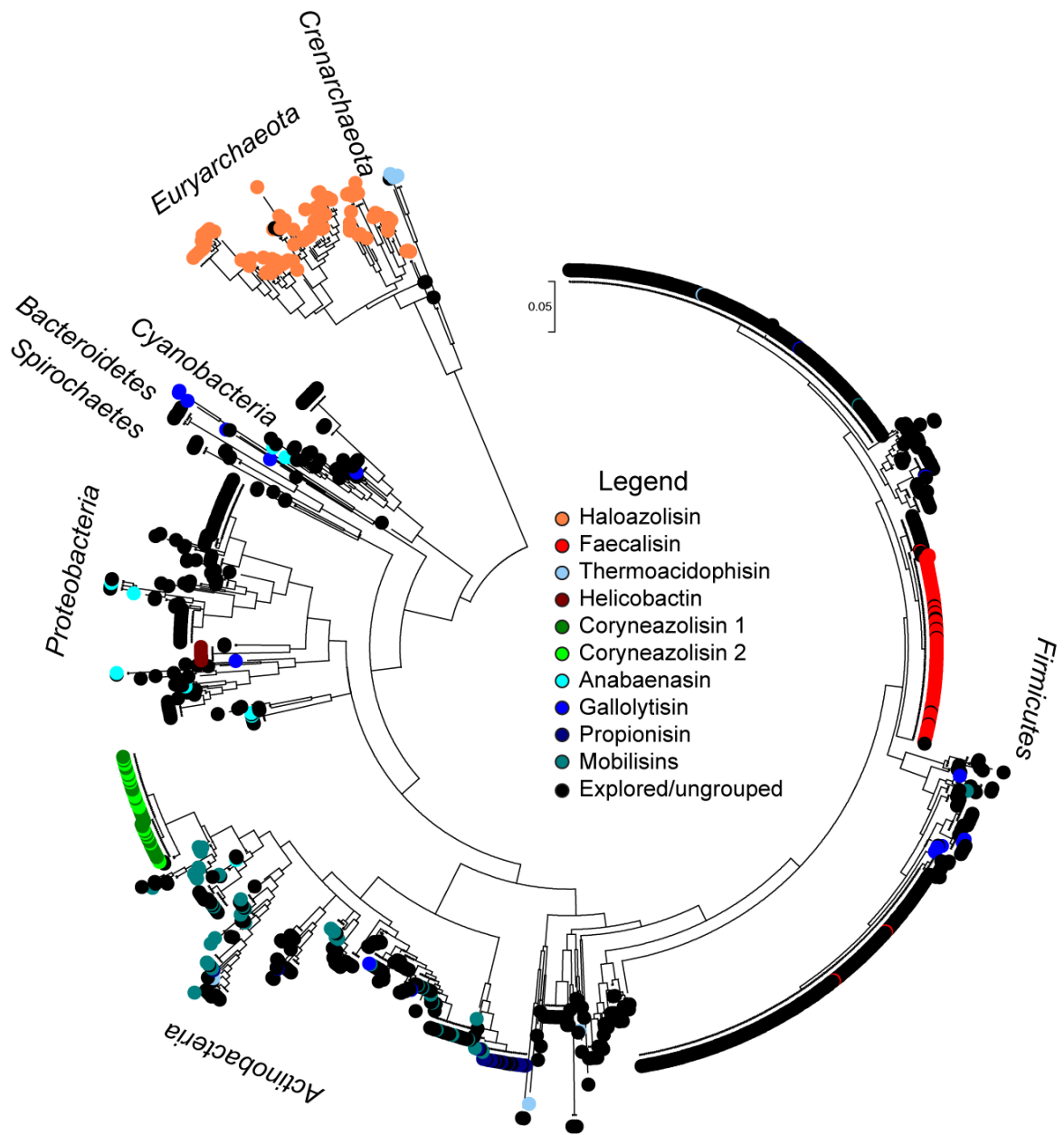


Figure 2.14 | Phylogenetic analysis of TOMM producers with uncharacterized clusters. A maximum likelihood tree was constructed using 16S sequences from all TOMM producers. This is the same tree produced in Figure 6, but with different TOMM classes mapped on with colored circles as represented in the legend. The phyla of the producing organisms are labeled around the tree. Most families of TOMMs appear to be produced within the same phylum; however, some are produced in multiple phyla.

2.9 References

1. Agarwal V, Pierce E, McIntosh J, Schmidt EW, and Nair SK. (2012) Structures of cyanobactin maturation enzymes define a family of transamidating proteases. *Chemistry & biology* 19: 1411-1422.
2. Arnison PG, Bibb MJ, Bierbaum G, Bowers AA, Bugni TS, Bulaj G, Camarero JA, Campopiano DJ, Challis GL, Clardy J, Cotter PD, Craik DJ, Dawson M, Dittmann E, Donadio S, Dorrestein PC, Entian KD, Fischbach MA, Garavelli JS, Goransson U, Gruber CW, Haft DH, Hemscheidt TK, Hertweck C, Hill C, Horswill AR, Jaspars M, Kelly WL, Klinman JP, Kuipers OP, Link AJ, Liu W, Marahiel MA, Mitchell DA, Moll GN, Moore BS, Muller R, Nair SK, Nes IF, Norris GE, Olivera BM, Onaka H, Patchett ML, Piel J, Reaney MJ, Rebuffat S, Ross RP, Sahl HG, Schmidt EW, Selsted ME, Severinov K, Shen B, Sivonen K, Smith L, Stein T, Sussmuth RD, Tagg JR, Tang GL, Truman AW, Vederas JC, Walsh CT, Walton JD, Wenzel SC, Willey JM, and van der Donk WA. (2013) Ribosomally synthesized and post-translationally modified peptide natural products: overview and recommendations for a universal nomenclature. *Natural product reports* 30: 108-160.
3. Bagley MC, Dale JW, Merritt EA, and Xiong X. (2005) Thiopeptide antibiotics. *Chemical reviews* 105: 685-714.
4. Belshaw PJ, Roy RS, Kelleher NL, and Walsh CT. (1998) Kinetics and regioselectivity of peptide-to-heterocycle conversions by microcin B17 synthetase. *Chemistry & biology* 5: 373-384.
5. Bent AF, Koehnke J, Houssen WE, Smith MC, Jaspars M, and Naismith JH. (2013) Structure of PatF from *Prochloron didemni*. *Acta crystallographica Section F, Structural biology and crystallization communications* 69: 618-623.
6. Betschel SD, Borgia SM, Barg NL, Low DE, and De Azavedo JC. (1998) Reduced virulence of group A streptococcal Tn916 mutants that do not produce streptolysin S. *Infection and immunity* 66: 1671-1679.
7. Blin K, Medema MH, Kazempour D, Fischbach MA, Breitling R, Takano E, and Weber T. (2013) antiSMASH 2.0--a versatile platform for genome mining of secondary metabolite producers. *Nucleic acids research* 41: W204-212.
8. Bowers AA, Walsh CT, and Acker MG. (2010) Genetic interception and structural characterization of thiopeptide cyclization precursors from *Bacillus cereus*. *Journal of the American Chemical Society* 132: 12182-12184.
9. Breil B, Borneman J, and Triplett EW. (1996) A newly discovered gene, *tfuA*, involved in the production of the ribosomally synthesized peptide antibiotic trifolixoxin. *Journal of bacteriology* 178: 4150-4156.

10. Crone WJK, Leeper FJ, and Truman AW. (2012) Identification and characterisation of the gene cluster for the anti-MRSA antibiotic bottromycin: expanding the biosynthetic diversity of ribosomal peptides. *Chemical science* 3: 3516-3521.
11. Donia MS, Ravel J, and Schmidt EW. (2008) A global assembly line for cyanobactins. *Nature chemical biology* 4: 341-343.
12. Donia MS, and Schmidt EW. (2011) Linking chemistry and genetics in the growing cyanobactin natural products family. *Chemical biology* 18: 508-519.
13. Doroghazi JR, and Metcalf WW. (2013) Comparative genomics of actinomycetes with a focus on natural product biosynthetic genes. *BMC genomics* 14: 611.
14. Dunbar KL, Chekan JR, Cox CL, Burkhart BJ, Nair SK, and Mitchell DA. (2014) Discovery of a new ATP-binding motif involved in peptidic azoline biosynthesis. *Nature chemical biology* 10: 823-829.
15. Dunbar KL, Melby JO, and Mitchell DA. (2012) YcaO domains use ATP to activate amide backbones during peptide cyclodehydrations. *Nature chemical biology* 8: 569-575.
16. Dunbar KL, and Mitchell DA. (2013) Insights into the mechanism of peptide cyclodehydrations achieved through the chemoenzymatic generation of amide derivatives. *Journal of the American Chemical Society* 135: 8692-8701.
17. Dunbar KL, and Mitchell DA. (2013) Revealing nature's synthetic potential through the study of ribosomal natural product biosynthesis. *ACS chemical biology* 8: 473-487.
18. Edgar RC. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC bioinformatics* 5: 113.
19. Edgar RC. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research* 32: 1792-1797.
20. Engelhardt K, Degnes KF, Kemmler M, Bredholt H, Fjaervik E, Klinkenberg G, Sletta H, Ellingsen TE, and Zotchev SB. (2010) Production of a new thiopeptide antibiotic, TP-1161, by a marine *Nocardiosis* species. *Applied and environmental microbiology* 76: 4969-4976.
21. Engelhardt K, Degnes KF, and Zotchev SB. (2010) Isolation and characterization of the gene cluster for biosynthesis of the thiopeptide antibiotic TP-1161. *Applied and environmental microbiology* 76: 7093-7101.

22. Finn RD, Clements J, and Eddy SR. (2011) HMMER web server: interactive sequence similarity searching. *Nucleic acids research* 39: W29-37.
23. Gomez-Escribano JP, Song LJ, Bibb MJ, and Challis GL. (2012) Posttranslational beta-methylation and macrolactamidation in the biosynthesis of the bottromycin complex of ribosomal peptide antibiotics. *Chemical science* 3: 3522-3525.
24. Goujon M, McWilliam H, Li W, Valentin F, Squizzato S, Paern J, and Lopez R. (2010) A new bioinformatics analysis tools framework at EMBL-EBI. *Nucleic acids research* 38: W695-699.
25. Haft DH. (2009) A strain-variable bacteriocin in *Bacillus anthracis* and *Bacillus cereus* with repeated Cys-Xaa-Xaa motifs. *Biology direct* 4: 15.
26. Haft DH, Basu MK, and Mitchell DA. (2010) Expansion of ribosomally produced natural products: a nitrile hydratase- and Nif11-related precursor family. *BMC biology* 8: 70.
27. Hao Y, Blair PM, Sharma A, Mitchell DA, and Nair SK. (2015) Insights into Methyltransferase Specificity and Bioactivity of Derivatives of the Antibiotic Plantazolicin. *ACS chemical biology*.
28. Hou Y, Tianero MD, Kwan JC, Wyche TP, Michel CR, Ellis GA, Vazquez-Rivera E, Braun DR, Rose WE, Schmidt EW, and Bugni TS. (2012) Structure and biosynthesis of the antibiotic bottromycin D. *Organic letters* 14: 5050-5053.
29. Houssen WE, Wright SH, Kalverda AP, Thompson GS, Kelly SM, and Jaspars M. (2010) Solution Structure of the Leader Sequence of the Patellamide Precursor Peptide, PatE(1-34). *Chembiochem* 11: 1867-1873.
30. Huo L, Rachid S, Stadler M, Wenzel SC, and Muller R. (2012) Synthetic biotechnology to study and engineer ribosomal bottromycin biosynthesis. *Chemistry & biology* 19: 1278-1287.
31. Ju KS, Doroghazi JR, and Metcalf WW. (2014) Genomics-enabled discovery of phosphonate natural products and their biosynthetic pathways. *Journal of industrial microbiology & biotechnology* 41: 345-356.
32. Kelly WL, Pan L, and Li C. (2009) Thiostrepton biosynthesis: prototype for a new family of bacteriocins. *Journal of the American Chemical Society* 131: 4327-4334.
33. Kobayashi Y, Ichioka M, Hirose T, Nagai K, Matsumoto A, Matsui H, Hanaki H, Masuma R, Takahashi Y, Omura S, and Sunazuka T. (2010) Bottromycin derivatives: efficient chemical modifications of the ester moiety and evaluation of anti-MRSA and anti-VRE activities. *Bioorganic & medicinal chemistry letters* 20: 6116-6120.

34. Koehnke J, Bent A, Houssen WE, Zollman D, Morawitz F, Shirran S, Vendome J, Nneoyiegbe AF, Trembleau L, Botting CH, Smith MC, Jaspars M, and Naismith JH. (2012) The mechanism of patellamide macrocyclization revealed by the characterization of the PatG macrocyclase domain. *Nature structural & molecular biology* 19: 767-772.
35. Lee J, Hao Y, Blair PM, Melby JO, Agarwal V, Burkhart BJ, Nair SK, and Mitchell DA. (2013) Structural and functional insight into an unexpectedly selective N-methyltransferase involved in plantazolicin biosynthesis. *Proceedings of the National Academy of Sciences of the United States of America* 110: 12954-12959.
36. Lee J, McIntosh J, Hathaway BJ, and Schmidt EW. (2009) Using marine natural products to discover a protease that catalyzes peptide macrocyclization of diverse substrates. *Journal of the American Chemical Society* 131: 2122-2124.
37. Lee SW, Mitchell DA, Markley AL, Hensler ME, Gonzalez D, Wohlrab A, Dorrestein PC, Nizet V, and Dixon JE. (2008) Discovery of a widely distributed toxin biosynthetic gene cluster. *Proceedings of the National Academy of Sciences of the United States of America* 105: 5879-5884.
38. Leikoski N, Liu L, Jokela J, Wahlsten M, Gugger M, Calteau A, Permi P, Kerfeld CA, Sivonen K, and Fewer DP. (2013) Genome mining expands the chemical diversity of the cyanobactin family to include highly modified linear peptides. *Chemistry & biology* 20: 1033-1043.
39. Letzel AC, Pidot SJ, and Hertweck C. (2014) Genome mining for ribosomally synthesized and post-translationally modified peptides (RiPPs) in anaerobic bacteria. *BMC genomics* 15: 983.
40. Li B, Sher D, Kelly L, Shi Y, Huang K, Knerr PJ, Joewono I, Rusch D, Chisholm SW, and van der Donk WA. (2010) Catalytic promiscuity in the biosynthesis of cyclic peptide secondary metabolites in planktonic marine cyanobacteria. *Proceedings of the National Academy of Sciences of the United States of America* 107: 10430-10435.
41. Li C, and Kelly WL. (2010) Recent advances in thiopeptide antibiotic biosynthesis. *Natural product reports* 27: 153-164.
42. Li YM, Milne JC, Madison LL, Kolter R, and Walsh CT. (1996) From peptide precursors to oxazole and thiazole-containing peptide antibiotics: microcin B17 synthase. *Science* 274: 1188-1193.
43. Liao R, Duan L, Lei C, Pan H, Ding Y, Zhang Q, Chen D, Shen B, Yu Y, and Liu W. (2009) Thiopeptide biosynthesis featuring ribosomally synthesized precursor peptides and conserved posttranslational modifications. *Chemistry & biology* 16: 141-147.

44. Maksimov MO, and Link AJ. (2014) Prospecting genomes for lasso peptides. *Journal of industrial microbiology & biotechnology* 41: 333-344.
45. Malcolmson SJ, Young TS, Ruby JG, Skewes-Cox P, and Walsh CT. (2013) The posttranslational modification cascade to the thiopeptide berninamycin generates linear forms and altered macrocyclic scaffolds. *Proceedings of the National Academy of Sciences of the United States of America* 110: 8483-8488.
46. Mavaro A, Abts A, Bakkes PJ, Moll GN, Driessen AJ, Smits SH, and Schmitt L. (2011) Substrate recognition and specificity of the NisB protein, the lantibiotic dehydratase involved in nisin biosynthesis. *The Journal of biological chemistry* 286: 30552-30560.
47. Maxson T, Deane CD, Molloy EM, Cox CL, Markley AL, Lee SW, and Mitchell DA. (2015) HIV Protease Inhibitors Block Streptolysin S Production. *ACS chemical biology*.
48. McIntosh JA, Donia MS, Nair SK, and Schmidt EW. (2011) Enzymatic basis of ribosomal peptide prenylation in cyanobacteria. *Journal of the American Chemical Society* 133: 13698-13705.
49. McIntosh JA, Donia MS, and Schmidt EW. (2010) Insights into heterocyclization from two highly similar enzymes. *Journal of the American Chemical Society* 132: 4089-4091.
50. McWilliam H, Li W, Uludag M, Squizzato S, Park YM, Buso N, Cowley AP, and Lopez R. (2013) Analysis Tool Web Services from the EMBL-EBI. *Nucleic acids research* 41: W597-600.
51. Medema MH, Blin K, Cimermancic P, de Jager V, Zakrzewski P, Fischbach MA, Weber T, Takano E, and Breitling R. (2011) antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. *Nucleic acids research* 39: W339-346.
52. Melby JO, Dunbar KL, Trinh NQ, and Mitchell DA. (2012) Selectivity, directionality, and promiscuity in peptide processing from a *Bacillus* sp. Al Hakam cyclodehydratase. *Journal of the American Chemical Society* 134: 5309-5316.
53. Melby JO, Li X, and Mitchell DA. (2014) Orchestration of enzymatic processing by thiazole/oxazole-modified microcin dehydrogenases. *Biochemistry* 53: 413-422.
54. Melby JO, Nard NJ, and Mitchell DA. (2011) Thiazole/oxazole-modified microcins: complex natural products from ribosomal templates. *Current opinion in chemical biology* 15: 369-378.

55. Metelev M, Serebryakova M, Ghilarov D, Zhao Y, and Severinov K. (2013) Structure of microcin B-like compounds produced by *Pseudomonas syringae* and species specificity of their antibacterial action. *Journal of bacteriology* 195: 4129-4137.
56. Milne JC, Roy RS, Eliot AC, Kelleher NL, Wokhlu A, Nickels B, and Walsh CT. (1999) Cofactor requirements and reconstitution of microcin B17 synthetase: a multienzyme complex that catalyzes the formation of oxazoles and thiazoles in the antibiotic microcin B17. *Biochemistry* 38: 4768-4781.
57. Mitchell DA, Lee SW, Pence MA, Markley AL, Limm JD, Nizet V, and Dixon JE. (2009) Structural and functional dissection of the heterocyclic peptide cytotoxin streptolysin S. *The Journal of biological chemistry* 284: 13004-13012.
58. Mohimani H, Kersten RD, Liu WT, Wang M, Purvine SO, Wu S, Brewer HM, Pasa-Tolic L, Bandeira N, Moore BS, Pevzner PA, and Dorrestein PC. (2014) Automated genome mining of ribosomal peptide natural products. *ACS chemical biology* 9: 1545-1551.
59. Molloy EM, Cotter PD, Hill C, Mitchell DA, and Ross RP. (2011) Streptolysin S-like virulence factors: the continuing saga. *Nature reviews Microbiology* 9: 670-681.
60. Molohon KJ, Melby JO, Lee J, Evans BS, Dunbar KL, Bumpus SB, Kelleher NL, and Mitchell DA. (2011) Structure determination and interception of biosynthetic intermediates for the plantazolicin class of highly discriminating antibiotics. *ACS chemical biology* 6: 1307-1313.
61. Morris RP, Leeds JA, Naegeli HU, Oberer L, Memmert K, Weber E, LaMarche MJ, Parker CN, Burren N, Esterow S, Hein AE, Schmitt EK, and Krastel P. (2009) Ribosomally synthesized thiopeptide antibiotics targeting elongation factor Tu. *Journal of the American Chemical Society* 131: 5946-5955.
62. Namikoshi M, and Rinehart KL. (1996) Bioactive compounds produced by cyanobacteria. *Journal of industrial microbiology & biotechnology* 17: 373-384.
63. Nizet V, Beall B, Bast DJ, Datta V, Kilburn L, Low DE, and De Azavedo JC. (2000) Genetic locus for streptolysin S production by group A streptococcus. *Infection and immunity* 68: 4245-4254.
64. Nunnery JK, Mevers E, and Gerwick WH. (2010) Biologically active secondary metabolites from marine cyanobacteria. *Curr Opin Biotech* 21: 787-793.
65. Oman TJ, and van der Donk WA. (2010) Follow the leader: the use of leader peptides to guide natural product biosynthesis. *Nature chemical biology* 6: 9-18.

66. Onaka H, Nakaho M, Hayashi K, Igarashi Y, and Furumai T. (2005) Cloning and characterization of the goadsporin biosynthetic gene cluster from *Streptomyces* sp. TP-A0584. *Microbiology* 151: 3923-3933.
67. Ortega MA, Hao Y, Zhang Q, Walker MC, van der Donk WA, and Nair SK. (2014) Structure and mechanism of the tRNA-dependent lantibiotic dehydratase NisB. *Nature* 517: 509-512.
68. Pappenheimer AM, Jr. (1977) Diphtheria toxin. *Annual review of biochemistry* 46: 69-94.
69. Pei J, Mitchell DA, Dixon JE, and Grishin NV. (2011) Expansion of type II CAAX proteases reveals evolutionary origin of gamma-secretase subunit APH-1. *Journal of molecular biology* 410: 18-26.
70. Piwowarska NA, Banala S, Overkleeft HS, and Sussmuth RD. (2013) Arg-Thz is a minimal substrate for the N(alpha),N(alpha)-arginyl methyltransferase involved in the biosynthesis of plantazolicin. *Chemical communications* 49: 10703-10705.
71. Schmidt EW, and Donia MS. (2009) Chapter 23. Cyanobactin ribosomally synthesized peptides--a case of deep metagenome mining. *Methods in enzymology* 458: 575-596.
72. Schmidt EW, Nelson JT, Rasko DA, Sudek S, Eisen JA, Haygood MG, and Ravel J. (2005) Patellamide A and C biosynthesis by a microcin-like pathway in *Prochloron didemni*, the cyanobacterial symbiont of *Lissoclinum patella*. *Proceedings of the National Academy of Sciences of the United States of America* 102: 7315-7320.
73. Scholz R, Molohon KJ, Nachtigall J, Vater J, Markley AL, Sussmuth RD, Mitchell DA, and Borriss R. (2011) Plantazolicin, a novel microcin B17/streptolysin S-like natural product from *Bacillus amyloliquefaciens* FZB42. *Journal of bacteriology* 193: 215-224.
74. Sivonen K, Leikoski N, Fewer DP, and Jokela J. (2010) Cyanobactins-ribosomal cyclic peptides produced by cyanobacteria. *Applied microbiology and biotechnology* 86: 1213-1225.
75. Tabata A, Nakano K, Ohkura K, Tomoyasu T, Kikuchi K, Whiley RA, and Nagamune H. (2013) Novel twin streptolysin S-like peptides encoded in the sag operon homologue of beta-hemolytic *Streptococcus anginosus*. *Journal of bacteriology* 195: 1090-1099.
76. Tamura K, Stecher G, Peterson D, Filipowski A, and Kumar S. (2013) MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Molecular biology and evolution* 30: 2725-2729.
77. Tocchetti A, Maffioli S, Iorio M, Alt S, Mazzei E, Brunati C, Sosio M, and Donadio S. (2013) Capturing linear intermediates and C-terminal variants during maturation of the thiopeptide GE2270. *Chemistry & biology* 20: 1067-1077.

78. Velasquez JE, and van der Donk WA. (2011) Genome mining for ribosomally synthesized natural products. *Current opinion in chemical biology* 15: 11-21.
79. Wever WJ, Bogart JW, Baccile JA, Chan A, Schroeder FC, and Bowers AA. (2015) Chemoenzymatic Synthesis of Thiazolyl Peptide Natural Products Featuring an Enzyme-Catalyzed Formal [4+2] Cycloaddition. *Journal of the American Chemical Society* 137(10): 3494-3497.
80. Yorgey P, Lee J, Kordel J, Vivas E, Warner P, Jebaratnam D, and Kolter R. (1994) Posttranslational Modifications in Microcin B17 Define an Additional Class of DNA Gyrase Inhibitor. *Proceedings of the National Academy of Sciences of the United States of America* 91: 4519-4523.
81. Yu X, Doroghazi JR, Janga SC, Zhang JK, Circello B, Griffin BM, Labeda DP, and Metcalf WW. (2013) Diversity and abundance of phosphonate biosynthetic genes in nature. *Proceedings of the National Academy of Sciences of the United States of America* 110: 20759-20764.
82. Yu Y, Zhang Q, and van der Donk WA. (2013) Insights into the evolution of lanthipeptide biosynthesis. *Protein science : a publication of the Protein Society* 22: 1478-1489.
83. Zhang Q, Yang X, Wang H, and van der Donk WA. (2014) High divergence of the precursor peptides in combinatorial lanthipeptide biosynthesis. *ACS chemical biology* 9: 2686-2694.
84. Zhang Q, Yu Y, Velasquez JE, and van der Donk WA. (2012) Evolution of lanthipeptide synthetases. *Proceedings of the National Academy of Sciences of the United States of America* 109: 18361-18366.

CHAPTER III: TOMM BIOSYNTHESIS IN ARCHAEA

I am grateful to D. Grogan (University of Cincinnati) for donating the strain MR31 and W. Metcalf (UIUC) for the donation of the fosmid used for library production as well as assistance in the library creation. I am grateful to R. Whitaker and C. Zhang (UIUC) for strains of *S. acidocaldarius* and *S. islandicus* as well as the knock out plasmid, reagents and assistance for the growth and gene deletion of *S. acidocaldarius*.

Abstract

Natural product biosynthesis in the domain archaea has been virtually unexplored, leaving their biosynthetic capabilities poorly characterized compared to bacteria and eukaryotes. The thiazole/oxazole-modified microcins (TOMMs) are posttranslationally modified peptides encoded in the domains bacteria and archaea. A trimeric enzymatic complex, from the TOMM biosynthetic gene cluster, is responsible for installing heterocycles on the TOMM peptides. These heterocycles are formed from cysteine, serine, and threonine residues of the precursor peptide(s). A putative TOMM cluster in the archaeal species *Sulfolobus acidocaldarius* was discovered through genome mining. As microbial genome sequences continue to accelerate, biosynthetic clusters remarkably similar to that of the *S. acidocaldarius* TOMM have been catalogued in other archaeal and bacterial species. A multipronged approach was used to identify and characterize the natural product(s) produced by these similar TOMMs. These characterizations gave fundamental insight into the biosynthesis and evolution of gene clusters within archaea. The natural product(s) were not identified, illuminating the difficulties inherent with genome guided natural product discovery methods. This work encouraged the creation of a novel genome guided technique to identify and isolate compounds which is discussed further in chapter 4.

3.1 Introduction

Antibiotics have been an important tool for innovation in both the basic science and medical fields (Falconer, *et al.* 2011, Peric-Concha, *et al.* 2003). However, with the sharp decline in the discovery of novel antibiotics during the last few decades, coupled with a rise in antibiotic resistant bacteria, it has become imperative to take alternate steps to discover new antibiotics

(Clatworthy, *et al.* 2007). Genomic sequencing has revealed a tremendous diversity in microbial species and their biosynthetic capabilities, which include a vast array of bioactive natural products. Although many bacterial natural products have been extensively studied (Clardy, *et al.* 2006, Walsh 2004), the prevalence of antimicrobial natural products from the domain archaea is still largely unknown. Characterization of gene encoded, archaeal, toxic polypeptides, termed archaeocins, remains in the early stages (O'Connor, *et al.* 2002). Halocins are archaeocins produced by halophilic euryarchaea and the few studied vary widely in both biochemical characteristics and biological activities (Haseltine, *et al.* 2001). Halocins are secreted into the media and range in size from 3.5 kDa (considered microhalocins) to 35 kDa (Cheung, *et al.* 1997, Price, *et al.* 2000). Five microhalocins have been structurally characterized, none of which undergo posttranslational modification (Shand 2008). The activity spectrum of halocins is broad, with toxins killing not only closely related haloarchaea, but also non-related hyperthermophilic, crenarchaeal *Sulfolobus* species (Haseltine, *et al.* 2001). Sulfolobin, a relatively large archaeocin (20 kDa), is produced by the *Sulfolobales* order (Ellen, *et al.* 2011, Leyva 2008, Prangishvili 2000). This archaeocidal toxin is encoded by two separate proteins that form a complex and associate with spherical vesicles which are secreted by the producer strain (Ellen, *et al.* 2011). The activity spectrum of sulfolobin is narrow, with only *S. solfataricus* P1, *S. shibatae* B12 (DSM 5389), and six strains of *S. islandicus* inhibited (Prangishvili 2000). Interestingly, these sulfolobins are extremely stable, resistant to high temperature and varying pH, as well as long term storage (Ellen, *et al.* 2011). A cypemycin-like biosynthetic gene cluster has also been described in the archaeal species *Haloterrigena turkmenica* DSM 5511, however the product of this cluster has not been experimentally investigated (Claesen, *et al.* 2010). Research on the archaeocins, especially those with unknown functions and activities, will lead to a better understanding of archaeal competition and secondary metabolism (O'Connor, *et al.* 2002). Since archaea employ eukaryotic-like cellular processes and can live in such varied

environments, some archaeal natural products may have distinct functions relative to those produced by bacteria. Bioactive, structurally complex molecules from thermoacidophiles (*i.e.* *S. acidocaldarius* TOMM) could potentially be pharmaceutically valuable, as they would be expected to have properties protecting them from both enzymatic and non-enzymatic degradation within patients.

3.1.1 Thiazole/oxazole-modified microcins

In the search for new bioactive compounds, genome sequencing has revealed a widely distributed biosynthetic gene cluster that posttranslationally modifies inactive peptides to yield bioactive natural products (Lee 2008, Melby, *et al.* 2011). This class of natural products has been termed the thiazole/oxazole modified microcin (TOMM) family. TOMMs are characterized by the installation of heterocycles that are derived from Cys, Ser and Thr residues in the precursor peptide (A). A trimeric complex containing a cyclodehydratase (C and D protein) and dehydrogenase (B), cooperatively act to install heterocycles onto precursor peptides (Figure 3.1A). The cyclodehydratase performs the cyclization of Cys and Ser/Thr residues to form thiazoline and (methyl)oxazoline heterocycles (Belshaw, *et al.* 1998, Dunbar, *et al.* 2012, Dunbar, *et al.* 2013, Li, *et al.* 1996, Melby, *et al.* 2012, Milne, *et al.* 1998, Milne, *et al.* 1999). A subset of, or sometimes all, “azoline” heterocycles are then further processed by a flavin mononucleotide (FMN)-dependent dehydrogenase, to produce the aromatic thiazole and (methyl)oxazole heterocycles (Melby, *et al.* 2014) (Fig. 3.1B). These heterocycles endow the precursor peptide with a rigidified structure necessary for activity. TOMM biosynthetic clusters often contain ancillary tailoring enzymes to increase structural complexity and transporters to export the fully functional natural product out of the cell.

The TOMM family is widely disseminated in both bacterial and archaeal species. Computational surveys have identified over 400 clusters in the genomes published in GenBank as

of 2012. Examples of studied TOMMs include microcin B17 (DNA gyrase inhibitor), streptolysin S (cytolysin), heterocycle-containing cyanobactins (eukaryotic cytotoxins), and the thiopeptides (ribosome inhibitors) (Melby, *et al.* 2011). Despite extensive characterization of some TOMMs, the vast majority have unknown structures and bioactivities (Figure 3.2). As of 2012, *Sulfolobus acidocaldarius* is one of three archaeal species which was discovered to contain a “complete” TOMM cluster (contains ‘A’ precursor and ‘BCD’ modifying proteins). There are 14 archaeal species that contain an incomplete cluster with only the ‘CD’ modifying enzymes. Thus, these clusters would be expected to produce compounds with exclusively azoline (not azole) heterocycles, unless a dehydrogenase encoded elsewhere in the genome is operative. Given our previous successes in studying TOMM biosynthetic clusters (Lee 2008, Mitchell 2009) and our current interests in defining the biosynthetic capabilities of the archaea, *S. acidocaldarius* provides a unique opportunity to explore this exciting area.

3.1.2 Sulfolobus acidocaldarius TOMM cluster

S. acidocaldarius is a well-known archaeal species that was first isolated from a hot spring in Yellowstone National Park (Brock 1972). *S. acidocaldarius* thrives in environments with temperatures ranging between 75 - 90 °C and a pH of 2-3. These living conditions make *S. acidocaldarius* one of the few thermoacidophiles known. While much is known about *S. acidocaldarius* relative to other archaeal species, our molecular understanding of this organism, and the archaea as a whole, pales in comparison to bacteria. Archaea are morphologically more similar to bacteria, but employ eukaryotic-like mechanisms for cellular processes (Whitman, *et al.* 1999). Other defining features of the domain archaea include a lack of peptidoglycan in their cell walls (*S. acidocaldarius* replaces peptidoglycan with a S-layer) and membranes composed of ether-linked lipids (Taylor, *et al.* 1982, Woese, *et al.* 1990). Not only are archaea globally underrepresented in microbiological research, but the challenge of culturing and creating genetic

systems in archaeal species has delayed the exploration of their natural product biosynthetic capabilities.

The predicted *S. acidocaldarius* TOMM biosynthetic gene cluster is comprised of eleven genes (Figure 3.3). Because of its similarity to other TOMM clusters, we have adopted the standard TOMM nomenclature for most of the encoded proteins. There are two predicted precursor peptides in the cluster that harbor four and six heterocyclizable residues, SaciA1 and SaciA2 (*Saci_0528* and unannotated, respectively) (Figure 3.3). The genes that encode for SaciA1 and SaciA2 are followed in the cluster by *saciBCD* (*Saci_0525*, *Saci_0527*, *Saci_0526*, respectively), which are orthologous to the thiazole/oxazole synthetase proteins in other TOMM clusters and are expected to carry out the same transformation. A predicted immunity gene is present, designated *saciF* (*Saci_0530*), which is weakly similar to suspected immunity proteins in other TOMM clusters. There are three ABC transporter genes within the cluster, *saciGHI* (*Saci_0532*, *Saci_0531* and *Saci_0529*, respectively). *SaciK* (*Saci_0523*) is a predicted transcriptional regulator, which belongs to the ArsR family of repressors. The last gene in the cluster, *saciJ* (*Saci_0524*), has an unknown function and is only homologous to an annotated gene in the *Sulfolobus islandicus* REY15A TOMM cluster. The predicted TOMM(s) from *S. acidocaldarius* are unique when compared to other archaeocins and will be the first extensively posttranslationally modified natural products studied in archaea.

3.1.3 Other related TOMM clusters

Our lab has shown, by using bioinformatics-guided chemotyping that TOMMs cluster by their putative function. For this analysis, we created a phylogenetic tree of the D proteins (Figure 3.2) (Melby, *et al.* 2011). This is a powerful analysis, as it allows the functional prediction of uncharacterized TOMMs based on their similarity to a relative with a known function. In the cases where no phylogenetic neighbors have a known function, no prediction can be made. The *S.*

acidocaldarius TOMM falls into this category, as it forms a new clade with other TOMMs of unknown function. A bioinformatics study using SaciBCD proteins was performed to determine TOMMs that will likely have similar functions and to find the evolutionary distribution of this cluster (Figure 3.4). Four other highly related clusters were discovered (Figure 3.3). The cluster from *Sulfolobus islandicus* REY15A was used in comparison with the *S. acidocaldarius* to determine the boundaries of these clusters. Interestingly, three other related clusters were discovered in bacterial species. The other clusters are in *Bacillus cereus* Rock3-44, *Thermoanaerobacter mathranii* subsp. *mathranii* str. A3, and *Actinomyces odonolyticus* F0309. These bacterial species have precursor peptides that likely function similarly to *S. acidocaldarius* and *S. islandicus* (Figure 3.3), and provide a possible pathway for acquisition of these clusters through horizontal gene transfer.

3.2 Multipronged Approach

To identify and characterize the *S. acidocaldarius* TOMM biosynthetic gene cluster and natural product(s) a multipronged approach was taken. Multiple genome guided techniques that had been previously used to study other ribosomally produced natural products were enlisted to elucidate the structure and biological activity of the *S. acidocaldarius* TOMM. The necessary *S. acidocaldarius* TOMM biosynthetic genes were cloned into *E. coli* expression vectors, expressed and purified. This work set the stage for *in vitro* reconstitution, structural characterization using MS, and functional characterization of the enzymes and natural product(s). For an *in vivo* approach, reverse transcriptase (RT)-PCR, a gene deletion, and fosmid-based heterologous expression was used to serve as a guide for the isolation of the *S. acidocaldarius* TOMM natural product(s). These techniques had all previously been used to identify other ribosomal natural products, however were unsuccessful when used to identify the TOMM from *S. acidocaldarius*. Often natural products are produced in extremely low levels or are not produced in laboratory culturing conditions and this is likely the reason the following techniques did not work in

isolating the *S. acidocaldarius* TOMM. However, these setbacks stimulated the creation of a new method that evaded this common difficulties resulting in the discovery of a different natural product (discussed in chapter 4).

3.3 Heterologous Expression and Purification of the *S. acidocaldarius* TOMM Proteins

In vitro assays provide an unparalleled method to study enzymatic reactions through the precise control of reaction conditions. To study the heterocycle formation within the two precursor peptides, SaciA1 and SaciA2, an *in vitro* synthetase reaction similar to previous reports (Lee 2008, Sinha Roy, *et al.* 1998) was attempted. To perform these experiments, the necessary genes (*saciA1*, *A2*, *B*, *C*, and *D*) were cloned from genomic DNA into the pET28-maltose binding protein (MBP) fusion *E. coli* expression vector. MBP was used for affinity purification of the proteins and to enhance their solubility/stability. The expression and purification of the *S. acidocaldarius* TOMM biosynthetic proteins (Figure 3.5) was successful, except for SaciD, which remained largely insoluble under all tested conditions and was unusable in the reconstitution of heterocycles. As shown with other systems, highly related non-cognate ‘BCD’ enzymes often have the ability to posttranslationally modify similar precursor peptides (Lee 2008, Mitchell 2009). Knowing this, attempts to clone the D protein from closely related TOMM clusters in *B. cereus* and *T. mathranii* were performed.

3.4 Heterologous Expression of the *Bacillus cereus* Rock3-44 TOMM Proteins

The *B. cereus* TOMM cluster was discovered via a bioinformatics search looking for closely related enzymes to the *S. acidocaldarius* TOMM cluster. The *B. cereus* D protein (RockD) is 58.4% similar to SaciD (Figures 3.4 and 3.6) and thus was considered as a substitute for SaciD for *in vitro* reactions. All of the essential TOMM proteins (RockA, B, C, D) from *B. cereus* were also cloned into the pET28-maltose binding protein (MBP) fusion *E. coli* expression vector. As with the *S. acidocaldarius* proteins, RockA was affinity purified using immobilized

amylose resin. Unfortunately, RockD was also insoluble and unable to install heterocycles onto the precursor peptides during *in vitro* reconstitution.

3.5 MS of Unmodified Precursor Peptides

Previously characterized TOMM natural products have been successfully studied using MS (Matrix-assisted laser desorption/ionization (MALDI)-time of flight (TOF) and electrospray ionization) techniques (Kelleher, *et al.* 1998). During the formation of each azoline ring (formally a cyclodehydration), 18 Da is lost from the parent peptide. Each subsequent dehydrogenation (-H₂) leads to an additional loss of 2 Da (Figure 3.1B). This loss in mass permits the use of MS-based assays for monitoring heterocycle formation on the precursor peptide. Using MALDI-TOF, the masses, of unmodified SaciA1, SaciA2, and RockA were measured (Figure 3.7). Tobacco Etch Virus (TEV) protease was used to cleave the precursor peptides from MBP leaving the final masses at 5207.9 Da, 4587.4 Da, and 4474.2 Da, respectively. Unfortunately, functional D protein was not obtained and therefore further characterization of heterocycle formation within the precursor peptides was never performed. Heterologous production of modification machinery can be very challenging, particularly when expressing proteins within a host that is dissimilar and therefore we identified other methods that could be used to characterize the natural product within the native host.

3.6 RT-PCR

Although reaction conditions can be carefully monitored and controlled during *in vitro* reconstitution studies, they do not always accurately represent physiological conditions within the producing organism. Combined with the possibility that other modifying enzymes are necessary, but not present during *in vitro* reconstitution assays, it is important to complement *in vitro* studies with *in vivo* data. Understanding this, I am pursued the isolation of the natural product(s) from the native host. RT-PCR was used to first demonstrate that the native producer was actively

transcribing the TOMM genes during laboratory cultivation (Figure 3.8). For this experiment, RNA was isolated from *S. acidocaldarius* cultures grown into late log phase, in liquid DT media (dextrin and tryptone, pH 3.5, 80°C). mRNA was then converted to cDNA using reverse transcriptase and the cDNA was probed for transcription with *S. acidocaldarius* TOMM specific primers. All genes within the cluster were transcribed. This indicated that there was a possibility of the natural product(s) being produced by *S. acidocaldarius* and therefore I pursued isolation from the native producer.

3.7 Gene Deletion Strains

Creation of a gene deletion strain has been previously used to facilitate the study and isolation of the mature TOMM natural product(s). By comparing HPLC and MS traces between a gene deletion strain and the parent strain, the molecular weight of the TOMM natural product(s) can be determined. The mass can aid in structural elucidation of the product because of the distinct weight change associated with each heterocycle formed. These gene deletions can also be coupled with the function of the peptides when employing activity assays. The complete maturation, and thus the activity, of the peptide should be lost when the deletion strains are tested.

A pyrimidine auxotroph strain of *S. acidocaldarius* was used to generate a deletion strain. A spontaneous mutation in the *pyrE* gene renders the strain MR31 deficient in pyrimidine biosynthesis, and therefore unable to replicate their DNA in the absence of supplemental uracil (Figure 3.9B). MR31 has an in-frame deletion and therefore the adjacent downstream gene, *pyrF*, is unaffected in transcription (Figure 3.9A). Disruption of *SaciC* was targeted with homologous recombination (Figure 3.9C) using a plasmid containing a *pyrE* gene from *S. islandicus* and its upstream and downstream elements, flanked by regions of *saciC*. Importantly, the *pyrE* gene in the plasmid is not identical to wild type *pyrE* gene, which prevents untargeted homologous recombination with the native *pyrE* gene. Mutants were selected for in media lacking uracil.

Mutants with a restored functional *pyrE* gene were able to grow in the absence of uracil supplementation. After transformation with the plasmid, two deletion strains were identified by PCR screening as these strains have an approximately 400 base pair increase in PCR amplicon size when amplifying with *saciC* primers, which anneal to the outer limits of the *saciC* gene. LC-MS was used to compare the gene deletion strain and the wild type strain in growth conditions identified using RT-PCR. These comparative analyses were unable to identify differences in the spectra between the two strains. Therefore it is likely that the natural product(s) are produced in vanishingly small amounts or are not produced at all during laboratory cultivation.

3.8 Fosmid Library Generation

Complimentary to the creation of the above deletion strain, the production of a fosmid that contains the *S. acidocaldarius* TOMM cluster was also created to facilitate the isolation of the natural product(s). In this sense, when using comparative LC-MS to identify the natural product(s) a peak gained in the fosmid-bearing strain is expected when compared to the parent strain. I created a fosmid library from *S. acidocaldarius* genomic DNA using the fosmid pJK050. After PCR screening approximately 2000 colonies, I identified and isolated four fosmid strains: two contain the entire TOMM cluster and two contain partial clusters. I used the fosmid strain pCOX6E6 for further studies. This fosmid contained the entire TOMM cluster, but has the smallest amount of excess *S. acidocaldarius* chromosomal DNA. Using RT-PCR, I verified all of the core TOMM genes (*saciA1, A2, B, C, D*) are expressed. As with the gene deletion strains, comparative studies with the empty fosmid did not lead to the identification of the natural product(s). Similar to the heterologous expression, these proteins (the D protein in particular) are likely not expressed in *E. coli* and therefore the natural product was not visualized.

3.9 Summary and Outlook

Natural product biosynthesis in the domain archaea has been overlooked. This project aimed to shed light on this unexplored area by studying the biosynthesis of a TOMM natural product from *Sulfolobus acidocaldarius*. This work had the potential to fundamentally shift the current perception that archaea are not capable of complex molecule biosynthesis. Furthermore, this TOMM biosynthetic gene cluster would represent the first posttranslationally modified natural product to be characterized in the domain archaea. The identification of the *S. acidocaldarius* family of TOMMs demonstrated the power of genome mining for the identification of novel biosynthetic gene clusters. However, characterization of this family verified the difficulty of natural product discovery, even when a gene cluster has been identified. Although genome mining has the potential to prevent identification of known compounds, techniques to identify compounds from a particular gene cluster still pose a difficult problem for researchers. Although many techniques have been created for genome based isolation, the characterization of the *S. acidocaldarius* TOMM natural products indicates the difficulties in identifying compounds from a particular cluster. Therefore I proceeded to identify novel and rapid techniques to correlated natural products with genome guided identification of biosynthetic gene clusters.

3.10 Experimental

3.10.1 Protein overexpression and purification

All proteins were overexpressed as tobacco etch virus (TEV) protease fusions to maltose-binding protein (MBP) and purified by amylose affinity chromatography as previously reported (Dunbar, *et al.* 2012). All characterizations of the proteins were performed with the tagged substrates, with the exception of the MS analysis of the precursor peptides SaciA1, SaciA2, and RockA.

3.10.2 MALDI-TOF mass spectrometric analysis of the precursor peptides

MALDI-TOF mass spectrometry was performed using a Bruker Daltonics UltrafleXtreme MALDI-TOF/TOF instrument operating in positive reflector mode. The instrument was calibrated before data acquisition using a commercial peptide calibration kit (AnaSpec – Peptide Mass Standard Kit). Analysis was carried out with Bruker Daltonics flexAnalysis software. The precursor peptides were incubated in the presence of TEV protease overnight to remove the fusion proteins. The peptides were purified using a ZipTip purification system and eluted with 10 μ L of 95% aqueous acetonitrile. 2 μ L of this peptide solution was mixed with 2 μ L of matrix solution (a saturated solution of α -cyano-4-hydroxycinnamic acid), spotted on the target, air dried and measured as stated above.

3.10.3 *Sulfolobus* cultivation conditions

S. acidocaldarius was cultivated in a tissue culture flask at 78°C without shaking. DT medium (pH 3.5), was used for cultivation. DT medium contains the following components (in 1 liter Milli-Q H₂O): 1x basal salts (CaCl₂ · 2H₂O, 0.1 g; K₂SO₄, 3.0 g; MgSO₄, 0.145 g; and NaH₂PO₄, 0.5 g), 20 μ L trace mineral stock solution (0.5% CoCl₂ · 6H₂O, 0.5% CuCl₂ · 2H₂O, 3.0% FeCl₃, 0.5% MnCl₂ · 4H₂O, and 0.5% ZnCl₂), 0.1% (wt/vol) dextrin, and 0.1% (wt/vol) EZMix-N-Z-amine A.

3.10.4 RT-PCR

Total RNA was isolated with the Qiagen RNeasy minikit. Cells (OD₆₀₀ 0.8) were harvested from DT medium and treated with RNAprotect bacterial reagent. Harvested cells were resuspended in 250 μ L of 10 mM Tris (pH 8.5). The isolation was then performed according to the manufacturer's instructions (without the lysozyme incubation). cDNA was prepared with commercially available reverse transcriptase PCR (RT-PCR) kits using 1 μ g of RNA and primers for all of the *S. acidocaldarius* TOMM genes.

3.10.5 *SaciC* gene deletion production

The *pyrEF* gene driven by its native promoter and terminator from *Sulfolobus solfataricus* P2 in a pUC19 backbone was graciously supplied by the Whitaker lab (UIUC). Homology arms for *saciC* were inserted upstream and downstream of the *PyrEF* element.

Sulfolobus acidocaldarius MR31 competent cells were prepared following the procedure described previously. Linearized plasmid (~1 µg) were transformed into the competent cells by electroporation. After electroporation, transformed cells were immediately regenerated in 800 µL incubation of DT medium supplemented with uracil. After 1 hour incubation, a portion of the cells were plated onto DT plates and a portion were added to 5 mL DT medium and incubated for 2 weeks prior to plating. Colonies were screened for the loss of *saciC* using gene specific primers.

3.10.6 Fosmid library generation

Genomic DNA for fosmid production was prepared from *Sulfolobus acidocaldarius* using cells grown in 200 mL of DT media to OD₆₀₀ 0.8. Harvested cells were resuspended in 10 mL of TE buffer (10 mM Tris, 1 mM EDTA, pH 8.0). 0.5 mL of 10% (w/v) sodium dodecyl sulfate as 50 µL of proteinase K was added and incubated for 1 h. The resulting lysate was sequentially extracted with buffer saturated phenol, chloroform, and isoamyl alcohol (25:24:1) and chloroform and isoamyl alcohol (24:1). Genomic DNA was precipitated from the aqueous layer by the addition of 0.7 volumes isopropanol, and washed twice with 70% (v/v) EtOH, dried and resuspended in water.

Fosmid construction was performed as previously reported (Eliot, *et al.* 2008). Positive clones were screened by PCR for the presence of the *Sulfolobus* gene cluster using primers for the outermost genes and primers for *saciA1*.

3.10.7 Liquid chromatography mass spectrometric comparisons

Sulfolobus strains (wild type and gene deletion) were grown to OD₆₀₀ 0.8 in DT media. *E. coli* strains (with or without the fosmid) were grown to OD₆₀₀ 1.0 in LB. Harvested cells were resuspended in 2 mL of 90% aqueous acetonitrile and agitated for 30s by vortex. After centrifugation (14,000 rpm, 5 min) the solution was analyzed by online HPLC (1100 series HPLC system; Agilent Technologies) coupled to a QTRAP 2000 mass spectrometer (Applied Biosystems). A sample of extract (30 µL) was separated by HPLC using a C18 column with a linear gradient of 5% to 95% acetonitrile with 0.1% formic acid in 30 min. MS analysis was performed in positive ion mode. Spectrums were then manually curated for differences in masses detected. Modification to this method including, varied extraction solvents (acetonitrile, MeOH and chloroform at various concentrations), as well as different HPLC gradient times ranging from 10 -75 minutes were used to try and identify the compounds.

3.11 Figures

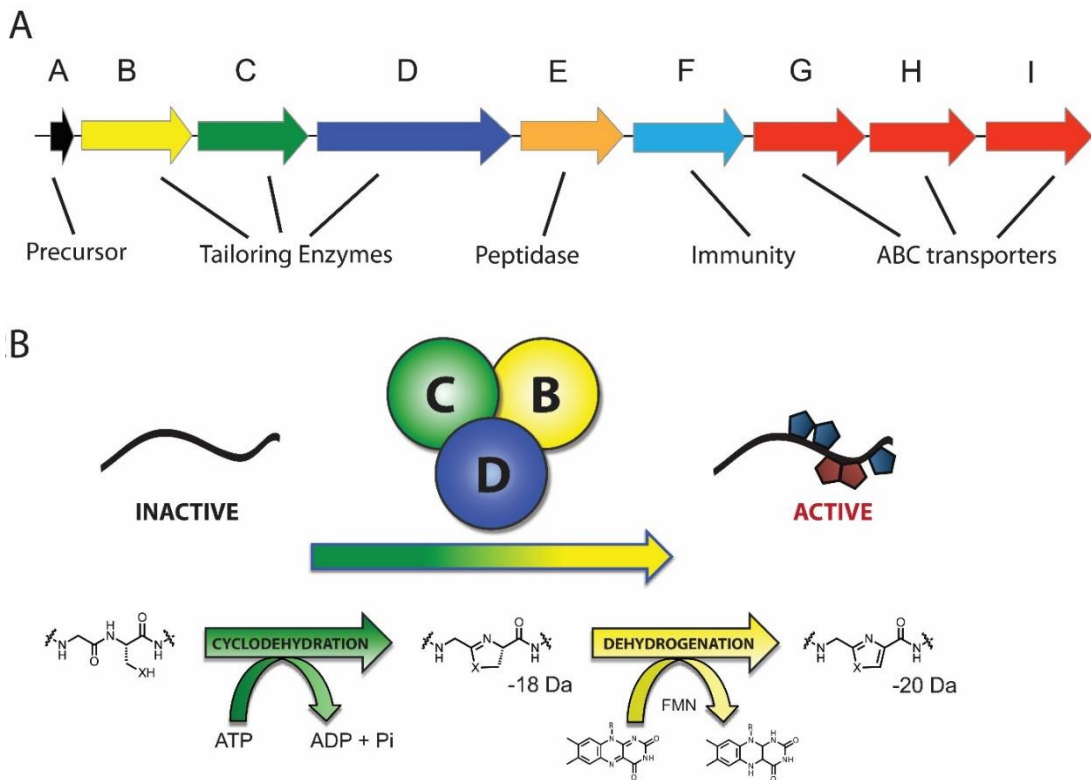


Figure 3.1 | Representative TOMM clusters and biosynthetic machinery. (A) Genetic organization of an example TOMM cluster. The precursor peptide (black) is modified by the trimeric BCD complex, consisting of a cyclodehydratase (C, green), dehydrogenase (B, yellow), and docking protein (D, blue). (B) Heterocycles are formed on the precursor peptide, converting the peptide into a bioactive natural product. Two enzymatic steps occur to modify the precursor peptides. First, a water molecule is removed from the peptide as the cyclodehydratase forms thiazoline/oxazoline rings. Secondly, a FMN-dependent dehydrogenase removes hydrogen to yield the thiazole/oxazole rings. Combined, the reactions lead to a molecular mass loss of 20 Da in the peptide (adapted from (Lee 2008)).

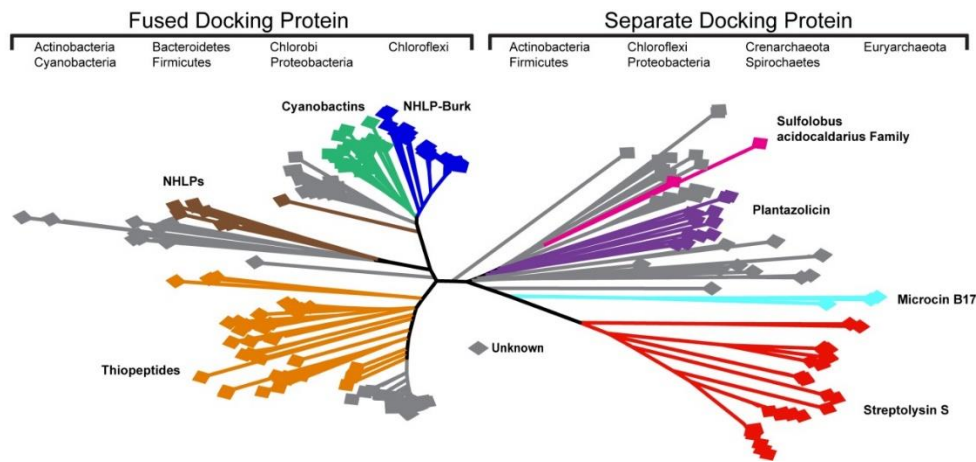


Figure 3.2 | The TOMM Family Tree. D proteins were used to construct a phylogenetic tree. For fused docking proteins, only the docking portion of the protein was used to create the tree. The docking proteins cluster based on the function of the precursor peptides they modify. The *S. acidocaldarius* TOMM cluster forms a clade with other TOMMs of unknown function (depicted in pink) (adapted from (Melby, *et al.* 2011)).

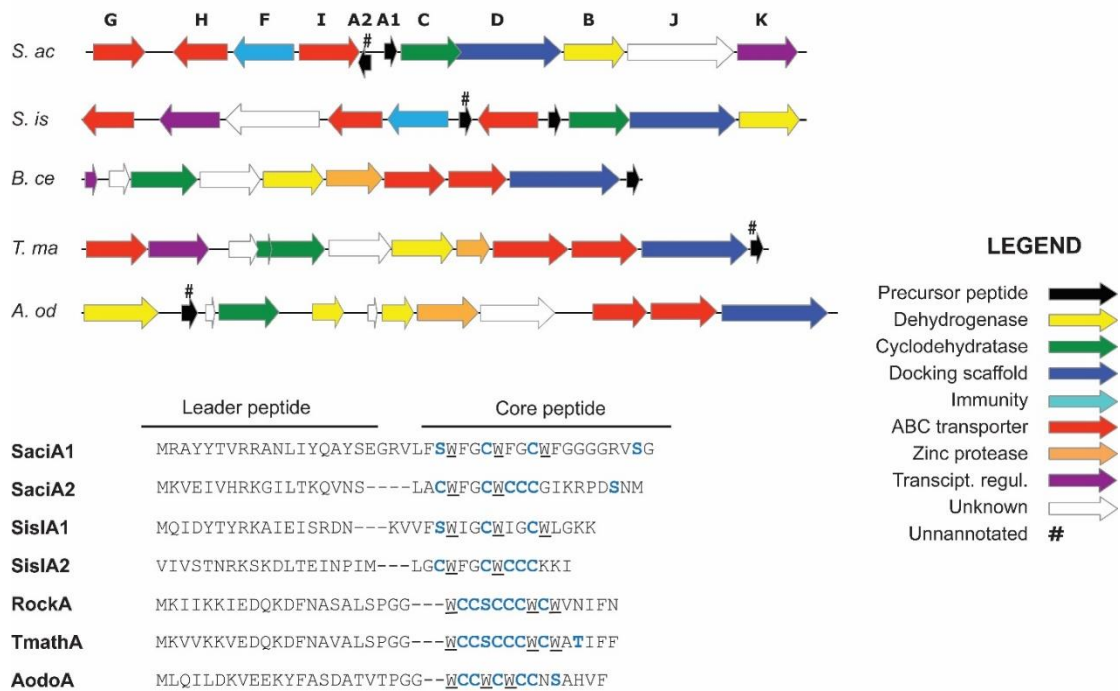


Figure 3.3 | The *S. acidocaldarius* TOMM family. A bioinformatics search discovered five similar TOMMs. The genetic organization of the clusters is depicted (above). The putative functions of the genes are listed in the legend. Note that the *S. acidocaldarius* and *S. islandicus* clusters contain two precursor peptides, one in the forward and one in the reverse direction. The precursor peptides have many potentially heterocyclized residues within the core (below). Blue residues indicate heterocyclizable residues. These peptides contain W (underlined) near heterocyclizable residues which could be necessary for function.

A

	SaciB	SislB	Rock 3-44B	TmatB	AodoB
SaciB	100	87	52.3	55.4	51
SislB	55.4	100	53.4	54.6	47.2
Rock 3-44B	18.3	18.6	100	90.7	42.7
TmatB	17.5	21.4	61.1	100	40.4
AodoB	21.2	20.8	16.1	15.7	100

B

	SaciC	SislC	Rock 3-44C	TmatC	AodoC
SaciC	100	63.1	47.7	50.8	49
SislC	36	100	36.9	51.7	41.2
Rock 3-44C	17.6	12.7	100	56.9	53.2
TmatC	14	14.4	34.6	100	52.2
AodoC	12.2	10.6	17.9	14.7	100

C

	SaciD	SislD	Rock 3-44D	TmatD	AodoD
SaciD	100	91.2	58.4	60.6	52.7
SislD	64.5	100	61.4	57.6	53.1
Rock 3-44D	21.2	21.2	100	92.5	64.5
TmatD	22.9	22.1	61.5	100	64.1
AodoD	21.4	20.3	26.1	27	100

Figure 3.4 | Similarity-identity matrices of related TOMM proteins. In blue are amino acid identity scores and in green are amino acid similarity scores, both obtained using ClustalW pairwise alignments. Scores were determined by comparing the similar or identical amino acids with the full protein sequence after alignment. **(A)** SaciB, dehydrogenase. **(B)** SaciC, cyclodehydratase. **(C)** SaciD, docking protein. Abbreviation of species are as follows: Saci, *Sulfolobus acidocaldarius*; Sisl, *Sulfolobus islandicus* REY15A; Rock 3-44, *Bacillus cereus* Rock3-44; Tmat, *Thermoanaerobacter mathranii* subsp. *mathranii* str. A3; Aodo, *Actinomyces odontolyticus* F0309.

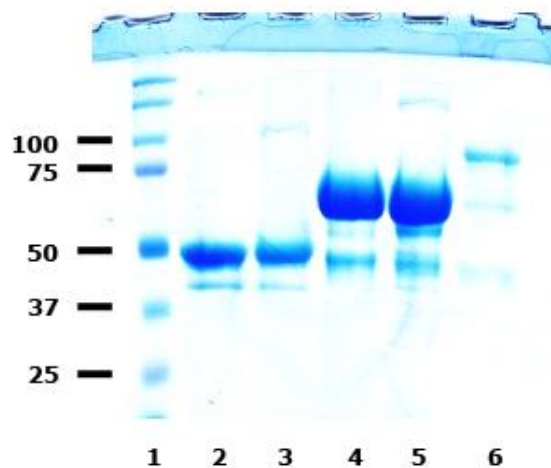


Figure 3.5 | Heterologous expression of *S. acidocaldarius* TOMM proteins. *S. acidocaldarius* TOMM biosynthetic proteins tagged with maltose binding protein and expressed in *E. coli* BL21 strain. Purified proteins were analyzed by SDS-PAGE and visualized by staining with Coomassie. Lane 1: Marker is in kDa. Lane 2-6: SaciA-D. The calculated masses and isolated yields are: MBP-SaciA1, 50.7 kDa, 6.4 mg/L; MBP-SaciA2, 50.1 kDa, 11.2 mg/L; MBP-SaciB, 71.9 kDa, 6.5 mg/L; MBP-SaciC, 73.0 kDa, 48 mg/L; MBP-SaciD, 96.1 kDa, 0.8 mg/L.


```

SaciD      VRLEEFINEFIGPVFGYVAEKYDN-VIITRLLNVVNYN-----ILESHIG-QGW 47
RockD      MSITILNKRKISEPASFEYEDFANYREMSSYLGYFNLIKKLDVVMAPSELPLYIGNCEF 60
           : : : : * . . : * : * : : * . * * * * * * * * * * : * * :

SaciD      IRSSIEVLTPPQVLPLLKYLNSVVPAGGKGFTEESLEGAIGEFLERFYGAFTLFDD-ES 106
RockD      MNINYFFNYLLRRTSMSVKLNESIFAGGKGFPLYKAICSSLGEAFERLMACLEYFIQKDN 120
           : . . . . : : * : : * * * * * * * * * * * * * * * * * * * * :

SaciD      IIFGRIGDLMTK--YEIFPITYKFFSTEQLQKL-LIFRDYNNENVMLS-LTRAKYYKGSVDV 162
RockD      VFLGTYKDIQEKGFKAHPSEIKSFSEEQFFDENFLFEEFDEETYTSWMIMEDLHSGEDI 180
           : : * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * :

SaciD      YIPVQITYLLNLIRPG-EDLIAYSTTGGLAFHKDFESAFIHGLLEYLERDTINISWISRI 221
RockD      YIPACLILMYYPKSDQEKRIYATSGGLTSHYLENHGIEHGLMEIIERHEINLSWYCKI 240
           * * * . : : : . . * . * . * : * * * * * * * * * * * * * * * * * * :

SaciD      PPKRIRLPERIRRKLKVLED---RNITCLSPGPEFKGLFVVGCLGFINDL---YVAGA 273
RockD      KPEEVIIDEIKHKQLAKYKDYIKEKNIRFFRHNVDQQNFHVITAMSFDDDMTKYSFNTGG 300
           * : : : * : : * * * * * * * * * * * * * * * * * * * * * * * * :

SaciD      GADITLEEAIRKAVFEVYQSISS-----FSKITDEEIKMARNLKP DYLTDFGLV 322
RockD      GISPDIEKAILASMEEYTAQVNNTRKIVYAPNWLTSKFSNGVLDVDEDDPRNFKTFYQA 360
           * . : * * * * * * * * * * * * * * * * * * * * * * * * * * * * * :

SaciD      PLYYSFVKTNFLVEYLRSLQSITYEELEREENLSYKSLIDIILGKNYIIYKDFDIEKYL 382
RockD      VSYYGLKSNQHKL DWYVKGKPRLLSEIR-KIQTNQNIREYVKEFNLEPFISRLGISEQF 419
           * * . : . . . : : . . . * : : : : : * : . : . * * * * * :

SaciD      GEGKLV RVIVPDLTPAHIPNL PFLGHKRYYTIRKEFNLSNVELFVEEVPVFPF 434
RockD      KNIYISKVYMKEFTPAFIAGVPVLGHEKYQEY-----LQDGSELNKEILPFP 466
           : : : * : : * * * * * * * * * * * * * * * * * * * * * * * * * * :

```

Figure 3.6 | Alignment with SaciD and RockD. A pairwise alignment of SaciD and RockD proteins was created using ClustalW. A Gonnet protein weight matrix was used with a gap opening penalty of 10 and a gap extension penalty of 0.1. Neighbor joining was used for clustering. The proteins are 58% similar.

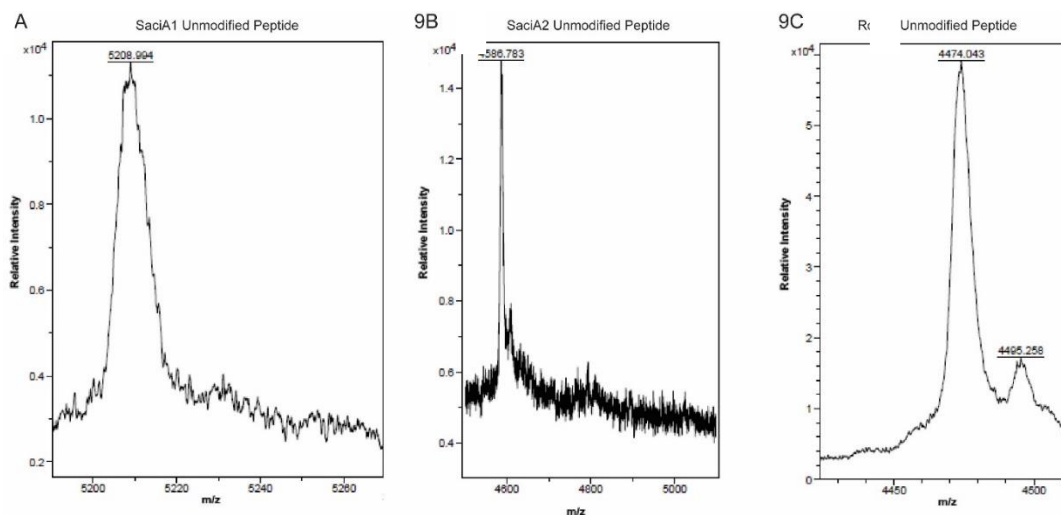


Figure 3.7 | MALDI-TOF of unmodified precursor peptides. MALDI-TOF analysis of the unmodified precursor peptides was performed in linear positive mode. The samples were treated with Tobacco Etch Virus protease to remove MBP. Leaving the expected weights of SaciA1, SaciA2 and RockA to be 5207.9 Da, 4587.4 Da, and 4474.2 Da, respectively. The samples were subjected to a Reverse-Phase Zip-Tip procedure using C18 Zip-Tips. The samples were eluted into 90% acetonitrile with 0.1% trifluoroacetic acid. Sinapinic acid was used as the matrix. **(A)** MALDI-TOF analysis of SaciA1 unmodified peptide. **(B)** MALDI-TOF analysis of SaciA2 unmodified peptide. **(C)** MALDI-TOF analysis of RockA unmodified peptide.

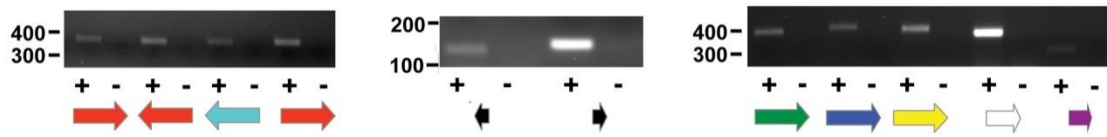


Figure 3.8 | RT-PCR of *S. acidocaldarius* TOMM mRNA. Total RNA was isolated from *S. acidocaldarius* grown in DT media in late log-phase. Reverse transcriptase (RT) was added to the samples indicated with +. In each case, the amplicons migrate with the anticipated size. All genes were verified to be transcribed. The ORF arrow coloring scheme is identical to Fig. 3.1.

3.12 References

1. Belshaw PJ, Roy RS, Kelleher NL, and Walsh CT. (1998) Kinetics and regioselectivity of peptide-to-heterocycle conversions by microcin B17 synthetase. *Chemistry & biology* 5: 373-384.
2. Brock TD. (1972) Sulfolobus: A New Genus of Sulfur-Oxidizing Bacteria Living at Low pH and High Temperature. *Archives of Microbiology* 84: 25.
3. Cheung J, Danna KJ, O'Connor EM, Price LB, and Shand RF. (1997) Isolation, sequence, and expression of the gene encoding halocin H4, a bacteriocin from the halophilic archaeon *Haloferax mediterranei* R4. *Journal of bacteriology* 179: 548-551.
4. Claesen J, and Bibb M. (2010) Genome mining and genetic analysis of cypemycin biosynthesis reveal an unusual class of posttranslationally modified peptides. *Proceedings of the National Academy of Sciences of the United States of America* 107: 16297-16302.
5. Clardy J, Fischbach MA, and Walsh CT. (2006) New antibiotics from bacterial natural products. *Nature biotechnology* 24: 1541-1550.
6. Clatworthy AE, Pierson E, and Hung DT. (2007) Targeting virulence: a new paradigm for antimicrobial therapy. *Nature chemical biology* 3: 541-548.
7. Dunbar KL, Melby JO, and Mitchell DA. (2012) YcaO domains use ATP to activate amide backbones during peptide cyclodehydrations. *Nature chemical biology* 8: 569-575.
8. Dunbar KL, and Mitchell DA. (2013) Insights into the mechanism of peptide cyclodehydrations achieved through the chemoenzymatic generation of amide derivatives. *Journal of the American Chemical Society* 135: 8692-8701.
9. Eliot AC, Griffin BM, Thomas PM, Johannes TW, Kelleher NL, Zhao H, and Metcalf WW. (2008) Cloning, expression, and biochemical characterization of *Streptomyces rubellomurinus* genes required for biosynthesis of antimalarial compound FR900098. *Chemistry & biology* 15: 765-770.
10. Ellen AF, Rohulya OV, Fusetti F, Wagner M, Albers SV, and Driessen AJ. (2011) The sulfolobacin genes of *Sulfolobus acidocaldarius* encode novel antimicrobial proteins. *Journal of bacteriology*.
11. Falconer SB, Czarny TL, and Brown ED. (2011) Antibiotics as probes of biological complexity. *Nature chemical biology* 415-423.

12. Haseltine C, Hill T, Montalvo-Rodriguez R, Kemper SK, Shand RF, and Blum P. (2001) Secreted euryarchaeal microhalocins kill hyperthermophilic crenarchaea. *Journal of bacteriology* 183: 287-291.
13. Kelleher NL, Belshaw PJ, and Walsh CT. (1998) Regioselectivity and chemoselectivity analysis of oxazole and thiazole ring formation by the peptide-heterocyclizing microcin B17 synthetase using high-resolution MS/MS. *Journal of the American Chemical Society* 120: 9716-9717.
14. Lee SW. (2008) Discovery of a widely distributed toxin biosynthetic gene cluster. *Proceedings of the national academy of science of the USA* 105: 6.
15. Leyva RSaK. (2008) *Archaea: New Models for Prokaryotic Biology*. Caister Academic Press, p. 248.
16. Li YM, Milne JC, Madison LL, Kolter R, and Walsh CT. (1996) From peptide precursors to oxazole and thiazole-containing peptide antibiotics: microcin B17 synthase. *Science* 274: 1188-1193.
17. Melby JO, Dunbar KL, Trinh NQ, and Mitchell DA. (2012) Selectivity, directionality, and promiscuity in peptide processing from a *Bacillus* sp. Al Hakam cyclodehydratase. *Journal of the American Chemical Society* 134: 5309-5316.
18. Melby JO, Li X, and Mitchell DA. (2014) Orchestration of enzymatic processing by thiazole/oxazole-modified microcin dehydrogenases. *Biochemistry* 53: 413-422.
19. Melby JO, Nard NJ, and Mitchell DA. (2011) Thiazole/oxazole-modified microcins: complex natural products from ribosomal templates. *Current opinion in chemical biology* 15(3): 396-378.
20. Milne JC, Eliot AC, Kelleher NL, and Walsh CT. (1998) ATP/GTP hydrolysis is required for oxazole and thiazole biosynthesis in the peptide antibiotic microcin B17. *Biochemistry* 37: 13250-13261.
21. Milne JC, Roy RS, Eliot AC, Kelleher NL, Wokhlu A, Nickels B, and Walsh CT. (1999) Cofactor requirements and reconstitution of microcin B17 synthetase: a multienzyme complex that catalyzes the formation of oxazoles and thiazoles in the antibiotic microcin B17. *Biochemistry* 38: 4768-4781.
22. Mitchell DA. (2009) Structural and Functional Dissection of the Heterocyclic Peptide Cytotoxin Streptolysin S. *The Journal of biological chemistry* 284: 9.

23. O'Connor EM, and Shand RF. (2002) Halocins and sulfolobocins: the emerging story of archaeal protein and peptide antibiotics. *Journal of industrial microbiology & biotechnology* 28: 23-31.
24. Peric-Concha N, and Long PF. (2003) Mining the microbial metabolome: a new frontier for natural product lead discovery. *Drug discovery today* 8: 1078-1084.
25. Prangishvili D. (2000) Sulfolobocins, Specific Proteinaceous Toxins Produced by Strains of the Extremely Thermophilic Archaeal Genus Sulfolobus. *Journal of bacteriology* 182: 4.
26. Price LB, and Shand RF. (2000) Halocin S8: a 36-amino-acid microhalocin from the haloarchaeal strain S8a. *Journal of bacteriology* 182: 4951-4958.
27. Shand RF. (2008) Archaeal Antimicrobials: An Undiscovered Country. *Archaea: New Models for Prokaryotic Biology*. Caister Academic Press, p11.
28. Sinha Roy R, Belshaw PJ, and Walsh CT. (1998) Mutational analysis of posttranslational heterocycle biosynthesis in the gyrase inhibitor microcin B17: distance dependence from propeptide and tolerance for substitution in a GSCG cyclizable sequence. *Biochemistry* 37: 4125-4136.
29. Taylor KA, Deatherage JF, and Amos LA. (1982) Structure of the S-layer of *Sulfolobus acidocaldarius*. *Nature* 299: 840-842.
30. Walsh CT. (2004) Polyketide and nonribosomal peptide antibiotics: modularity and versatility. *Science* 303: 1805-1810.
31. Whitman WB, Pfeifer F, Blum P, and Klein A. (1999) What Archaea Have to Tell Biologist. *Genetics* 152: 1245-1248.
32. Woese CR, Kandler O, and Wheelis ML. (1990) Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proceedings of the National Academy of Sciences of the United States of America* 87: 4576-4579.

CHAPTER IV: NUCLEOPHILIC 1,4-ADDITIONS FOR NATURAL PRODUCT DISCOVERY

This chapter was taken from Cox, Tietz, Sokolowski, Melby, Doroghazi and Mitchell (Cox, *et al.* 2014) and is reproduced with permission from ACS Chemical Biology.

I am grateful to W. van der Donk and D. Eastburn (University of Illinois at Urbana-Champaign, UIUC) for donating strains, in addition to M. Burke, S. Blanke, and P. Orlean (UIUC) for miscellaneous reagents. I thank L. Zhu for NMR assistance. Lastly, I am thankful to K. Sokolowski for assistance with the mass spectrometry and labeling, J. Melby for assisting in the structure elucidation using mass spectrometry, and to J. Tietz for assisting in the purification of cyclothiazomycin C and structure determination using NMR.

Abstract

Natural products remain an important source of drug candidates, but the difficulties inherent to traditional isolation, coupled with unacceptably high rates of compound rediscovery, limit the pace of natural product detection. Here we describe a reactivity-based screening method to rapidly identify exported bacterial metabolites that contain dehydrated amino acids (*i.e.* carbonyl- or imine-activated alkenes), a common motif in several classes of natural products. Our strategy entails the use of a commercially available thiol, dithiothreitol, for the covalent labeling of activated alkenes by nucleophilic 1,4-addition. Modification is easily discerned by comparing mass spectra of reacted and unreacted cell surface extracts. When combined with bioinformatic analysis of putative natural product gene clusters, targeted screening and isolation can be performed on a prioritized list of strains. Moreover, known compounds are easily dereplicated, effectively eliminating superfluous isolation and characterization. As a proof of principle, this labeling method was used to identify known natural products belonging to the thiopeptide, lanthipeptide, and linaridin classes. Further, upon screening a panel of only 23 actinomycetes, we discovered and characterized a novel thiopeptide antibiotic, cyclothiazomycin C.

4.1 Introduction

Bacteria have historically been a rich reservoir of architecturally complex natural products exhibiting antibiotic activity (Newman, *et al.* 2012). However, the traditional approach

to natural product discovery—bioassay-guided isolation of compounds from extracts—is limited by high rates of compound rediscovery (Lewis 2013). As such, the potential value of novel natural products to advance the treatment of disease, and in particular to address the issue of antibiotic resistance (Fischbach, *et al.* 2009), warrants the development of alternative strategies to discover novel compounds. The advent of widely available genome sequences makes bioinformatics-driven methods increasingly appealing, since the enzymatic machinery responsible for natural product biosynthesis can be readily identified (Deane, *et al.* 2013, Velásquez, *et al.* 2011). Consequently, a number of strategies have emerged that aid in connecting biosynthetic gene clusters to their products, including selective enzymatic derivatization (Gao, *et al.* 2013), chemoselective enrichment (Odendaal, *et al.* 2011), mass spectrometry-based network analysis (Nguyen, *et al.* 2013), and PCR prioritization among others. Another approach to address the innovation gap in natural product discovery is to utilize the intrinsic chemical reactivity of functional groups that are enriched in a target class of metabolites. Here, we report the development of a reactivity-based screening method to identify, isolate, dereplicate and characterize novel natural products using a combination of bioinformatics and a simple chemical probe for modifying a reactive functional group (Figure 4.1).

The dehydrated amino acids (DHAAs) dehydroalanine and dehydrobutyrine are frequently found in natural products (Gersch, *et al.* 2012), including thiopeptides (Kelly, *et al.* 2009), lanthipeptides (Yu, *et al.* 2013, Zhang, *et al.* 2012), and linaridins (Claesen, *et al.* 2010, Claesen, *et al.* 2011, Komiyama, *et al.* 1993), among others (Figure 4.2). We thus envisioned DHAAs serving as a useful chemical handle for the discovery of natural products. It has been demonstrated that thiol nucleophiles participate in 1,4-addition into α,β -unsaturated carbonyl/imine DHAAs under mild conditions to yield covalent thioether adducts (Figure 4.1A) (Bonauer, *et al.* 2006). This reactivity has been exploited previously in the chemical modification of thiostrepton (Myers, *et al.* 2010, Schoof, *et al.* 2009, Schoof, *et al.* 2010), the mapping of

Ser/Thr-phosphorylation in proteins (Wells, *et al.* 2002), the design of solid-phase capture resins (Tseng, *et al.* 2005), and the identification of lanthipeptides (Li, Girard, *et al.* 2012). Thus, we sought to employ this well-established, reliable chemistry as part of a novel tandem bioinformatics/reactivity-based screening effort.

Many classes of DHAA-bearing natural products are ribosomally produced, rendering them ideal for genome-guided discovery. The availability of genome sequences has revealed a tremendous biosynthetic capability among diverse microbial species (Challis 2008). It has become apparent that even well-characterized bacteria harbor the potential to produce an abundance of yet-uncharacterized natural products (Bentley, *et al.* 2002). To overcome the burden of rediscovery (Watve, *et al.* 2001), bioinformatics can be used to preselect bacterial strains for screening to only include the organisms with the theoretical capacity to produce a particular type of natural product. However, even with the bioinformatics identification of promising biosynthetic gene clusters, the detection and isolation of the resultant natural products often proves to be difficult given that the products of most biosynthetic pathways are present in extremely low quantities (if present at all) during laboratory cultivation. Accordingly, a broadly applicable companion strategy to genome mining that would allow the determination of whether a natural product of interest is produced at a detectable level would be valuable. We thus reasoned that a combination of bioinformatics and reactivity-based screening (*i.e.* nucleophilic 1,4-addition to DHAAs) would streamline natural product discovery efforts.

4.2 Rationale and Overview of a New Natural Product Discovery Method

Herein we have utilized the combination of bioinformatics and nucleophilic 1,4-addition chemistry for the rapid labeling, discovery, and dereplication of DHAA-containing natural products (Figure 4.1B) by reactivity-based screening. Our discovery pipeline begins with a bioinformatic survey for strains of *Actinobacteria* predicted to be capable of producing a DHAA-

containing natural product. (Figure 4.1B, Step 1, *vide infra* for specifics on the bioinformatics-based strain prioritization). After cultivation, the exported metabolites from the prioritized *Actinobacteria* are extracted with organic solvent using a non-lytic procedure (see Methods). A portion of this cell-surface extract then undergoes treatment with dithiothreitol (DTT) in the presence of base. DTT was chosen as the thiol probe owing to its low cost and ubiquity in natural product discovery laboratories. If reactive DHAA moieties are present in the cell-surface extract, the resulting DTT adducts increase the mass of the exported metabolite by multiples of 154.0 Da (Figure 4.1B, Step 2). Differential mass spectrometry between the unreacted control and the DTT-reacted extracts readily identifies the compounds containing DHAA within a pre-determined mass range. The molecular mass, number of DTT additions, and analysis of tandem mass spectra, combined with the initial bioinformatic prediction of DHAA-containing natural products, permits a rapid determination of compound novelty. At this step, every DTT-labeled compound can be analyzed, irrespective of whether the mass corresponds to a predicted biosynthetic gene cluster. Known compounds are removed from further analysis at this step, leaving only compounds with a high probability of novelty for further structural and functional characterization, which is considerably more time-consuming (Figure 4.1B, Step 3). To determine if the above proposed discovery pipeline was viable, we sought to discover a novel DHAA-containing thiopeptide via bioinformatic prioritization and reactivity-based screening utilizing nucleophilic 1,4-addition chemistry.

4.3 Validation of the DTT-Labeling Strategy

With the ultimate goal of using the above-described DTT-labeling method to discover a new natural product, we first sought to establish an operationally simple route to rapidly screen organic extracts for compounds of interest. We utilized two DHAA-containing natural products, thiostrepton and geobacillin I, for method development and validation.

Thiostrepton, whose biosynthetic gene cluster was identified in 2009 (Kelly, *et al.* 2009), is a thiopeptide produced by *Streptomyces azureus* ATCC 14921 (among others) (Donovick, *et al.* 1955). Notably, the highly-modified scaffold of thiostrepton contains four DHAA sites where labeling can occur: three dehydroalanine residues and one dehydrobutyrine (Figure 4.3A) (Hensens, *et al.* 1983). To test the method, reactions were conducted using commercially-obtained thiostrepton, DTT, and either diisopropylethylamine (DIPEA) or no base at 23 °C for 16 h in a 1:1 mixture of chloroform and methanol. The authentic thiostrepton standard and the DTT-reacted samples were then subjected to matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF MS) analysis. The peaks corresponding to unmodified thiostrepton (m/z 1664.4 Da) were supplanted in the DTT-reacted sample by peaks corresponding to the addition of multiple DTT labels. The tertiary adduct was the most prominent, suggesting the successful addition of DTT into the reactive alkenes (Figure 4.4). The addition of DIPEA enhanced the DTT-labeling reaction. Other bases, including triethylamine and 1,8-diazabicycloundec-7-ene (DBU), were tested and labeling occurred similarly to the reactions using DIPEA. A range of DIPEA concentrations were tested (10–50 mM) and the extent of labeling did not greatly vary. Therefore all further experiments employed 10 mM DIPEA.

To confirm DTT-labeling of thiostrepton could be observed by MALDI-TOF MS in the context of a more complex biological mixture, we subjected an organic cell-surface extract of *S. azureus* ATCC 14921 (thiostrepton producer) to the above labeling reaction. Analogous to the pure thiostrepton sample, comparison of the crude extract with the DTT-labeled extraction again showed the appearance of multiple DTT adducts, this time with the tetra-adduct being the primary species; a higher extent of labeling was seen here due to the larger relative excess of the labeling reagents in the context of a biological extract (Figure 4.3B). Although thiostrepton contains only 4 reactive DHAA sites, a minor 5th adduct was observed in both the commercially available and extracted samples, presumably from reaction with another electrophilic site.

Thiostrepton possesses an additional alkene that is conjugated to pyridine within the quinaldic acid moiety; we hypothesize that addition of DTT may have occurred at this site, given the literature precedent for addition of thiols to aromatic-conjugated alkenes (Ranu, *et al.* 2007). Importantly, the appearance of this low-intensity ion does not complicate detection or interpretation of the labeled analyte.

Lanthipeptides are ribosomally synthesized and post-translationally modified peptide natural products (RiPPs) that are easily identified using bioinformatics and frequently contain DHAAAs (Arnison, *et al.* 2013, Velásquez, *et al.* 2011, Yu, *et al.* 2013, Zhang, *et al.* 2012). To test if the reactivity-based screening method could also be used to identify other classes of natural products in varied bacterial extracts, we attempted to label the lanthipeptide geobacillin I. Geobacillin I, a nisin analogue, is produced by *Geobacillus sp.* M10EXG (Figure 4.5A) (Garg, *et al.* 2013, Garg, *et al.* 2012). Upon subjecting an organic cell-surface extract from *Geobacillus sp.* M10EXG to our labeling conditions, a mass corresponding to 2 DTT adducts was prominently observed; a third adduct was visible but of very low intensity (Figure 4.5B). Only two reactive DHAA sites are present in geobacillin I: a dehydroalanine and a dehydrobutyrine. However, transient DHAA sites occur in the biosynthesis of the lanthionine rings, which are formed by intramolecular 1,4-addition of cysteines to DHAAAs (Yu, *et al.* 2013). We hypothesize, accordingly, that a small percentage of the geobacillin present in the extract may have an unformed lanthionine ring, leaving a free reactive site available for DTT-labeling. Again, even under stoichiometrically forcing conditions, this extract adduct was of only minor abundance and thus did not interfere with compound detection or analysis.

4.4 Bioinformatics Guided Strain Prioritization

Like lanthipeptides, thiopeptides are RiPPs and the biosynthetic genes responsible for their production are often clustered, rendering them identifiable by sequence similarity searching.

From the perspective of the present study, we sought to prioritize bacterial strains for subsequent screening based on the presence of biosynthetic genes capable of installing DHAAAs (often misleadingly annotated as “lantibiotic dehydratases”) (Zhang, *et al.* 2012). These genes, however, can be found in a variety of other natural product gene clusters and not exclusively in thiopeptide clusters. Therefore, we first identified clusters that encode for the YcaO cyclodehydratase protein that is necessary for the biosynthesis of all thiazole/oxazole-modified microcin natural products, of which thiopeptides can be broadly categorized. Strains containing a YcaO cyclodehydratase were analyzed further for the local co-occurrence of genes encoding a “lantibiotic dehydratase” (for the production of DHAAAs) and a thiopeptide-like precursor peptide (Figure 4.6A) (Li, Qu, *et al.* 2012). 130 unique strains of recently sequenced (in-house) *Actinobacteria* from the Northern Regional Research Laboratory collection (NRRL), which is curated by the Agricultural Research Service under the supervision of the U.S. Department of Agriculture (USDA/ARS), were predicted to have the genetic capacity to produce a DHAA-containing thiopeptide (Figure 4.1B). The precursor peptide sequences from these clusters were then used to estimate the masses of the final natural products for dereplication and characterization purposes (Figure 4.6B). These strains were then subjected to reactivity-based screening with DTT and DIPEA to discover a novel thiopeptide.

4.5 MS-Based Screening of Prioritized Strains

Twenty-three of the prioritized strains with novel precursor peptide sequences were selected for screening by DTT-labeling (Figure 4.7). We first noticed a sample containing 1–2 DTT adducts on an exported metabolite with a mass of $[M+H]^+$, m/z 1855.0 Da. While we were intentionally blind to which of the *Actinobacteria* strains were undergoing analysis, after labeling we established that this particular extract originated from *Streptomyces griseus* subsp. *griseus*, and the labeled mass did not correlate with the expected mass of the predicted thiopeptide from this strain. However, *Streptomyces griseus* subsp. *griseus* is a known producer of grisemycin, of

which the mass of the labeled natural product did correlate (Figure 4.8A-B). MS/MS fragmentation analysis yielded a seven amino acid sequence tag confirming the identity of the compound as grisemycin (Figure 4.8C) (Claesen, *et al.* 2011). The labeling and identification of grisemycin, a member of the linaridin class of natural products, further validated our reactivity-based screen while also highlighting the usefulness of bioinformatic integration to rapidly dereplicate known compounds.

The organic cell-surface extract from a separate sample contained a compound ($[M+H]^+$, m/z 1486.3 Da) that underwent labeling to contain primarily three DTT adducts (Figure 4.9A). This mass correlated well with the predicted mass of a hypothetical thiopeptide from NRRL strain WC-3908. The thiopeptide gene cluster from WC-3908 was similar to the gene clusters responsible for the production of the thiopeptides cyclothiazomycin A, originally termed 5102-I (Wang, *et al.* 2010), and cyclothiazomycin B (Figure 4.9B). The core region of the precursor peptide (*i.e.* the portion that undergoes enzymatic tailoring to yield the mature natural product) (Arnison, *et al.* 2013, Melby, *et al.* 2011) from WC-3908 differed by two amino acids from the precursor peptides of cyclothiazomycin A and B (Figure 4.9C). Accordingly, we designated the WC-3908 thiopeptide cyclothiazomycin C. Given that the structures of cyclothiazomycin A and B have been reported (Aoki, Ohtsuka, Itezono, *et al.* 1991, Aoki, Ohtsuka, Yamada, *et al.* 1991, Hashimoto, *et al.* 2006), we could accurately predict the structure of cyclothiazomycin C, which was in agreement with the labeling results (Figure 4.9D).

4.6 Verification of the Cyclothiazomycin C Structure

Prior to detailed structural characterization, cyclothiazomycin C was purified by MPLC and HPLC (Figure 4.10). The mass spectrum of purified cyclothiazomycin C revealed an $[M+H]^+$ ion of m/z 1486.3309 Da (Figure 4.11A), supporting the molecular formula for the predicted structure of cyclothiazomycin C ($C_{60}H_{67}N_{19}O_{13}S_7$). Analysis of the collision-induced dissociation (CID) mass spectrum corroborated the amino acid sequence of the precursor peptide, strongly

connecting the predicted gene cluster to the mature natural product (Figure 4.11B). NMR spectroscopy was then used to confirm the predicted structure of cyclothiazomycin C (Supplemental Figures 7–8). Bond connectivity was established using ^1H - ^1H COSY, ^1H - ^1H TOCSY, ^1H - ^{13}C HSQC, and ^1H - ^{13}C HMBC experiments. Chemical shifts were assigned from this information and by comparison to the reported values for cyclothiazomycin B (Hashimoto, *et al.* 2006). Due to the spectral similarity to cyclothiazomycin B, we have assigned the stereochemistry of cyclothiazomycin C analogously to the reported compound.

4.7 Conservation Analysis of the Cyclothiazomycin C Biosynthetic Gene Cluster

To provide additional evidence that the thiopeptide gene cluster from WC-3908 was responsible for the production of cyclothiazomycin C, conservation analysis was performed with the cyclothiazomycin A, B, and C (putative) gene clusters. The cyclothiazomycin A biosynthetic genes derived from *Streptomyces hygroscopicus* subsp. *jinggangensis* 5008 while the cyclothiazomycin B genes were from *Streptomyces mobaraensis*. A subset of the genes predicted for the production of cyclothiazomycin B (Wang, *et al.* 2010) was conserved among the three clusters (Figure 4.9B). All three clusters contain a short open reading frame, here designated *ctmA*, encoding the precursor peptide. *CtmD* encodes a “fused” TOMM cyclodehydratase (E1 ubiquitin-activating enzyme/MccB-like and YcaO domains), which implicates CtmD in the formation of thiazolines (Dunbar, *et al.* 2012, Dunbar, *et al.* 2013). *CtmB* encodes a flavin mononucleotide-dependent protein, putatively responsible for the dehydrogenation of the thiazolines to thiazoles (Zhang, *et al.* 2014). *CtmE* and *ctmF* encode homologs of a split lanthipeptide dehydratase, which performs the dehydration of serine and threonine to dehydroalanine and dehydrobutyrine (Yu, *et al.* 2013, Zhang, *et al.* 2012). Like all thiopeptides, cyclothiazomycin C has a substituted 6-membered, nitrogen-containing central heterocycle (here a pyridine). In the case of cyclothiazomycins A and B, the pyridine moiety is likely formed by the gene product of *ctmG*, given the homology to *tclM*, which has been implicated in the formal

[4+2] cycloaddition reaction during thiocillin biosynthesis (Figure 4.12) (Bowers, *et al.* 2010). For cyclothiazomycin C, a gene with high similarity to *ctmG* from the cyclothiazomycin A and B clusters is present, but distantly located in the genome, indicating that the cyclothiazomycin C gene cluster is fragmented. Interestingly, *ctmG* from WC-3908 is found directly next to a gene duplication of *ctmF*, which is suggestive of paralogous duplication (Figure 4.12). *CtmI*, which is present in all three clusters, encodes a ThiF-like protein. ThiF-like proteins have been implicated in the biosynthesis of thiamine diphosphate in *E. coli*. However, the function of ThiF-like proteins in the context of TOMM biosynthesis remains to be established. Other local genes include *ctmH*, which is a LuxR-type regulatory gene and *ctmJK*, which are omitted from the cyclothiazomycin A and C clusters and have no known function (Figure 5). We further note that the genes flanking the conserved region are highly disparate between the three clusters (Figure 4.13). This subset of genes, *ctmA-G* and *ctmI* from *Streptomyces hygrosopicus* subsp. *jinggangensis* 5008 were recently shown to be regulated by the LuxR-type regulatory gene *ctmH*. Furthermore, the deletion of *ctmA*, *ctmD*, *ctmF*, and *ctmG* abolished the production of cyclothiazomycin A (Zhang, *et al.* 2014). These data further support the gene cluster prediction for cyclothiazomycin C from WC-3908.

4.8 Assessment of Cyclothiazomycin Bioactivity

Previous reports on cyclothiazomycins A and B describe a wide range of bioactivities, including renin inhibition (Aoki, Ohtsuka, Yamada, *et al.* 1991), RNA polymerase inhibition (Hashimoto, *et al.* 2006), and antifungal activity (Mizuhara, *et al.* 2011). We found that purified cyclothiazomycin C exhibited growth inhibitory action toward several Gram-positive (*Firmicutes*) bacteria but was inactive against all tested Gram-negative (*Proteobacteria*) organisms (Table 1). The greatest inhibitory activity was observed towards the genus *Bacillus*. Based on prior reports, we decided to also evaluate if cyclothiazomycin C exhibited growth inhibitory action toward a variety of fungal strains, but none was observed.

To further clarify cyclothiazomycin bioactivity, we obtained a cyclothiazomycin B producer, strain NRRL B-3306, and purified cyclothiazomycin B in a manner analogous to that employed for cyclothiazomycin C (Figures 4.14-4.15). As above, we assessed cyclothiazomycin B for antibiotic and antifungal activity. Cyclothiazomycin B also had the greatest inhibitory activity towards the genus *Bacillus*, with little to no activity against a panel of Gram-negatives and fungal strains (Table 4.1). This activity does not align with previous reports (Hashimoto, *et al.* 2006, Mizuhara, *et al.* 2011); however, additional fungal strains will need to be tested to more concretely establish cyclothiazomycin spectrum of activity. The antibiotic activity of cyclothiazomycin B and C are similar to known thiopeptides, which act as translation inhibitors by binding to either the 50S subunit or EF-Tu (Just-Baringo, *et al.* 2014). It is possible that the cyclothiazomycins act in a similar manner but the determination of the precise mode of action will require further exploration.

4.9 Summary and Outlook

In summary, we have described a reactivity-based screening method to conveniently identify natural products containing dehydrated amino acids (DHAAs). This method employs ubiquitous reagents and instrumentation, making it a broadly accessible strategy for natural product discovery. Three characteristics make the nucleophilic 1,4-addition labeling procedure operationally straightforward: (a) anhydrous solvents are unnecessary, meaning the reaction is performed under ambient atmosphere; (b) the reagents employed are common in most laboratories and easily handled; and (c) the large excess of labeling reagent relative to the substrate means that precise stoichiometric calculations for each reaction are unnecessary. Although under these excess labeling conditions we often observe minor peaks related to non-DHAA labeling, these species never convoluted spectral interpretation. Including a rapidly dereplicated example, we validated the use of nucleophilic 1,4-additions for natural product discovery with the labeling of three previously characterized natural products: thiostrepton,

grisemycin, and geobacillin I. This reactivity based screen was combined with bioinformatics to increase the rate of discovery, even with low abundance products. Often, natural products are present only at trace quantities. By capitalizing on the remarkable sensitivity of mass spectrometry, the compound(s) to be discovered do not need to be present at bioactive concentrations, they only need to be detectable upon labeling. After screening the organic extracts of only 23 *Actinobacteria*, we report on a new thiopeptide, cyclothiazomycin C. The structure of cyclothiazomycin C was established through MS and NMR, along with confirmed bioactivity towards Gram-positive bacteria. When compared to traditional bioassay-guided isolation, which can require many thousands of samples to be screened to find new compounds, our discovery rate (1 out of 23 strains) highlights the potential of this tandem strategy. With the substantial rise of available genomic sequences, we anticipate that the combination of bioinformatics and simple chemoselective reactivity-based labeling will provide a powerful tool to identify novel natural products, while dramatically reducing the time invested on the unfruitful rediscovery of known compounds.

4.10 Experimental

4.10.1 Preparation of cell extracts for screening

Actinomycete strains were grown in 10 mL of MS medium (1 L contains 20 g mannitol, 20 g roasted soy flour) at 30 °C for 7 d. Exported metabolites were extracted from the cultures using 2 mL of *n*-BuOH at room temperature. For thiostrepton production, *Streptomyces azureus* was grown in 10 mL of ISP4 medium (1 L contains 10 g soluble starch, 1 g K₂HPO₄, 1 g MgSO₄, 1 g NaCl, 2 g Na₂SO₄, 2 g CaCO₃, 1 mg FeSO₄, 1 mg ZnSO₄ heptahydrate, 1 mg MnCl₂ heptahydrate) for 7 d at 30 °C. Thiostrepton was extracted with 1 mL of CHCl₃ at 23 °C. Both extracts were agitated for 1 min by vortex, submitted to centrifugation (4000 × *g*, 5 min), and the organic layer was removed from the intact, harvested cells. For geobacillin I production, *Geobacillus* sp. M10EXG was grown on modified LB agar (1 L contains 10 g casein enzymatic

hydrolysate, 5 g yeast extract, 5 g NaCl and 10 g agar) at 50 °C for 60 h. Cells were removed from the plates with 10 mL of 70% aq. *i*-PrOH and agitated by rocking for 24 h at 23 °C. The intact cells were then removed from the extract by centrifugation (4000 × *g*, 5 min). An aliquot (1 μL) of the extract was then mixed with 9 μL of sat. α-cyano-4-hydroxycinnamic acid (CHCA) matrix solution in 1:1 MeCN/H₂O containing 0.1% trifluoroacetic acid (TFA). 1 μL was spotted onto a MALDI plate for subsequent MALDI-TOF MS analysis.

4.10.2 DTT-labeling

For commercially-obtained thiostrepton (Calbiochem, 99%), a 20 μL volume of 10.5 mM thiostrepton, 500 mM DTT, and 10 mM DIPEA in 1:1 CHCl₃/MeOH was allowed to react at 23 °C for 16 h. For the no base reaction, thiostrepton and DTT were added similarly to above and MeOH (without DIPEA) was added to establish a 1:1 CHCl₃/MeOH. The sample was then analyzed for DTT incorporation by MALDI-TOF MS (see below). For thiostrepton produced by *Streptomyces azureus* (and thus labeling occurred in the context of the crude cell-surface extract), 14 μL of the extract was mixed with DTT (in MeOH) and DIPEA (in MeOH) to generate a final volume of 20 μL with a final concentration of 500 mM DTT and 10 mM DIPEA, in 7:3 CHCl₃/MeOH and the mixture was allowed to proceed for 16 h at 23 °C. An aliquot (1 μL) of the extract was then mixed with 9 μL of sat. α-cyano-4-hydroxycinnamic acid (CHCA) matrix solution in 1:1 MeCN/H₂O containing 0.1% TFA. 1 μL was spotted onto a MALDI plate for subsequent MALDI-TOF MS analysis.

4.10.3 Bioinformatics based strain prioritization

A previously reported profile Hidden Markov Model and the program HMMER were used to identify the YcaO cyclodehydratase (Pfam PF02624) (Doroghazi, *et al.* 2013). The local genomic region (10 open reading frames on either side of the YcaO gene) was analyzed manually

for the presence of a “lantibiotic dehydratase” gene and a putative precursor peptide. Only strains with the presence of all three genes were taken forward for reactivity-based screening.

4.10.4 MALDI-TOF mass spectrometric analysis

MALDI-TOF mass spectrometry was performed using a Bruker Daltonics UltrafleXtreme MALDI-TOF/TOF instrument operating in positive reflector mode. The instrument was calibrated before data acquisition using a commercial peptide calibration kit (AnaSpec – Peptide Mass Standard Kit). Analysis was carried out with Bruker Daltonics flexAnalysis software. All spectra were processed by smoothing and baseline subtraction.

4.10.5 Isolation of cyclothiazomycin C

WC-3908 was grown in 10 mL of ATCC 172 medium at 30 °C for 48 h. 300 µL of the culture was spread onto 15 cm plates (*ca.* 75 mL of solid ATCC medium). The plates were then incubated for 7 d at 23 °C. A razor blade was used to remove the bacterial lawn from the solid medium. The bacterial growth from 14 plates (~1 L of medium) was extracted with *n*-BuOH (500 mL) for 24 h at 23 °C. The extract was then filtered through Whatman filter paper and allowed to evaporate under nitrogen before being redissolved in 3:1 pyridine:water (*ca.* 3 mL) and transferred to a 50 mL conical tube. The resulting solution was clarified by centrifugation, to remove insoluble debris (4000 × *g*, 5 min). The supernatant was then injected onto a reverse-phase C18 silica column (TeleDyne Isco 5.5 g C18 Gold cartridge) and purified by MPLC (gradient elution from 20-95% MeOH/10 mM aq. NH₄HCO₃). Fractions containing the desired product (as determined by MALDI-TOF MS; [M+H] *m/z* = 1486) were combined and immediately concentrated by rotary evaporation. The resulting residue was dissolved in 3:1 pyridine/water (*ca.* 0.5 mL), transferred to a microcentrifuge tube, centrifuged (15000 × *g*, 5 min), filtered (0.2 µm polyethersulfone syringe filter), and further purified by HPLC. Semi-preparative HPLC employed a Thermo Scientific Betasil C18 column (100 Å; 250 × 10 mm; 5

μm particle size) operating at 4.0 mL min^{-1} on a PerkinElmer Flexar LC system using Flexar Manager software. Solvent A was $10 \text{ mM aq. NH}_4\text{HCO}_3$. Solvent B was MeOH. Cyclothiazomycin C was purified by isocratic elution at 72% B, typically eluting 19.5 min after initiation of the HPLC run (alternatively, the elution time was ~ 12 min when 75% B was used). HPLC progress was monitored by photodiode array (PDA) UV-Vis detection. Fractions corresponding to the desired product (as determined by UV-Vis and MALDI-TOF MS) were immediately concentrated under rotary evaporation or under a stream of N_2 gas. The resulting residue was suspended in water (*ca.* 1 mL), assisted by vortex mixing and sonication. The suspended product was flash-frozen in liquid N_2 and lyophilized for >24 h to give purified cyclothiazomycin C as a white to off-white powder. Purity was determined by analytical HPLC [Thermo Scientific Betasil C18 column (100 \AA ; $250 \times 4.6 \text{ mm}$; $5 \mu\text{m}$ particle size) operating at 1.0 mL min^{-1} using the same solvents] and NMR. Isolated yield ranged from 10-90 $\mu\text{g/plate}$ (15 cm diameter).

4.10.6 Isolation of cyclothiazomycin B

NRRL strain B-3306 was grown in a fashion identical isolation conditions for WC-3908. Cyclothiazomycin B ($[\text{M}+\text{H}] m/z = 1528$) was also purified in the same manner as cyclothiazomycin C, except that HPLC purification employed 75% B (retention time typically *ca.* 17 min). After lyophilization, an off-white powder was obtained. Purity was determined by analytical HPLC [Thermo Scientific Betasil C18 column (100 \AA ; $250 \times 4.6 \text{ mm}$; $5 \mu\text{m}$ particle size) operating at 1.0 mL min^{-1} using the same solvents]; identity was determined by high-resolution mass spectrometry. Isolated yield was approximately 13 $\mu\text{g/plate}$ (15 cm diameter).

4.10.7 FT-MS/MS analysis of cyclothiazomycin B and C

The purified cyclothiazomycins were dissolved in 80% aq. MeCN with 0.1% formic acid. Samples were directly infused using a 25 μL Hamilton gas-tight syringe (cyclothiazomycin C) or

an Advion Nanomate 100 (cyclothiazomycin B), into a ThermoFisher Scientific LTQ-FT hybrid linear ion trap, operating at 11T (calibrated weekly). The FT-MS was operated using the following parameters: minimum target signal counts, 5,000; resolution, 100,000; m/z range detected, dependent on target m/z ; isolation width (MS/MS), 5 m/z ; normalized collision energy (MS/MS), 35; activation q value (MS/MS), 0.4; activation time (MS/MS), 30 ms. Data analysis was conducted using the Qualbrowser application of Xcalibur software (Thermo-Fisher Scientific).

4.10.8 NMR spectroscopy of cyclothiazomycin C

NMR spectra were recorded on a Varian NMR System 750 MHz narrow bore magnet spectrometer (VNS750NB employing a 5 mm Varian $^1\text{H}[^{13}\text{C}/^{15}\text{N}]$ PFG X, Y, Z probe) or a Varian Unity Inova 500 MHz narrow bore magnet spectrometer (UI500NB employing a 5 mm Varian $^1\text{H}[^{13}\text{C}/^{15}\text{N}]$ PFG Z probe). Spectrometers were operated at 750 MHz and 500 MHz, respectively, for ^1H detection, and 188 MHz for indirect ^{13}C detection. Carbon resonances were assigned via indirect detection (HSQC and HMBC experiments). Resonances were referenced internally to the most downfield solvent peak (8.74 ppm, pyridine). Default Varian pulse sequences were employed for ^1H , COSY, DQF-COSY, TOCSY, HSQC, HMBC, and ROESY experiments. Samples were prepared by dissolving approximately 3-7 mg of cyclothiazomycin C (HPLC-purified and lyophilized) in pyridine- d_5 /D $_2$ O (3:1, 600 μL). Pyridine- d_5 (99.94% D) and D $_2$ O (99.9% D) were obtained from Cambridge Isotope Laboratories (Andover, MA). Samples were held at 25 $^\circ\text{C}$ during acquisition.

4.10.9 Analysis of NMR data

Assigned resonances are shown in tabular form and directly on the structure within Supplemental Figure 7. Due to the solvent employed (3:1 pyridine- d_5 /D $_2$ O), exchangeable peaks (*i.e.* N-H, O-H) were not detected. The corresponding ^1H resonances of the analogous locations in

cyclothiazomycin B1 (reported previously (Hashimoto, *et al.* 2006)) are also given in Supplemental Figure 7 for comparison. Resonances were assigned by 2D NMR spectroscopy, as well as by comparison to the reported spectra of cyclothiazomycin B1 (Hashimoto, *et al.* 2006).

4.10.10 Evaluation of cyclothiazomycin B and C antibiotic activity

Bacillus subtilis strain 168, *Bacillus anthracis* strain Sterne, *E. coli* MC4100, and *Pseudomonas putida* KT2440 were grown to stationary phase in 10 mL of Luria-Bertani broth (LB) at 37 °C. *Staphylococcus aureus* USA300 (methicillin-resistant), *Enterococcus faecalis* U503 (vancomycin-resistant), and *Listeria monocytogenes* strain 4b F2365 were grown to stationary phase in 10 mL brain-heart infusion (BHI) medium at 37 °C. *Neisseria sicca* ATCC 29256 was grown to stationary phase in 5 mL of gonococcal broth at 37 °C. The cultures were adjusted to an OD₆₀₀ of 0.013 in the designated medium before being added to 96-well microplates. Successive two-fold dilutions of cyclothiazomycin C or cyclothiazomycin B (standard solution: 5 mg mL⁻¹ in DMSO) were added to the cultures (0.5–64 µg mL⁻¹). As a control, kanamycin was added to samples of *E. coli*, *B. subtilis*, *B. anthracis*, *P. putida*, *L. monocytogenes*, and *N. sicca* with dilutions from 1–32 µg mL⁻¹. Gentamycin was used as a control for *S. aureus* and *E. faecalis*. As a negative control, an equal volume of DMSO lacking antibiotic was used. Plates were covered and incubated at 37 °C for 12 h with shaking. The minimum inhibitory concentration (MIC) reported is the value that suppressed all visible growth.

4.10.11 Evaluation of cyclothiazomycin B and C antifungal activity

Saccharomyces cerevisiae, *Talaromyces stipitatus*, and *Aspergillus niger* were grown for 36 h in 2 mL of YPD medium (1 L contains 10 g yeast extract, 20 g Peptone and 20 g Dextrose) at 30 °C. *Fusarium virguliforme* was grown for 7 d on potato dextrose agar at 30 °C. Spores were isolated and a suspension of 10⁶ spores in potato dextrose broth was added to the 96-well microplate. *S. cerevisiae* cultures were adjusted to an OD₆₀₀ of 0.013 in the designated medium

before being added to 96-well microplates. *T. stipitatus*, and *A. niger* were not diluted prior to adding to the 96-well microplate. Successive two-fold dilutions of cyclothiazomycin C and cyclothiazomycin B (standard solution: 5 mg mL⁻¹ in DMSO) were added to the cultures (0.5–64 µg mL⁻¹). As a positive control, amphotericin B was added to the cultures with dilutions from 0.5–8 µg mL⁻¹. An equal volume of DMSO was used as a negative control. Plates were covered and incubated at 30 °C for 36 h for *T. stipitatus*, *A. niger*, and *S. cerevisiae* or 60 h for *F. virguliforme* with shaking. The minimum inhibitory concentration (MIC) reported is the value that suppressed all visible growth.

4.11 Figures and Tables

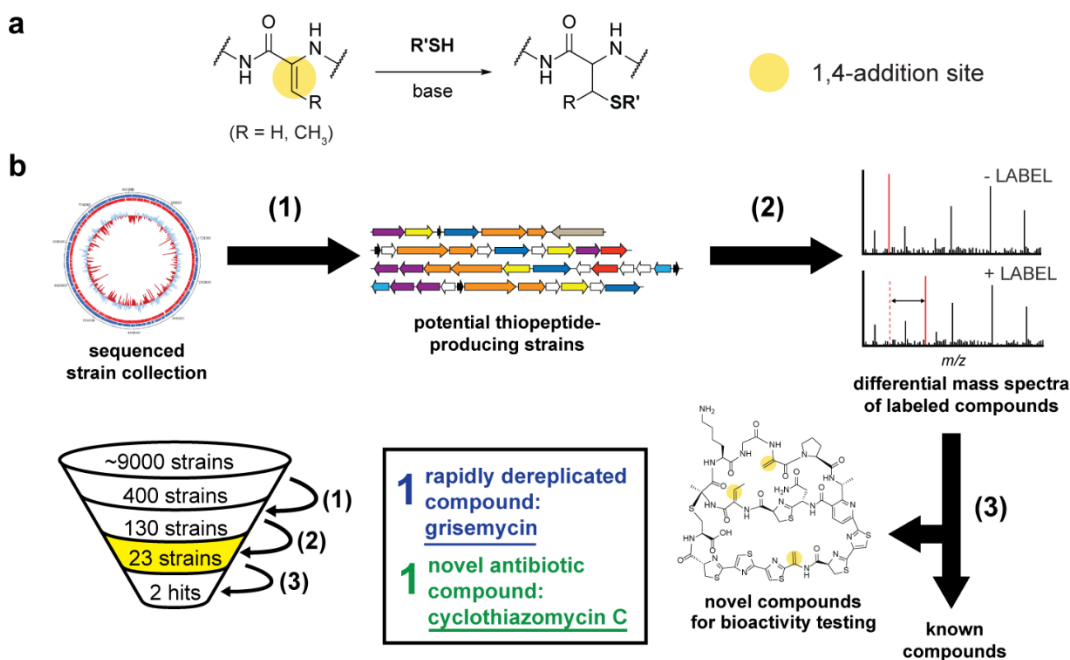


Figure 4.1 | Strategy for natural product discovery by bioinformatics prioritization and nucleophilic 1,4-addition chemistry. (A) Reaction scheme for the thiol (DTT/DIPEA) labeling method with 1,4-addition sites indicated with yellow circles. (B) Work flow for the bioinformatics-based strain prioritization, subsequent DTT-labeling, and MS screening (reactivity-based screening). (1) Prediction of DHAA-containing thiopeptide biosynthetic gene clusters from 400 in-house sequenced genomes (all from the USDA ARS *Actinobacteria* collection, which totals ~9000 unique strains). More information on strain prioritization is given in Supplemental Figure 3. (2) DHAA-containing thiopeptide biosynthetic gene clusters that are reactive towards nucleophilic 1,4-additions (by DTT/DIPEA) are identified by differential mass spectrometry. (3) Compound isolation and characterization after dereplication. Compounds are dereplicated, taking only potentially novel compounds through the time-consuming characterization steps. Of the 400 sequenced genomes, 130 strains were prioritized, 23 strains were screened, 1 compound was rapidly dereplicated, and 1 compound was predicted to be novel and thus further characterized.

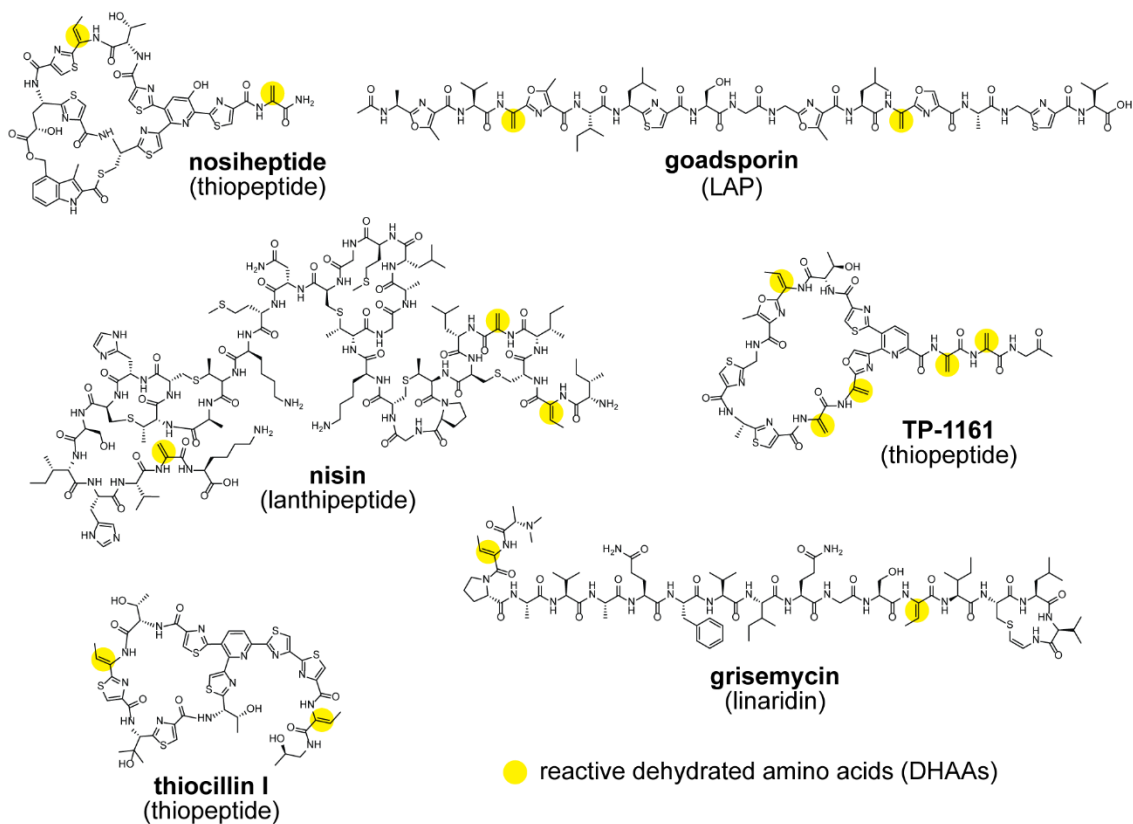


Figure 4.2 | Representative natural products bearing dehydrated amino acids (DHAAs). Structures of example molecules that contain DHAAs suitable for nucleophilic addition are shown. The sites of potential nucleophilic reactivity (*i.e.* the DHA alkenes, often in the form of an α,β -unsaturated carbonyl) are indicated with yellow circles. LAP, linear azol(in)e-containing peptide.

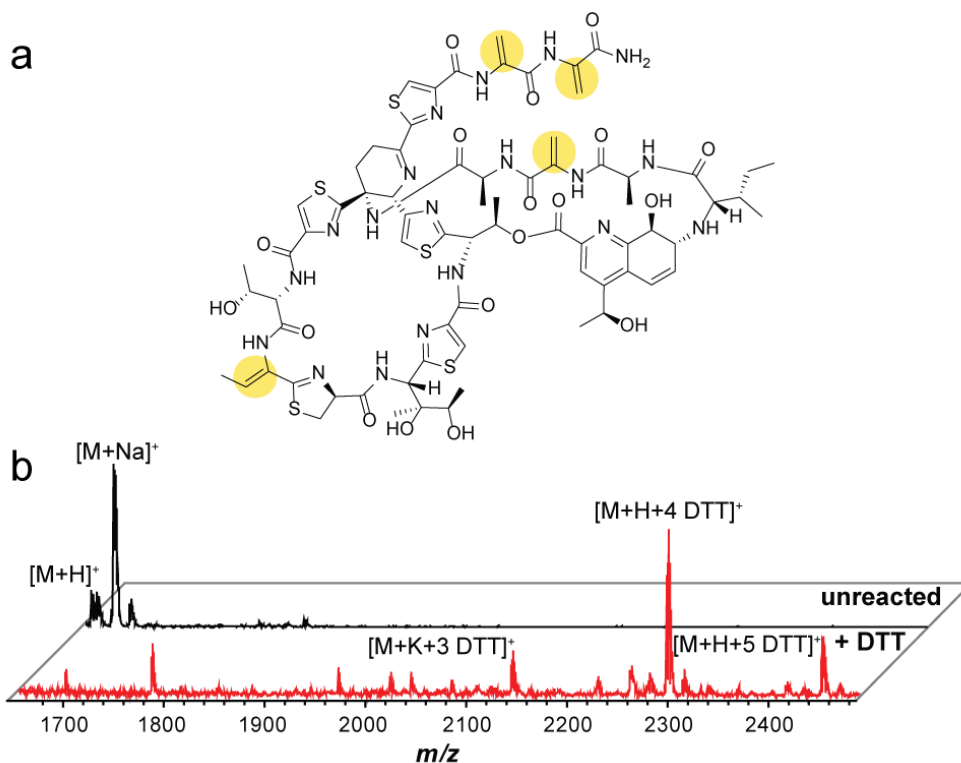


Figure 4.3 | DTT-labeling of thiostrepton as a proof of principle. (A) Structure of thiostrepton with DHAA's suitable for nucleophilic addition highlighted with yellow circles **(B)** MALDI-TOF MS of thiostrepton labeling performed in the context of an organic, cell-surface extract of *Streptomyces azureus* ATCC 14921. The black spectrum (top) is an unreacted control while the red spectrum (bottom) resulted from DTT-labeling. Thiostrepton was visibly labeled by 1-5 DTT moieties, with the 4 DTT adduct being the majority product.

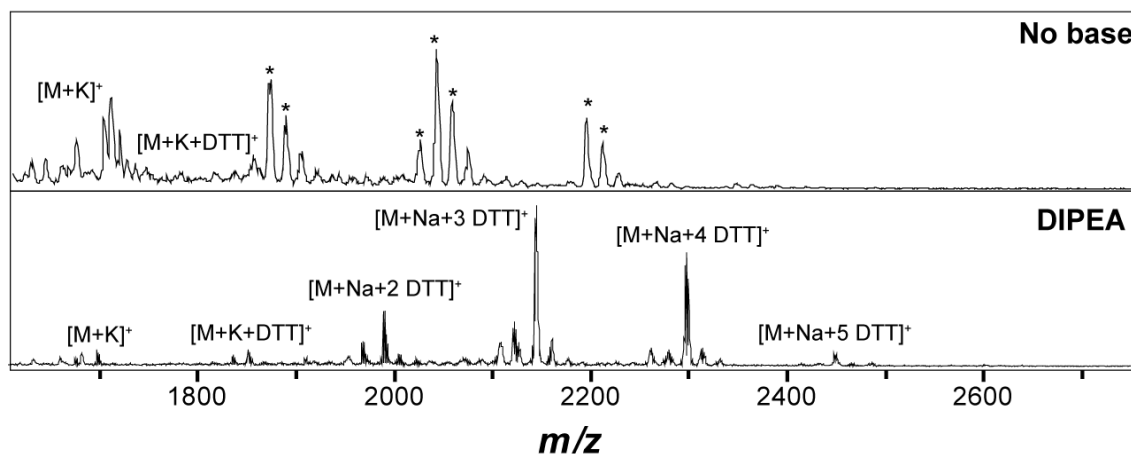


Figure 4.4 | Base-dependence of the DTT-labeling reaction. MALDI-TOF MS of pure (commercially-obtained) thiostrepton reacted with DTT in the presence of diisopropylethylamine (DIPEA) (top), or no base (bottom). Thiostrepton was visibly labeled with 1-5 DTT moieties. * denotes peaks not corresponding to DTT labeling.

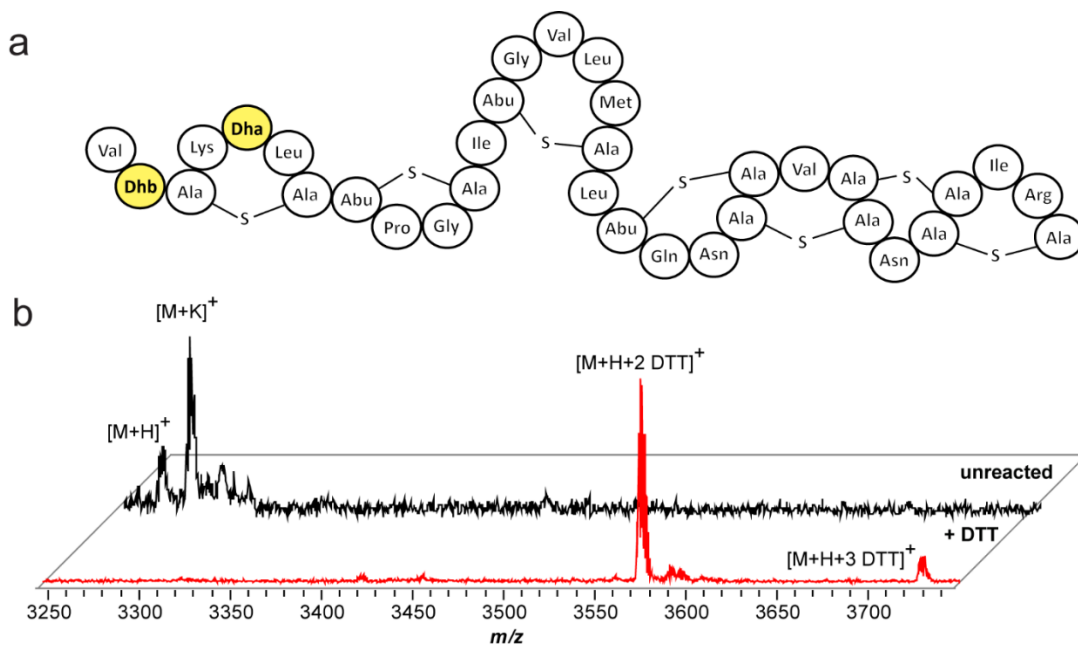


Figure 4.5 | DTT-labeling of geobacillin I in a cell-surface extract. (A) Structure of geobacillin I. Dha, dehydroalanine; Dhb, dehydrobutyrine; Ala-S-Ala, lanthionine; Abu-S-Ala, β -methyl-lanthionine. **(B)** Nucleophilic labeling with DTT of geobacillin I within the context of the organic extract of *Geobacillus sp.* M10EXG. Mass spectra of crude unlabeled extract (black spectrum, top) and DTT-labeled material (red spectrum, bottom) are shown. Extent of labeling with DTT is indicated on the bottom spectrum (2 DTT adducts are clearly observed, with the third being a very low intensity ion).

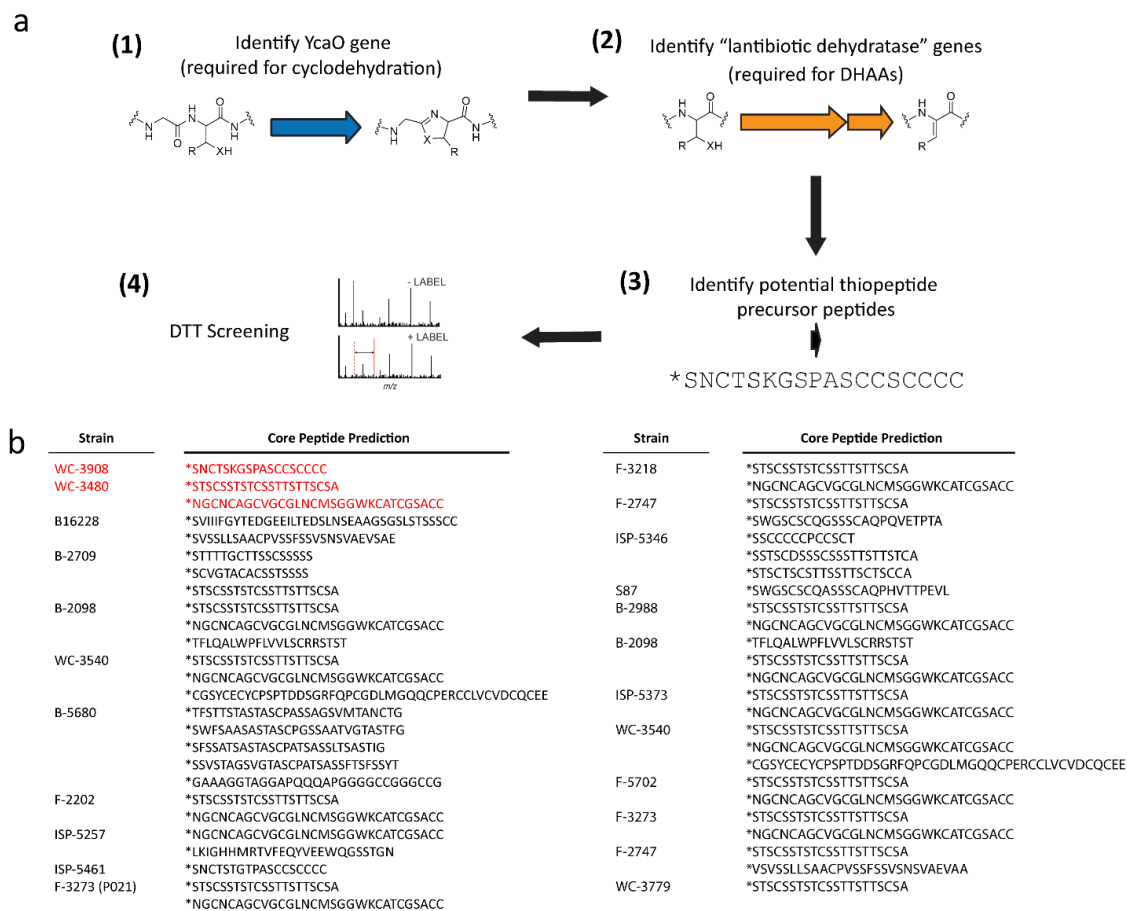


Figure 4.6 | Bioinformatic prioritization of strains. (A) Bioinformatics prioritization schematic. 1) A list is populated with strains encoding a thiazole/oxazole-modified microcin (TOMM) cyclodehydratase “YcaO” necessary for the heterocyclization of select Cys, Ser, and Thr, residues. 2) The list of strains is then trimmed to only contain strains that also harbor a “lantibiotic” dehydratase in close proximity (within 10 open reading frames on either side) to the YcaO protein. 3) TOMM-like precursor peptides from the trimmed list are then identified, and the mass of the final natural product is predicted for use in the dereplication process. 4) If strains make it through steps 1-3, reactivity-based screening with DTT is utilized to identify natural products of interest. (B) Predicted core regions of the precursor peptides identified in the 23 strains prioritized and screened using the DTT labeling method. Highlighted in red are the precursor peptides predicted from WC-3908 (the producer of cyclothiazomycin C) and WC-3480 (the producer of grisemycin).

Figure 4.7

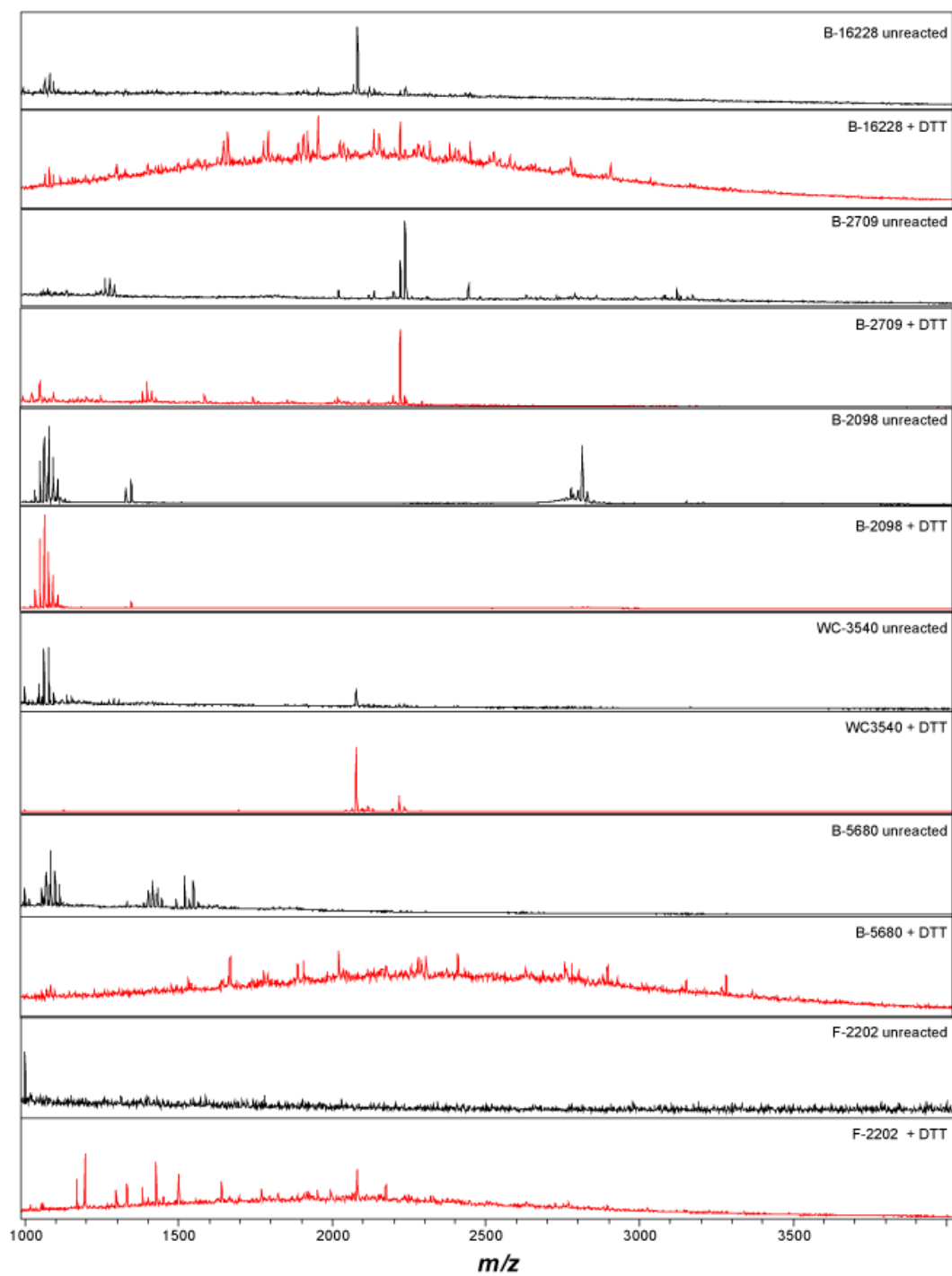


Figure 4.7 (continued)

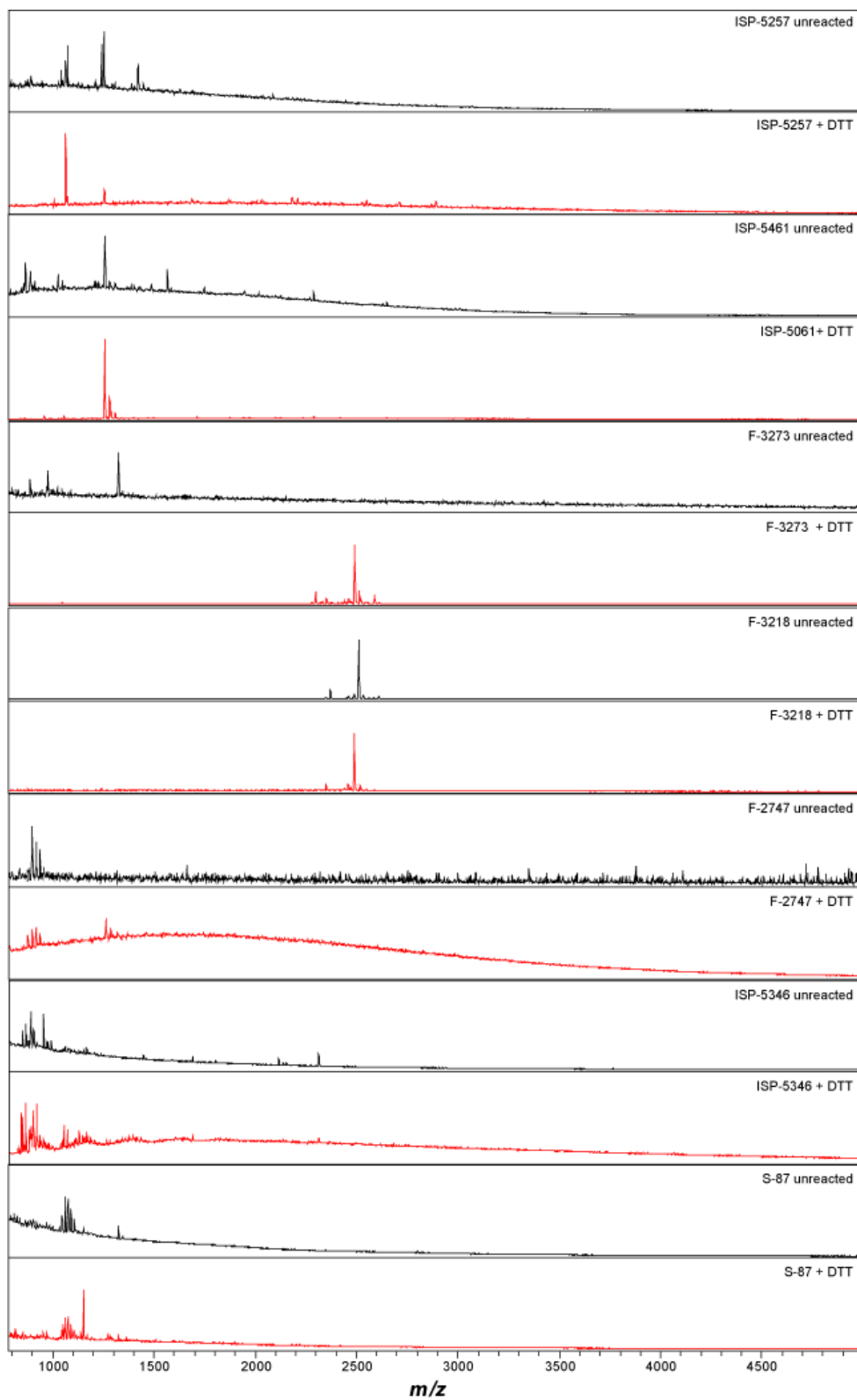


Figure 4.7 (continued)

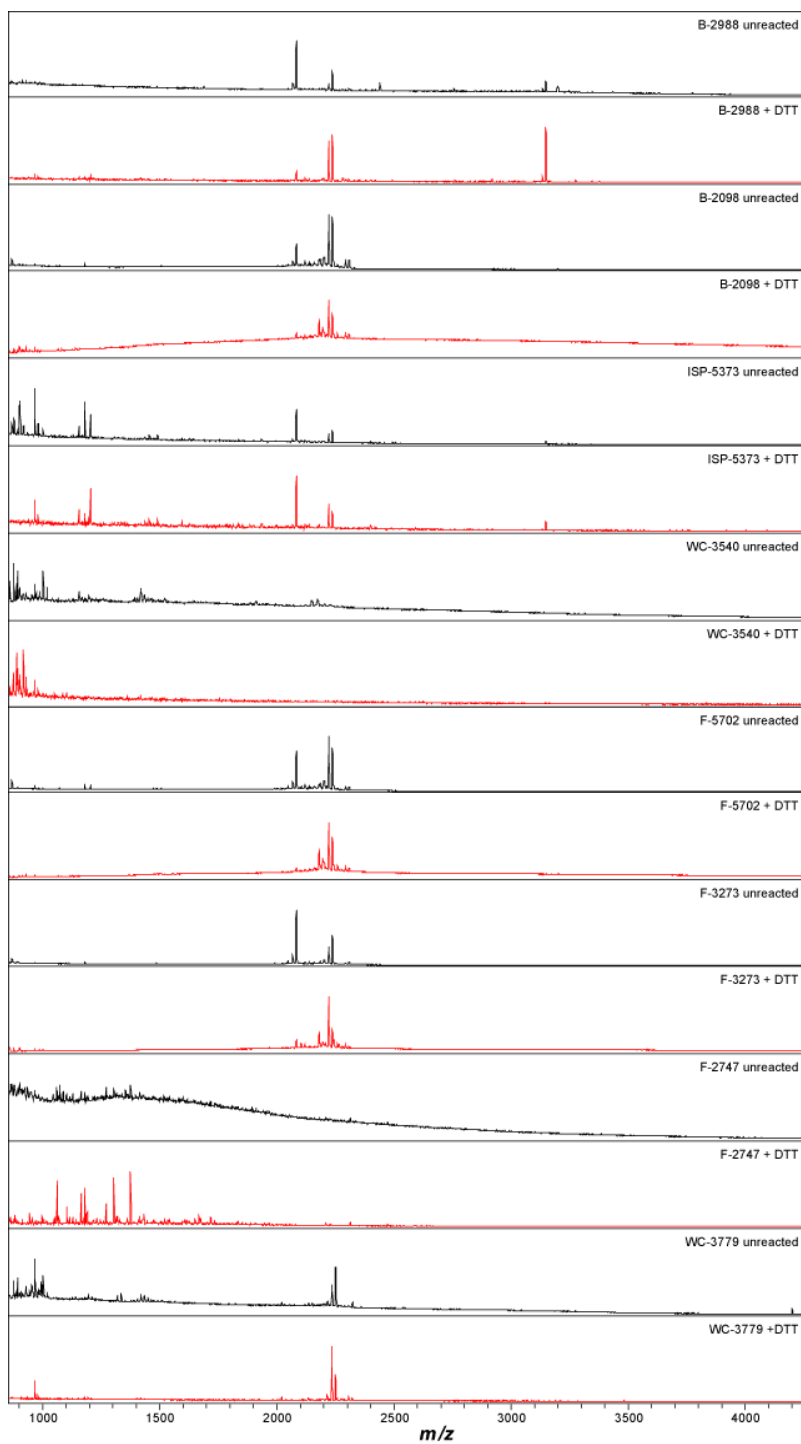


Figure 4.7 (continued) | Mass spectra of strains screened by the DTT labeling method. Mass spectrometry data (m/z 900 – 4200 Da) is shown for all strains screened except *Streptomyces griseus* subsp. *griseus* and WC-3908 (shown as figures 4 and 5 in the main text, respectively). The mass spectra of the unreacted organic cell-surface extracts are shown in black with the corresponding DTT-reacted extracts in red. Each spectrum is labeled according to the strain designation (NRRL identifier) and whether or not DTT/DIPEA was added. NRRL, Northern Regional Research Laboratory collection, which is curated by the Agricultural Research Service under the supervision of the U.S. Department of Agriculture (USDA/ARS).

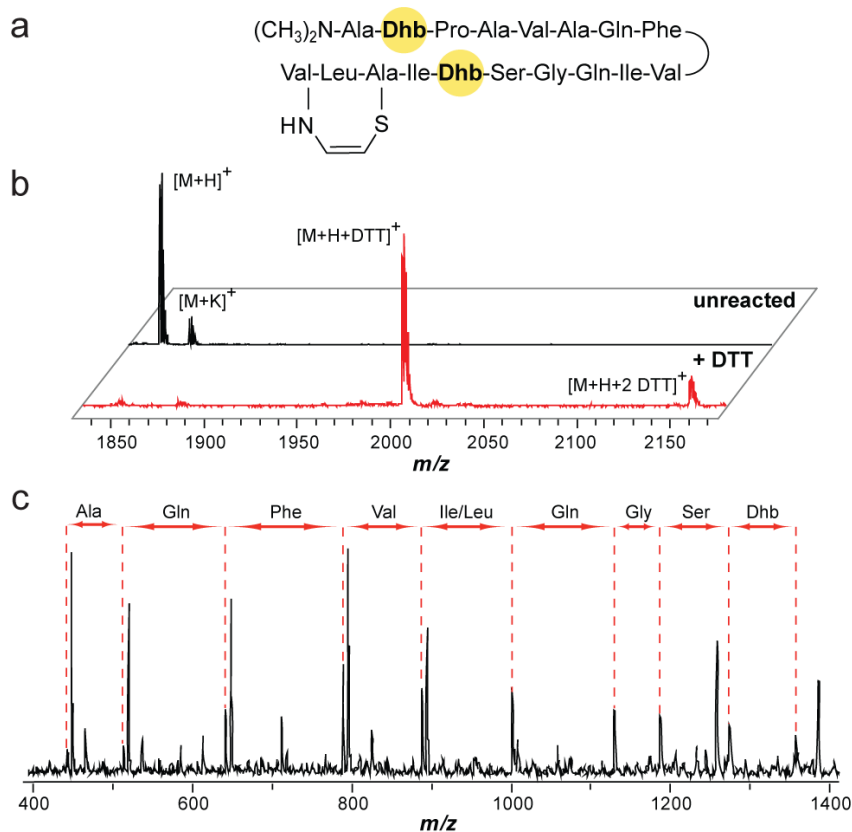


Figure 4.8 | Grisemycin DTT-labeling and dereplication. (A) Structure of grisemycin. Dhb, Dehydrobutyrine. (B) MALDI-TOF MS analysis of unreacted grisemycin (black spectrum, top) and DTT-labeled grisemycin (red spectrum, bottom) from an organic, cell-surface extract showing 1-2 DTT adducts. (C) MS/MS analysis of grisemycin with the discerned sequence tag listed above the spectrum.

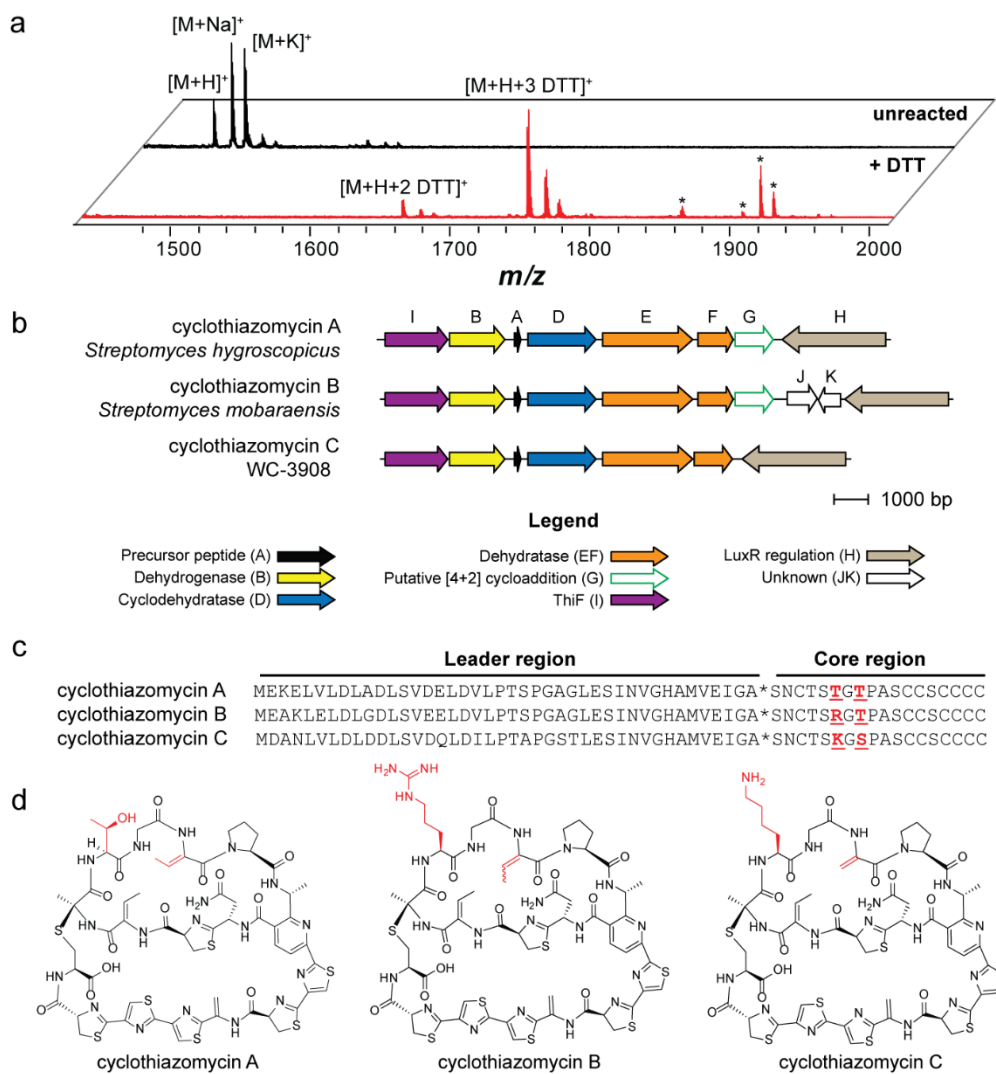


Figure 4.9 | Identification, genetics, and structure of cyclothiazomycin C. (A) MALDI-TOF MS analysis showing spectra of unreacted (black spectrum, top) and DTT-labeled (red spectrum, bottom) extracts of WC-3908, the producer of cyclothiazomycin C. *, peaks do not correspond to DTT-labeled cyclothiazomycin C. (B) Conserved open-reading frames from each of the three cyclothiazomycin gene clusters (precise cluster boundaries are not yet established). Genes are color-coded with proposed functions given in the legend. The strain used for the comparison of cyclothiazomycin A is *Streptomyces hygroscopicus* subsp. *jinggangensis* 5008 and cyclothiazomycin B is *Streptomyces mobaraensis*. (C) Precursor peptide sequences of cyclothiazomycins A, B, and C. Highlighted in red are residues that differ in the core region of the peptide. The asterisk denotes the leader peptide cleavage site. (D) Structures of cyclothiazomycins A, B, and C.

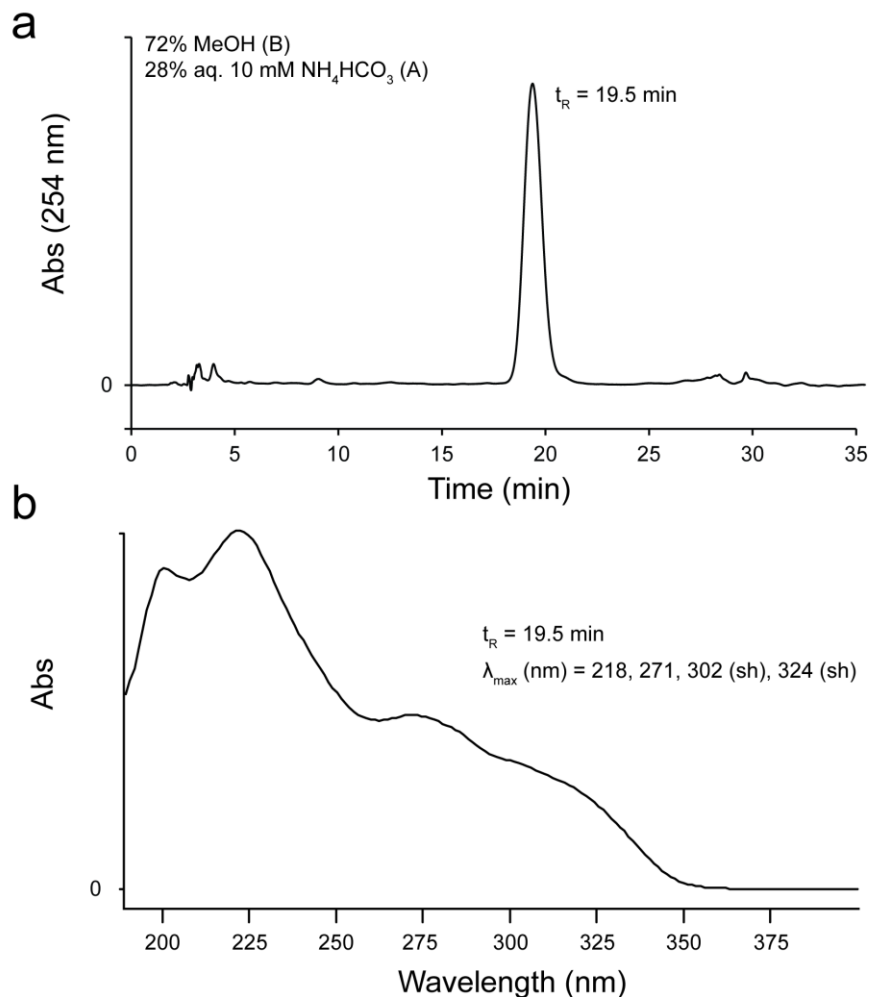


Figure 4.10 | HPLC trace and UV spectrum of cyclothiazomycin C. (A) A sample (spatula tip) of purified cyclothiazomycin C was dissolved in 50% MeOH (B)/aq. 10 mM NH₄HCO₃ (A) (100 μ L). An aliquot (20 μ L) was analyzed by HPLC (isocratic 72% B for 35 min). Photodiode array (PDA) detection was used to monitor absorbance (abs) from 190-400 nm. A blank injection was also run and subtracted from the cyclothiazomycin C chromatogram; the resulting spectrum with UV monitoring at 254 nm is shown. (B) Cyclothiazomycin C exhibits UV absorbance consistent with that reported for cyclothiazomycin A and B1/B2 (Aoki, Ohtsuka, Yamada, *et al.* 1991, Hashimoto, *et al.* 2006). A UV spectrum (PDA) from the HPLC trace at 19.5 min is shown (sh, shoulder).

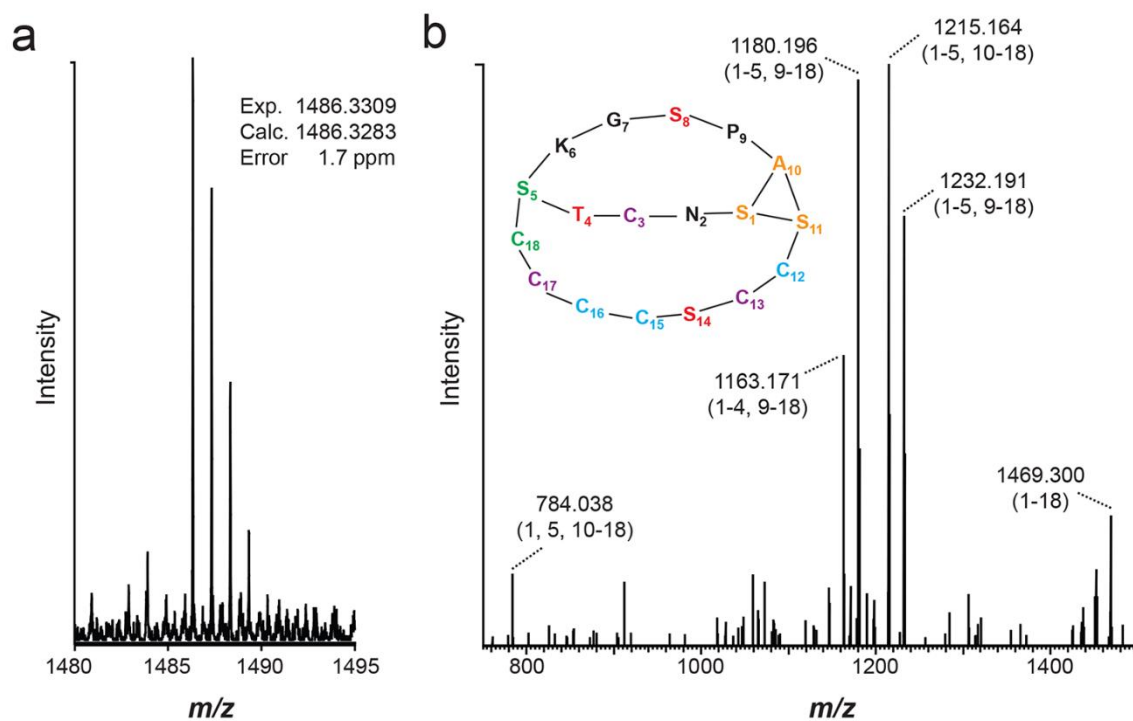


Figure 4.11 | High resolution Fourier transform mass spectrometry (FT-MS) analysis of cyclothiazomycin C. (A) The m/z scan of purified cyclothiazomycin C showed an ion in the 1^+ charge state with an observed isotopic m/z value with < 2 ppm error from the calculated value for cyclothiazomycin C. (B) CID spectrum of m/z 1486. The monoisotopic mass values are given for assigned peak predictions. The number ranges given below the mass values refer to a shorthand notation describing predicted fragments of cyclothiazomycin C. A key for the shorthand notation for the structure of cyclothiazomycin C is given in pictorial format using single letter codes for the amino acids, the residue's N to C position, and lines depicting molecular connectivity within the mature structure. The colors used for the shorthand notation depict the modification present at a particular residue. Purple, thiazoline moieties; green, thioether linkage; cyan, thiazole moieties; red, dehydrated amino acids; orange, pyridine moiety; black, unmodified amino acids.

Figure 4.13

A. Gene Neighborhood Homologs

Gene	Cyclothiazomycin A (<i>Streptomyces hygroscopicus</i> subsp. <i>jinggangensis</i> 5008)	Cyclothiazomycin B (<i>Streptomyces mobaraensis</i>)	Cyclothiazomycin C
	Homolog (Species / Accession number)	Homolog (Species / Accession number)	Homolog (Species / Accession number)
1	Ferredoxin reductase (<i>Streptomyces albulus</i> / WP_020929831.1)	ArsR family transcriptional regulator (<i>Nocardiopsis</i> <i>halophila</i> / WP_017541891.1)	Anti-sigma b factor RsbV (<i>Streptomyces</i> <i>rimosus</i> / WP_003980648.1)
2	Ferredoxin reductase (<i>Streptomyces albulus</i> / WP_016575819.1)	Helix-turn-helix protein (<i>Streptomyces purpureus</i> / WP_019884370.1)	Anti-sigma b factor antagonist RsbV (<i>Streptomyces</i> <i>rimosus</i> / WP_003980649.1)
3	Copper domain protein (<i>Streptomyces</i> sp. FR1 / YP_008995649.1)	Secreted amidase (<i>Streptomyces</i> <i>bingchenggensis</i> BCW-1 / YP_004958967.1)	Anti-anti-sigma regulatory factor (<i>Streptomyces</i> <i>rimosus</i> / WP_003980650.1)
4	Isoprenylcysteine carboxyl methyltransferase (<i>Frankia</i> sp. CN3 / WP_007507509.1)	NAD binding lipoprotein (<i>Streptomyces viridosporus</i> / WP_016827948.1)	Hypothetical (<i>Streptomyces</i> <i>rimosus</i> / WP_003980651.1)
5	Hypothetical protein (<i>Rhodococcus</i> sp. DK17 / WP_016881545.1)	Hypothetical protein (<i>Streptomyces</i> sp. MspMP-M5 / WP_018538980.1)	Acyl-CoA dehydrogenase (<i>Streptomyces</i> sp. FxanaC1 / WP_018088719.1)
6	Hypothetical protein (<i>Rhodococcus</i> sp. DK17 / WP_016881544.1)	Hypothetical protein (<i>Streptomyces</i> sp. 351MFTsu5.1 / WP_020141300.1)	PadR-like family transcriptional regulator (<i>Streptomyces</i> <i>rimosus</i> / WP_003980654.1)

Figure 4.13 (continued)

7	Transposase (<i>Streptomyces coelicolor</i> A3(2) / NP_639869.1)	Putative oxidoreductase (<i>Sinorhizobium meliloti</i> SM11 / YP_005719396.1)	Protein serine/threonine phosphatase (<i>Streptomyces rimosus</i> / WP_003982697.1)
8	Hypothetical protein (<i>Streptomyces albulus</i> / WP_016576055.1)	MFS transporter (<i>Streptomyces roseochromogenes</i> / WP_023546609.1)	Cytochrome P450 FAS1 (<i>Streptomyces rimosus</i> / WP_003980657.1)
9	Hypothetical protein (<i>Streptomyces</i> sp. PsTaAH-124 / WP_018569508.1)	Transglycosylase-associated protein (<i>Streptomyces rimosus</i> / WP_003987332.1)	Aminoglycoside phosphotransferase (<i>Streptomyces cattleya</i> / YP_004920848.1)
10	Hypothetical protein (<i>Streptomyces hygrosopicus</i> subsp. <i>jinggangensis</i> 5008 / YP_006249961.1)	Chitinase (<i>Streptomyces</i> / WP_018524268.1)	ABC transporter (<i>Streptomyces rimosus</i> / WP_003980658.1)

<i>CtmI</i>	Hypothetical protein (<i>Streptomyces mobaraensis</i> / WP_004943462.1)	Putative peptidase containing docking domain (<i>Streptomyces hygrosopicus</i> / ACS50133.1)	Putative peptidase containing docking domain (<i>Streptomyces hygrosopicus</i> / ACS50133.1)
<i>CtmB</i>	NADH oxidase (<i>Streptomyces mobaraensis</i> / WP_004943465.1)	NADH oxidase (<i>Streptomyces hygrosopicus</i> / ACS50132.1)	NADH oxidase (<i>Streptomyces mobaraensis</i> / WP_004943465.1)

Figure 4.13 (continued)

CtmA	Hypothetical protein (<i>Streptomyces mobaraensis</i> / WP_004943468.1)	Hypothetical protein (<i>Streptomyces hygrosopicus</i> / YP_006249964.1)	Hypothetical (<i>Streptomyces</i> <i>mobaraensis</i> / WP_004943468.1)
CtmD	Cyclodehydratase (<i>Streptomyces mobaraensis</i> / WP_004943471.1)	Putative docking protein (<i>Streptomyces hygrosopicus</i> / ACS50130.1)	Hypothetical protein (<i>Streptomyces</i> <i>mobaraensis</i> / WP_004943471.1)
CtmE	Lantibiotic dehydratase (<i>Streptomyces mobaraensis</i> / WP_004943473.1)	Lantibiotic dehydratase (<i>Streptomyces hygrosopicus</i> / ACS50129.1)	Lanthionine biosynthesis protein (<i>Streptomyces</i> <i>mobaraensis</i> / WP_004943473.1)
CtmF	Lantibiotic biosynthesis protein (<i>Streptomyces</i> <i>mobaraensis</i> / WP_004943476.1)	Lantibiotic biosynthesis protein (<i>Streptomyces</i> <i>hygrosopicus</i> / YP_006249967.1)	Lantibiotic biosynthesis protein (<i>Streptomyces</i> <i>mobaraensis</i> / WP_004943476.1)
CtmG	Hypothetical protein (<i>Streptomyces mobaraensis</i> / WP_004943480.1)	Hypothetical protein (<i>Streptomyces hygrosopicus</i> / YP_006249968.1)	Hypothetical protein (<i>Streptomyces</i> <i>mobaraensis</i> / WP_004943480.1)
CtmJ		Hypothetical protein (<i>Streptomyces auratus</i> / WP_006606582.1)	
CtmK		Hypothetical protein (<i>Streptomyces auratus</i> / WP_006606583.1)	

Figure 4.13 (continued)

CtmH	LuxR transcriptional regulator (<i>Streptomyces mobaraensis</i> / WP_004943488.1)	LuxR transcriptional regulator (<i>Streptomyces hygroscopicus</i> / ACS50126.1)	LuxR family transcriptional regulator (<i>Streptomyces hygroscopicus</i> / ACS50126.1)
11	Hypothetical protein (<i>Streptomyces</i> sp. FR1 / YP_008995424.1)	Integral membrane protein (<i>Streptomyces avermitilis</i> MA-4680 / NP_828391.1)	Short chain dehydrogenase (<i>Streptomyces auratus</i> / WP_006608143.1)
12	Ethyl tert-butyl ether degradation (<i>Streptomyces griseoaurantiacus</i> / WP_006139146.1)	Hypothetical protein (<i>Streptomyces griseoflavus</i> / WP_004934717.1)	Hypothetical protein (<i>Streptomyces rimosus</i> / WP_003980661.1)
13	MFS transporter (<i>Streptomyces griseoaurantiacus</i> / WP_006139145.1)	LysR family transcriptional regulator (<i>Streptomyces himastatinicus</i> / WP_009713438.1)	Hypothetical protein (<i>Streptomyces rimosus</i> / WP_003980662.1)
14	DNA binding protein (<i>Streptomyces griseoaurantiacus</i> / WP_006139144.1)	Ornithine carbamoyltransferase (<i>Patulibacter medicamentivorans</i> / WP_007570660.1)	Methylated DNA protein cysteine S-methyltransferase (<i>Streptomyces rimosus</i> / WP_003980663.1)
15	AraC family transcriptional protein (<i>Streptomyces albulus</i> / WP_020930180.1)	Hypothetical protein (<i>Streptomyces mobaraensis</i> / WP_004943504.1)	Ricin superfamily (<i>Streptomyces rimosus</i> / WP_003980666.1)
16	Hypothetical protein (<i>Streptomyces hygroscopicus</i> subsp. <i>jinggangensis</i> 5008 / YP_006249975.1)	Hypothetical protein (<i>Amycolatopsis nigrescens</i> / WP_020673411.1)	Oxidoreductase domain protein (<i>Streptomyces rimosus</i> / WP_003980667.1)

Figure 4.13 (continued)

17	Transposase (<i>Streptomyces lividans</i> / WP_003974545.1)	TetR family transcriptional regulator (<i>Streptomyces clavuligerus</i> / WP_003956021.1)	ArsR family transcriptional regulator (<i>Streptomyces rimosus</i> / WP_003980668.1)
18	Transposase (<i>Streptomyces</i> sp. Mg1 / WP_008736550.1)	Hypothetical protein (<i>Streptomyces sulphureus</i> / WP_019547651.1)	30S ribosomal protein (<i>Thauera terpenica</i> / WP_021250332.1)
19	Transposase (<i>Streptomyces violaceusniger</i> Tu 4113/ YP_004800126.1)	Hypothetical protein (<i>Streptomyces aurantiacus</i> / WP_016638852.1)	Ribosome-inactivating protein (<i>Streptomyces scabiei</i> 87.22 / YP_003494282.1)
20	Oxidoreductase (<i>Streptomyces viridochromogenes</i> / WP_003995212.1)	Hypothetical protein (<i>Streptomyces thermolilacinus</i> / WP_023588110.1)	Ricin superfamily (<i>Streptomyces</i> sp. Mg1 / WP_008735505.1)

Figure 4.13 (continued)**B. Cyclothiazomycin C protein identities**

Protein (cyclothiazomycin C)	Top BLAST hit / accession number	% Identity	Residues (aligned/total)
CtmI (WC3908_03952)	<i>Streptomyces hygrosopicus</i> / ACS50133.1	60	379/634
CtmB (WC3908_03953)	<i>Streptomyces mobaraensis</i> / WP_004943465.1	61	324/531
CtmA (WC3908_03954)	<i>Streptomyces mobaraensis</i> / WP_004943468.1	78	43/55
CtmD (WC3908_03955)	<i>Streptomyces mobaraensis</i> / WP_004943471.1	72	455/636
CtmE (WC3908_03956)	<i>Streptomyces mobaraensis</i> / WP_004943473.1	58	500/858
CtmF (WC3908_03957)	<i>Streptomyces mobaraensis</i> / WP_004943476.1	57	192/334
CtmG (WC3908_00024)	<i>Streptomyces mobaraensis</i> / WP_004943480.1	68	265/397
CtmH (WC3908_03958)	<i>Streptomyces hygrosopicus</i> / ACS50133.1	46	429/924

Figure 4.13 (continued) | Gene similarities for the cyclothiazomycin biosynthetic gene clusters. (A) Genes surrounding the conserved portion of the cyclothiazomycin biosynthetic gene clusters were used as query sequences to identify homologs via BLAST searching. Genes 1-10 represent the genes upstream of the conserved cluster with 1 being the farthest from *ctmI*. *CtmI* – *H* are the conserved genes in the clusters (Figure 3b, NCBI accession number KJ651958) and are highlighted in gray. Red denotes *ctmG* from the cyclothiazomycin C producer that is conserved, but not present in the gene cluster, but rather elsewhere in the genome. Genes 11-20 lie downstream of the conserved region. (B) BLAST results using the conserved genes from the cyclothiazomycin C gene cluster as query sequences. The best match returned by BLAST and the percent identities are given.

Table 4.1 | Antimicrobial activity of cyclothiazomycin B and C toward a panel of diverse bacteria and fungi.

Species ^a	MIC ^b , cyclothiazomycin B	MIC ^b , cyclothiazomycin C
<i>Bacillus anthracis</i>	1	1
<i>Bacillus subtilis</i>	2	4
<i>Enterococcus faecalis</i>	32	32-64
<i>Listeria monocytogenes</i>	8	16
<i>Staphylococcus aureus</i>	4	16
<i>Escherichia coli</i>	64	>64
<i>Neisseria sicca</i>	>64	>64
<i>Pseudomonas putida</i>	>64	>64
<i>Aspergillus niger</i>	>64	>64
<i>Fusarium virguliforme</i>	64	>64
<i>Saccharomyces cerevisiae</i>	64	>64
<i>Talaromyces stipitatus</i>	64	>64

^aThe top five species are Gram positive bacteria from the *Firmicutes* phylum. The next three species are Gram negative bacteria from the *Proteobacteria* phylum. The lowest 4 species are fungi from the *Ascomycota* phylum. ^bAll minimum inhibitory concentrations (MIC) were determined by the microbroth dilution method and are presented in µg/mL.

4.12 References

1. Aoki M, Ohtsuka T, Itezono Y, Yokose K, Furihata K, and Seto H. (1991) Structure of cyclothiazomycin, a unique polythiazole-containing peptide with renin inhibitory activity. Part 1. Chemistry and partial structures of cyclothiazomycin. *Tetrahedron letters* 32: 217-220.
2. Aoki M, Ohtsuka T, Yamada M, Ohba Y, Yoshizaki H, Yasuno H, Sano T, Watanabe J, Yokose K, and Seto H. (1991) Cyclothiazomycin, a novel polythiazole-containing peptide with renin inhibitory activity. Taxonomy, fermentation, isolation and physico-chemical characterization. *Journal of antibiotics (Tokyo)* 44: 582-588.
3. Arnison PG, Bibb MJ, Bierbaum G, Bowers AA, Bugni TS, Bulaj G, Camarero JA, Campopiano DJ, Challis GL, Clardy J, Cotter PD, Craik DJ, Dawson M, Dittmann E, Donadio S, Dorrestein PC, Entian KD, Fischbach MA, Garavelli JS, Goransson U, Gruber CW, Haft DH, Hemscheidt TK, Hertweck C, Hill C, Horswill AR, Jaspars M, Kelly WL, Klinman JP, Kuipers OP, Link AJ, Liu W, Marahiel MA, Mitchell DA, Moll GN, Moore BS, Muller R, Nair SK, Nes IF, Norris GE, Olivera BM, Onaka H, Patchett ML, Piel J, Reaney MJ, Rebuffat S, Ross RP, Sahl HG, Schmidt EW, Selsted ME, Severinov K, Shen B, Sivonen K, Smith L, Stein T, Sussmuth RD, Tagg JR, Tang GL, Truman AW, Vederas JC, Walsh CT, Walton JD, Wenzel SC, Willey JM, and van der Donk WA. (2013) Ribosomally synthesized and post-translationally modified peptide natural products: overview and recommendations for a universal nomenclature. *Natural product reports* 30: 108-160.
4. Bentley SD, Chater KF, Cerdeno-Tarraga AM, Challis GL, Thomson NR, James KD, Harris DE, Quail MA, Kieser H, Harper D, Bateman A, Brown S, Chandra G, Chen CW, Collins M, Cronin A, Fraser A, Goble A, Hidalgo J, Hornsby T, Howarth S, Huang CH, Kieser T, Larke L, Murphy L, Oliver K, O'Neil S, Rabinowitsch E, Rajandream MA, Rutherford K, Rutter S, Seeger K, Saunders D, Sharp S, Squares R, Squares S, Taylor K, Warren T, Wietzorrek A, Woodward J, Barrell BG, Parkhill J, and Hopwood DA. (2002) Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2). *Nature* 417: 141-147.
5. Bonauer C, Walenzyk T, and König B. (2006) α,β -Dehydroamino Acids. *Synthesis* 2006: 1-20.
6. Bowers AA, Acker MG, Koglin A, and Walsh CT. (2010) Manipulation of thiocillin variants by prepeptide gene replacement: structure, conformation, and activity of heterocycle substitution mutants. *Journal of the American Chemical Society* 132: 7519-7527.
7. Challis GL. (2008) Genome mining for novel natural product discovery. *Journal of medicinal chemistry* 51: 2618-2628.
8. Claesen J, and Bibb M. (2010) Genome mining and genetic analysis of cypemycin biosynthesis reveal an unusual class of posttranslationally modified peptides. *Proceedings of the National Academy of Sciences of the United States of America* 107: 16297-16302.

9. Claesen J, and Bibb MJ. (2011) Biosynthesis and regulation of grisemycin, a new member of the linaridin family of ribosomally synthesized peptides produced by *Streptomyces griseus* IFO 13350. *Journal of bacteriology* 193: 2510-2516.
10. Cox CL, Tietz JI, Sokolowski K, Melby JO, Doroghazi JR, and Mitchell DA. (2014) Nucleophilic 1,4-additions for natural product discovery. *ACS chemical biology* 9: 2014-2022.
11. Deane CD, and Mitchell DA. (2013) Lessons learned from the transformation of natural product discovery to a genome-driven endeavor. *Journal of industrial microbiology & biotechnology* 41(2): 315-331.
12. Donovan R, Pagano JF, Stout HA, and Weinstein MJ. (1955) Thiostrepton, a new antibiotic. I. In vitro studies. *Antibiotic Annual* 3: 554-559.
13. Doroghazi JR, and Metcalf WW. (2013) Comparative genomics of actinomycetes with a focus on natural product biosynthetic genes. *BMC genomics* 14: 611.
14. Dunbar KL, Melby JO, and Mitchell DA. (2012) YcaO domains use ATP to activate amide backbones during peptide cyclodehydrations. *Nature chemical biology* 8: 569-575.
15. Dunbar KL, and Mitchell DA. (2013) Insights into the mechanism of peptide cyclodehydrations achieved through the chemoenzymatic generation of amide derivatives. *Journal of the American Chemical Society* 135: 8692-8701.
16. Fischbach MA, and Walsh CT. (2009) Antibiotics for emerging pathogens. *Science* 325: 1089-1093.
17. Gao J, Ju K-S, Yu X, Velásquez JE, Mukherjee S, Lee J, Zhao C, Evans BS, Doroghazi JR, Metcalf WW, and van der Donk WA. (2013) Use of a Phosphonate Methyltransferase in the Identification of the Fosfazinomycin Biosynthetic Gene Cluster. *Angewandte chemie, international edition*.
18. Garg N, Oman TJ, Andrew Wang TS, De Gonzalo CV, Walker S, and van der Donk WA. (2013) Mode of action and structure-activity relationship studies of geobacillin I. *Journal of antibiotics (Tokyo)*.
19. Garg N, Tang W, Goto Y, Nair SK, and van der Donk WA. (2012) Lantibiotics from *Geobacillus thermodenitrificans*. *Proceedings of the National Academy of Sciences of the United States of America* 109: 5241-5246.
20. Gersch M, Kreuzer J, and Sieber SA. (2012) Electrophilic natural products and their biological targets. *Natural product reports* 29: 659-682.

21. Hashimoto M, Murakami T, Funahashi K, Tokunaga T, Nihei K, Okuno T, Kimura T, Naoki H, and Himeno H. (2006) An RNA polymerase inhibitor, cyclothiazomycin B1, and its isomer. *Bioorganic medical chemistry* 14: 8259-8270.
22. Hensens OD, Albers-Schonberg G, and Anderson BF. (1983) The solution conformation of the peptide antibiotic thiostrepton: a ¹H NMR study. *Journal of antibiotics (Tokyo)* 36: 799-813.
23. Just-Baringo X, Albericio F, and Alvarez M. (2014) Thiopeptide antibiotics: retrospective and recent advances. *Marine drugs* 12: 317-351.
24. Kelly WL, Pan L, and Li C. (2009) Thiostrepton biosynthesis: prototype for a new family of bacteriocins. *Journal of the American Chemical Society* 131: 4327-4334.
25. Komiyama K, Otaguro K, Segawa T, Shiomi K, Yang H, Takahashi Y, Hayashi M, Otani T, and Omura S. (1993) A new antibiotic, cypemycin. Taxonomy, fermentation, isolation and biological characteristics. *Journal of antibiotics (Tokyo)* 46: 1666-1671.
26. Lewis K. (2013) Platforms for antibiotic discovery. *Nature reviews Drug discovery* 12: 371-387.
27. Li J, Girard G, Florea BI, Geurink PP, Li N, van der Marel GA, Overhand M, Overkleeft HS, and van Wezel GP. (2012) Identification and isolation of lantibiotics from culture: a bioorthogonal chemistry approach. *Organic & Biomolecular Chemistry* 10: 8677-8683.
28. Li J, Qu X, He X, Duan L, Wu G, Bi D, Deng Z, Liu W, and Ou HY. (2012) ThioFinder: a web-based tool for the identification of thiopeptide gene clusters in DNA sequences. *PloS one* 7: e45878.
29. Melby JO, Nard NJ, and Mitchell DA. (2011) Thiazole/oxazole-modified microcins: complex natural products from ribosomal templates. *Current opinion in chemical biology* 15: 369-378.
30. Mizuhara N, Kuroda M, Ogita A, Tanaka T, Usuki Y, and Fujita K. (2011) Antifungal thiopeptide cyclothiazomycin B1 exhibits growth inhibition accompanying morphological changes via binding to fungal cell wall chitin. *Bioorganic medical chemistry* 19: 5300-5310.
31. Myers CL, Hang PC, Ng G, Yuen J, and Honek JF. (2010) Semi-synthetic analogues of thiostrepton delimit the critical nature of tail region modifications in the control of protein biosynthesis and antibacterial activity. *Bioorganic medical chemistry* 18: 4231-4237.
32. Newman DJ, and Cragg GM. (2012) Natural products as sources of new drugs over the 30 years from 1981 to 2010. *Journal of natural products* 75: 311-335.

33. Nguyen DD, Wu CH, Moree WJ, Lamsa A, Medema MH, Zhao XL, Gavilan RG, Aparicio M, Atencio L, Jackson C, Ballesteros J, Sanchez J, Watrous JD, Phelan VV, van de Wiel C, Kersten RD, Mehnaz S, De Mot R, Shank EA, Charusanti P, Nagarajan H, Duggan BM, Moore BS, Bandeira N, Palsson BO, Pogliano K, Gutierrez M, and Dorrestein PC. (2013) MS/MS networking guided analysis of molecule and gene cluster families. *Proceedings of the National Academy of Sciences of the United States of America* 110: E2611-E2620.
34. Odendaal AY, Trader DJ, and Carlson EE. (2011) Chemoselective enrichment for natural products discovery. *Chemical science* 2: 760-764.
35. Ranu BC, and Mandal T. (2007) Water-promoted highly selective anti-Markovnikov addition of thiols to unactivated alkenes. *Synlett* 925-928.
36. Schoof S, Baumann S, Ellinger B, and Arndt H-D. (2009) A Fluorescent Probe for the 70 S-Ribosomal GTPase-Associated Center. *ChemBioChem* 10: 242-245.
37. Schoof S, Pradel G, Aminake MN, Ellinger B, Baumann S, Potowski M, Najajreh Y, Kirschner M, and Arndt H-D. (2010) Antiplasmodial Thiostrepton Derivatives: Proteasome Inhibitors with a Dual Mode of Action. *Angewandte chemie, International edition* 49: 3317-3321.
38. Tseng H-C, Ovaas H, Wei NJC, Ploegh H, and Tsai L-H. (2005) Phosphoproteomic Analysis with a Solid-Phase Capture-Release-Tag Approach. *Chemical biology* 12: 769-777.
39. Velásquez JE, and van der Donk WA. (2011) Genome mining for ribosomally synthesized natural products. *Current opinion chemical biology* 15: 11-21.
40. Wang J, Yu Y, Tang K, Liu W, He X, Huang X, and Deng Z. (2010) Identification and analysis of the biosynthetic gene cluster encoding the thiopeptide antibiotic cyclothiazomycin in *Streptomyces hygroscopicus* 10-22. *Appl Environmental microbiology* 76: 2335-2344.
41. Watve MG, Tickoo R, Jog MM, and Bhole BD. (2001) How many antibiotics are produced by the genus *Streptomyces*? *Arch Microbiology* 176: 386-390.
42. Wells L, Vosseller K, Cole RN, Cronshaw JM, Matunis MJ, and Hart GW. (2002) Mapping Sites of O-GlcNAc Modification Using Affinity Tags for Serine and Threonine Post-translational Modifications. *Molecular cellular Proteomics* 1: 791-804.
43. Yu Y, Zhang Q, and van der Donk WA. (2013) Insights into the evolution of lanthipeptide biosynthesis. *Protein science : a publication of the Protein Society* 22: 1478-1489.

44. Zhang Q, Ortega M, Shi Y, Wang H, Melby JO, Tang W, Mitchell DA, and van der Donk WA. (2014) Structural investigation of ribosomally synthesized natural products by hypothetical structure enumeration and evaluation using tandem MS. *Proceedings of the National Academy of Sciences of the United States of America* 111: 12031-12036.

45. Zhang Q, Yu Y, Velasquez JE, and van der Donk WA. (2012) Evolution of lanthipeptide synthetases. *Proceedings of the National Academy of Sciences of the United States of America* 109: 18361-18366.

APPENDIX A: FURTHER PUBLICATIONS WITH MINOR CONTRIBUTIONS

A.1 Discovery of a New ATP-Binding Motif Involved in Peptidic Azoline Biosynthesis

This chapter was reprinted with permission from Dunbar, Chekan, Cox, Burkhart, Nair, and Mitchell (Dunbar, *et al.* 2014).

I aided in the collection, interpretation and figure generation for the bioinformatics analysis of the YcaO superfamily. This included supplemental figures 5 and 6.

Discovery of a new ATP-binding motif involved in peptidic azoline biosynthesis

Kyle L Dunbar^{1,2,6}, Jonathan R Chekan^{3,6}, Courtney L Cox^{2,4}, Brandon J Burkhardt^{1,2}, Satish K Nair^{2,3,5*} & Douglas A Mitchell^{1,2,4*}

Despite intensive research, the cyclodehydratase responsible for azoline biogenesis in thiazole/oxazole-modified microcin (TOMM) natural products remains enigmatic. The collaboration of two proteins, C and D, is required for cyclodehydration. The C protein is homologous to E1 ubiquitin-activating enzymes, whereas the D protein is within the YcaO superfamily. Recent studies have demonstrated that TOMM YcaOs phosphorylate amide carbonyl oxygens to facilitate azoline formation. Here we report the X-ray crystal structure of an uncharacterized YcaO from *Escherichia coli* (Ec-YcaO). Ec-YcaO harbors an unprecedented fold and ATP-binding motif. This motif is conserved among TOMM YcaOs and is required for cyclodehydration. Furthermore, we demonstrate that the C protein regulates substrate binding and catalysis and that the proline-rich C terminus of the D protein is involved in C protein recognition and catalysis. This study identifies the YcaO active site and paves the way for the characterization of the numerous YcaO domains not associated with TOMM biosynthesis.

The YcaO family of proteins currently comprises nearly 5,000 members distributed across the bacterial and archaeal domains. Disparate functions have been ascribed to members of this family, which is sometimes referred to as DUF181 (DUF, domain of unknown function). In *E. coli*, the deletion or overexpression of the eponymous YcaO protein (Ec-YcaO; Fig. 1a) suggested that it potentiates the methylthiolation of ribosomal protein S12 and influenced biofilm formation, respectively^{1,2}. However, a molecular explanation for these observations is currently unavailable. Another YcaO-associated activity is the ATP-dependent cyclodehydration of serine, threonine and cysteine residues to azoline heterocycles, which is the defining modification of TOMM natural products (Fig. 1a,b)³. TOMMs display diverse structures and activities^{3,4}, with some implicated in bacterial pathogenesis⁵, making the ~1,000 bioinformatically identifiable TOMM YcaO proteins noteworthy members of the larger superfamily.

Although the TOMM YcaO domain was first implicated in cyclodehydration reactions in the mid-1990s⁶, its exact role remains unclear^{7,8}. The function of the TOMM YcaO (D protein) is intimately linked to members of the E1 ubiquitin-activating enzyme family (C protein) found in canonical TOMM biosynthetic clusters^{6,9,10}. Underscoring this linked function, roughly half of known TOMM clusters express C and D as a single polypeptide^{9,10}. Studies on both fused and unfused cyclodehydratases have demonstrated that these domains are necessary and sufficient for TOMM azoline formation^{7,11}. Consequently, the C-D complex is referred to as the TOMM cyclodehydratase (or alternatively, heterocyclase^{8,10}). As early studies on the cyclodehydratase were unable to observe activity from either protein in isolation^{9,12}, the respective contributions of C and D were inferred by bioinformatics. Given the ATP dependence of the reaction¹³ and the homology of C to members of the E1 ubiquitin-activating superfamily¹⁰, which includes other ATP-using enzymes (for example, MccB, ThiF and MoeB)¹⁴, it was assumed that C was responsible for cyclodehydration, whereas the uncharacterized YcaO (D protein) played a regulatory or scaffolding role^{9,10}.

In 2012, research from our group challenged these assignments with the characterization of the TOMM cyclodehydratase from *Bacillus* sp. Al Hakam (Balh; Fig. 1a)⁷. This YcaO protein (BalhD) displayed ATP-dependent cyclodehydratase activity in the absence of the cognate C protein (BalhC); however, BalhC potentiated cyclodehydration by nearly 1,000-fold. Considering that the C protein had been implicated in precursor peptide recognition in streptolysin S biosynthesis¹⁵ and that BalhC dictated the regio- and chemoselectivity of the Balh cyclodehydratase¹⁶, we hypothesized that the YcaO contained the active site residues, whereas the C protein was responsible for binding the peptide substrate. Although YcaO proteins lack recognizable ATP-binding motifs, the presence of one or more YcaOs in the bottromycin and trifolitinol biosynthetic clusters, which lack recognizable C proteins, supports these functional assignments (Supplementary Results, Supplementary Fig. 1)^{17–21}.

Although the above studies assigned a putative activity to TOMM YcaOs, a molecular understanding of cyclodehydratase catalysis remained elusive. Recently, the X-ray crystal structure of a fused cyclodehydratase was reported (TruD; Protein Data Bank (PDB) code 4BS9), providing what is to our knowledge the first structural glimpse of a TOMM cyclodehydratase⁸. The C domain adopted the expected E1 fold, whereas the YcaO fold was unique. As the structure lacked both the ATP and peptide substrates, no information regarding substrate engagement and catalysis could be gleaned. However, the lack of structural homology of YcaO to known ATP-binding proteins led to the reassertion that the C domain was responsible for ATP binding and carbonyl activation, whereas the YcaO domain catalyzed the requisite nucleophilic attack⁸.

Here, we report the structure of a non-TOMM YcaO from *E. coli* in various nucleotide-bound and nucleotide-free forms and demonstrate that the most conserved residues in YcaOs comprise a previously uncharacterized ATP-binding motif. We show that these ATP-binding residues are critical for catalysis in TOMM YcaOs using BalhD as a model cyclodehydratase. Further, we identify the active site of TOMM cyclodehydratases and demonstrate that the



¹Department of Chemistry, University of Illinois at Urbana-Champaign, Urbana, Illinois, USA. ²Institute for Genomic Biology, University of Illinois at Urbana-Champaign, Urbana, Illinois, USA. ³Department of Biochemistry, University of Illinois at Urbana-Champaign, Urbana, Illinois, USA. ⁴Department of Microbiology, University of Illinois at Urbana-Champaign, Urbana, Illinois, USA. ⁵Center for Biophysics and Computational Biology, University of Illinois at Urbana-Champaign, Urbana, Illinois, USA. ⁶These authors contributed equally to this work. *e-mail: douglasrm@illinois.edu or snair@uiuc.edu

conserved, proline-rich C termini are involved in active site organization and C protein binding. Our results strongly support a model where ATP use is a universal feature of YcaOs (TOMM and non-TOMM) and where TOMM C proteins recognize the peptide substrate and potentiate the activity of the cognate YcaO.

RESULTS

E. coli YcaO hydrolyzes ATP to AMP and PP_i

With the ATP-dependent cyclodehydratase activity of BalhD previously established⁷, we attempted to locate the ATP-binding site in BalhD by 8-azido-ATP cross-linking; however, these experiments were unsuccessful. Furthermore, the TruD crystal structure did not reveal an obvious ATP-binding site⁸, and BalhD was refractory to numerous crystallization attempts. We reasoned that because TOMM YcaOs evolved to interact with their cognate C proteins, working with a non-TOMM (with no C protein partner) might alleviate the previously encountered challenges. The local genomic environment of Ec-YcaO does not contain an E1 homolog (i.e., TOMM C protein; Fig. 1a), and Ec-YcaO is not known to interact with an E1 homolog, making it an attractive candidate for structural and biochemical characterization. Ec-YcaO was cloned into a tobacco etch virus (TEV) protease-cleavable maltose-binding protein (MBP)-fusion vector and expressed in *E. coli* (Supplementary Fig. 2). Although the function of Ec-YcaO was unknown, characterized cyclodehydratases hydrolyze ATP in the absence of peptide substrates⁷. Consequently, we measured the ATPase activity of Ec-YcaO using an established purine nucleoside phosphorylase assay²². This assay revealed that Ec-YcaO indeed hydrolyzed ATP, preferentially generating AMP and PP_i (Supplementary Fig. 3). Although ATP hydrolysis was slow, perhaps because the native substrate was not present, Ec-YcaO displayed a K_M for ATP of ~80 μM (Supplementary Fig. 3), comparable to that of several characterized cyclodehydratases^{21,13}.

Crystallization of Ec-YcaO

The structure of nucleotide-free Ec-YcaO (containing a mercurial salt for phasing) was determined to a Bragg limit of 2.63 Å and revealed a circularly symmetric homodimer in the asymmetric unit (Supplementary Table 2). The overall structure consists of an N-terminal YcaO domain of ~400 residues and a 150-residue C-terminal domain resembling a tetratricopeptide repeat that mediates dimerization. A structure-based comparison against the PDB revealed similarity solely with TruD, the only other solved YcaO structure (r.m.s. deviation of 3.1 Å over 279 aligned Cα atoms)²³, confirming that YcaOs constitute a new structural fold (Supplementary Fig. 4).

To identify the ATP-binding site, we determined the structure of Ec-YcaO in complex with multiple nucleotides. Co-crystallization of Ec-YcaO with ATP produced an AMP-bound structure (2.25 Å), suggesting that *in situ* hydrolysis had occurred (Fig. 2a). To clarify the residues involved in ATP binding, we also determined the co-structure of Ec-YcaO with α,β-methyleneadenosine 5'-triphosphate (AMPCPP, a nonhydrolyzable ATP analog). These three structures facilitated the characterization of the ATP-binding site in the YcaO superfamily.

Structural characterization of ATP-binding in Ec-YcaO

Analysis of the 2.25-Å resolution AMP-bound and the 3.29-Å resolution AMPCPP-bound co-crystal structures revealed that the adenine ring is recognized by Glu191

and Asn187 via interactions through the N7 nitrogen and between Ser16 and the exocyclic N6 (Fig. 2b). Additionally, Lys9 resides above one face of the adenine ring, whereas Ala70, Ser71 and Gly74 are found within an α-helix that extends below the adenine and ribose rings, forming a hydrophobic surface. The ribose within the AMP- and AMPCPP-bound structures is oriented perpendicular to the adenine ring, with Ser184 and Glu78 coordinating the 2'- and 3'-hydroxyls, respectively (Fig. 2c,d). Although Ser71 coordinates the α-phosphate in both structures, Arg286 coordinates to the α-phosphate in only the AMP-bound form (Fig. 2c,d). To our surprise, two Mg²⁺ ions are found in the nucleotide-binding pocket in both structures. In the AMP structure, Glu199 and Glu78 ligate one Mg²⁺ ion, and Glu290 and Glu75 bind the second (Fig. 2c). The Mg²⁺ ions are coordinated in a similar fashion in the AMPCPP structure with the subtle difference that Glu202, rather than Glu75, coordinated the second Mg²⁺ (Fig. 2d). This slight change in the coordination of the second Mg²⁺ ion positions the metal ions on opposite sides of the β- and γ-phosphates. Furthermore, Arg203 coordinates the γ-phosphate of AMPCPP (Fig. 2d). The interactions between Ec-YcaO and AMPCPP are summarized in Figure 2e.

The ATP-binding site is conserved in TOMM YcaOs

Using the nucleotide-bound structures of Ec-YcaO, we established the conservation of the ATP-binding residues across the superfamily. First, we generated a Cytoscape sequence similarity network²⁴ of all of the YcaO members in InterPro (IPR003776)²⁵. During assembly, redundant sequences were removed, leaving ~2,000 sequences in the network. Although the sequence of the TOMM precursor peptide dictates the structure of the natural product, there is also a strong correlation between TOMM structure and the sequence similarity of the cognate YcaO³. For the network, all of the YcaO sequences were manually annotated on the basis of neighboring genes. YcaOs were categorized as being involved in TOMM biosynthesis if there was a gene encoding a recognizable C protein in the local region (~10 kb on either side of the *ycaO* gene) or if the protein had an experimentally verified link to a known TOMM (for example, bottromycin^{17–20} and trifolitin²¹). The remaining YcaOs were separated into two other categories, non-TOMM YcaOs (for example, Ec-YcaO) and TfuA-associated non-TOMM YcaOs. The latter were found within 10 kb of a gene encoding for the protein TfuA, which is implicated in trifolitin biosynthesis²⁶. Whenever possible, TOMM YcaOs were further subdivided by expected structural class. On the basis of these classifications, we determined that an expectation value of 10⁻⁸⁰ gave an optimal separation of YcaO sequences into isofunctional clusters (Supplementary Fig. 5).

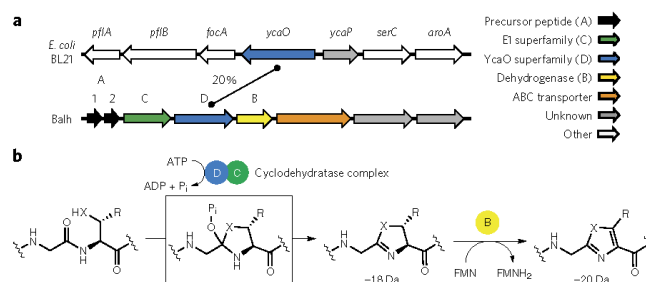


Figure 1 | YcaO gene clusters and characterized roles of YcaO proteins. **(a)** The local genomic environment for Ec-YcaO and BalhD is depicted along with the percentage amino acid identity for the YcaOs. Gene assignments are shown. **(b)** Azole heterocycles in TOMM natural products are installed by the successive action of a cyclodehydratase (C and D proteins) and a flavin mononucleotide (FMN)-dependent dehydrogenase (B protein). The cumulative mass change for each step is shown below the modification. X = S or O; R = H or CH₃.

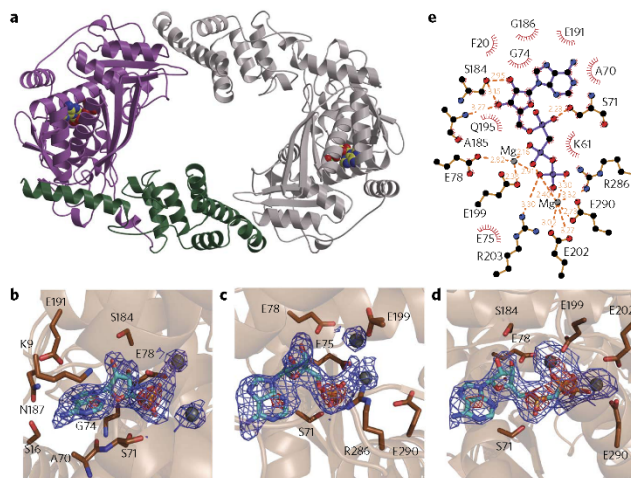


Figure 2 | Structure of Ec-YcaO and ATP-binding residues of Ec-YcaO. (a) Structure of the Ec-YcaO homodimer with one monomer in purple (ATP-binding domain) and green (tetrapeptide repeat domain) and the other monomer in gray. AMP is shown as spheres in both monomers. (b,c) Orthogonal views of the Ec-YcaO AMP-bound structure in the vicinity of the active site, showing residues responsible for adenine and ribose binding (b) and Mg^{2+} and P_i binding (c). AMP and Mg^{2+} ions are cyan and gray, respectively. The superimposed difference Fourier maps are contoured at levels of 2.0σ (blue) and 8.0σ (red). (d) AMP-PP_i (cyan) and Mg^{2+} (gray) bound in the Ec-YcaO active site coordinates with superimposed difference Fourier maps contoured at levels of 2.0σ (blue) and 6.0σ (red). Residues responsible for P_i , ribose and Mg^{2+} binding are indicated. (e) A ligand interaction diagram for the AMP-PP_i-bound structure is shown. Putative hydrogen bonds are shown in orange with distances indicated, and red arcs denote hydrophobic interactions. Due to slight differences in residue orientation in the monomer subunits, only a subset of the interactions is displayed for clarity.

Using the sequence similarity network as a guide, 349 of the ~2,000 members in the nonredundant network were selected from across all of the clusters, and a maximum likelihood tree was generated (Supplementary Fig. 6). Among these 349 were all of the singletons, defined as divergent family members not grouping with any other YcaO at an e -value of 10^{-80} . Using this diversity-maximized tree, a sequence logo for each of the regions involved in ATP binding was generated using WebLogo (Fig. 3a)²⁷. The logos clearly demonstrate that the ATP-binding pocket is highly conserved in the YcaO family across all three groups (i.e., TOMM, non-TOMM and IfuA-associated non-TOMM). The ATP-binding residues were found to be the most conserved feature in the YcaO superfamily (Fig. 3b). This is in stark contrast to TOMM C proteins, which lack the ATP-binding residues conserved in all of the characterized non-TOMM E1 ubiquitin-activating superfamily members¹⁴ (Supplementary Fig. 7). Furthermore, the conservation in the ATP-binding residues is maintained in all of the characterized TOMM YcaOs (Supplementary Figs. 8 and 9), suggesting that the previously reported carbonyl activation mechanism is likely to be a universal biosynthetic feature^{7,16}.

The Ec-YcaO ATP-binding site is vital for BalhD activity

Because the native substrate of Ec-YcaO is unknown, we validated the ATP-binding residues by conducting structure-function studies on BalhD. An alignment of BalhD and Ec-YcaO permitted the mapping of the nucleotide- and Mg^{2+} -binding residues onto BalhD (Supplementary Figs. 4 and 9). Subsequently, an alanine

mutagenesis scan was performed on the polar residues of BalhD predicted to bind ATP. Every mutation was well tolerated in terms of protein yield and stability (Supplementary Fig. 2). The effect on heterocycle formation on BalhA1 (the peptide substrate) by the mutant BalhD proteins, in the presence of BalhC, was monitored in a 16-h endpoint assay (Supplementary Fig. 10). Of the 11 mutated residues in the ATP-binding pocket, four were able to convert BalhA1 to the previously reported penta-azoline species⁷, three showed intermediate levels of processing (two to four heterocycles), and the remaining four generated no heterocyclic products within the limit of detection (Table 1). To quantify the effect of each mutation to BalhC and BalhD activity, the rate of ATP hydrolysis was monitored using the K_M concentration of BalhA1 (15 μ M) and a concentration of ATP that would be saturating for wild-type BalhD (3 mM). Mutants unable to cyclize BalhA1, even after extended reaction times, displayed no detectable ATP hydrolysis over the assay background (Supplementary Fig. 11). Likewise, mutants that installed five azolines on BalhA1 in the endpoint assay had the highest ATP hydrolysis rates. These data are congruent with our earlier work showing that ATP hydrolysis is tightly coupled to heterocycle formation⁷. The YcaO mutations examined here did not appear to disrupt this feature of TOMM cyclodehydration.

Although mutation of the BalhD ATP-binding pocket reduced cyclodehydratase activity, an alternative interpretation of the above data could be that these mutations interfered with the association of BalhC and BalhD. A unique feature of the Balh cyclodehydratase is that BalhD is catalytically active in the absence of BalhC⁷. This permitted the use of BalhD-only activity measurements to determine whether the alanine mutations affected the intrinsic cyclodehydratase activity. Heterocycle formation endpoint assays (16 h) were again conducted, but this time with 50 μ M BalhA1 and 25 μ M BalhD mutant to account for the expected ~1,000-fold drop in catalytic activity in the absence of BalhC⁷. The resultant mass spectra confirmed that the decrease in cyclodehydration arose from a perturbation in BalhD activity (Table 1 and Supplementary Fig. 12).

For all of the BalhD mutants with measureable cyclodehydratase activity, we obtained the Michaelis-Menten kinetic parameters for BalhA1 and ATP (Table 1 and Supplementary Fig. 13). Every mutation negatively affected the observed k_{cat} (k_{obs}), indicating that the selected residues were of catalytic importance. Apart from K281A, and to a lesser extent S72A, all of the ATP-binding site mutations of BalhD substantially increased the K_M for ATP. In contrast, the only mutant in this series to substantially raise the K_M for BalhA1 was R198A (Table 1).

Four BalhD mutants (i.e., BalhD^{E76A}, BalhD^{E79A}, BalhD^{E194A} or BalhD^{E197A}) did not exhibit detectable cyclodehydratase activity. Potential explanations include an inability to bind the substrates (BalhA1 and/or ATP) or hydrolyze ATP or a structural perturbation with these mutants. Previous work demonstrated that BalhC and BalhD hydrolyze ATP slowly in the absence of BalhA1 (ref. 7). In reactions with wild-type BalhD, addition of BalhC potentiates the rate of ATP hydrolysis by 2.5-fold over an additive rate of both proteins; however, when the four inactive mutants were assayed, no

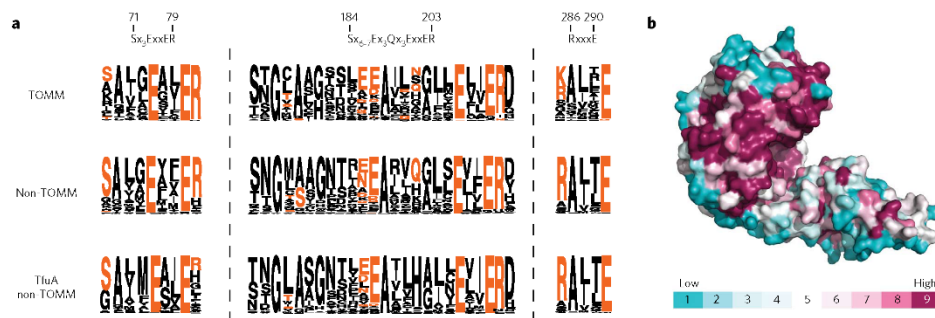


Figure 3 | Conservation of the Ec-YcaO ATP-binding residues in the superfamily. (a) WebLogo frequency plots for the ATP- and Mg²⁺-binding residues of YcaO domains for each subclass (TOMM, non-TOMM and TfuA non-TOMM). Owing to the high level of diversity in the sequences, WebLogos for the N-terminal ATP-binding residues were not generated. The ATP-binding motif identified in Ec-YcaO is shown above each of the ATP-binding regions with the number representing the residue in Ec-YcaO and conserved residues in orange. (b) YcaO superfamily sequence conservation was mapped onto the structure of Ec-YcaO, which highlights strong conservation of the ATP-binding region.

potentiation was observed (Supplementary Fig. 14). This suggested that the lack of BalhD activity was due to a structural perturbation or the inability to bind or hydrolyze ATP. Unfortunately, attempts to directly measure ATP binding or a secondary structure perturbation of BalhD with isothermal titration calorimetry or CD spectroscopy, respectively, were problematic owing to the solubility characteristics of BalhD. However, we reasoned that the latter could be assayed indirectly by monitoring the interaction between BalhC and a mutant BalhD through a competition assay and size-exclusion chromatography. Although all of the BalhD mutants were able to associate with the BalhC (Supplementary Fig. 15), BalhD^{E76A} and BalhD^{E197A} did so with reduced affinity, suggesting that these mutations affected the BalhC-BalhD interaction surface. Conversely, the wild type-like affinity that BalhD^{R79A} and BalhD^{E197A} displayed for BalhC suggested that these mutants were inactive owing to an inability to bind or hydrolyze ATP.

The BalhD C terminus affects BalhC binding and catalysis

In addition to the conserved ATP-binding site, TOMM YcaO proteins have a highly conserved, proline-rich C terminus. In the most pronounced cases, the final five residues of the YcaO are P_xP_xP (Supplementary Fig. 16)⁹. The proline-rich C terminus is not conserved in non-TOMM or TfuA-associated YcaO domains

(Supplementary Fig. 16), implicating the motif in either C protein recognition or cyclodehydratase activity. This hypothesis is supported by the observation that the C terminus of TruD is in close proximity to the YcaO ATP-binding site and is surface-accessible (Supplementary Fig. 16). We first interrogated the importance of this motif by truncating five residues from the BalhD C terminus. This minor perturbation abolished the catalytic activity of BalhC and BalhD (Table 2 and Supplementary Fig. 17). Removing the C-terminal three residues of BalhD produced an identical result, and removal of the C-terminal residue of BalhD (BalhD^{P429*}, where the asterisk represents a stop codon) decreased activity by > 100-fold (Table 2 and Supplementary Fig. 17). Similarly, extending the C terminus by a single amino acid (BalhD P_xP_xP_G), or deleting two amino acids upstream of the P_xP_xP motif (BalhD Δ418-419; Δ2 AA), also resulted in inactive cyclodehydratases (Table 2 and Supplementary Fig. 17).

To establish whether altering the BalhD C terminus affected the interaction with BalhC, we assessed the ability of BalhC to potentiate the background ATPase activity of BalhD mutants lacking detectable activity. Potentiation was not observed in any case (Supplementary Fig. 18), indicating that mutants of the BalhD P_xP_xP motif had lost the ability to bind or hydrolyze ATP or to bind

Table 1 | Mutations to the ATP-binding pocket of BalhD decrease cyclodehydratase activity.

Mutation	Ring formation ^a		BalhA1 kinetics			ATP kinetics		
	CD	D only	k_{obs} (min ⁻¹)	K_M (μM) ^b	k_{obs}/K_M (M ⁻¹ s ⁻¹)	k_{obs} (min ⁻¹)	K_M (μM) ^b	k_{obs}/K_M (M ⁻¹ s ⁻¹)
WT	5; 100%	2-3; 45%	12.9 ± 0.4	16 ± 2	13,000	12.2 ± 0.3	240 ± 20	850
S72A	5; 100%	2; 40%	8.3 ± 0.2	11 ± 1	12,500	8.0 ± 0.4	360 ± 60	370
E76A	0; 0%	0; 0%	ND	ND	ND	ND	ND	ND
E79A	0; 0%	0; 0%	ND	ND	ND	ND	ND	ND
R80A	2-4; 63%	0-1; 10%	0.79 ± 0.05	16 ± 5	823	0.88 ± 0.02	620 ± 50	24
Q186A	5; 100%	0-2; 10%	3.9 ± 0.1	12 ± 1	5,400	7.5 ± 0.3	2,500 ± 200	50
N190A	5; 100%	0-2; 15%	5.1 ± 0.2	27 ± 3	3,150	4.7 ± 0.2	920 ± 110	85
E194A	0; 0%	0; 0%	ND	ND	ND	ND	ND	ND
E197A	0; 0%	0; 0%	ND	ND	ND	ND	ND	ND
R198A	2-4; 40%	0; 0%	3.3 ± 0.2	50 ± 8	1,100	2.4 ± 0.2	760 ± 65	52
K281A	5; 100%	2-3; 45%	7.9 ± 0.2	16 ± 2	8,200	6.1 ± 0.1	230 ± 20	440
E286A	2-4; 64%	0; 0%	0.49 ± 0.03	27 ± 3	272	0.42 ± 0.01	1,200 ± 100	6

^a% processing = $(5P_2 - 4P_3 + 3P_4 + 2P_5 - 1P_6)/G$, where P_i is the percentage of the substrate with i number of azolines. The number of heterocycles formed in the assay is listed in parentheses.

^bApparent K_M . Error on the Michaelis-Menten parameters represents the s.d. from the regression analysis. ND, not determined.

Table 2 | Mutations to the C terminus of BalhD disrupt catalysis.

Mutation	Ring formation ^a		BalhA1 kinetics			ATP kinetics		
	CD	D only	k_{obs} (min ⁻¹)	K_M (μM) ^b	k_{obs}/K_M (M ⁻¹ s ⁻¹)	k_{obs} (min ⁻¹)	K_M (μM) ^b	k_{obs}/K_M (M ⁻¹ s ⁻¹)
WT	5; 100%	1-3; 45%	12.9 ± 0.4	16 ± 2	13,000	12.2 ± 0.3	240 ± 20	850
P425*	0; 0%	0; 0%	ND	ND	ND	ND	ND	ND
P427*	0; 0%	0; 0%	ND	ND	ND	ND	ND	ND
P429*	1-2; 36%	0; 0%	0.19 ± 0.03	80 ± 20	31	0.21 ± 0.05	110 ± 10	32
$\Delta 2$ AA	0; 0%	0; 0%	ND	ND	ND	ND	ND	ND
PxPxPG	0; 0%	0; 0%	ND	ND	ND	ND	ND	ND
H426A	0-2; 25%	0; 0%	ND	ND	ND	ND	ND	ND
P427G	3-5; 80%	0; 0%	0.43 ± 0.04	35 ± 8	204	0.41 ± 0.02	700 ± 100	10
F428A	3-4; 80%	0; 0%	0.70 ± 0.02	22 ± 2	530	0.69 ± 0.03	480 ± 60	24
P429G	4; 80%	0-2; 6%	4.7 ± 0.1	14 ± 1	5,600	5.0 ± 0.1	75 ± 4	1,100
GxGxG	0-3; 45%	0; 0%	ND	ND	ND	ND	ND	ND

^a% processing = $(5P_5 + 4P_4 + 3P_3 + 2P_2 + 1P_1)/5$, where P_x is the percentage of the substrate with x number of azolines. The number of heterocycles formed in the assay is listed in parentheses.

^bApparent K_M . Asterisk indicates stop codon. Error on the Michaelis-Menten parameters represents the s.d. from the regression analysis. ND, not indicated.

BalhC. We next assessed the ability of each PxPxP mutant to bind BalhC. Using a combination of size-exclusion chromatography and a competition assay, all of the truncations to the BalhD PxPxP motif were shown to have decreased affinity to BalhC (Supplementary Fig. 19). Moreover, the order of heterocycle formation was dysregulated in BalhD^{P429*}, reminiscent of wild-type BalhD reactions lacking BalhC (Supplementary Fig. 20)¹⁶.

Intrigued by the loss of activity observed upon extending or truncating the C terminus, we next investigated the importance of the amino acid composition of the BalhD PxPxP motif (PHPPF₁₂₉). As with the truncations, any mutation to the five C-terminal residues of BalhD decreased cyclodehydratase activity (Table 2 and Supplementary Fig. 21). The decrease in activity ranged from ~2.5-fold (BalhD^{P429G}) to 100-fold (BalhD^{H426A}), with severity diminishing the closer the mutation was to the C terminus. This result was consistent with the observation that the C-terminal residues of TruD are located in a channel leading to the active site. As with the PxPxP truncations, every mutant tested, apart from BalhD^{P428A}, displayed a decreased affinity for BalhC (Supplementary Fig. 22). Consistent with this observation, mutation of the terminal amino acid (P429G) resulted in an aberrant order of heterocycle formation (Supplementary Fig. 23). Increasing the flexibility of the PHPPF motif by substituting it with GHGFG, yielded an inactive cyclodehydratase (Table 2 and Supplementary Fig. 21).

Although these results implicated the C terminus of BalhD in BalhC recognition, a decrease in BalhC affinity could not explain the data in its entirety. For example, both BalhD^{P425*} and BalhD^{P429*} displayed reduced interactions with BalhC, but only BalhD^{P425*} was catalytically inactive (Table 2 and Supplementary Fig. 19). Furthermore, BalhD PxPxPG showed a wild-type level of interaction with BalhC despite being catalytically inactive. Given these results, we tested the activity of each mutant in the absence of BalhC. Analogous to the mutations to the ATP-binding pocket, mutations to the PxPxP motif affected the intrinsic activity of BalhD (Supplementary Fig. 24). For all tractable BalhD mutants, BalhA1 and ATP Michaelis-Menten kinetic curves were obtained for the mutant BalhC–BalhD complexes. Although the largest effects were on k_{obs} , the mutations also affected the K_M for ATP and BalhA1 (Table 2).

Unlike the ATP-binding mutants, the changes to K_M for the two substrates were similar in the PxPxP mutants, suggesting that the C terminus is involved in active site organization and catalysis, not substrate binding. Furthermore, the importance of the YcaO C terminus seems to be general for TOMM biosynthesis, given that the cyclodehydratase activity of McbD (microcin B17 YcaO protein) was also abolished when the C terminus was truncated (Supplementary Fig. 25).

BalhC regulates BalhD ATP-binding and catalysis

To further characterize the C terminus of BalhD, we generated a derivative containing a C-terminal His₆ tag (PA₃LEH₆; where P is the last residue of wild-type BalhD). As expected from earlier experiments with BalhD PxPxPG, addition of the longtag abolished heterocycle formation (Supplementary Fig. 26). Although heterocycle formation is stoichiometric with ATP hydrolysis for wild-type cyclodehydratase⁷, the C-terminal His₆-tagged BalhD displayed robust ATPase activity, irrespective of the presence of BalhA1 (Fig. 4a). Moreover, this high level of unproductive ATP hydrolysis was potentiated by the addition of BalhC to the same extent observed with wild-type BalhD (2.5-fold increase; see Supplementary Fig. 14), indicating that the His₆ tag did not interfere with BalhC recognition (Fig. 4a). With a BalhD derivative displaying robust BalhC-independent ATPase activity in hand, we evaluated the role of BalhC on ATP use by BalhD by obtaining ATP Michaelis-Menten kinetics parameters for BalhD–A₃LEH₆ alone and in complex with BalhC (Fig. 4b).

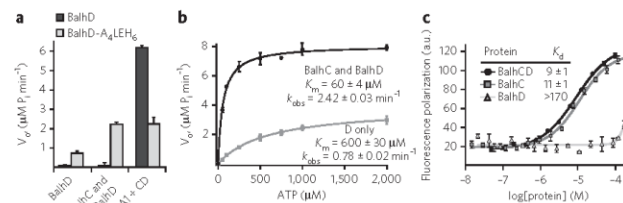


Figure 4 | BalhC modulates ATP binding and hydrolysis by BalhD and is responsible for leader peptide binding. (a) ATP hydrolysis rates measured by the PNP assay. (b) ATP kinetic curves for BalhD–A₃LEH₆ with and without BalhC. Error bars represent the s.d. from the mean ($n = 3$), and error on the Michaelis-Menten parameters represents the s.d. from the regression analysis. (c) A fluorescent polarization curve for fluorescein-labeled BalhA1 leader peptide recognition by BalhC and BalhD is displayed. Error bars represent the s.e.m. of three independent titrations. Errors on the K_D values represent the error from curve fitting. a.u., arbitrary units.

These data indicate that the addition of BalhC modulates BalhD activity by increasing the k_{obs} and decreasing the K_M for ATP.

TOMM C proteins provide leader peptide binding

Precursor peptide recognition in the majority of ribosomally synthesized post-translationally modified peptides occurs in a bipartite fashion. An N-terminal sequence (leader peptide) serves as the recognition sequence by the modification enzymes, and the C-terminal sequence (core peptide) contains the sites of post-translational modification^{4,28}. We previously demonstrated that the regioselectivity and directionality of BalhD azoline formation is dependent on BalhC¹⁶. Thus, we hypothesized that BalhC was responsible for presenting the core peptide to BalhD, most likely by engaging the leader peptide. However, the identification of BalhD mutants with a perturbed K_M for BalhA1 suggested that the YcaO domain might have a role in substrate recognition. To assess the role of the C and D proteins in peptide substrate recognition, a fluorescein-labeled BalhA1 leader peptide was used to monitor binding to BalhC and BalhD by fluorescence polarization. Owing to a very weak, potentially nonspecific, association with BalhD, the K_d toward the BalhA1 leader peptide could not be obtained. In contrast, BalhC displayed a K_d of 11 μ M (Fig. 4c), near the previously measured K_M for BalhA1 of 16 μ M²⁹. Moreover, the addition of BalhD did not significantly alter BalhC's affinity for the BalhA1 leader peptide ($P > 0.05$). Consequently, these data support a model where (i) BalhD does not engage the leader region of BalhA1 and (ii) the elevation in K_M value of select BalhD mutants for BalhA1 is due to a decreased affinity for the core region of BalhA1.

DISCUSSION

We have discovered that Ec-YcaO contains a new ATP-binding fold. Given the steric and electrostatic complimentary requirements for binding ATP, the YcaO strategy is reminiscent of that of other structurally characterized ATP-binding proteins. For example, the Lys- α -helix 'sandwich' involved in adenine recognition is similar to the conserved arginine and glycine motif found in class I amino acyl tRNA synthetases³⁰. Furthermore, select members of the ATP-grasp and PurM families have been shown to bind ATP through the use of multiple divalent cations^{31,32}; however, in these proteins, the Mg²⁺ ions are coordinated to all three phosphates and not just the β - and γ -phosphates. As these similarities in ATP binding occur despite a lack of structural and primary homology between YcaO and all of the other known ATP-binding proteins, this represents an example of convergent evolution in ATP-binding domains.

The ATP-binding residues are the most highly conserved motifs in the YcaO superfamily and, appropriately, represent a prominent signature for the hidden Markov model that bioinformatically defines the YcaO family (IPR003776). Our extensive bioinformatics analysis, X-ray crystallographic data on Ec-YcaO and biochemical characterization of BalhD confirm that ATP use is a conserved feature in the superfamily. In spite of the low level of overall similarity between Ec-YcaO and BalhD, we were able to demonstrate that the YcaO ATP-binding motif was critical for cyclodehydratase activity. Although the mutations affected BalhD activity to differing extents, the impact of mutating a particular residue on Balh cyclodehydratase activity was proportional to the level of conservation within the YcaO family (Fig. 3a).

During our analysis of the YcaO ATP-binding motif, we observed a marked difference between the TOMM and non-TOMM YcaO domains. TOMM YcaOs (D proteins) almost invariably harbor proline-rich C-termini, with PxPxP most often serving as the terminal five residues of the protein. Though the widespread nature of the PxPxP motif had been previously recognized³, before this work, it was unclear whether this motif had any role in TOMM biogenesis. Our data indicate that the C terminus of TOMM YcaOs assists in both C protein recognition and cyclodehydration. It is rare for the

C terminus of an enzyme to be important for catalysis³³⁻³⁷. In fact, the terminal regions are often highly sequence-variable within a protein family. Notably, the C-terminal proline content of a YcaO has powerful predictive value. If present, the YcaO is quite probably involved in TOMM biosynthesis. This tentative assignment can be confidently made even without knowledge of the flanking genes. As such, we hypothesize that a subset of the 249 (~8%) non-TOMM YcaOs that contain a proline-rich C terminus may actually be stand-alone TOMM YcaOs (akin to the bottromycin YcaOs).

Previously, BalhC was shown to potentiate BalhD via an unknown mechanism⁷. The current study indicates that this potentiation occurs via two distinct mechanisms. First, the serendipitous identification of a C-terminal His-tagged construct of BalhD with robust ATP hydrolysis (BalhD-A₁LEH₁) allowed us to show that the presence of BalhC increases k_{obs} and lowers the K_M for ATP. Although our data suggest that C protein potentiation occurs via allosteric activation, follow-up studies will be required to validate this hypothesis. Second, our data demonstrate that BalhC is responsible for binding the leader peptide of BalhA1, thus efficiently bringing the substrate in close proximity to the BalhD active site. This result is in accord with a previous study implicating SagC in leader peptide recognition during streptolysin S biosynthesis¹⁵. In further support of a general role for TOMM C proteins in peptide substrate binding, the 'C portion' (homologous to E1 ubiquitin-activating enzymes) of TruD has an MCB-like N-terminal 'peptide clamp'³⁸, which is responsible for leader peptide binding in microcin C7 biosynthesis³⁸. Combined with the fact that TOMM C proteins lack the ATP-binding site that is conserved in all of the characterized non-TOMM E1 ubiquitin-activating family members, all lines of evidence suggest that TOMM C proteins engage the leader peptide while simultaneously potentiating the carbonyl activation chemistry of their cognate YcaO domain (D protein; Supplementary Fig. 27).

It is not yet clear how stand-alone TOMM YcaO proteins (i.e., for bottromycin and trifolixotoxin production) perform cyclodehydrations in the absence of a C protein. Given the diversity between these stand-alone and canonical (C protein-containing) TOMM YcaOs, we envision that multiple solutions to the substrate recognition problem could exist. For example, it is possible that these biosynthetic pathways use an unidentified companion protein to bind the precursor peptide. Alternatively, these YcaO proteins may have evolved to bind a specific motif within the core peptide and modify the substrate a single time. Of note is the fact that the bottromycin and trifolixotoxin stand-alone YcaO domains are each predicted to install a single heterocycle¹⁷⁻²¹. This is in stark contrast to canonical TOMMs that process a wide array of core peptides, often at numerous locations^{3,29,39}. Such promiscuity is common in other ribosomally synthesized post-translationally modified peptides and most likely accounts for the existence of leader peptides (i.e., the modification enzymes can be specific for motifs within the leader peptide but promiscuous on the core once the enzyme-substrate complex is formed)^{4,28}. Further work, including the reconstitution of a stand-alone YcaO, will be required to test these claims.

The capacity to bind ATP (or possibly other nucleotide triphosphates) seems to be ubiquitous in the YcaO superfamily, but it remains unclear whether the TOMM cyclodehydratase-like direct activation of carbonyls is a universal feature. It is intriguing that YcaOs have recently been implicated in the formation of thioamides⁴⁰ and macroimidine rings¹⁷⁻²⁰, as both of these modifications could conceivably occur through carbonyl activation. In addition to providing major insight into the mechanics of TOMM cyclodehydration, the results presented here provide an initial framework to explore the elusive functions of the 4,000 uncharacterized non-TOMM YcaOs.

Received 14 March 2014; accepted 19 June 2014;
published online 17 August 2014

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. PDB. Coordinates for apo-YcaO, AMP-bound YcaO and AMPCPP-bound YcaO structures were deposited under accession codes 4Q84, 4Q86 and 4Q85, respectively.

References

- Strader, M.B. *et al.* A proteomic and transcriptomic approach reveals new insight into β -methylthiolation of *Escherichia coli* ribosomal protein S12. *Mol. Cell. Proteomics* **10**, M11005199 (2011).
- Tenorio, E. *et al.* Systematic characterization of *Escherichia coli* genes/ORFs affecting biofilm formation. *FEMS Microbiol. Lett.* **225**, 107–114 (2003).
- Melby, J.O., Nard, N.J. & Mitchell, D.A. Thiazole/oxazole-modified microcins: complex natural products from ribosomal templates. *Curr. Opin. Chem. Biol.* **15**, 369–378 (2011).
- Arnison, P.G. *et al.* Ribosomally synthesized and post-translationally modified peptide natural products: overview and recommendations for a universal nomenclature. *Nat. Prod. Rep.* **30**, 108–160 (2013).
- Molloy, E.M., Cotter, P.D., Hill, C., Mitchell, D.A. & Ross, R.P. Streptolysin S-like virulence factors: the continuing saga. *Nat. Rev. Microbiol.* **9**, 670–681 (2011).
- Li, Y.M., Milne, J.C., Madison, L.L., Kolter, R. & Walsh, C.T. From peptide precursors to oxazole and thiazole-containing peptide antibiotics: microcin B17 synthase. *Science* **274**, 1188–1193 (1996).
- Dunbar, K.L., Melby, J.O. & Mitchell, D.A. YcaO domains use ATP to activate amide backbones during peptide cyclodehydrations. *Nat. Chem. Biol.* **8**, 569–575 (2012).
- Koehnke, I. *et al.* The cyanobactin heterocyclase enzyme: a processive adenylase that operates with a defined order of reaction. *Angew. Chem. Int. Ed. Engl.* **52**, 13991–13996 (2013).
- Lee, S.W. *et al.* Discovery of a widely distributed toxin biosynthetic gene cluster. *Proc. Natl. Acad. Sci. USA* **105**, 5879–5884 (2008).
- Schmidt, E.W. *et al.* Patellamide A and C biosynthesis by a microcin-like pathway in *Prochloron didemni*, the cyanobacterial symbiont of *Lissoclitum patella*. *Proc. Natl. Acad. Sci. USA* **102**, 7315–7320 (2005).
- McIntosh, J.A. & Schmidt, E.W. Marine molecular machines: heterocyclization in cyanobactin biosynthesis. *ChemBioChem* **11**, 1413–1421 (2010).
- Milne, J.C. *et al.* Cofactor requirements and reconstitution of microcin B17 synthetase: a multienzyme complex that catalyzes the formation of oxazoles and thiazoles in the antibiotic microcin B17. *Biochemistry* **38**, 4768–4781 (1999).
- Milne, J.C., Eliot, A.C., Kelleher, N.L. & Walsh, C.T. ATP/GTP hydrolysis is required for oxazole and thiazole biosynthesis in the peptide antibiotic microcin B17. *Biochemistry* **37**, 13250–13261 (1998).
- Schulman, B.A. & Harper, J.W. Ubiquitin-like protein activation by E1 enzymes: the apex for downstream signalling pathways. *Nat. Rev. Mol. Cell Biol.* **10**, 319–331 (2009).
- Mitchell, D.A. *et al.* Structural and functional dissection of the heterocyclic peptide cytotoxin streptolysin S. *J. Biol. Chem.* **284**, 13004–13012 (2009).
- Dunbar, K.L. & Mitchell, D.A. Insights into the mechanism of peptide cyclodehydrations achieved through the chemoenzymatic generation of amide derivatives. *J. Am. Chem. Soc.* **135**, 8692–8701 (2013).
- Huo, L., Rachid, S., Stadler, M., Wenzel, S.C. & Müller, R. Synthetic biotechnology to study and engineer ribosomal bottromycin biosynthesis. *Chem. Biol.* **19**, 1278–1287 (2012).
- Hou, Y. *et al.* Structure and biosynthesis of the antibiotic bottromycin D. *Org. Lett.* **14**, 5050–5053 (2012).
- Gomez-Escribano, J.P., Song, L., Bibb, M.J. & Challis, G.L. Posttranslational β -methylation and macrolactamization in the biosynthesis of the bottromycin complex of ribosomal peptide antibiotics. *Chem. Sci.* **3**, 3522–3525 (2012).
- Crone, W.J.K., Leeper, F.J. & Truman, A.W. Identification and characterization of the gene cluster for the anti-MRSA antibiotic bottromycin: expanding the biosynthetic diversity of ribosomal peptides. *Chem. Sci.* **3**, 3516–3521 (2012).
- Breil, B.T., Ladden, P.W. & Triplett, E.W. DNA sequence and mutational analysis of genes involved in the production and resistance of the antibiotic peptide trifolitoxin. *J. Bacteriol.* **175**, 3693–3702 (1993).
- Webb, M.R. A continuous spectrophotometric assay for inorganic phosphate and for measuring phosphate release kinetics in biological systems. *Proc. Natl. Acad. Sci. USA* **89**, 4884–4887 (1992).
- Holm, L. & Rosenstrom, P. Dali server: conservation mapping in 3D. *Nucleic Acids Res.* **38**, W545–W549 (2010).
- Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).
- Hunter, S. *et al.* InterPro in 2011: new developments in the family and domain prediction database. *Nucleic Acids Res.* **40**, D306–D312 (2012).
- Breil, B., Borneman, J. & Triplett, E.W. A newly discovered gene, *ftfA*, involved in the production of the ribosomally synthesized peptide antibiotic trifolitoxin. *J. Bacteriol.* **178**, 4150–4156 (1996).
- Crooks, G.E., Hon, G., Chandonia, J.M. & Brenner, S.E. WebLogo: a sequence logo generator. *Genome Res.* **14**, 1188–1190 (2004).
- Oman, T.J. & van der Donk, W.A. Follow the leader: the use of leader peptides to guide natural product biosynthesis. *Nat. Chem. Biol.* **6**, 9–18 (2010).
- Melby, J.O., Dunbar, K.L., Trinh, N.Q. & Mitchell, D.A. Selectivity, directionality, and promiscuity in peptide processing from a *Bacillus* sp. Al Hakam cyclodehydratase. *J. Am. Chem. Soc.* **134**, 5309–5316 (2012).
- Denesiouk, K.A. & Johnson, M.S. When fold is not important: a common structural framework for adenine and AMP binding in 12 unrelated protein families. *Proteins* **38**, 310–326 (2000).
- Zhang, Y., Morar, M. & Ealick, S.E. Structural biology of the purine biosynthetic pathway. *Cell. Mol. Life Sci.* **65**, 3699–3724 (2008).
- Fawaz, M.V., Topper, M.E. & Firestone, S.M. The ATP-grasp enzymes. *Bioorg. Chem.* **39**, 185–191 (2011).
- Rakus, J.E. *et al.* Evolution of enzymatic activities in the enolase superfamily: d-mannanase dehydratase from *Novosphingobium aromaticivorans*. *Biochemistry* **46**, 12896–12908 (2007).
- Chen, W., Biswas, T., Porter, V.R., Tsodikov, O.V. & Garneau-Tsodikova, S. Unusual regiospecificity of acetyltransferase E1s, a cause of drug resistance in XDR TB. *Proc. Natl. Acad. Sci. USA* **108**, 9804–9808 (2011).
- Selvy, P.E., Lavrier, R.R., Lindsley, C.W. & Brown, H.A. Phospholipase D: enzymology, functionality, and chemical modulation. *Chem. Rev.* **111**, 6064–6119 (2011).
- Bhatnagar, R.S., Futterer, K., Waksman, G. & Gordon, J.L. The structure of myristoyl-CoA:protein N-myristoyltransferase. *Biochim. Biophys. Acta* **1441**, 162–172 (1999).
- Climie, S.C., Carreras, C.W. & Santi, D.V. Complete replacement set of amino acids at the C terminus of thymidylate synthase: quantitative structure-activity relationship of mutants of an enzyme. *Biochemistry* **31**, 6032–6038 (1992).
- Regni, C.A. *et al.* How the MccB bacterial ancestor of ubiquitin E1 initiates biosynthesis of the microcin C7 antibiotic. *EMBO J.* **28**, 1953–1964 (2009).
- Donia, M.S. *et al.* Natural combinatorial peptide libraries in cyanobacterial symbionts of marine ascidians. *Nat. Chem. Biol.* **2**, 729–735 (2006).
- Izawa, M., Kawasaki, T. & Hayakawa, Y. Cloning and heterologous expression of the thioviridamide biosynthesis gene cluster from *Streptomyces olivoviridis*. *Appl. Environ. Microbiol.* **79**, 7110–7113 (2013).

Acknowledgments

We are grateful to C. Deane and K. Taylor for the generation of select BalhD mutants and J. Melby for assistance with collecting MS/MS data. This work was supported by the US National Institutes of Health (NIH) (1R01 GM071942 to D.A.M., 1R01 GM102602 to S.K.N. and ZT32 GM070421 to K.L.D., B.J.B. and J.R.C.). Additional support was from the Harold R. Snyder Fellowship (University of Illinois at Urbana-Champaign (UIUC) Department of Chemistry to K.L.D.), the Robert C. and Carolyn J. Springborn Endowment (UIUC Department of Chemistry to J.R.C.), the National Science Foundation Graduate Research Fellowship (DGE-1144245 to B.J.B.) and the University of Illinois Distinguished Fellowship (UIUC Graduate College to J.R.C.) The Bruker UltraflexTreme MALDI TOF/TOF mass spectrometer was purchased in part with a grant from the NIH-National Center for Research Resources (S10 RR027109 A).

Author contributions

Experiments were designed by D.A.M., S.K.N., K.L.D., J.R.C., C.L.C. and B.J.B. and were performed by K.L.D., J.R.C., C.L.C. and B.J.B. The manuscript was written by D.A.M., K.L.D. and J.R.C. with critical editorial input from S.K.N., C.L.C. and B.J.B. The overall study was conceived and managed by D.A.M. with S.K.N. overseeing all aspects of protein structure determination.

Competing financial interests

The authors declare no competing financial interests.

Additional information

Supplementary information is available in the online version of the paper. Reprints and permissions information is available online at <http://www.nature.com/reprints/index.html>. Correspondence and requests for materials should be addressed to S.K.N. or D.A.M.



ONLINE METHODS

General methods. Unless otherwise specified, all chemicals were purchased from Sigma or Fisher Scientific. DNA sequencing was performed by either the Roy J. Carver Biotechnology Center (UIUC) or ACGT Inc. Restriction enzymes were purchased from New England BioLabs (NEB). Pfu Turbo was purchased from Agilent, and dNTPs were purchased from either NEB or GenScript. Oligonucleotide primers were synthesized by either Integrated DNA Technologies (IDT) or Eurofins MWG Operon. Fluorescein-labeled BalhA1 leader peptide was purchased from GenScript as an N-terminal FITC-Ahx (fluorescein isothiocyanate and aminoethyl linker) conjugate with a single glycine spacer. Unless otherwise stated, all of the proteins and substrates were used as MBP fusions to circumvent solubility issues. A table of the peptide substrates used in this study can be found in **Supplementary Table 3**.

Cloning of MBP-YcaO. Ec-YcaO was amplified by PCR from *E. coli* BL21 cells using the forward and reverse primers listed in **Supplementary Table 1**. Polymerase reactions were carried out with Pfu Turbo, and the amplified product was digested with BamHI and NotI following a gel extraction. The digested gene was PCR purified and ligated into an appropriately digested, modified pET28 vector containing a TEV protease-cleavable, N-terminal MBP tag.

Preparation of BalhD PxPxP mutants. BalhD was amplified by PCR from the previously described pET28-BalhD plasmid³⁹ using the primers listed in **Supplementary Table 1**. Polymerase reactions were performed with Pfu Turbo, and the amplified product was digested with BamHI and NotI following gel extraction. The digested gene was PCR purified and ligated into an appropriately digested, modified pET28 vector containing a TEV protease cleavable, N-terminal MBP tag.

Site-directed mutagenesis. Site-directed mutagenesis of BalhD and McbD was carried out using the QuikChange method according to the manufacturer's instructions. All of the mutagenesis primers are listed in **Supplementary Table 1**.

Overexpression and purification of MBP-tagged proteins. All proteins were purified with amylose resin (NEB) according to previously described procedures².

Multiple sequence alignments. Alignments were made with Clustal Omega using the standard parameters⁴¹.

Cytoscape sequence similarity network. A sequence similarity network was created using the Enzyme Function Initiative Enzyme Similarity Tool (EFI-EST; <http://www.enzymefunction.org/>)⁴². Sequences from the YcaO superfamily (InterPro number IPR003776)⁴³ were used for the analysis. The network was constructed at an expectation-value (e-value) of 10⁻⁶⁰. Networks were visualized by Cytoscape using the organic layout⁴⁴. Sequences with 100% identity were visualized as a single node in the network.

Maximum likelihood phylogenetic analysis. A set of YcaO sequences representing the full diversity of the YcaO family was selected on the basis of the Cytoscape sequence similarity network (**Supplementary Fig. 5**). This included at least one protein from each cluster and all singletons (proteins that fail to cluster with any other YcaO) for a total of 349 proteins (~17% of the non-redundant Cytoscape sequence similarity network). The phylogenetic analysis was performed with Molecular Evolutionary Genetics Analysis (MEGA5)⁴⁵. An amino acid sequence alignment was created using the standard parameters of ClustalW⁴⁴ and a maximum likelihood phylogenetic tree was created using standard parameters in MEGA5.

ATP-binding site conservation. The ligand interaction network for AMPCPP was generated using LigPlot plus⁴⁶ using the standard parameters. A WebLogo²⁷ frequency plot for the ATP-binding motif was generated from a Clustal Omega alignment of all of the sequences from the specified family using the standard parameters. The conservation map for the YcaO family was generated by aligning 150 unique YcaO sequences with at least 35% sequence similarity to Ec-YcaO (e-value < 10⁻⁹) and mapping the resulting conservation data onto the Ec-YcaO structure using ConSurf⁴⁶. The structure-based YcaO alignment was generated using Clustal Omega⁴¹ and ALINE⁴⁷, and the structural overlay of YcaO and TruD was generated using PyMOL version 1.5 (Schrödinger).

Proline-rich C terminus analysis. The proline content of the C-termini of all of the YcaOs in the Cytoscape sequence similarity network was determined. Proteins were deemed to have a proline-rich (P-rich) C terminus if at least 4 of the final 30 residues were proline. In the most pronounced cases, the terminal six residues of the YcaO contain a PxPxP motif. Proteins were identified as containing a PxPxP C terminus if they contained a PxP motif in the final 6 residues and at least 3 Pro in the final 30 residues.

Ec-YcaO crystallization. Purified MBP-tagged Ec-YcaO was treated with TEV protease for 18 h at 4 °C. Successful cleavage was confirmed by SDS-PAGE, and Ec-YcaO was subsequently separated from the His₆-tagged MBP by subtraction Ni²⁺ affinity chromatography. Fractions with a purity of at least 90% were combined and purified by size-exclusion chromatography using a GE Superdex 200 column equilibrated in a buffer containing 300 mM KCl and 20 mM HEPES, pH 7.5. Fractions were collected and concentrated for crystallization screening experiments. The apo-Ec-YcaO crystals were obtained by sitting drop crystallization using a mother liquor containing 1.8 M ammonium citrate, pH 7.0, and 8 mg ml⁻¹ Ec-YcaO in a 1:1 ratio. Incubation at 4 °C yielded crystals after several days. Soaking with 1 mM PCMBAs for 4 h before vitrification was used to generate phases. Immediately before vitrification, crystals were soaked in a cryoprotectant containing 1.8 M ammonium citrate, pH 7.0, and 30% trehalose. AMP Ec-YcaO crystals were grown using hanging drop vapor diffusion with a mother liquor of 18% PEG 8,000, 0.1 M magnesium acetate and 0.1 M sodium cacodylate, pH 6.5. An Ec-YcaO concentration of 8 mg ml⁻¹ and a substrate concentration of 1 mM ATP and 1 mM MgCl₂ was used for the formation of crystals at 4 °C. Directly before vitrification, crystals were immersed in a cryoprotectant containing the mother liquor supplemented with 30% MPD. AMPCPP Ec-YcaO crystals were grown and frozen in an identical manner with the exception of using 6 mg/ml Ec-YcaO, 1 mM AMPCPP instead of 1 mM ATP and 20% PEG 8,000 instead of 18% PEG 8,000. All of the data were collected at LS-CAT sector 21 of the Argonne Nation Labs Advanced Photon Source at 100 K using wavelengths of 0.97872 Å for the apo and AMP-bound structures and 0.97857 Å for the AMPCPP structure.

Ec-YcaO structure solution and refinement. Collected data were integrated and scaled using HKL2000 or autoProc⁴⁸. The PCBMA-soaked apo-Ec-YcaO crystals were used as a source of anomalous signal in SAD phasing using the PHENIX software suite⁴⁹. Automated building of the structure was accomplished by the arp/wARP server⁵⁰. Manual refinement was performed using COOT⁵¹ and REFMAC5 (ref. 52). For the AMP- and AMPCPP-containing structures, the apo-Ec-YcaO was used as a search model to obtain phases using the PHENIX software suite. PHENIX was also used for automated building of the structures. Manual refinement was again performed using COOT and REFMAC5. Final Ramachandran statistics as determined by PROCHECK⁵² are as follows: 97.2% favored, 2.8% allowed and 0.0% outliers for the apo structure; 97.8% favored, 2.2% allowed and 0.0% outliers for the AMP-bound structure; and 96.3% favored, 3.7% allowed and 0.0% outliers for the AMPCPP-bound structure.

Endpoint heterocycle formation assays. For the reactions with the BalhD mutants, 50 μM MBP-BalhA1 was mixed with either 2 μM MBP-BalhC/D (CD activity) or 25 μM MBP-BalhD (D-only activity) in synthetase buffer (50 mM Tris (pH 7.5), 125 mM NaCl, 10 mM DTT, 20 mM MgCl₂ and 3 mM ATP). The MBP tags were removed using 0.05 mg/ml of TEV protease and reactions were carried out for 18 h at 25 °C. Reactions with the Mcb enzymes were carried out identically except that the dehydrogenase (McbC) was also added to the reaction and proteins were cleaved with 0.1 μg ml⁻¹ thrombin protease (from bovine plasma). Samples were desalted via C18 ZipTip (Millipore) according to the manufacturer's instructions and analyzed on a Bruker Daltonics UltrafleXtreme MALDI-TOF spectrometer. Spectra were obtained in positive reflector mode using α-cyano-4-hydroxycinnamic acid as the matrix.

PNP-based kinetic studies. Substrate processing kinetics for the BalhD mutants were determined using a previously described purine nucleoside phosphorylase (PNP)-coupled assay²². For BalhA1 kinetic experiments, variable concentrations of MBP-BalhA1 (1–120 μM) were reacted with 1–10 μM MBP-tagged BalhC and BalhD, and ATP was held constant at 3 mM. ATP kinetic experiments were carried out in an identical fashion except that MBP-BalhA1 was fixed at 80 μM, and variable concentrations of ATP (0.1–5 mM)

were used. Although this does not provide a saturating level of BalhA1 for all mutants, the K_M for ATP does not change with varied BalhA1 concentration (Supplementary Fig. 13). For BalhD-A4LEH6 ATP kinetic assays, the rate of ATP hydrolysis was measured for 3 μ M MBP-tagged BalhD-A4LEH6 with and without 3 μ M MBP-tagged BalhC. Reactions were carried out in triplicate. Regression analyses to obtain the kinetic parameters for both substrates were carried out with IGOR Pro version 6.12 (WaveMetrics).

BalhC potentiation assays. The effect of BalhC on the background ATPase activity of BalhD was measured using the PNP phosphate detection assay. Given the slow rate of ATP hydrolysis without the presence of the precursor peptide, the background ATPase activity was measured using 15 μ M MBP-tagged BalhD with and without MBP-tagged BalhC. Identical conditions were utilized for all BalhD mutants except BalhD-A4LEH6. For BalhD-A4LEH6, reactions were carried out with 1 μ M of the indicated MBP-tagged enzymes and 100 μ M BalhA1 (where applicable). Reactions were carried out in triplicate.

Fluorescence polarization. Equilibrium BalhC and BalhD fluorescence polarization binding assays were performed at 25 °C in non-binding surface, 384-black-well polystyrene microplates (Corning) and measured using a FilterMax F5 multi-mode microplate reader (Molecular Devices) with default settings. For each titration, protein was serially diluted into binding buffer (50 mM HEPES, pH 7.5, 300 mM NaCl, 2.5% (v/v) glycerol, 0.5 mM TCEP), mixed with 10 nM fluorescein-labeled BalhA1 leader peptide (FP-BalhA1-LP) and equilibrated for 15 min with shaking before measurement. Data from three independent titrations were background subtracted and fitted using a nonlinear dose response curve in OriginPro9 (OriginLab).

Size-exclusion chromatography. A 200- μ l sample containing 25 μ M MBP-tagged BalhC or BalhD was prepared in cleavage buffer (20 mM Tris (pH 7.5), 500 mM NaCl, 10% glycerol (v/v), 0.5 mM tris(2-carboxyethyl)phosphine (Gold Biotechnology)) and treated with 0.05 mg ml⁻¹ TEV protease for 12 h at 4 °C. The amount of protein in a BalhC or BalhD complex was assessed on a Flexar HPLC (PerkinElmer) equipped with an analytical Yarra SEC-3000 (300 \times 4.6 mm, Phenomenex) equilibrated with cleavage buffer. Peaks of interest were collected, and their composition was determined via a Coomassie stained 12% SDS-PAGE gel. The approximate molecular weights were determined by generating a standard curve with a 12- to 200-kDa molecular weight standard kit (Sigma). Control runs were also performed in which one of the two proteins was omitted. Chromatograms were analyzed using Flexar Manager (PerkinElmer).

Mutant BalhD IC_{50} determination. 1 μ M MBP-tagged BalhC and BalhD were mixed in synthetase buffer (50 mM Tris (pH 7.5), 125 mM NaCl, 10 mM DTT, 20 mM MgCl₂ and 3 mM ATP) with 0.25–20 μ M mutant BalhD protein. Reactions were initiated by the addition of 15 μ M MBP-BalhA1, and progress was measured using the PNP phosphate detection assay. All of the reactions were performed in triplicate. IC_{50} values were calculated with IGOR Pro version 6.12 (WaveMetrics).

BalhC K_M for active BalhD mutants. The affinity of catalytically active BalhD mutants for BalhC was determined using the PNP phosphate detection assay and a previously described procedure¹². 25 μ M MBP-BalhA1 was mixed with 1 μ M MBP-BalhD in synthetase buffer, and reactions were initiated by the addition of 0.15–4 μ M BalhC. All of the reactions were performed in triplicate. Regression analyses to obtain kinetic parameters for BalhC were carried out with IGOR Pro version 6.12 (WaveMetrics).

Heterocycle localization via FT-MS/MS. 50 μ M MBP-BalhA1 was modified by 2 μ M MBP-tagged BcerB, BalhC and BalhD in synthetase buffer for 18 h at 25 °C. Proteins were digested with 0.02 mg ml⁻¹ sequencing grade trypsin (Promega) in 50 mM NH₄CO₃ (pH 8.0) for 30 min at 37 °C before the sample was quenched via the addition of formic acid to a final concentration of 10% (v/v), and precipitate was removed via centrifugation at 17,000g. FT-MS/MS analysis was carried out as previously described¹⁶.

41. Stevers, F. *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **7**, 539 (2011).
42. Atkinson, H.J., Morris, J.H., Ferrin, T.E. & Babbitt, P.C. Using sequence similarity networks for visualization of relationships across diverse protein superfamilies. *PLoS ONE* **4**, e4345 (2009).
43. Tamura, K. *et al.* MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol. Biol. Evol.* **28**, 2731–2739 (2011).
44. Larkin, M.A. *et al.* Clustal W and Clustal X version 2.0. *Bioinformatics* **23**, 2947–2948 (2007).
45. Laskowski, R.A. & Swindells, M.B. LigPlot+: multiple ligand-protein interaction diagrams for drug discovery. *J. Chem. Inf. Model.* **51**, 2778–2786 (2011).
46. Ashkenazy, H., Erez, E., Martz, E., Pupko, T. & Ben-Tal, N. ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. *Nucleic Acids Res.* **38**, W529–W533 (2010).
47. Bond, C.S. & Schuttelkopf, A.W. ALINE: a WYSIWYG protein-sequence alignment editor for publication-quality alignments. *Acta Crystallogr. D Biol. Crystallogr.* **65**, 510–512 (2009).
48. Vonrhein, C. *et al.* Data processing and analysis with the autoPROC toolbox. *Acta Crystallogr. D Biol. Crystallogr.* **67**, 293–302 (2011).
49. Adams, P.D. *et al.* PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 213–221 (2010).
50. Langer, G., Cohen, S.X., Lamzin, V.S. & Perrakis, A. Automated macromolecular model building for X-ray crystallography using ARP/wARP version 7. *Nat. Protoc.* **3**, 1171–1179 (2008).
51. Emsley, P., Lohkamp, B., Scott, W.G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 486–501 (2010).
52. Vagin, A.A. *et al.* REFMAC5 dictionary: organization of prior chemical knowledge and guidelines for its use. *Acta Crystallogr. D Biol. Crystallogr.* **60**, 2184–2195 (2004).
53. Laskowski, R.A., MacArthur, M.W., Moss, D.S. & Thornton, J.M. PROCHECK: a program to check the stereochemical quality of protein structures. *J. Appl. Crystallogr.* **26**, 283–291 (1993).

Supplementary Information for:

Discovery of a new ATP-binding motif involved in peptidic azoline biosynthesis

Kyle L. Dunbar^{1,2,†}, Jonathan R. Chekan^{3,†}, Courtney L. Cox^{2,4}, Brandon J. Burkhart^{1,2}, Satish K. Nair^{2,3,5*}, and Douglas A. Mitchell^{1,2,4*}

¹Department of Chemistry, University of Illinois at Urbana-Champaign, Urbana, Illinois, USA. ²Institute for Genomic Biology, University of Illinois at Urbana-Champaign, Urbana, Illinois, USA. ³Department of Biochemistry, University of Illinois at Urbana-Champaign, Urbana, Illinois, USA. ⁴Department of Microbiology, University of Illinois at Urbana-Champaign, Urbana, Illinois, USA. ⁵Center for Biophysics and Computational Biology, University of Illinois at Urbana-Champaign, Urbana, Illinois, USA.

[†]These authors contributed equally to this work

*e-mail: douglasm@illinois.edu, phone: 1-217-333-1345, fax: 1-217-333-0508

*e-mail: snair@uiuc.edu, phone: 1-217-333-0641, fax: 1-217-244-5858

Supplementary Results

Table of Contents:

Supplementary Table 1: Oligonucleotide primers used in this study

Supplementary Table 2: Crystallographic statistics for Ec-YcaO structures

Supplementary Table 3: List of peptides used in this study

Supplementary Figure 1: The biosynthetic gene clusters for trifolitoxin and bottromycin do not contain a recognizable TOMM C protein

Supplementary Figure 2: Coomassie-stained SDS-PAGE gel of proteins described in this study

Supplementary Figure 3: Ec-YcaO hydrolyzes ATP to AMP and pyrophosphate

Supplementary Figure 4: Structural homology between Ec-YcaO and TruD

Supplementary Figure 5: Cytoscape sequence similarity network for the YcaO superfamily

Supplementary Figure 6: Diversity oriented Maximum likelihood tree for the YcaO family

Supplementary Figure 7: The canonical E1 domain ATP-binding site is not conserved in TOMM C proteins

Supplementary Figure 8: The ATP-binding pocket is conserved in characterized TOMM YcaOs

S1

Supplementary Figure 9: Multiple sequence alignment of Ec-YcaO with diverse TOMM YcaOs

Supplementary Figure 10: Mutations to the BalhD ATP-binding site affect heterocycle formation

Supplementary Figure 11: Mutations to the BalhD ATP-binding site affect ATP hydrolysis

Supplementary Figure 12: Mutations to the ATP-binding site that decrease cyclodehydratase activity also affect BalhD-only activity

Supplementary Figure 13: The K_M for ATP does not depend on the concentration of BalhA1

Supplementary Figure 14: Mutations of the Mg^{2+} -binding residues prevent ATPase potentiation by BalhC

Supplementary Figure 15: Effect of the ATP-binding site mutations on BalhC binding

Supplementary Figure 16: The *C*-terminal PxPxP motif conserved in TOMM YcaOs is near the ATP-binding site in TruD

Supplementary Figure 17: The PxPxP motif of BalhD is critical for cyclodehydratase activity

Supplementary Figure 18: Mutations that relocate the *C*-terminus of BalhD prevent ATPase potentiation by BalhC

Supplementary Figure 19: *C*-terminal truncations of BalhD affect BalhC binding

Supplementary Figure 20: Removal of the *C*-terminal residue of BalhD dysregulates the order of azole formation

Supplementary Figure 21: Mutagenesis of the *C*-terminus of BalhD affects cyclodehydratase activity

Supplementary Figure 22: The sequence of the *C*-terminus of BalhD is important for binding BalhC

Supplementary Figure 23: BalhD P429G displays an altered order of heterocycle formation

Supplementary Figure 24: Mutations to the *C*-terminus of BalhD also affect BalhC-independent catalysis

Supplementary Figure 25: The conserved TOMM YcaO *C*-terminus is critical for activity in the microcin B17 cyclodehydratase

Supplementary Figure 26: A *C*-terminal His₆ tag on BalhD abolishes cyclodehydratase activity

Supplementary Figure 27: An updated model for azoline formation by TOMM cyclodehydratases

S2

Supplementary Table 1. Oligonucleotide primers used in this study. Mutations are listed in parentheses, asterisks denote stop codons, and endonuclease cut sites are listed when appropriate. The *430A mutation indicates that the BalhD stop codon was mutated to alanine. This resulted in an additional 12 residues to the C-terminus of BalhD (-A₄LEH₆). F, forward primer; R, reverse primer.

Primer	Sequence	Cut Site
YcaO F	AAGGATCCATGACGCCAAACATTTATCCCCGGCAAAG	BamHI
YcaO R	TTGCGGCCGCTTATTTTGCCCAAGAAATGCTGCTTTGGCGCGC	NotI
McbD (M392*) F	GAGAAATCAAAGTAGGTACCATTTCCATAAAAGCTTGGCGGCCG	
McbD (M392*) R	GGGAATGGTACCTACTTTGATTCTCTGACTTTAATACCGTCCC	
BalhD F	CCGGATCCATGGGTATACAGAATGC	BamHI
BalhD (S72A) F	CAGCACTAATAGCGGCAGTTGGAGAAATCTTGAGCGTTATTGC	
BalhD (S72A) R	TCTCCAACATGCGCTTATTAGTGTGATTACAGAGAATC	
BalhD (E76A) F	CTAATATCGGCAGTTGGAGCAATCTTGAGCGTTATTGC	
BalhD (E76A) R	GCAATAACGCTCAAGAAATGCTCCAACTGCCGATATTAG	
BalhD (E79A) F	GGCAGTTGGAGAAATCTTTGCGGTTATTGCTCATGTTATC	
BalhD (E79A) R	GATAACATGAGCAATAACGCGCAAGAATTTCTCCAACCTGCC	
BalhD (R80A) F	CGGCACTTGGAGAAATCTTGAGGCTTATTGCTCATGTTATCTAAATA	
BalhD (R80A) R	TATTTAGATAACATGAGCAATAAGCCTCAAGAATTTCTCCAACCTGCCG	
BalhD (Q186A) F	CTACAAGGTTGGCAGCAATAGAAAACGCGGCACCTAGAATG	
BalhD (Q186A) R	GTTTTCTATTGCTGCCAACCTTTGAGAACCTGTTGCTAAACCAG	
BalhD (N190A) F	GGTTCTACAAGGTTGCAAGCAATAGAAGCCGCGGCACTAGAAT	
BalhD (N190A) R	ATTCTAGTGCCGCGCTTCTATTGCTTGAACCTTTGAGAAC	
BalhD (E194A) F	TAGAAAACGCGGCACCTAGCATGTATAGAAAGAGACGC	
BalhD (E194A) R	GCCTCTCTTTCTATACATGCTAGTGCCCGCTTTTCTA	
BalhD (E197A) F	CGCGGCCTAGAAATGTATAGCAAGAGACGCGATTAT	
BalhD (E197A) R	ATAATCGCGTCTCTTGTCTATACATTTAGTGCCCGC	
BalhD (R198A) F	AATGTATAGAAGCAGACGCGATTATGATCACATGGTTAAATG	
BalhD (R198A) R	ATAATCGCGTCTGCTTCTATACATTTAGTGCCCGCTTTTCTATTG	
BalhD (K281A) F	CGATCCTCTTATAGCAATGGCGGAGCTTTAATGGAGACG	
BalhD (K281A) R	CGTCTCCATTAAGCTCCCGCAATGCTATAAGAGGATCG	
BalhD (E286A) F	GAAGGGAGCTTTAATGGCGACGTTGGCAAGTCTAA	
BalhD (E286A) R	TTAGACTTGCCAACGTCGCCATTAAGCTCCCTTC	
BalhD (P425*) F	GGATATGCACCAGCAAAAAGCTTTTAAATAAGAATTAGCATCCATTTCCGTAAGCG	
BalhD (P425*) R	CGCTTACGGAAATGGATGCTAATTTCTTATTAAGAATTTGCTGGTGCATATCC	
BalhD (P427*) F	AAAGCTTTTAAATAAGAATCCACATTAGTTTCCGTAAGCCGCGCACTCG	
BalhD (P427*) R	CGATGCGCGCCGCTTACGGAAACTAATGTTGATTTTATTAAGAATTT	
BalhD (P429*) F	AGCTTTTAAATAAGAATCCACATCCATTTTAGTAAGCGGCGCACTC	
BalhD (P429*) R	GAATGCGCGCGCTTACTAAAATGGATGTGGATTTTATTAAGAAT	
BalhD (P429G) R	GTGCGGCCGCTTACCCAATGGATGTGG	NotI
BalhD (*430A) R	GTGCGGCCGCTGCCGGAATGGATGTGG	NotI
BalhD (PxGxG) R	AGCGGCCGCTTACCCAATCCATGTGGATTTCTTATTAAGAATTTGCTGGTGCATATCC	NotI
BalhD (GxGxG) R	AGCGGCCGCTTACCCAATCCATGTCCATTTCTTATTAAGAATTTGCTGGTGCATATCC	NotI
BalhD (P427G) R	TGCGGCCGCTTACGGAAATCCATGTGGATTTCTTATTAAGAATTTGCTGGTGCATATCC	NotI
BalhD (F428A) R	TGCGGCCGCTTACGGAGCTGGATGTGGATTTCTTATTAAGAATTTGCTGGTGCATATCC	NotI
BalhD (H426A) R	AAAGCGGCCGCTTACGGAAATGGAGCTGGATTTCTTATTAAGAATTTGCTGGTGCATATCC	NotI
BalhD (PxPxPG) R	TTGCGGCCGCTTATCCGGAATGGATGTGGATTTCTTATTAAGAATTTGCTGGTGC	NotI
BalhD (AA418-K419) F	CATGGGATATGACCCAGCAAAAATAAGAATCCACATCCATTTTC	
BalhD (AA418-K419) R	GAAATGGATGTGGATTTCTTATTTTGTGCTGGTGCATATCCCAATG	

Supplementary Table 2. Crystallographic statistics for Ec-YcaO structures.

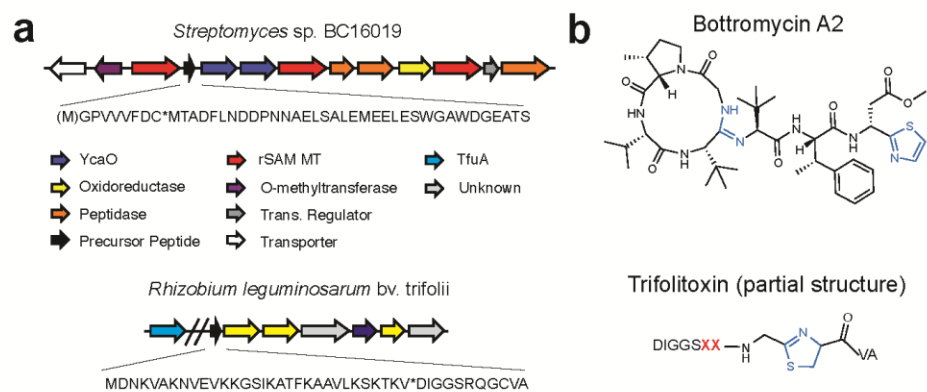
	Ec-YcaO APO	Ec-YcaO AMP	Ec-YcaO AMP CPP
Data collection			
Space group	P2 ₁ 2 ₁ 2 ₁	P1	P1
Cell dimensions <i>a, b, c</i> (Å)	68.96, 140.33, 163.49	111.25, 112.95, 132.89	110.28, 112.40, 130.68
α, β, γ (°)	90.0, 90.0, 90.0	89.94, 73.51, 77.23	89.40, 73.62, 77.62
Resolution (Å)	50.0-2.65 (2.70-2.65)	127.14-2.25 (2.26-2.25)	125.20-3.29 (3.30-3.29)
<i>R</i> _{sym}	9.4 (68.2)	10.6 (71.0)	17.6 (59.3)
<i>I</i> / σ <i>I</i>	16.5 (2.3)	10.6 (2.0)	7.0 (2.1)
Completeness (%)	99.5 (99.6)	93.1 (66.1)	99.0 (98.2)
Redundancy	6.0 (6.1)	3.8 (3.9)	2.5 (2.5)
Refinement			
Resolution (Å)	50.0-2.64	127.14-2.25	125.20-3.29
No. reflections	43,214	251,531	84,143
<i>R</i> _{work} / <i>R</i> _{free}	0.1628/0.2185	0.2178/0.2500	0.1896/0.2377
No. atoms			
Protein	9,027	36,574	35,815
Nucleotide+Metal	4	182	174
Water	512	1,987	432
<i>B</i> -factors			
Protein	43.95	40.03	64.89
Nucleotide		29.62	51.83
Metal	88.88	25.93	23.18
Water	39.21	33.71	33.68
R.m.s. deviations			
Bond lengths (Å)	0.0141	0.0119	0.0113
Bond angles (°)	1.5489	1.4810	1.5215

Each dataset was derived from one crystal.

Supplementary Table 3. List of peptides used in this study. Residues known to be cyclized *in vitro* or *in vivo* are colored orange. In McbA, the orange underlined serine (CSN) is the site of the ninth heterocycle that is installed *in vitro* and found as a minor microcin B17 species¹. A caret and an asterisk denote putative and known leader peptide cleavage sites, respectively. FP-BalhA1-LP is a synthetically prepared reagent designed for fluorescent polarization (FP) studies. The reagent contains a fluorescein label (installed via the amine reactive isothiocyanate, FITC) linked to an aminohexanoic-glycine spacer, which in turn is linked to the predicted leader peptide (LP) of BalhA1.

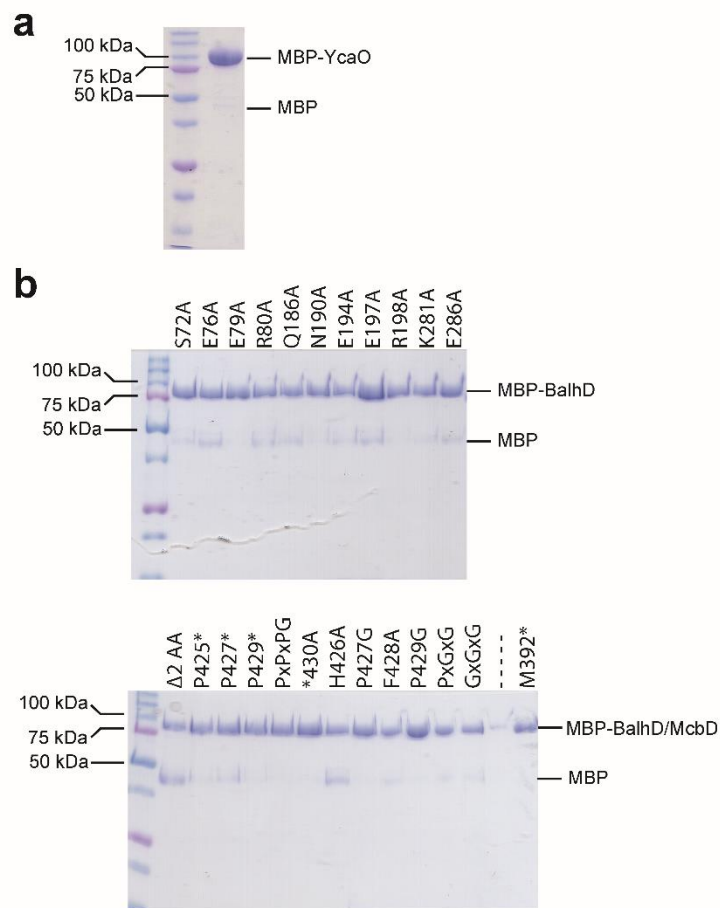
Peptide	Sequence
BalhA1	MEQKKILDIKLTETGKINYAHKPDD^SGCAGCMGCAGGTGCAGTGCIGQGVWKKCSGK
FP-BalhA1-LP	FITC-Ahx-GMEQKKILDIKLTETGKINYAHKPDD
McbA	MELKASEFGVVLSDALKLSRQSPLG*VGIGGGGGGGGGSCGGQGGGCGGCSNGCSGGNGGSGGSGSHI

Supplementary Figure 1. The biosynthetic gene clusters for trifolitoxin and bottromycin do not contain a recognizable TOMM C protein. (a) The gene clusters for bottromycin A2 and trifolitoxin biosynthesis are displayed along with the amino acid sequence of the precursor peptide. Although each cluster contains at least one YcaO homolog, neither cluster contains a recognizable TOMM C protein (homolog of the E1 ubiquitin-activating enzyme superfamily)²⁻⁶. Asterisks indicate the leader and follower peptide cleavage sites in the trifolitoxin and bottromycin precursor peptides, respectively. (b) The structure of bottromycin A2 and the partial structure of trifolitoxin are displayed. Post-translational modifications presumably installed by the YcaO proteins encoded in each cluster are colored blue. The red Xs in the trifolitoxin structure denote an uncharacterized post-translational modification involving Arg and Gln⁷.



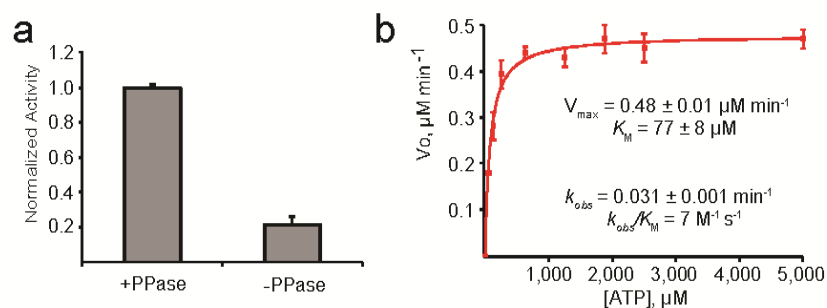
S6

Supplementary Figure 2. Coomassie-stained SDS-PAGE gel of proteins described in this study. MBP-tagged proteins were purified by amylose affinity chromatography. Coomassie-stained 12% SDS-PAGE gels for MBP-Ec-YcaO (**a**) and the YcaO mutants described in this study (**b**) are shown. Masses for pertinent bands of the molecular marker are shown. The faint bands appearing just below the 50 kDa marker band likely result from proteolysis of MBP fusion partner. Asterisks denote stop codons. The dashes (---) indicate an empty lane.



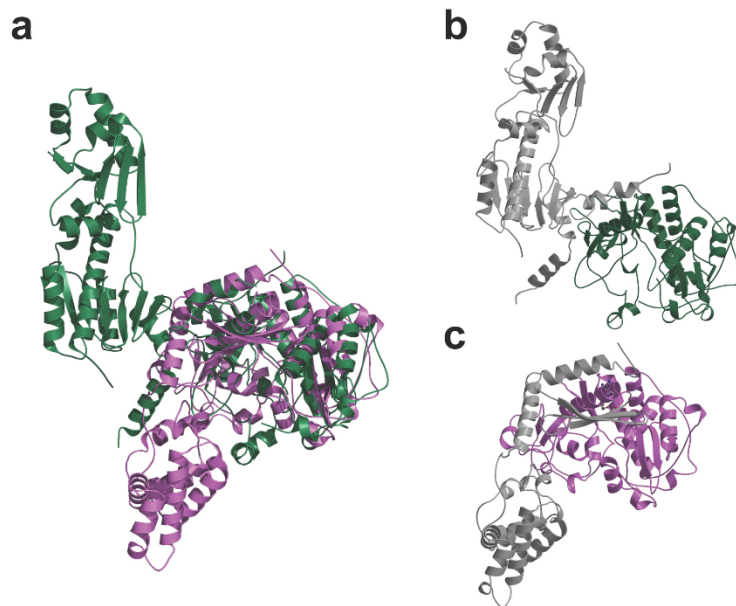
S7

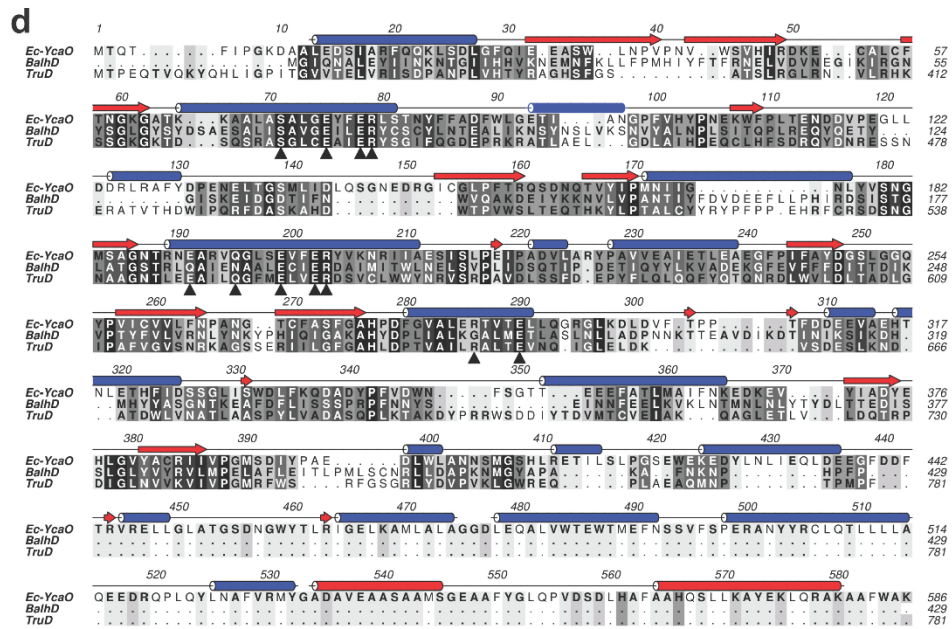
Supplementary Figure 3. Ec-YcaO hydrolyzes ATP to AMP and pyrophosphate. (a) The ATPase activity of Ec-YcaO was screened using the purine nucleoside phosphorylase (PNP)-coupled assay⁸ with and without 1 unit of pyrophosphatase (PPase). The addition of PPase increased the rate of chromophore production in the assay by 5-fold after the signal amplification achieved during PP_i cleavage is taken into account. These data indicate that ATP is preferentially hydrolyzed to AMP and pyrophosphate by Ec-YcaO *in vitro*. This result is corroborated by the observation that co-crystallization of Ec-YcaO with ATP yielded an AMP-bound structure (Fig. 2A). Error bars represent the standard deviation from the mean (n=3). (b) An ATP kinetic curve was obtained for Ec-YcaO using the PPase-supplemented PNP assay. Error bars represent the standard deviation from the mean (n=3). Regression analyses to obtain Michaelis-Menten kinetic parameters were carried out in IGOR Pro version 6.12 (Wavemetrics). The error on the kinetic parameters represents the standard deviation from the curve fitting.



S8

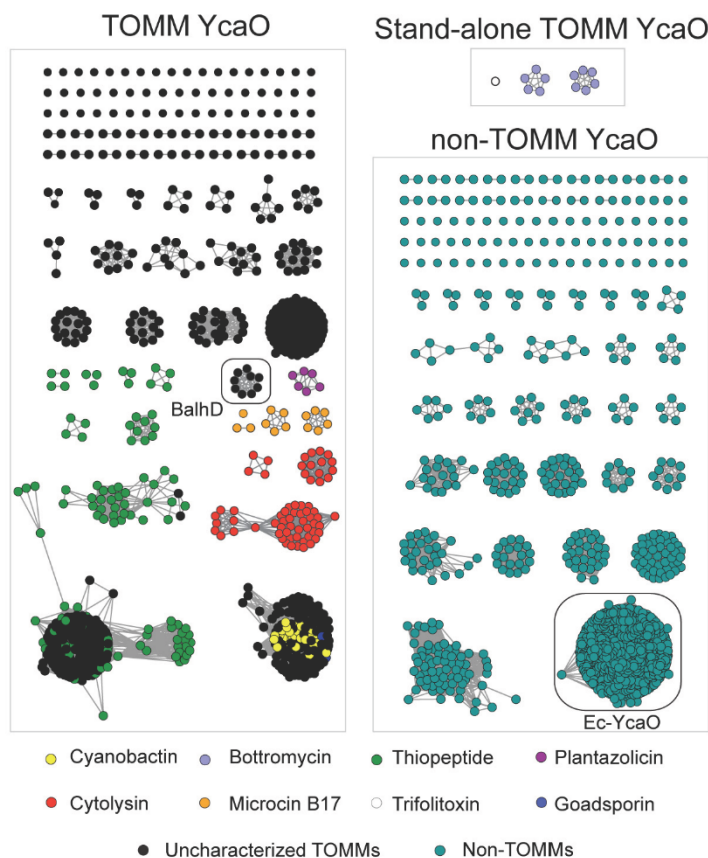
Supplementary Figure 4. Structural homology between Ec-YcaO and TruD. (a) Structural alignment of TruD (green) and Ec-YcaO (purple). The alignment demonstrates that the ATP-binding region of Ec-YcaO is also conserved in TruD. (b) TruD structure with the conserved domain shown in green. (c) Ec-YcaO structure with the conserved region shown in purple. (d) Structure-based sequence alignment of YcaO domain-containing proteins with the secondary structure of Ec-YcaO superimposed. Residues mutated in BalhD for this study are denoted by black triangles. A summary of the BalhD mutants made, along with their predicted roles in ATP recognition is provided.





Mutation	Rationale	Mutation	Rationale
S72A	α -PO ₄ binding	E194A	Mg 1 binding (AMPCPP)
E76A	α -PO ₄ binding, Mg 2 binding (AMP), catalytic	E197A	Mg 2 binding
E79A	Mg 1 binding	R198A	γ -PO ₄ binding
R80A	Active site organization, catalytic	K281A	α -PO ₄ binding (AMP), Mg 2 binding (AMPCPP)
Q186A	Adenine binding	E286A	Mg 2 Binding
N190A	3'-OH binding, active site organization		

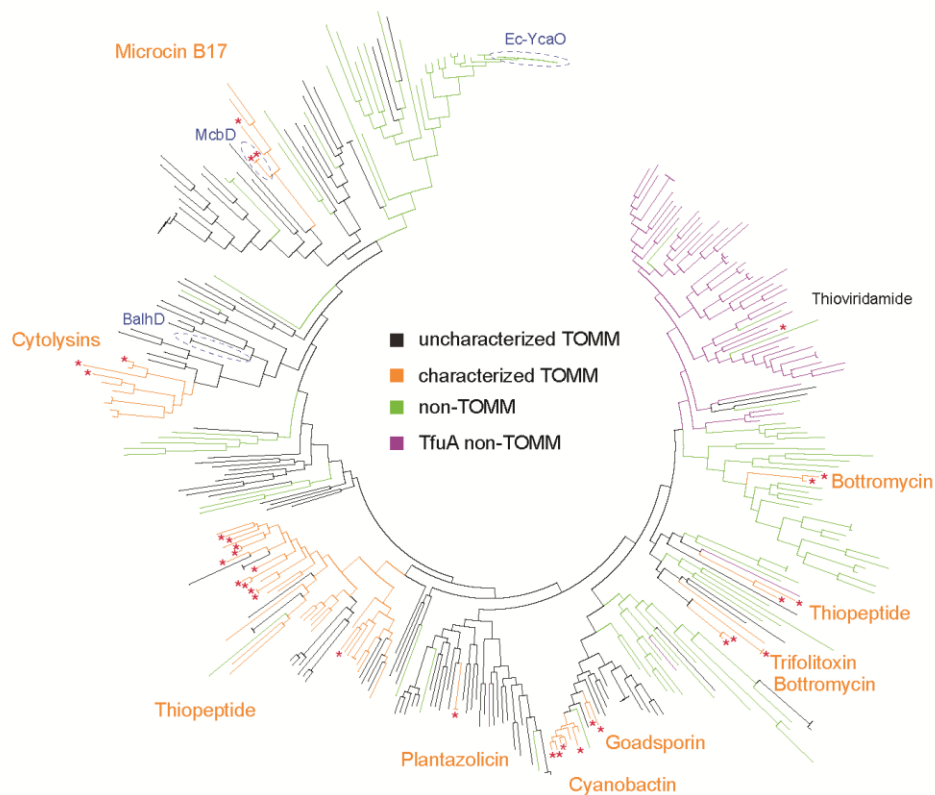
Supplementary Figure 5. Cytoscape sequence similarity network for the YcaO superfamily. A similarity network for non-redundant YcaO sequences is displayed. Each node represents a unique YcaO, while lines between nodes exist if the proteins bear significant similarity (in this case, a BLAST expectation value lower than 10^{-80}). At this e-value, YcaO proteins from characterized TOMM biosynthetic clusters primarily form isofunctional clusters. TOMM YcaOs are defined as having a bioinformatically identifiable C protein within 10 kb of the YcaO protein. Stand-alone TOMM YcaOs are proteins found in a characterized TOMM biosynthetic clusters that lack a C protein (e.g. bottromycin and trifolitoxin). It is noteworthy that many non-TOMM YcaOs appear to exist in RiPP biosynthetic gene clusters, indicating that additional stand-alone YcaO cyclodehydratases may exist. Nodes are colored according to the legend and the clusters containing Ec-YcaO and BalhD are indicated.



S11

Nature Chemical Biology: doi:10.1038/nchembio.1608

Supplementary Figure 6. Diversity oriented Maximum likelihood tree for the YcaO family. A diversity-maximized tree for the YcaO superfamily is displayed. Red asterisks denote YcaOs associated with clusters of functionally or structurally characterized natural products, while the proteins used in this study are circled and colored blue. The members of this tree were used to generate WebLogo⁹ frequency plots for the conservation of the ATP-binding residues in TOMM, non-TOMM, and TfuA-associated non-TOMM subclasses of the YcaO superfamily (**Fig. 3A**).



S12

b

```
Trunkamide (TruD)      EKGYLTVAPSLSEV-----AATWSELGIAPSVVAGLEQFTV-TTA-----GGIGREGIVAMLAALLEEAGIQVSDP
Arthrospiramide (ArtD) KRGFLLEKKELPSML-----ALLCHLQVSOQAQFERLQSLKLVTKSL-----GSLPQQD---LNLNLSLQVQV---
Teneucyclamide (TenD)  EKGYLTAATPDLSPFV-----AATWTELGIAFTVAAGLQGLQPTL-TTV-----GEMISEVTVAAALATALADHGIFVQNA
Trichamide (TriA)     TKGYLTAATPDLSPFA-----AATWELLNVEPQAFSDCLQNTVYLSV-----GETTDL---LLESLAVGIGTFW
Cyanothecamide (CyaD)  TKGYLSEPVVEITPEA-----AATWGLLKVPSFAVQCLOQTQVYVPSL-----SHIITEG---LINSKAVGIKALSM
682765 (MccB)         KENFFIIPCEYHNSTENRYSNPLHYQSYGANPVLVQOKLKNKAVVILKCGGIGCHHVSIVLANSIGIEILLNDQIEMNLTROVLPSEDDVGKMEVEIKRELLKRNSEISVSEIAL
414234 (ThiF)         -----NRDRD*MRYSRQILL-----DDIALDGGKLLDSQVLLIGLGGQLTPAALYLAAGVGTLLVLAODDVHLSNLRQILPTTDEDIDRFSSQVSRQLTQIANPDIQLTALQ
157160305 (HoeB)     -----MAELSDQELRYRQILL-----RGFDFDGGALKRDSKVLVYVGGGLCAASQYLSAGAVGHTLLPDTVLSNLRQILHSDATVQGRFVESARDALARINPHIALTPVNA

Trunkamide (TruD)      EE-PEAEKAGDSTAGLAVVLTDOYLQPELAAINKEALRQQQFWLVFVSGSILMGLPVPVCEFCMMHLLAQRLQGNREVEASVLOQRRAIGER-----N-GQN
Arthrospiramide (ArtD) -----RDEGLMILITDYLWFLSLENINQQALSKHFWLLVFKGHLAWIGLFDQDKGMMHLLAQRLQGNREVEFTIIRQREKIS-----
SD-I-----GSSAALIVLTDQYLQPELAAINQALQSQQTWLWLVFVGVVWLVLPVVFQKQMSLHRLRGNREVEASVLRQKGAQGER-----N-CQQ
EN-----VPLTSDRRKALVVLTDQYLQPELKEINKTALSASQFWLAKPVGCVLWLGPIPEFENTGMMELSQLRGNREVEATVLRQKEFSVVS-----SAS
Cyanothecamide (CyaD)  DG----ELEFPFHSLVILTDOYLQPELSEINQIALKQQQWLLIKPVGTFIPWLGPIFHSKTCMMELAQRLRGNREVEASVLRQKYSFPKSPKTPSSQGGVGGQEEERSNRSNGHKK
682765 (MccB)         NINDFTLRFVPEADIVVSGMSEF-FSLIMWNYCVRAMQPIINAGYVHDAVGFVPLVSKGTEGKRVVADLSSKEEN-IDRKLK-----
414234 (ThiF)         RLFGALKKAVARADVTLKCTDM-NATROEINAACVALNTELTASAVFGGQILVLTFFWEGCYRLLWPN---QEPER-----
157160305 (HoeB)     LDDAEIALAIALBHDLLKCTDM-VAVRNQLNAGCFAKVPVLSGCAAIRHEGQITVFTYQDGEVYRLLSRLF---GENAL-----

Trunkamide (TruD)      KMGAVSCLPTARATLFTLQGLQAAATEIAKHWKRLNAIAPCTARFPTLACKIFTFNQITLLEKAPLSRRPQPTTCORET---
Arthrospiramide (ArtD) -----LISFLGFSSTIQTVLMTAHEVFKMIQGN-----ORLEGLMITYOTLNLQDHLVRFKPFSSGYSL---
Teneucyclamide (TenD)  G-RVSEILPTARATLFTLQALQAAATEIAKVIQVQVMTAPGATLFFLDGKVIETFWOTLDSHLLKRPQPTTCDFE---
Trichamide (TriA)     KRVDGCLFPFPVFLFTLQGLMILITTEVAKHIVKSG----VENKFPFLKGVVTFNQITFTTEILSEKFPKPKDSEK---
Cyanothecamide (CyaD)  NFIQTECLPFGAALASLETALMATTETAKWIKQST----EETAKFPLEKRAITFOQNKLEQNHILTRPQPSGDSNF---
682765 (MccB)         ---INSRFK-P-ATFA*VNVVAALCAADVIFKIGKYSF-----LSLNRIGIWS-DEIKIHSQNMGRSPVSVCCNRM---
414234 (ThiF)         ---KRT-A-GVVG*VGVVHTLQALEIKLLSGIETP-----AG-ELR--LFDGKSSQ-CSLALRRASGFPVGGSHADPV
157160305 (HoeB)     ---TCVE-A-GVMA*LIQVIGSQMREIKLLAGYKRP-----ASGKIV--NRDANTQFRELKRNPFVSVCCN---
```

Supplementary Figure 8. The ATP-binding pocket is conserved in characterized TOMM YcaOs. All of the YcaO sequences from the indicated TOMM subclasses were aligned with Clustal Omega¹⁰ and the multiple sequence alignments were used to generate sequence logos (WebLogos)⁹ of the regions involved in ATP binding. The size of the letter is proportional to its level of conservation. The ATP-binding motif identified in Ec-YcaO is displayed above each of the ATP-binding regions. Residues that are similar to the ATP-binding motif identified in Ec-YcaO are colored orange. Due to the high level of diversity in the sequences, WebLogos for the *N*-terminal ATP-binding residues could not be generated. The number of sequences represented in each WebLogo is displayed in parentheses. The ATP-binding motif of Ec-YcaO is conserved in diverse TOMM YcaOs.



S15

Supplementary Figure 9. Multiple sequence alignment of Ec-YcaO with diverse TOMM YcaOs. Sequences of TOMM YcaOs from characterized natural product clusters were aligned with Ec-YcaO. The C domain of the naturally occurring CD fusion proteins (TruD, GodD, TsrH) was manually removed prior to alignment. Residues implicated in ATP- and Mg²⁺-binding in Ec-YcaO are highlighted in orange. The conserved proline rich C-terminus of TOMM YcaOs is colored blue. A sequence identity matrix is displayed to show the divergence of the proteins in the alignment. The alignment indicates that the ATP-binding site is under intense selective pressure. McdB (microcin B17), YcaO (Ec-YcaO), TruD (trunkamide), GodD (goadsporin), TsrH (thiostrepton), SagD (streptolysin S), BalhD (unknown), PznD (plantazolicin).

```

McdB -----MINVYS-----NLSAWPATMAMSPK-----LNRNMPTFSQIWDYERITPASAAGETLKSIGQAGICEYFER
Ec-YcaO MTQTFIPGDAALDEIARFQQLKSLDLGFQIEEA-----SWLNPNVNVVSHIRDKEC-----ALCFTNGKG-ATKKAALASALGEYFER
TruD  HFTSDGGHRAMTPEQIVQRYQHLIGPITGVVTELVRI SDPANPLVHTYRAGHSFGSAT-S-----LRGL-----RNVLRHKSXSGKG-KTDSQSRASGLCEAIER
GodD  --SSDGGYHSRPFQEQVARYEHLISPVVTGPI SNLVKVPLEVE-GLHTYTAGQNFAPMAN--ANDL-----RAGLRSCSAGKG-MSDAQAKASALGAEIER
TsrH  -----RFPFAAR-----FAE-----RLEREYLDGWSGVTRSAAVGDRAVLPSTQVRVPTVWGPFDEIAIGRA-EDYAGARPAALIEGLER
SagD  -----MLYYPSFNHIFDELKSLSGNRTGILNQSQ-V--PVCNHPHDVLKSTIQGM-PD-----YHKQFIGELSQVSYHII-GYGS-Y-YEEALIKYLGESIER
BalhD -----MG-----IQ-----NALEYIINKNTGIIHHVK-N-EMN--FKLLFPMHYFTFRNE--LVDVNEGKIRGNYSGL-GYSY-SDAESALISAVGELER
PznD  -----MSKWTTLDESXKVLDSLADPVVGIYRREY-R--KLR-DVDDVW-GHSYGAG-GT-----DSKILFGVP-SNSYNG-----GGG-DNAAAAARLAAIGETVER
: * : **

McdB RHF--FNEIVT-----GGQKTYE--MMPSSAAKAFTE-----AFFQISSLTRDEIITHKFKTVRAFNLFSLEQQEIP
Ec-YcaO LSTNYFFADFVLGGETIANGPF-----VHYPNKWFPLTENDVPEGLLDRRAFYPENELTGSMLIDLQSGNEDR-----GICGLPFTQSDNQTVYIP
TruD  YSGIFQGDPR--KR--ATLAEGLD-AIHPEQCLHFSDRQYDN--RESSNERA-----TVTHDWIQRFDASK-----AHDWTPVNSLTEQTHKYLIP
GodD  YSGLFHGDAR--RT--ARYADLPAEVIHPNAVHLYSEAQFRD--RAEWNARP-----SHPHWVADLDPNA-----EVENSPVNSLTEQTHKFLP
TsrH  YAGWHCGG--R--DPVRFASAEASADAEPGGGGPDRPAAS--DAVVDPRSL-LHPEE-AYGQPGFEYTPYAPE-----PTGWAAYASALTRRTLVLP
SagD  YATVIAGDLS--DRIVYASYNELKLLHKV--MPLYLQVFTQ-E-----Q-----IALS-CDLQMMCKMVTEND-----VLGWKVCMPFFDEAEMVLP
BalhD YCSCYLN-TEA--LTK--NSYNSLVKSNVYAL--NPLSITQPLR-E-----Q-----YQETYGISKEIDGDT-----IPNWNQKDEIYKKNVLPV
PznD  YSAAYLP-FDS--DVVCFGSQKELEKENRFRIGYKDWELYP-----T-----Q-----FKDPIPPHEVWNENT-----KLNWRKGYAKTNEEVWTP
: * : **

McdB AVIIALDNIT-AADDLKFYPRDRTCGCSFHGSLNDIATEGSLCEFMETQSLLLYLWQKANTEISSEIVTGI-NHI DEILLALRSEGDIRIFDITLPGAPGHA-
Ec-YcaO MNLIIGN-----LYVSNMGSAQTRNEARVQGLSEVPERYVKNRI IAESISLPEIPADVLARYPAVVEAIELEAEGFPIFAYDGSLSGGQYFVIL-
TruD  TALCYRYFF-----PPEHRFCRSDSNGNAAGTLEAIIQGFMLVERDSVCLWVYNRVSRPAVDLSSFDPEP-YFLQLQQFYQVQRNRLWVLDLTLADLGI FAF-
GodD  TSGLIFYAHH--FKNRYQAGANSNGCAAGTSLDAVLQGFMLVERDSVCLWVYNRVSRPAVDLSSFDPEP-YFLQLQQFYQVQRNRLWVLDLTLADLGI FSV-
TsrH  FHWVYGATRRPPT-GPRVYVENSNGCALGSGTTEALLAALLEVSRDAFLCAMLSCPTLEIDLGRLTGE--AARAVRVRHRTIGRELRARFALGADFPVVA-
SagD  AQMLCVYKTMETVGERRIIPGFSYCTASHKTELAAMCNSLIEYQIDSMMLSWYTKKCPKIIIVDDPDI-E-VILEEARLGGKSLDIIIPIDMTVGEDNPLYT
BalhD ANTIYFDVDE-EFL-LPHRDSISGLTAGSTRLOAIEAALCEIENDAIMTWLNELSVPLIDSOI PDE-TIQVYLVKADEKGFVEVFFDITTDIKVPTVY-
PznD  GQLIHIANHT-EPWGDNTNIGHATSNGLACGTFVAAISGLFETTERDAFMTWYNKLSLPQIDVNSPRL-K-RFYERYKPTSLRLHLVDMTVFSGIPSV-
: * : **

McdB -VLTLYGTKNKISRIKYSTGLSVANSLLKALCKSVVLLNQSICL-HNPLIGGY-----TDDDIIDSYQRHFMSCN-----KYESPTDLCENTVLLS
Ec-YcaO -CVVLFNPAANGT--CFASGAPH--DFGVALERTVTELLQGRGLKDLDFV-TPPTFDDEEVAEHTN-----LETHFIDS--S-----GLISDWLQKQADYFPV
TruD  -VGVSNRKAQSSERILGFGAHL--DPTVAIIRALTEVNIQIGLEL--DKVSD-----E-SLKND-A--TDWLNVNATLAASPYV
GodD  -AAVSVRRDKPAQDILFALGAHF--DVEVAIGRALSMMNQFLPAV--IGMKA-----DGSG-----TYAYNDPD-Q--LRHWRTATTENQYVLP
TsrH  -LLVSTAEPPDLFATLVTAGSGL--TVERALLGAVHMAASAPVNTVEPQRRRAEL--EAALDDPGLVRRMEDHALVGALE--ARWFSEFLDQTPGGV
SagD  FGIILKKNYDEGPFVILFGVQAGL--DPKHALLIGIMASAYSYYNLLYQKASLANIECEPEFL--DLDNSVYFYAHPKQDQKHNKAFEPILSGEVLSS-
BalhD -FVLVRLNLYNKYHIQIGAKAHY--DPLIAIKGALMETLAS--LNLADPNPKTTEAVD IKDTINIKSIKDHMHYASGN--TKEAFDILISSFR--
PznD  -LAVVRNPHNTNAPFAIGAASSY--SIERACEAAIECMYT--RTVWVKTQRLGNALVNADPNRDNINSFEDHIKLYAGTD--IVKEADFLTNSSTLVD
: * : **

McdB D-----DVKLTLEENIT-SDTNLLNYLQOISDNIFVYARERVNSLVW-YTKIVSPDFP-----LHMNNSGAINNKIYHTGDGI---KVRES
Ec-YcaO -----D--WNFSGTTEEFATLMAIFNKEDKEVYIA-----DYEHLGVYACRIIVPGMSDIYPA-----EDLWLANS-MGS-----HLRE
TruD  ADASQPLKTAKDYPRKMSDDIYDVTMCVEIAKQAGLETLVLDOTRPDI--GLNVVUVIIVPGMR-----FWSRFGSGRLYDVVFK-LGWREQLAEAGMNP
GodD  P-AREGPRRTKADFYRYESTDVRDDVLRARIIVREHGHEMLVLDOTRVDI--GLPVVRVIVPGMR-----HFWRPFAAGRLFDVPLE-LGWDRKPSAEALNP
TsrH  -----PGPAAPDGLTPSGDVAADLAAMLAARROGDGVVVVDHTTSELDRGLRCVKAIVPGTPMPTFGHRHRLPPATLRAFARHRTDVFVEFSPPEVRH
SagD  -----DLEDHSGKDKEDLKTLLAYAKKVPNAVFLDITPEALEKGYVTVRLMPELEMCIPAF-PFANHPRM-----FGG-----VTNA
BalhD -----PFNNYSEINNFEELK--VKLNTMNLNLYTYDLTTEDISSGLYVYRVLMPELAFLEITL--PMLSCNRLDAPKN-MGYAPAA--KAFNK
PznD  -----VNQKSFDDSSPDVWNSLN--DHYEKGFVAFTDLTSPDIKDGAGYVVKSIIPGFKPIDVLYRGRMLGGERILRHSYD-LGLVDPPTIETLNP
: * : **

```

```

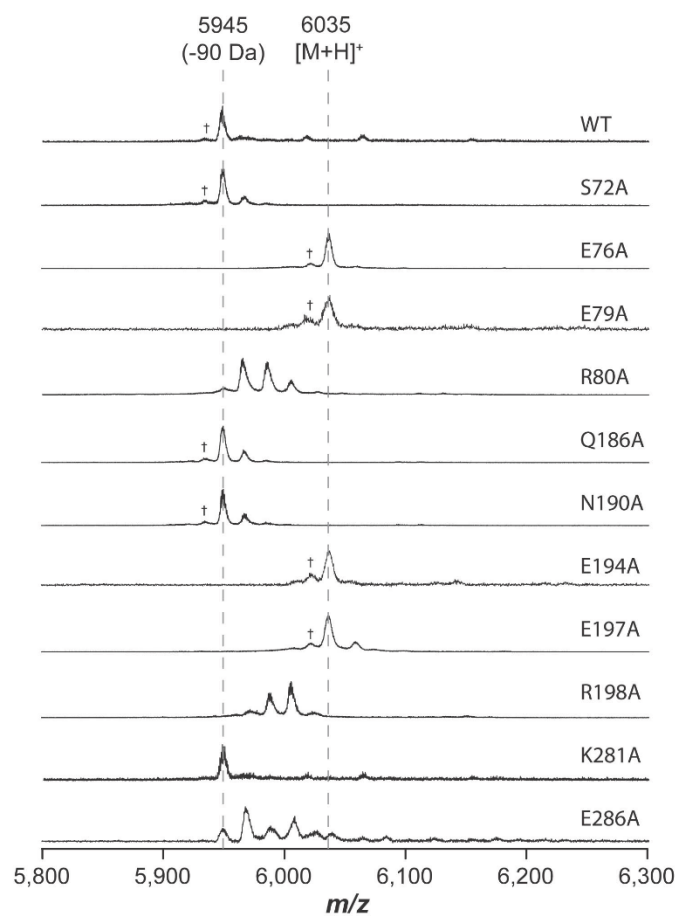
McdB KMVFFP-----
Ec-YcaO TILSLPGSEWEKEDYLNLEQLDEE
TruD  TPMFP-----
GodD  IAMFL-----
TsrH  EPHFP-----
SagD  FVHPMP-----
BalhD NPHFP-----
PznD  VPHFP-----
:

```

	McdB	YcaO	TruD	GodD	TsrH	SagD	BalhD	PznD
McdB	100	-	-	-	-	-	-	-
YcaO	16	100	-	-	-	-	-	-
TruD	12	20	100	-	-	-	-	-
GodD	12	18	47	100	-	-	-	-
TsrH	16	18	20	22	100	-	-	-
SagD	16	45	20	19	18	100	-	-
BalhD	18	18	19	20	20	25	100	-
PznD	14	19	21	19	23	22	26	100

Supplementary Figure 10. Mutations to the BalhD ATP-binding site affect heterocycle formation.

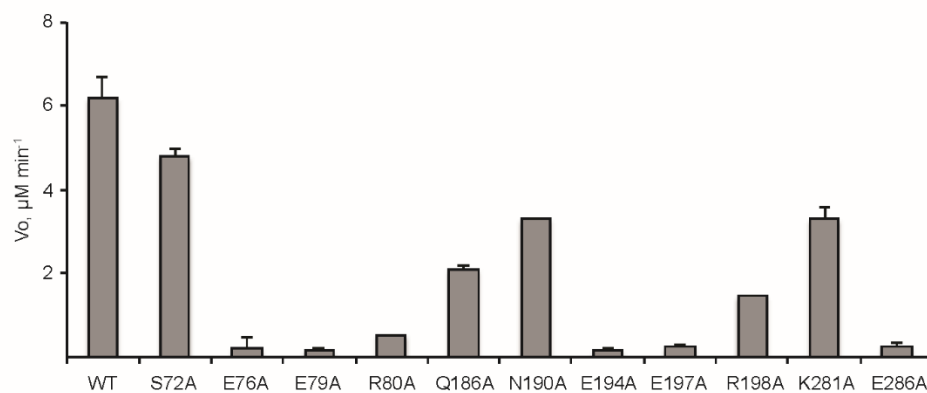
A MALDI-TOF MS spectral overlay for BalhA1 treated with BalhC and either wild-type (WT) or mutant BalhD is displayed. Many of the mutations to the ATP-binding residues, which are listed on the right, decreased cyclodehydratase activity. The level of processing is summarized in **Table 1**. †, laser-induced deamination. A loss of 90 Da indicates the formation of 5 azolines, a full *in vitro* processed BalhA1 substrate¹¹.



S17

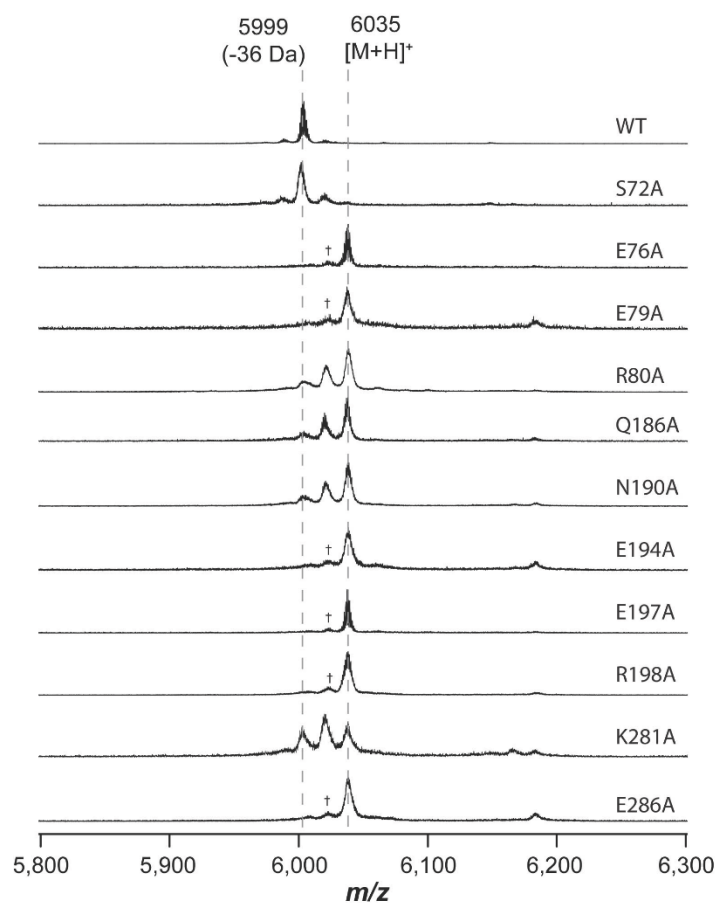
Nature Chemical Biology: doi:10.1038/nchembio.1608

Supplementary Figure 11. Mutations to the BalhD ATP-binding site affect ATP hydrolysis. The rate of ATP hydrolysis was measured with BalhC and either wild-type (WT) or mutant BalhD using the PNP assay. Error represents the standard deviation from the mean ($n \geq 3$). Mutations showing the slowest rate of ATP hydrolysis also installed fewer heterocycles on BalhA1 in an endpoint assay (Supplementary Fig. 10).



S18

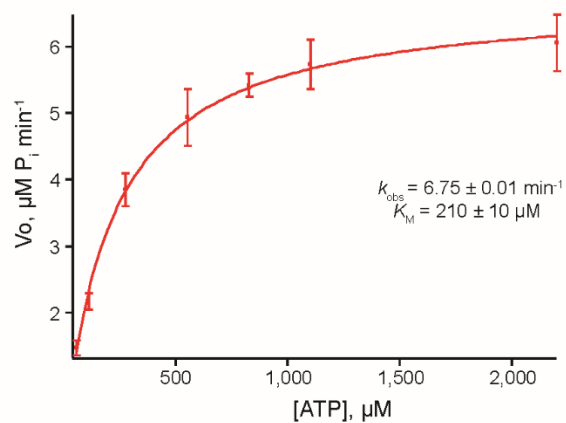
Supplementary Figure 12. Mutations to the ATP-binding site that decrease cyclodehydratase activity also affect BalhD-only activity. MALDI-TOF MS spectra for BalhA1 treated with wild-type (WT) or mutant BalhD is displayed. Apart from BalhD S72A, all of the mutants displayed a decreased level of D-only processing consistent with the rate of ATP hydrolysis measured in the presence of the C protein (Supplementary Fig. 11). The level of processing is summarized in Table 1. †, laser-induced deamination.



S19

Nature Chemical Biology: doi:10.1038/nchembio.1608

Supplementary Figure 13. The K_M for ATP does not depend on the concentration of BalhA1. Due to solubility limitations, saturating concentrations of BalhA1 were not obtainable for all of the BalhD mutants. To determine if the K_M for ATP changed at non-saturating concentrations of BalhA1, an ATP kinetic curve was carried out at the K_M for BalhA1 (15 μM). The resultant K_M is within error of the previously reported K_M for ATP, $240 \pm 20 \mu\text{M}$, obtained with a saturating concentration of BalhA1¹⁸. Error on the Michaelis-Menten parameters represents the standard deviation from the regression analysis.



S20

A.2 HIV Protease Inhibitors Block Streptolysin S Production

This chapter was reprinted with permission from Maxson, Deane, Molloy, Cox, Markley, Lee, and Mitchell (Maxson, *et al.* 2015).

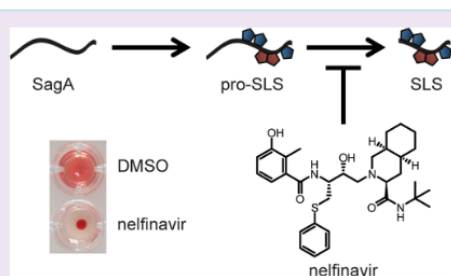
I aided in the collection and interpretation of the bioinformatic analysis for the number of CaaX-like proteases located within TOMM biosynthetic gene clusters.

HIV Protease Inhibitors Block Streptolysin S Production

Tucker Maxson,[†] Caitlin D. Deane,^{‡,§} Evelyn M. Molloy,[‡] Courtney L. Cox,^{‡,§} Andrew L. Markley,^{||} Shaun W. Lee,[⊥] and Douglas A. Mitchell^{*,†,‡,§}[†]Department of Chemistry, University of Illinois at Urbana–Champaign, Urbana, Illinois, United States[‡]Institute for Genomic Biology, University of Illinois at Urbana–Champaign, Urbana, Illinois, United States[§]Department of Microbiology, University of Illinois at Urbana–Champaign, Urbana, Illinois, United States[⊥]Department of Chemical and Biological Engineering, University of Wisconsin–Madison, Madison, Wisconsin, United States^{||}Department of Biological Sciences, Eck Institute for Global Health, University of Notre Dame, Notre Dame, Indiana, United States

Supporting Information

ABSTRACT: Streptolysin S (SLS) is a post-translationally modified peptide cytotoxin that is produced by the human pathogen *Streptococcus pyogenes*. SLS belongs to a large family ofazole-containing natural products that are biosynthesized via an evolutionarily conserved pathway. SLS is an important virulence factor during *S. pyogenes* infections, but despite an extensive history of study, further investigations are needed to clarify several steps of its biosynthesis. To this end, chemical inhibitors of SLS biosynthesis would be valuable tools to interrogate the various maturation steps of both SLS and biosynthetically related natural products. Such chemical inhibitors could also potentially serve as antivirulence therapeutics, which in theory may alleviate the spread of antibiotic resistance. In this work, we demonstrate that FDA-approved HIV protease inhibitors, especially nelfinavir, block a key proteolytic processing step during SLS production. This inhibition was demonstrated in live *S. pyogenes* cells and through *in vitro* protease inhibition assays. A panel of 57 nelfinavir analogs was synthesized, leading to a series of compounds with improved anti-SLS activity while illuminating structure–activity relationships. Nelfinavir was also found to inhibit the maturation of otherazole-containing natural products, namely those involved in listeriolysin S, clostridiolysin S, and plantazolicin production. The use of nelfinavir analogs as inhibitors of SLS production has allowed us to begin examining the proteolysis event in SLS maturation and will aid in further investigations of the biosynthesis of SLS and related natural products.



The ribosomally synthesized and post-translationally modified peptides (RiPPs) comprise a rapidly expanding class of natural products that includes a wide variety of structural modifications.¹ These modifications impart RiPPs with diverse activities, giving rise to a range of products from antibacterials^{2–4} to anticancer agents.⁵ The installation ofazole and/or azoline heterocycles is one such modification common to many RiPPs, forming a subclass of natural products called the thiazole/oxazole-modified microcins (TOMMs).⁶ The azoles are biosynthesized by the cyclodehydration and subsequent dehydrogenation of cysteine, serine, and threonine residues to form thiazole and (methyl)oxazole rings on the C-terminal portion, or “core,” of a ribosomally produced precursor peptide.⁶ Theazole-containing peptides will often undergo further processing, including the proteolytic removal of the N-terminal “leader” portion of the peptide and export of the mature product.⁷ Although recent discoveries have shed light on the mechanism ofazole formation,^{8–10} the proteolytic processing step of most TOMMs has yet to be investigated.

Streptolysin S (SLS), a key virulence factor of *Streptococcus pyogenes*, is one such TOMM whose biosynthesis is incompletely understood.¹¹ *S. pyogenes* is the causative agent of diseases ranging in severity from pharyngitis to necrotizing fasciitis¹² and is a major global health burden, causing over 600 million infections and 500 000 deaths annually.¹³ SLS is the cytolytic toxin responsible for the classic β -hemolytic phenotype when *S. pyogenes* is grown on blood agar¹⁴ and has been shown to be critical to pathogenesis in mammalian infection models.^{15–17} Although a few strains of non- β -hemolytic, pathogenic *S. pyogenes* have been described, such as the Lowry strain,¹⁸ the vast majority of *S. pyogenes* isolates produce SLS.¹⁹ The toxin is biosynthesized by a nine-gene biosynthetic operon that encodes the precursor peptide (*sagA*), cyclodehydratase and dehydrogenase enzymes (*sagBCD*), a

Received: October 17, 2014

Accepted: February 10, 2015

monocytogenes, *Clostridium botulinum*, and *Staphylococcus aureus*; Figure 1A,B).

In this study, we identified inhibitors of SLS biosynthesis in *S. pyogenes* by searching for compounds that block an essential proteolytic maturation step. This proteolysis event has been proposed to be performed by SagE, a putative peptidase with homology to a large family of proteases referred to as the CaaX proteases and bacteriocin-processing enzymes (CPBP; often confusingly annotated as abortive infection proteins, Abi), which includes the eukaryotic type II CaaX proteases as well as prokaryotic proteins with putative bacteriocin-related functions.^{28,30} The type II CaaX proteases are involved in the processing of a number of C-terminally prenylated proteins in eukaryotes and have been much more thoroughly studied than their prokaryotic counterparts.^{31–33} Type II CaaX proteases were initially believed to be cysteine proteases,³¹ but the presumed catalytic cysteine was later proven to be unnecessary for activity.³³ In contrast, conserved glutamate and histidine residues were shown to be important for activity, leading to the hypothesis that type II CaaX proteases were metalloproteases.³³ Many of the prokaryotic members of the CPBP family, including SagE, have been annotated as immunity proteins due to the role in bacteriocin self-immunity of the family members in *Lactobacillus plantarum* and *L. sakei*.³⁵ However, the production of viable allelic exchange mutants of *sagE* and the homologue in *Listeria monocytogenes* (*lspI*) indicates that SagE may not be serving an immunity role or may be redundant with other uncharacterized immunity mechanisms.^{14,36} Thus, it is plausible that the prokaryotic CPBPs actually act as proteases and that this function is exploited to provide self-immunity in certain cases. The results presented herein support a role for SagE as a protease through the discovery and characterization of a family of small molecule inhibitors of SLS biosynthesis. Appealingly, these inhibitors were identified through the repurposing of existing drugs by an examination of known off-target effects. This approach facilitated the rapid identification of lead compounds without the need to perform expensive and laborious high-throughput screens, as well as aiding in the synthesis of analogs by leveraging previous work on the compounds.^{37–39}

RESULTS

Evidence for the Role of SagE as a Protease.

Reconstitution of SagE *in vitro* would be the most direct means of testing for the predicted protease activity and for the screening of inhibitors. Unfortunately, numerous attempts to heterologously express SagE proved unsuccessful, and alternative assays were developed to address this issue. We attribute much of the difficulty in expressing SagE to the predicted transmembrane nature of the protein; topology modeling of SagE with SPOCTOPUS⁴⁰ predicted five transmembrane helices and an N-terminal signal peptide (Figure S1). With this in mind, crude membranes from *S. pyogenes* were prepared from total cellular lysates by ultracentrifugation. The resultant samples were then assessed for proteolytic activity toward ³⁵S-labeled SagA, generated through *in vitro* transcription/translation. Robust proteolysis of SagA to the predicted molecular weight²² was observed after treatment with the membrane fraction (Figure 2A). SagA processing in the whole cell lysate and supernatant fractions was much less extensive and often not observable (Figure S2). To test the substrate specificity of proteolysis, a mutant version of SagA with residues Ala20, Gly22, and Gly23 mutated to leucines (SagA-VLPLL) was

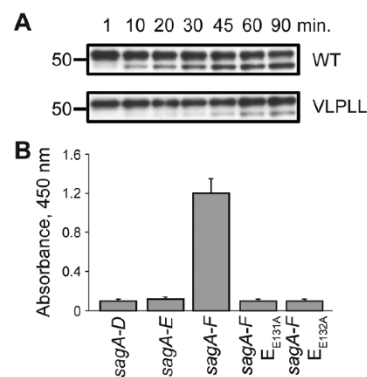


Figure 2. Role of SagE in the processing of pre-SLS. (A) Purified *S. pyogenes* membranes display proteolytic activity toward SagA. Both wild type (WT) SagA and, to a reduced extent, the predicted cut site mutant SagA-VLPLL function as cleavage substrates. (B) Lytic activity of *E. coli* expressing the SLS biosynthetic machinery after a 2 h induction. The hemolytic activity of extracted SLS on erythrocytes is measured by a colorimetric readout of hemoglobin release.

generated (Figure 1B). These residues are directly N-terminal to the predicted leader peptide cleavage site and were expected to be important for protease recognition. Pro21 was left intact to avoid inducing a drastic structural change on the peptide. When treated with isolated *S. pyogenes* membranes, SagA-VLPLL was processed at a rate slower than wild-type SagA (Figure 2A), suggesting that the cleavage was performed by a membrane protease specifically recognizing the SagA cleavage site (other membrane-bound proteases may also be minor contributors to the observed proteolysis).

To provide additional evidence for the role of SagE, a previously reported multiplasmid-based expression system for the generation of SLS in *Escherichia coli* was used. In that study, a lytic entity was generated by the expression of a maltose-binding protein (MBP)-tagged SagA in conjunction with SagB-D from pETDuet vectors after extended induction times.⁴¹ SLS produced in this system could be extracted with bovine serum albumin (BSA) and applied to blood using procedures long established for *S. pyogenes*.¹² The heterologously produced SLS was presumably exported by generalized transporters after nonspecific proteolysis of the MBP tag or was released via *E. coli* cell death following buildup of the toxin. Using this system, no lytic activity was observable after a considerably shorter induction time (2 h) from a strain expressing only SagA-D; however, *E. coli* expressing SagA-F was highly lytic after 2 h (Figure 2B), indicating that SagE expedites SLS maturation. The additional inclusion of SagF was required for this effect, although its functional role is unknown. SagF shares no homology to any known proteins but has previously been demonstrated to be necessary for SLS biosynthesis¹⁴ and may aid in the folding or localization of the other modification machinery. To support the role of SagE as a protease involved in SLS maturation, two highly conserved glutamate residues (Glu131 and Glu132) that are critical for activity in mammalian CPBP family members^{30,33} were mutated to alanine, resulting in the complete loss of the observed lytic activity.

C

DOI: 10.1021/acs500843r
ACS Chem. Biol. XXXX, XXX, XXX–XXX

Aspartyl Protease Inhibitors Block SagA Proteolysis. A small panel of general mechanism-based protease inhibitors was screened for inhibition of SagA leader proteolysis using the membrane cleavage assay with *S. pyogenes* membranes. This assay was preferred over the *E. coli* multiplasmid system to avoid potential issues with membrane penetration or general toxicity of the inhibitors. From the panel, only the aspartyl protease inhibitor pepstatin displayed inhibitory activity (Figure S3), which was unexpected given that SagE bears no similarity to known aspartyl proteases and that members of the type II CaaX protease family were previously hypothesized to function as zinc-dependent metalloproteases.³³ However, inhibition of a metalloprotease by aspartyl protease inhibitors has literature precedent, as several inhibitors of HIV protease were found to also inhibit the type I CaaX protease ZMPSTE24, which is a known zinc metalloprotease.^{43–45} Lipodystrophy is a possible side effect of treatment with certain HIV protease inhibitors and also occurs from genetic deficiencies in ZMPSTE24, leading to the discovery of this off-target effect for these drugs.⁴³ Although type I and II CaaX proteases do not share sequence similarity, they redundantly process some of the same substrates and share similar substrate-binding site architectures.^{46,47} Given this, we reasoned that HIV protease inhibitors might be repurposed as inhibitors of SLS production.

HIV Protease Inhibitors Block SLS Production. A whole-cell assay in *S. pyogenes* based on the extraction of SLS with BSA was used to screen a panel of nine FDA-approved HIV protease inhibitors (Figure S4). Nelfinavir, ritonavir, saquinavir, and lopinavir were found to inhibit the production of SLS when tested at 50 μM , while indinavir, amprenavir, atazanavir, and darunavir did not inhibit SLS production (Figure 3A). Interestingly, tipranavir caused significant growth suppression in *S. pyogenes*, and treated cultures never reached late exponential phase (Figure S5), which is when SLS becomes detectable *in vitro*.⁴⁸ The efficacies of the HIV protease inhibitors shown to inhibit SLS production in the initial screen were evaluated by determining 50% inhibition concentration (IC_{50}) values. The IC_{50} of nelfinavir (6 μM) was much lower than those of ritonavir (35 μM), saquinavir (25 μM), and lopinavir (25 μM). Owing to its greater potency, nelfinavir (Figure 3B) was selected for further study.

Nelfinavir was evaluated in the membrane proteolysis assay to determine if the observed loss of β -hemolysis was due to inhibition of the proteolytic processing step of SLS maturation. As expected, treatment with nelfinavir greatly reduced the proteolytic activity toward SagA contained within *S. pyogenes* membranes (Figure 3C, the membrane fraction used in this experiment was prepared on a different day than the fraction used in Figure 2 and displayed much higher activity). Evidence that nelfinavir was not drastically perturbing normal cellular function came from the observation that the growth rates of *S. pyogenes* treated with nelfinavir were identical to the DMSO control (Figure S5). Additionally, minimum inhibitory concentration (MIC) testing revealed no growth inhibition up to the highest concentration tested (64 μM) for a range of bacterial species (Table S1). Transmission electron microscopy (TEM) was used to examine the morphology of *S. pyogenes* treated with nelfinavir, with no apparent changes compared to the control sample (Figure S6). The transcription levels of a panel of virulence factor genes, as assessed by qRT-PCR, were also not significantly impacted (Table S2). Notably, the levels of *sagA* and *sagB* expression were unchanged. These data indicate that nelfinavir inhibited peptide processing directly,

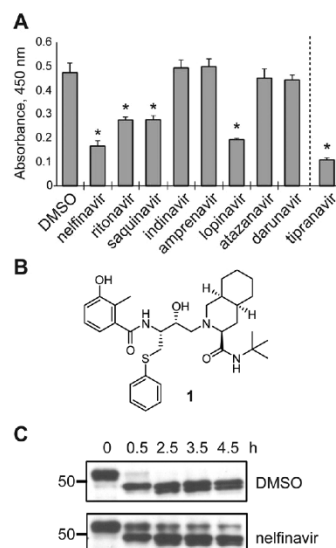


Figure 3. Inhibition of SLS production. (A) Hemolytic activity of SLS extracts from *S. pyogenes* treated with HIV protease inhibitors. Tipranavir is shown separated by a dashed line due to significant growth effects during treatment (Figure S5). Asterisks indicate a P value <0.01 relative to the DMSO control. (B) The structure of nelfinavir (1). (C) The proteolytic effect of *S. pyogenes* membranes treated with nelfinavir on MBP-SagA, relative to a DMSO-treated control.

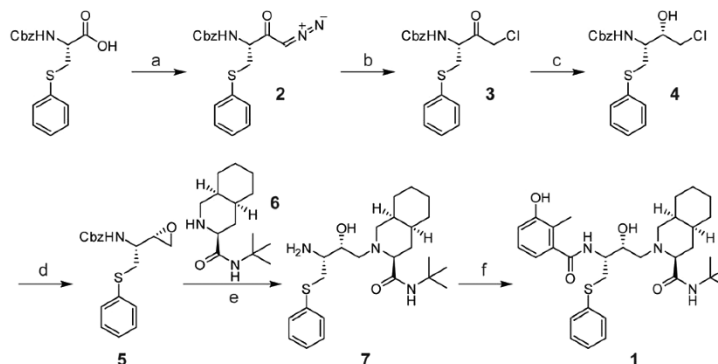
rather than through transcriptional regulation or by significantly perturbing other cellular processes.

Structure–Activity Relationships. A series of nelfinavir analogs was synthesized to gain a better understanding of the structure–activity relationships (SAR) for inhibition of SLS maturation. If nelfinavir were to interact with the SLS leader peptidase in a manner analogous to HIV protease, the secondary hydroxyl group would function as a tetrahedral intermediate mimic, as is the case for many aspartyl and metalloprotease inhibitors (Figure S7). Synthetic routes to nelfinavir have been thoroughly explored, and a route that allowed facile derivatization was adapted for this study (Scheme 1).³⁷ The efficacy of each analog at 25 μM was evaluated using the hemolysis assay (Table 1 and Table S3).

Nelfinavir has peptidomimetic features, including an *S*-phenyl-ethyl group intended to replace the side chain of phenylalanine commonly found in the P1 position of the HIV protease substrate. Since the putative P1 position of SagA is a much smaller glycine residue (Figure 1B, Gly23), we surmised that smaller substituents at this location on nelfinavir might improve the observed anti-SLS activity. Instead, removal of the side chain (8) abolished detectable activity. It is possible that the cleavage site is closer to the N-terminus, however, and that Pro21 or an aliphatic residue preceding it resides in the P1 position. Accordingly, analogs mimicking these amino acids (9, 10) were synthesized, but these compounds did not exhibit detectable inhibitory activity. Even retention of the phenyl ring

D

DOI: 10.1021/acs500843r
ACS Chem. Biol. XXXX, XXX, XXX–XXX

Scheme 1. Synthesis of Nelfinavir⁴⁴

^a(a) *t*-BuOCOCl, Et₃N, THF; CH₂N₃, Et₂O (62%). (b) HCl, Et₂O. (c) NaBH₄, THF (62% over two steps). (d) KOH, EtOH (91%). (e) 6, KOH, IPA, 80 °C (84%). (f) 3-hydroxy-2-methylbenzoic acid, EDC, HOBt, THF (66%).

in a phenylalanine mimic (11) was insufficient to maintain activity. Conversely, a tryptophan mimic (12) was inhibitory, albeit weakly, suggesting that a relatively large group in that position may be necessary for activity.

The lack of detectable activity with the series of P1 position analogs prevented the rigorous establishment of SAR. Modifications to other portions of the molecule were prepared in order to enhance the inhibitory activity to address this pitfall. As installation of the benzamide is the final step of the synthesis (Scheme 1), preparation of analogs at this location was convenient. Initial analogs revealed that neither the hydroxyl nor the methyl groups (13–15) were important for activity but that the ring itself was necessary (16). Replacement of the ring with bulkier naphthyl and cyclohexyl groups (17, 18) provided >5-fold increases in potency, indicating that this part of the pharmacophore may reside in a hydrophobic pocket not fully occupied by the single planar ring. However, increasing the size of this moiety to an anthracene (19) greatly reduced activity. In general, electron-deficient rings had improved activity relative to electron-rich rings (Table S3), although hydrophobicity appeared to be a more significant contributing factor toward potency.

Given the enhanced activity of the naphthylamide-bearing compound (17), this group was incorporated into the collection of P1 position analogs (Table 1; Table S3), increasing the activity of these analogs to detectable levels. The tryptophan mimic with the naphthylamide (20) displayed considerably increased potency relative to the 3-hydroxy-2-methylbenzamide analog (12). Within the naphthylamide series, the phenylalanine and leucine mimics (21, 22) were equivalently active to nelfinavir, while the glycine mimic (23) displayed weak SLS-inhibitory activity. The proline mimic (24) remained devoid of activity, possibly due to structural perturbations enforced by the ring. These data support a trend of larger substituents imparting higher activity that was foreshadowed by the initial SAR.

To further probe the pseudo-P1 position of nelfinavir, the stereochemical configurations of both the side chain and the secondary hydroxyl group were varied. Surprisingly, inverting stereocenters with this series of analogs (25–27) did not result

in large changes in activity. Many of the derivatives were still highly potent, suggesting that nelfinavir may not inhibit SagE through mimicking the proteolytic tetrahedral intermediate (Figure S7). This is further supported by the retention of activity in acetylated derivatives (28, 29). Overall, the sum of the SAR analysis resulted in the development of an analog with significantly improved potency (17, IC₅₀ = 1 μM). Additionally, the activity trends led us to believe that nelfinavir might serve as an inhibitor for the protease in other TOMM natural product biosynthetic clusters in which the precursor peptides do not contain a predicted Gly-Gly cleavage motif (Figure 1B).

Biosynthesis Inhibition of Other TOMMs. SLS is the best-studied member of a group of related cytolytins produced by a number of bacteria, including pathogens such as *L. monocytogenes* and *C. botulinum*.^{11,20} The SLS-like biosynthetic gene clusters in these strains are highly similar to that of *S. pyogenes* and include orthologs of *sagE* (Figure 1A, Figure S8). The SLS-like toxin from *L. monocytogenes*, listeriolysin S (LLS), is known to be expressed during oxidative stress; therefore, a strain with LLS under the control of a constitutive promoter was used in the blood lysis assay to determine if nelfinavir could also inhibit LLS production.⁴⁹ The strain was deficient in the production of an unrelated cytolytin, listeriolysin O (LLO), to ensure any hemolysis observed derived from LLS production. When treated with nelfinavir, this strain produced significantly less LLS (Figure 4A). The LLO⁻/LLS⁻ strain of *L. monocytogenes* was included as a negative control to demonstrate the observed hemolysis was indeed LLS-dependent (Figure 4A).

Similar to SLS and LLS, clostridiolysin S (CLS) is a hemolysin from *C. botulinum* as well as from certain strains of *C. sporogenes*, which are nearly identical to *C. botulinum* but do not produce botulinum toxin.²¹ The presence of the CLS cluster in an unsequenced strain of *C. sporogenes* (ATCC 19404) known to be hemolytic on blood agar was confirmed by PCR amplification of *clsC* and *clsD* (the cyclodehydratase genes). When *C. sporogenes* was grown in the presence of nelfinavir, the production of CLS was significantly reduced (Figure 4B). The inhibition of not only SLS but also LLS and

E

DOI: 10.1021/acs500843r
ACS Chem. Biol. XXXX, XXX, XXX–XXX

Table 1. Relative Activity of Nelfinavir Analogs for SLS Inhibition

$R^1 = A =$ $\text{Core} = B =$

Compound	R ¹	Core	Relative Activity ^a	Compound	R ¹	Core	Relative Activity ^a
nelfinavir (1)			++	19			+
8			-	20			+++
9			-	21			++
10			-	22			++
11			-	23			+
12			+	24			-
13			++	25			++
14			++	26			+++
15			++	27			+++
16			+	28			+++
17			+++	29			++
18			+++				

^aThe activity of each analog is reported qualitatively relative to nelfinavir due to variability in commercial blood lots and extraction effectiveness. Inhibitory activity that is >3-fold that of nelfinavir is designated as (+++); activity that is equal to nelfinavir is (++); detectable activity that is <3-fold that of nelfinavir is (+); nondetectable activity is denoted (-).

F

DOI: 10.1021/cb500843r
ACS Chem. Biol. XXXX, XXX, XXX–XXX

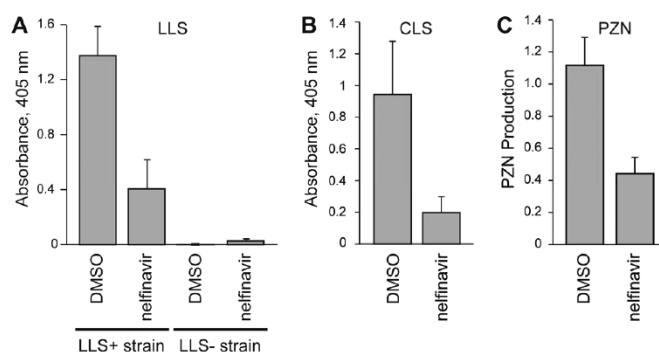


Figure 4. Inhibition of the biosynthesis of other TOMMs by nelfinavir. (A) Lytic activity of extracts from an LLS-producing strain (LLS+) of *L. monocytogenes* treated with nelfinavir relative to a DMSO control ($n = 3$). Extracts from a separate strain with *llsA* deleted (LLS-) were also used as a control. (B) Lytic activity of extracts from a CLS-producing strain of *C. sporogenes* treated with nelfinavir relative to a DMSO control ($n = 3$). (C) Production of PZN by *B. amyloliquefaciens* treated with nelfinavir relative to a DMSO control ($n = 4$). Nelfinavir was used at 50 μM in all cases. Excluding LLS- negative control, P values < 0.05 were obtained for all nelfinavir-treated samples relative to the corresponding DMSO positive controls.

CLS production suggests that nelfinavir would likely inhibit the production of additional TOMM cytolytins.

TOMM biosynthetic gene clusters are widespread among bacteria and archaea, and the products of the vast majority of these clusters have not been structurally or functionally characterized.⁶ A bioinformatic analysis of TOMM clusters revealed that 22% (328 out of 1520 as of October 2014) contained a CPBP family member within the cluster, many of which are predicted or known to have noncytolytic products. The presence of a CPBP member in these clusters led us to postulate that nelfinavir would also inhibit TOMM production in these cases. Plantazolicin (PZN) is a TOMM produced by *Bacillus amyloliquefaciens* FZB42 with highly selective antibacterial activity against *Bacillus anthracis* and has a *sagE*-like gene (*bamE*) in its biosynthetic gene cluster (Figure 1A, Figure S8).⁵⁰ The structure of PZN has been determined; thus the precise cleavage site is known (after Ala27, Figure 1B), although no biochemical studies have directly linked BamE to leader peptide cleavage. Methanolic surface extracts from *B. amyloliquefaciens* were analyzed by liquid chromatography–mass spectrometry, and the nelfinavir-treated cultures were found to produce significantly less PZN than a DMSO control (Figure 4C). Unlike the cytolytins, PZN can be readily observed and quantified by mass spectrometry, which permitted confirmation of the nelfinavir dose-dependent inhibition (Figure S9). The nelfinavir-dependent inhibition in divergent organisms of additional TOMM cytolytins, as well as a functionally distinct antimicrobial TOMM, not only supports the assignment of the CPBP protein being responsible for leader peptide removal during maturation but also suggests that nelfinavir, and analogs thereof, could be generally useful for inhibition of TOMM production in a large number of hitherto uncharacterized clusters.

DISCUSSION

In this work, the FDA-approved HIV protease inhibitor nelfinavir was repurposed as the first small molecule inhibitor of SLS production in *S. pyogenes*, displaying low micromolar activity. Nelfinavir was identified as a lead compound by

leveraging the extensive basic and clinical research data accumulated on the effects of the drug. Lipodystrophy, a known side effect of nelfinavir and several other HIV protease inhibitors, had been previously linked to the off-target inhibition of the CaaX protease ZMPSTE24. We surmised that the HIV protease inhibitors would also inhibit SagE due to its homology with CaaX proteases, allowing us to rapidly identify a lead compound for the inhibition of SLS production. This strategy for lead identification negated the need for high-throughput screening or for a crystal structure of the target for *in silico* and structure-based design. Utilizing a drug with synthetic routes that have been thoroughly explored also greatly accelerated the creation of analogs for SAR efforts that yielded compound 17, with an improved IC_{50} value of 1 μM .

Nelfinavir and related compounds are inhibitors of SLS biosynthesis, most likely through inhibition of proteolytic cleavage of the protoxin by the CPBP family member SagE. Although *in vitro* reconstitution was unsuccessful, the necessity for SagE during SLS production in the multiplasmid expression system in *E. coli* provides considerable evidence for its role in proteolytic processing. Like many CPBP members, SagE is commonly referred to as an immunity protein in the literature,^{14,15} but inhibition by nelfinavir did not have any effect on the growth of *S. pyogenes*, providing evidence that SagE is not involved in self-immunity. Alternatively, compensatory mutations that abolish SLS production may arise when SagE is inactivated, as has been previously suggested.¹⁵ The original annotation of SagE as an immunity protein stems from its similarity to PlnP from *Lactobacillus plantarum*. PlnP and several related proteins are found downstream of bacteriocin structural genes in *L. plantarum* and have been shown to provide immunity to the antibacterial effect of the respective bacteriocins.³⁵ Yet, unlike these bacteriocins, SLS has not been demonstrated to possess any antibacterial activity against intact *S. pyogenes* cells. A large buildup of intracellular SLS might result in toxicity, but this would be expected to affect any *S. pyogenes* strains in which the transport machinery is inactivated as well, which has not been observed.¹⁴ Furthermore, treatment of SagA with the cyclization machinery (SagBCD) *in vitro*

G

DOI: 10.1021/acs500843r
ACS Chem. Biol. XXXX, XXX, XXX–XXX

results in a lytic entity without cleavage of the leader peptide, indicating that while proteolysis is required for cellular export, it is unnecessary for lytic activity.²⁰ These observations lead us to conclude that the principal function of SagE is to proteolytically mature SLS.

In addition to experimentally supporting a biochemical role for SagE, we have also addressed the mechanism of proteolysis and the probable inhibition by nelfinavir. CPBP family members have been postulated to function through a zinc metalloprotease mechanism based on the presence of two glutamates, two histidines, and an asparagine residue that are conserved across the family (Figure S8).³³ Mutation of these residues in the eukaryotic type II CaaX protease Ras-converting enzyme (Rce1) has also demonstrated that these residues are critical for activity.³³ We found that Glu131 and Glu132 were critical for the activity of SagE. Thus, it was initially unexpected that proteolysis was inhibited by pepstatin, a general aspartyl protease inhibitor, but not by bestatin, a metalloprotease inhibitor (Figure S3). However, a recent report detailing the crystal structure of Rce1 from *Methanococcus maripaludis* provides compelling evidence that this family of proteases actually functions through a novel glutamate-dependent mechanism (Figure S10).⁴⁷ The authors hypothesized that a glutamate residue extending into the active site is responsible for deprotonating a water molecule, activating it for nucleophilic attack. Given the similarities between this proposed mechanism and the mechanism of aspartyl proteases, it is perhaps unsurprising that a CaaX protease homologue would be inhibited by aspartyl protease inhibitors, including nelfinavir.

In addition to increasing the potency of inhibition with compound 17, our synthetic effort also yielded information on how nelfinavir may interact with SagE. The SAR analysis revealed that a rather large side chain in the pseudo P1 position of the structure was required for potent activity, with the original *S*-phenyl-ethyl group displaying the greatest inhibition. This result was not anticipated, given that the P1 residue in the SagA substrate is putatively glycine. One possible explanation for this discrepancy is that the catalytic site architecture of SagE is highly conserved with the type II CaaX proteases, which normally cleave substrates with a prenylated cysteine in the P1 position. In this case, nelfinavir would be an ideal fit for the active site, as the core of the molecule closely resembles a cysteine with a hydrophobic group appended. An alternative explanation is that nelfinavir does not bind in the active site in the expected fashion (with the secondary hydroxyl interacting with the catalytic residues, Figure S7) or does not bind in the active site at all. These possibilities are supported by the potent activity of several analogs with different stereochemical configurations, as these molecules would likely be forced into different conformations that do not allow the same favorable interactions with active site residues. In this scenario, binding of nelfinavir to the membrane protease may be driven by hydrophobic interactions. This explanation would also account for the attenuated activity of saquinavir, which is nearly identical to nelfinavir except for the presence of a more hydrophilic group at the benzamide position (Figure S4, XLogP3 values from PubChem for nelfinavir and saquinavir are 5.7 and 4.2, respectively). Comparisons to the other HIV protease inhibitors do not yield much additional information due to their low similarity to nelfinavir, as reflected in the Tanimoto similarity coefficients (Table S4).

Unequivocal confirmation of the bacterial target of nelfinavir (and analogs) was unfortunately not possible in the present study, which can be attributed to the technical challenges inherent to the study of integral membrane proteins. Nelfinavir also likely interacts with multiple targets in *S. pyogenes*, as the drug is known to have multiple off-target effects in humans.⁴³ A significant body of work in mammalian cell lines has demonstrated that nelfinavir displays promiscuous activity, such as interruption of Akt signaling and inhibition of the proteasome.⁵¹ Target promiscuity likely also exists in bacteria, so possible interactions of nelfinavir with additional targets cannot be ruled out. Thus, nelfinavir may be inhibiting additional participants that indirectly result in the inhibition of SLS production in a mechanism unrelated to proteolysis. Further targets of nelfinavir may also exist that do not result in observable phenotypes. However, nelfinavir also inhibited the biosynthesis of other natural products that include a CPBP family member in the gene cluster (i.e., LLS, CLS, and PZN). The fact that this inhibition occurred in a range of disparate bacterial species decreases the probability that another protease is responsible and provides substantial, albeit indirect, support that SagE is the primary target of nelfinavir in *S. pyogenes*. Finally, the HIV protease inhibitors found to inhibit SLS production parallel those capable of inhibiting the human CaaX protease ZMPSTE24 (Figure S4),^{44,45} providing additional evidence that nelfinavir and its analogs inhibit SLS production by blocking the action of SagE.

Conclusion. Despite their prevalence, prokaryotic members of the CPBP family have not yet been thoroughly investigated. Many of the family members are incorrectly annotated or have a predicted function based solely on distant homology. The discovery of nelfinavir as an inhibitor of CPBPs has provided evidence that SagE functions as a protease and will aid in the assignment of functions to other family members, including those in other human pathogens such as *S. aureus*. Additionally, nelfinavir is the first reported inhibitor of the production of SLS and related toxins. Nelfinavir and improved analogs will provide new tools to investigate toxin function without the need to create genetic deletions while also allowing for temporal control over toxin production. Reversible control of SLS production with nelfinavir analogs will also help to clarify the precise contribution of SLS to virulence in *in vivo* models of infection and may open the door to the development of virulence-targeting strategies for the control of *S. pyogenes* infections. Finally, the chemical knockdown effect of nelfinavir can be utilized for the discovery of natural products from the 22% of TOMM gene clusters that contain a CPBP family member, potentially accelerating the structural and functional characterization of these compounds.

■ ASSOCIATED CONTENT

Supporting Information

Materials and methods (contains list of primers used), hydrophathy plot of SagE, cleavage of Flag-tagged SagA with the *S. pyogenes* membrane fraction, inhibition of SagA proteolysis by general-mechanism based inhibitors, structures of the FDA-approved HIV protease inhibitors, growth effects of HIV protease inhibitors on *S. pyogenes*, minimum inhibitory concentrations of nelfinavir for various bacteria, transmission electron microscopy of nelfinavir-treated *S. pyogenes*, changes in virulence factor expression during nelfinavir treatment, the aspartyl and metalloprotease common intermediate, relative activity of additional nelfinavir analogs for SLS inhibition,

H

DOI: 10.1021/acs00843r
ACS Chem. Biol. XXXX, XXX, XXX–XXX

amino acid sequence alignment of select bacterial CPBPs, dose-dependent inhibition of PZN by nelfinavir, proposed mechanism for glutamate-dependent proteases, Tanimoto similarity scores for the FDA-approved HIV protease inhibitors, and compound synthesis. This material is available free of charge via the Internet at <http://pubs.acs.org>.

AUTHOR INFORMATION

Corresponding Author

*Phone: 1-217-333-1345. Fax: 1-217-333-0508. E-mail: douglasm@illinois.edu.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We would like to thank P. Cotter (Teagasc, Moorepark Food Research Center, Fermoy, Co. Cork, Ireland) for the gift of engineered *L. monocytogenes* strains. We would like to thank B. White (University of Illinois at Urbana–Champaign, Urbana, IL 61801, USA) for the use of his anaerobic chamber. All FDA-approved HIV protease inhibitors were obtained through the NIH AIDS Research and Reference Reagent Program, Division of AIDS (NIAID). Electron microscopy was carried out in part in the Frederick Seitz Materials Research Laboratory Central Research Facilities, University of Illinois. This research was supported in part by the NIH Director's New Innovator Award Program (DP2 OD008463, to D.A.M.). T.M. and C.D.D. were supported in part by fellowships from the Department of Chemistry at the University of Illinois at Urbana–Champaign and the NIH Chemical Biology Interface Training Program (T32 GM070421). E.M.M. was formerly funded by the Irish Research Council for Science, Engineering and Technology through University College Cork, Ireland, and received travel-related funding from the UK Society for General Microbiology and Science Foundation Ireland.

REFERENCES

- (1) Arison, P. G., Bibb, M. J., Bierbaum, G., Bowers, A. A., Bugni, T. S., Bulaj, G., Camarero, J. A., Campopiano, D. J., Challis, G. L., Clardy, J., Cotter, P. D., Craik, D. J., Dawson, M., Dittmann, E., Donadio, S., Dorrestein, P. C., Entian, K. D., Fischbach, M. A., Garavelli, J. S., Goransson, U., Gruber, C. W., Haft, D. H., Hemscheidt, T. K., Hertweck, C., Hill, C., Horswill, A. R., Jaspars, M., Kelly, W. L., Klinman, J. P., Kuipers, O. P., Link, A. J., Liu, W., Marahiel, M. A., Mitchell, D. A., Moll, G. N., Moore, B. S., Muller, R., Nair, S. K., Nes, I. F., Norris, G. E., Olivera, B. M., Onaka, H., Patchett, M. L., Piel, J., Reaney, M. J., Rebuffat, S., Ross, R. P., Sahl, H. G., Schmidt, E. W., Selsted, M. E., Severinov, K., Shen, B., Sivonen, K., Smith, L., Stein, T., Sussmuth, R. D., Tagg, J. R., Tang, G. L., Truman, A. W., Vederas, J. C., Walsh, C. T., Walton, J. D., Wenzel, S. C., Willey, J. M., and van der Donk, W. A. (2013) Ribosomally synthesized and post-translationally modified peptide natural products: overview and recommendations for a universal nomenclature. *Nat. Prod. Rep.* 30, 108–160.
- (2) Bagley, M. C., Dale, J. W., Merritt, E. A., and Xiong, X. (2005) Thiopetide antibiotics. *Chem. Rev.* 105, 685–714.
- (3) Chatterjee, C., Paul, M., Xie, L., and van der Donk, W. A. (2005) Biosynthesis and mode of action of lantibiotics. *Chem. Rev.* 105, 633–684.
- (4) Maksimov, M. O., Pan, S. J., and James Link, A. (2012) Lasso peptides: structure, function, biosynthesis, and engineering. *Nat. Prod. Rep.* 29, 996–1006.
- (5) Sivonen, K., Leikoski, N., Fewer, D. P., and Jokela, J. (2010) Cyanobactin-ribosomal cyclic peptides produced by cyanobacteria. *Appl. Microbiol. Biot.* 86, 1213–1225.
- (6) Melby, J. O., Nard, N. J., and Mitchell, D. A. (2011) Thiazole/oxazole-modified microcins: complex natural products from ribosomal templates. *Curr. Opin. Chem. Biol.* 15, 369–378.
- (7) Oman, T. J., and van der Donk, W. A. (2010) Follow the leader: the use of leader peptides to guide natural product biosynthesis. *Nat. Chem. Biol.* 6, 9–18.
- (8) Dunbar, K. L., Melby, J. O., and Mitchell, D. A. (2012) YcaO domains use ATP to activate amide backbones during peptide cyclodehydrations. *Nat. Chem. Biol.* 8, 569–575.
- (9) Dunbar, K. L., and Mitchell, D. A. (2013) Insights into the mechanism of peptide cyclodehydrations achieved through the chemoenzymatic generation of amide derivatives. *J. Am. Chem. Soc.* 135, 8692–8701.
- (10) Melby, J. O., Li, X., and Mitchell, D. A. (2014) Orchestration of enzymatic processing by thiazole/oxazole-modified microcin dehydrogenases. *Biochemistry* 53, 413–422.
- (11) Molloy, E. M., Cotter, P. D., Hill, C., Mitchell, D. A., and Ross, R. P. (2011) Streptolysin S-like virulence factors: the continuing saga. *Nat. Rev. Microbiol.* 9, 670–681.
- (12) Cunningham, M. W. (2000) Pathogenesis of group A streptococcal infections. *Clin. Microbiol. Rev.* 13, 470–511.
- (13) Carapetis, J. R., Steer, A. C., Mulholland, E. K., and Weber, M. (2005) The global burden of group A streptococcal diseases. *Lancet Infect. Dis.* 5, 685–694.
- (14) Nizet, V., Beall, B., Bast, D. J., Datta, V., Kilburn, L., Low, D. E., and De Azavedo, J. C. (2000) Genetic locus for streptolysin S production by group A streptococcus. *Infect. Immun.* 68, 4245–4254.
- (15) Datta, V., Myskowski, S. M., Kwinn, L. A., Chiem, D. N., Varki, N., Kansal, R. G., Koth, M., and Nizet, V. (2005) Mutational analysis of the group A streptococcal operon encoding streptolysin S and its virulence role in invasive infection. *Mol. Microbiol.* 56, 681–695.
- (16) Betschel, S. D., Borgia, S. M., Barg, N. L., Low, D. E., and De Azavedo, J. C. (1998) Reduced virulence of group A streptococcal Tn916 mutants that do not produce streptolysin S. *Infect. Immun.* 66, 1671–1679.
- (17) Fontaine, M. C., Lee, J. J., and Kehoe, M. A. (2003) Combined contributions of streptolysin O and streptolysin S to virulence of serotype M5 *Streptococcus pyogenes* strain Manfredo. *Infect. Immun.* 71, 3857–3865.
- (18) James, L., and McFarland, R. B. (1971) An epidemic of pharyngitis due to a nonhemolytic group A streptococcus at low air force base. *N. Engl. J. Med.* 284, 750–752.
- (19) Yoshino, M., Murayama, S. Y., Sunaoshi, K., Wajima, T., Takahashi, M., Masaki, J., Kurokawa, I., and Ubukata, K. (2010) Nonhemolytic *Streptococcus pyogenes* isolates that lack large regions of the sag operon mediating streptolysin S production. *J. Clin. Microbiol.* 48, 635–638.
- (20) Lee, S. W., Mitchell, D. A., Markley, A. L., Hensler, M. E., Gonzalez, D., Wohlrab, A., Dorrestein, P. C., Nizet, V., and Dixon, J. E. (2008) Discovery of a widely distributed toxin biosynthetic gene cluster. *Proc. Natl. Acad. Sci. U.S.A.* 105, 5879–5884.
- (21) Gonzalez, D. J., Lee, S. W., Hensler, M. E., Markley, A. L., Daresh, S., Mitchell, D. A., Bandeira, N., Nizet, V., Dixon, J. E., and Dorrestein, P. C. (2010) Clostridiolysin S, a post-translationally modified biotoxin from *Clostridium botulinum*. *J. Biol. Chem.* 285, 28220–28228.
- (22) Bernheimer, A. W. (1967) Physical Behavior of Streptolysin S. *Bacteriol.* 93, 2024–2025.
- (23) Jack, R. W., Tagg, J. R., and Ray, B. (1995) Bacteriocins of gram-positive bacteria. *Microbiol. Rev.* 59, 171–200.
- (24) Todd, E. W. (1938) The differentiation of two distinct serological varieties of streptolysin, streptolysin O and streptolysin S. *Pathol. Bacteriol.* 47, 423–445.
- (25) Marmorek, A. (1895) Le streptocoque et le sérum antistreptococcique. *Ann. Inst. Pasteur* 9, 593–620.
- (26) Rasko, D. A., and Sperandio, V. (2010) Anti-virulence strategies to combat bacteria-mediated disease. *Nat. Rev. Drug Discovery* 9, 117–128.

- (27) Cegelski, L.; Marshall, G. R.; Eldridge, G. R.; and Hultgren, S. J. (2008) The biology and future prospects of antiviral therapies. *Nat. Rev. Microbiol.* 6, 17–27.
- (28) Baruch, M.; Belotserkovsky, L.; Hertzog, B. B.; Ravins, M.; Dov, E.; McIver, K. S.; Le Breton, Y. S.; Zhou, Y.; Cheng, C. Y.; and Hanski, E. (2014) An extracellular bacterial pathogen modulates host metabolism to regulate its own sensing and proliferation. *Cell* 156, 97–108.
- (29) Pei, J., and Grishin, N. V. (2001) Type II CAAX prenyl endopeptidases belong to a novel superfamily of putative membrane-bound metalloproteases. *Trends Biochem. Sci.* 26, 275–277.
- (30) Pei, J.; Mitchell, D. A.; Dixon, J. E.; and Grishin, N. V. (2011) Expansion of type II CAAX proteases reveals evolutionary origin of gamma-secretase subunit APH-1. *J. Mol. Biol.* 410, 18–26.
- (31) Bergo, M. O.; Ambroziak, P.; Gregory, C.; George, A.; Otto, J. C.; Kim, E.; Nagase, H.; Casey, P. J.; Balmain, A.; and Young, S. G. (2002) Absence of the CAAX endoprotease Rce1: effects on cell growth and transformation. *Mol. Cell. Biol.* 22, 171–181.
- (32) Bergo, M. O.; Wahlstrom, A. M.; Fong, L. G.; and Young, S. G. (2008) Genetic analyses of the role of RCE1 in RAS membrane association and transformation. *Methods Enzymol.* 438, 367–389.
- (33) Plummer, L. J.; Hildebrandt, E. R.; Porter, S. B.; Rogers, V. A.; McCracken, J.; and Schmidt, W. K. (2006) Mutational analysis of the ras converting enzyme reveals a requirement for glutamate and histidine residues. *J. Biol. Chem.* 281, 4596–4605.
- (34) Dolence, J. M.; Steward, L. E.; Dolence, E. K.; Wong, D. H.; and Poulter, C. D. (2000) Studies with recombinant *Saccharomyces cerevisiae* CaaX prenyl protease Rce1p. *Biochemistry* 39, 4096–4104.
- (35) Kjos, M.; Snipen, L.; Salehian, Z.; Nes, L. F.; and Diep, D. B. (2010) The Abi Proteins and Their Involvement in Bacteriocin Self-Immunity. *J. Bacteriol.* 192, 2068–2076.
- (36) Clayton, E. M.; Hill, C.; Cotter, P. D.; and Ross, R. P. (2011) Real-time PCR assay to differentiate *Listeria monocytogenes* S-positive and -negative strains of *Listeria monocytogenes*. *Appl. Environ. Microb.* 77, 163–171.
- (37) Kaldor, S. W.; Kalish, V. J.; Davies, J. F., 2nd; Shetty, B. V.; Fritz, J. E.; Appelt, K.; Burgess, J. A.; Campanale, K. M.; Chirgadze, N. Y.; Clawson, D. K.; Dressman, B. A.; Hatch, S. D.; Khalil, D. A.; Kosa, M. B.; Lubbehusen, P. P.; Muesing, M. A.; Patick, A. K.; Reich, S. H.; Su, K. S.; and Tatlock, J. H. (1997) Viracept (nelfinavir mesylate, AG1343): a potent, orally bioavailable inhibitor of HIV-1 protease. *J. Med. Chem.* 40, 3979–3985.
- (38) Albizati, K. F.; Babu, S.; Birchler, A.; Busse, J. K.; Fugett, M.; Grubbs, A.; Haddach, A.; Pagan, M.; Potts, B.; Remarchuk, T.; Rieger, D.; Rodriguez, R.; Shanley, J.; Szendroi, R.; Tibbetts, T.; Whitten, K.; and Borer, B. C. (2001) A synthesis of the HIV-protease inhibitor nelfinavir from D-tartaric acid. *Tetrahedron Lett.* 42, 6481–6485.
- (39) Ma, D.; Zou, B.; Zhu, W.; and Xu, H. D. (2002) A short synthesis of the HIV-protease inhibitor nelfinavir via a diastereoselective addition of ammonia to the alpha,beta-unsaturated sulfoxide derived from (R)-glyceraldehyde acetonide. *Tetrahedron Lett.* 43, 8511–8513.
- (40) Viklund, H.; Bernsel, A.; Skwark, M.; and Elofsson, A. (2008) SPOCTOPUS: a combined predictor of signal peptides and membrane protein topology. *Bioinformatics* 24, 2928–2929.
- (41) Markley, A. L.; Jensen, E. R.; and Lee, S. W. (2012) An *Escherichia coli*-based bioengineering strategy to study streptolysin S biosynthesis. *Anal. Biochem.* 420, 191–193.
- (42) Alouf, J. E. (1980) Streptococcal toxins (streptolysin O, streptolysin S, erythrogenic toxin). *Pharmacol. Ther.* 11, 661–717.
- (43) Caron, M.; Auclair, M.; Sterlingot, H.; Kornprobst, M.; and Capeau, J. (2003) Some HIV protease inhibitors alter lamin A/C maturation and stability, SREBP-1 nuclear localization and adipocyte differentiation. *AIDS* 17, 2437–2444.
- (44) Coffinier, C.; Hudon, S. E.; Farber, E. A.; Chang, S. Y.; Hrycyna, C. A.; Young, S. G.; and Fong, L. G. (2007) HIV protease inhibitors block the zinc metalloproteinase ZMPSTE24 and lead to an accumulation of prelamin A in cells. *Proc. Natl. Acad. Sci. U.S.A.* 104, 13432–13437.
- (45) Coffinier, C.; Hudon, S. E.; Lee, R.; Farber, E. A.; Nubumori, C.; Miner, J. H.; Andres, D. A.; Spielmann, H. P.; Hrycyna, C. A.; Fong, L. G.; and Young, S. G. (2008) A potent HIV protease inhibitor, darunavir, does not inhibit ZMPSTE24 or lead to an accumulation of farnesyl-prelamin A in cells. *J. Biol. Chem.* 283, 9797–9804.
- (46) Quigley, A.; Dong, Y. Y.; Pike, A. C.; Dong, L.; Shrestha, L.; Berridge, G.; Stansfeld, P. J.; Sansom, M. S.; Edwards, A. M.; Bountra, C.; von Delft, F.; Bullock, A. N.; Burgess-Brown, N. A.; and Carpenter, E. P. (2013) The structural basis of ZMPSTE24-dependent laminopathies. *Science* 339, 1604–1607.
- (47) Manolaridis, L.; Kulkarni, K.; Dodd, R. B.; Ogasawara, S.; Zhang, Z.; Bineva, G.; O'Reilly, N.; Hanrahan, S. J.; Thompson, A. J.; Cronin, N.; Iwata, S.; and Barford, D. (2013) Mechanism of farnesylated CAAX gamma processing by the intramembrane protease Rce1. *Nature* 504, 301–305.
- (48) Bernheimer, A. W. (1949) Formation of a bacterial toxin (streptolysin S) by resting cells. *J. Exp. Med.* 90, 373–392.
- (49) Cotter, P. D.; Draper, L. A.; Lawton, E. M.; Daly, K. M.; Groeger, D. S.; Casey, P. G.; Ross, R. P.; and Hill, C. (2008) Listeriolysin S, a novel peptide haemolysin associated with a subset of lineage I *Listeria monocytogenes*. *PLoS Pathog.* 4, e1000144.
- (50) Molohon, K. J.; Melby, J. O.; Lee, J.; Evans, B. S.; Dunbar, K. L.; Bumpus, S. B.; Kelleher, N. L.; and Mitchell, D. A. (2011) Structure determination and interception of biosynthetic intermediates for the plantazolicin class of highly discriminating antibiotics. *ACS Chem. Biol.* 6, 1307–1313.
- (51) Gantt, S.; Casper, C.; and Ambinder, R. F. (2013) Insights into the broad cellular effects of nelfinavir and the HIV protease inhibitors supporting their role in cancer treatment and prevention. *Curr. Opin. Oncol.* 25, 495–502.

A.3 Undecaprenyl Diphosphate Synthase Inhibitors: Antibacterial Drug Leads

This chapter was reprinted with permission from Sinko, Wang, Zhu, Feixas, Cox, Mitchell, Oldfield, and McCammon (Sinko, *et al.* 2014).

I aided in the collection, interpretation and figure generation for the MIC and synergistic activity of the compounds.

Undecaprenyl Diphosphate Synthase Inhibitors: Antibacterial Drug Leads

William Sinko,^{*,†,‡} Yang Wang,[§] Wei Zhu,[§] Yonghui Zhang,[§] Ferran Feixas,[‡] Courtney L. Cox,^{||,⊥} Douglas A. Mitchell,^{§,||,⊥} Eric Oldfield,[§] and J. Andrew McCammon^{‡,§,#}

[†]Pharmacology Department, University of California San Diego, La Jolla, California 92093-0365, United States

[‡]Department of Chemistry & Biochemistry, Department of Pharmacology, and NSF Center for Theoretical Biological Physics, University of California San Diego, La Jolla, California 92093-0365, United States

[§]Department of Chemistry, University of Illinois, Urbana, Illinois 61801, United States

^{||}Department of Microbiology, University of Illinois, Urbana, Illinois 61801, United States

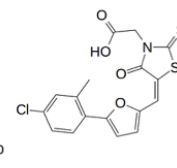
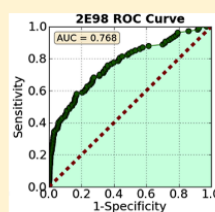
[⊥]Institute for Genomic Biology, University of Illinois, Urbana, Illinois 61801, United States

[#]Howard Hughes Medical Institute, University of California San Diego, La Jolla, California 92093-0365, United States

Supporting Information

ABSTRACT: There is a significant need for new antibiotics due to the rise in drug resistance. Drugs such as methicillin and vancomycin target bacterial cell wall biosynthesis, but methicillin-resistant *Staphylococcus aureus* (MRSA) and vancomycin-resistant *Enterococci* (VRE) have now arisen and are of major concern. Inhibitors acting on new targets in cell wall biosynthesis are thus of particular interest since they might also restore sensitivity to existing drugs, and the *cis*-prenyl transferase undecaprenyl diphosphate synthase (UPPS), essential for lipid I, lipid II, and thus, peptidoglycan biosynthesis, is one such target. We used 12 UPPS crystal structures to validate virtual screening models and then assayed 100 virtual hits (from 450,000 compounds) against UPPS from *S. aureus* and *Escherichia coli*.

The most promising inhibitors ($IC_{50} \sim 2 \mu M$, $K_i \sim 300$ nM) had activity against MRSA, *Listeria monocytogenes*, *Bacillus anthracis*, and a vancomycin-resistant *Enterococcus* sp. with MIC or IC_{50} values in the 0.25–4 $\mu g/mL$ range. Moreover, one compound (**1**), a rhodanine with close structural similarity to the commercial diabetes drug epalrestat, exhibited good activity as well as a fractional inhibitory concentration index (FICI) of 0.1 with methicillin against the community-acquired MRSA USA300 strain, indicating strong synergism.



INTRODUCTION

The need for new antibiotics has arisen due to the widespread resistance to current drugs.¹ Despite this need, the antibiotic pipeline in the past few decades has been relatively dry in terms of new antibacterial classes when compared with progress against other diseases.² One strategy to fight bacterial resistance is to inhibit enzymes that are not the targets of current antibiotics but, instead, act in the same pathways as existing drugs since this might enable the restoration of drug sensitivity via combination therapy. Undecaprenyl diphosphate synthase (UPPS) is one such target. The undecaprenyl diphosphate product (UPP) is essential for bacterial cell growth because of its role in the formation of bacterial cell wall peptidoglycan,^{1,3} Scheme 1, and it is not produced by humans.^{2,4}

SmithKline Beecham screened their compound collection against UPPS but reported no chemically tractable low micromolar hits.⁵ Novartis pursued tetramic and tetroneic acids and dihydropyridin-2-ones, but noted issues associated with human serum albumin binding and a lack of in vivo activity.^{6,7} Previously, we reported several potent UPPS inhibitors together with X-ray crystallographic (or modeled)

binding modes for a variety of chemical classes including lipophilic bisphosphonates,⁸ phthalic acids,⁹ diketo acids,¹⁰ anthranilic acids, benzoic acids,^{11,12} aryl phosphonates, bis-amines, and bis-amidines.¹² The most promising of these compounds, a bis-amidine, was shown to have potent activity in biochemical assays, in cellular assays, and in a murine model of MRSA infection.¹²

Since UPPS must bind multiple substrates (IPP, FPP, or more elongated prenyl-PP intermediates) and many inhibitors are to some degree substrate mimics, it is common to observe numerous inhibitors simultaneously bound to UPPS, with up to 4 binding sites being occupied.⁸ However, it is unclear whether inhibitory activity is due to binding to one specific site or to multiple sites. It has been shown that some inhibitors occupy only site 4, an allosteric site distant from the catalytic center, while others bind to site I, the substrate binding site,¹² complicating docking studies and, regardless of the inhibitor-binding mode, the flexibility of UPPS creates challenges for

Received: March 24, 2014

Published: May 14, 2014

Scheme 1. Undecaprenyl Diphosphate Synthase Reaction and Relationship of UPP to Bacterial Cell Wall Biosynthesis

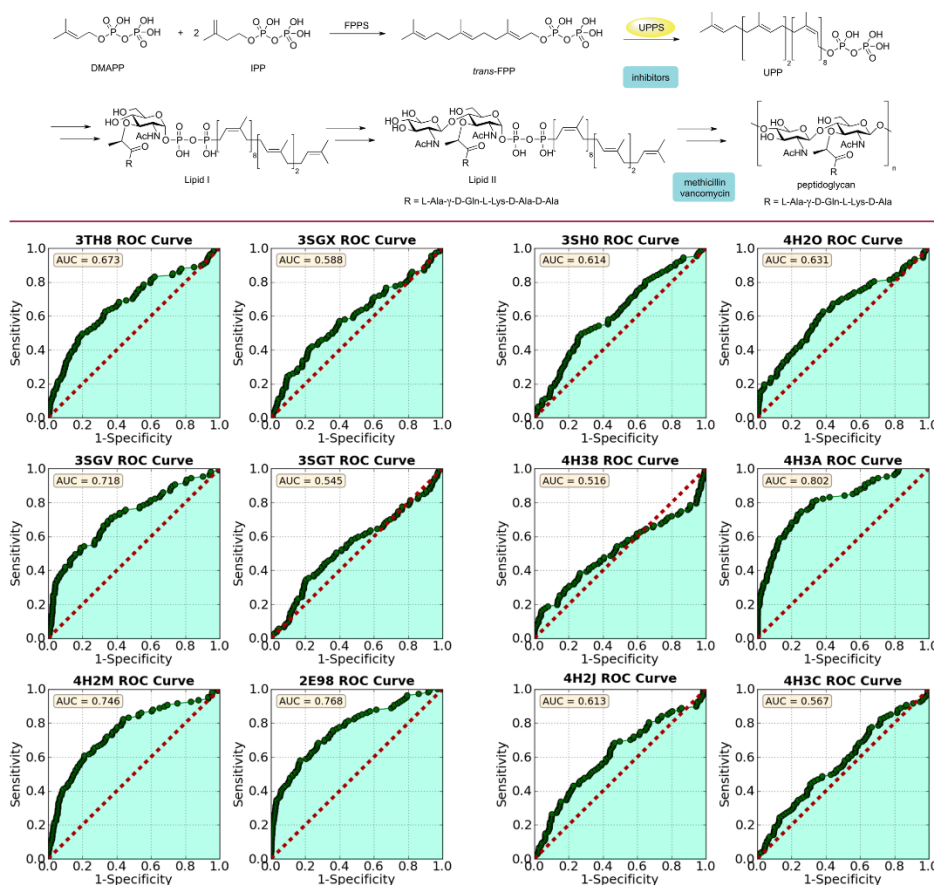


Figure 1. AUC/ROC curves for 12 EcUPPS crystal structures. 4H3A and 2E98 were chosen for further study.

virtual screening. Here, to help reduce these problems we employed the 12 crystallographic structures described in previous work^{8,12} to select those that provided maximal enrichment in retrospective virtual screening studies. We then made prospective predictions using these structures, leading to novel UPPS inhibitors, some with promising antibacterial activity.

METHODS AND MATERIALS

Computational Aspects. Following the methods described in previous work,¹² we docked 112 known UPPS inhibitors having IC_{50} values $<100 \mu M$, together with 1000 decoys from the Schrödinger decoy collection (having an average molecular weight of 400 Da), to *Escherichia coli* UPPS (hereafter, EcUPPS). Docking was performed using the Glide^{13–15} program, and compounds were ranked by their Glide XP score. The proteins were prepared by stripping water and

ligand molecules, capping, and neutralizing any unsolved loops, followed by preparation with the Schrödinger protein preparation wizard using standard parameters.¹⁶ After docking, compounds were ranked by their docking score, and then area under the curve (AUC) analyses were performed. Retrospective enrichment was quite good for 2/12 structures (PDB codes 2E98 and 4H3A), so we docked into these structures for the prospective studies (Figure 1). 2E98 is an EcUPPS X-ray structure containing four lipophilic bisphosphonates (BPH-629; $IC_{50} \sim 300$ nM), which bind to sites 1–4, one inhibitor to each site.⁸ 4H3A is an EcUPPS structure containing a diketone acid inhibitor (BPH-1330) which has a $2 \mu M$ IC_{50} , and the inhibitor binds (in the solid state) only to site 4.^{10,12} These structures thus have significant differences: only site 4 is occupied in 4H3A, while in 2E98, all four sites are occupied and the protein is in a “wide-open” conformation (Figure 2).

To find new inhibitors, we began with a library of ~450,000 commercially available compounds, the ChemBridge Experimental

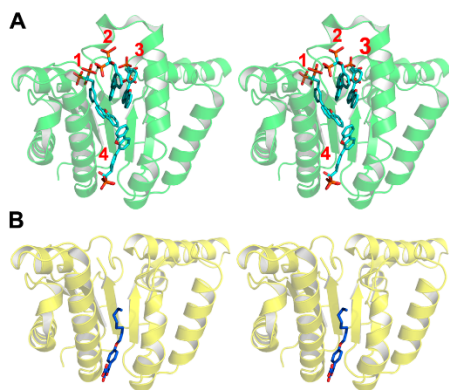


Figure 2. Stereo presentation of the X-ray structures chosen for further virtual screening from docking and ROC analysis. (A) 2E98 showing all four inhibitor binding sites. (B) 4H3A showing one inhibitor bound to site 4.

Library. The library was filtered to exclude compounds that had undesired, toxic or reactive functional groups; known promiscuous binders; MW >460 Da or MW <250 Da; more than 4 chiral centers; polar surface area (PSA) >150 Å² or PSA <50 Å²; number of rotatable bonds >10; or clogP >5 or clogP ≤2. Salts were also removed. Next, the selected compounds (~100,000) were loaded into Schrödinger's virtual screening workflow, where they were prepared with Ligprep and then docked using the filtering procedure for efficiency and only retaining the top 20% of compounds in the two rapid, initial docking modules HTVS (top 20% retained) and SP (top 20% retained). Finally, Glide XP was used to assign a final docking score to each molecule. AUC analyses on active and decoy data sets were previously performed using the Glide XP module, however, this was impractical for the large filtered ChemBridge library. Therefore, we relied on HTVS and SP modules to provide early filtering before employing the more time intensive XP protocol.

We then extracted the ~400 top scoring compounds (docking score less than -7 kcal/mol). Binary Molprint2D fingerprints were generated using Canvas, and 40 clusters were generated using K-means clustering.^{17–19} Of these 40 clusters, the top scoring compounds from each cluster were visually inspected and a representative was chosen from each cluster, resulting in a final list of 100 compounds. These were purchased from ChemBridge (ChemBridge Corporation, San Diego, CA) and then assayed for UPPS inhibition activity. Three out of the 100 compounds were UPPS inhibitors. Similarity searches based on these active compounds were then performed using PubChem and SciFinder, and additional compounds were obtained and tested.

■ ENZYME AND CELL GROWTH INHIBITION ASSAYS

Protein Expression and Purification. EcUPPS and SaUPPS were expressed and purified as described previously.⁸ Molecular weights and purities were verified by mass spectrometry and SDS-PAGE, respectively.

UPPS Inhibition Screening. The UPPS inhibition assays were carried out as described previously.⁸ Briefly, the condensation of FPP with IPP catalyzed by UPPS was monitored by using a continuous spectrophotometric assay²⁰ in 96-well plates with 200 μL reaction mixtures containing 400 μM 2-amino-6-mercapto-7-methylpurine ribonucleoside (MESG), 350 μM IPP, 35 μM FPP, 20 mM Tris-HCl buffer

(pH 7.5), 0.01% v/v Triton X-100, and 1 mM MgCl₂. The IC₅₀ values were obtained by fitting the inhibition data to a rectangular hyperbolic dose–response function using GraphPad PRISM 4.0 software (Graphpad Software, San Diego, CA). The IC₅₀ values for the most active hits were verified using a radiometric assay²¹ with 2.5 μM FPP, 25 μM [³H]IPP, and 0.01% v/v Triton X-100.

Cell Strains. *Bacillus subtilis* subsp. *subtilis* (ATCC 6051), *Escherichia coli* (ATCC 29425), and *Saccharomyces cerevisiae* (ATCC 208352) were purchased from the American Type Culture Collection. *Bacillus subtilis* strain 168, *Bacillus anthracis* strain Sterne, *Listeria monocytogenes* strain 4b F2365, *Staphylococcus aureus* USA300 (methicillin-resistant), *E. coli* MC400, *Pseudomonas putida*, and *Enterococcus faecalis* U503 (vancomycin-resistant) were from our laboratory strain collection.²²

***E. coli* ATCC 29425 Growth Inhibition Assay.** IC₅₀ values for *E. coli* growth inhibition were determined by using a microbroth dilution method. A 16 h culture of *E. coli* was diluted 50-fold into fresh Luria–Bertani (LB) broth and grown to an OD₆₀₀ of ~0.4. The culture was then diluted 500-fold into fresh LB medium, and 100 μL was inoculated into a 96-well flat bottom culture plate (Corning Inc., Corning, NY). The starting concentration of each compound was 0.3 mM, and this was 2-fold serially diluted. Plates were incubated for 3 h at 37 °C to midexponential phase. A 3-(4,5-dimethylthiazol-2-yl)-2,5-diphenyltetrazolium bromide (MTT) cell proliferation assay (ATCC) was then carried out to obtain bacterial viability dose–response curves. Ten microliters of MTT reagent was added into each well, followed by incubation for 2–4 h until a purple precipitate was visible. Then, 100 μL of detergent reagent was added and plates were further incubated in the dark at 23 °C for 2 h. The absorbance was recorded at 570 nm. A nonlinear regression analysis was then carried out using Origin 6.1. For each inhibitor, two independent experiments were performed and the IC₅₀ values found were averaged.

***B. subtilis* ATCC 6051 Growth Inhibition Assay.** A 16 h culture of *B. subtilis* was diluted 50-fold into fresh Luria–Bertani (LB) broth and incubated to an OD₆₀₀ of ~0.4. The culture was then diluted 500-fold into fresh LB medium, and 100 μL was inoculated into a 96-well flat bottom culture plate (Corning Inc., Corning, NY). The starting concentration of each compound was 0.5 mM and was then serially diluted. Plates were incubated for 12–16 h at 37 °C. The absorbance was recorded at 570 nm. A nonlinear regression analysis was carried out on the data obtained using Origin 6.1. For each inhibitor, two independent experiments were performed and the IC₅₀ values found were averaged.

***S. cerevisiae* Growth Inhibition Assay.** The protocol was the same as with the *B. subtilis* assay protocol except that YPD medium was used and the 96-well plate was incubated for 36 h instead of 12–16 h.

Evaluation of 1 and 4 Inhibitory Activity and Synergy. *B. subtilis* strain 168, *B. anthracis* strain Sterne, *E. coli* MC4100, and *P. putida* were grown to stationary phase in 10 mL of LB broth at 37 °C. *S. aureus* USA300 (methicillin-resistant), *E. faecalis* U503 (vancomycin-resistant), and *L. monocytogenes* strain 4b F2365 were grown to stationary phase in 10 mL of brain–heart infusion (BHI) medium at 37 °C. The cultures were adjusted to an OD₆₀₀ of 0.016 in the designated medium before being added to 96-well microplates. Successive 2-fold dilutions of compounds 1 and 4 were added to the cultures (0.25–64 μg mL⁻¹). As a control, kanamycin (1–32 μg mL⁻¹) was added to samples of *E. coli*, *B. subtilis*, *B. anthracis*, *P. putida*,

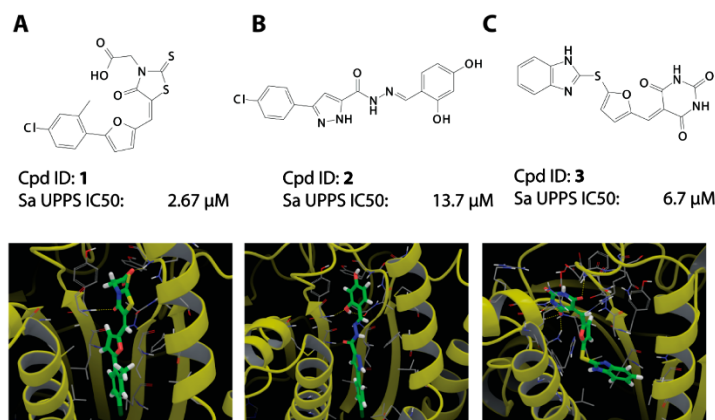


Figure 3. Three new classes of UPPS inhibitors discovered via virtual screening: (A) chemical structure and computed docking mode of compound 1, a rhodanine derivative; (B) chemical structure and docking mode of compound 2, a resorcinol derivative; (C) same for compound 3, a barbiturate.

and *L. monocytogenes*. Gentamycin was used as a control for *S. aureus* and *E. faecalis* with dilutions from 1 to 64 μg mL⁻¹. As a negative control, an equal volume of DMSO lacking antibiotic was used. Plates were covered and incubated at 37 °C for 16 h with shaking. The minimum inhibitory concentration (MIC) that suppressed at least 99% of bacterial growth was established based on culture turbidity in the microbroth dilution assay. The assay was repeated in three replicates, and values were averaged, or a range was reported. Isobolograms were carried out as previously described.^{10,23}

■ INHIBITOR CHARACTERIZATION

The purities of the key compounds investigated, obtained from ChemBridge (1 and 4), were determined by high-performance liquid chromatography and structures verified by NMR spectroscopy and high resolution mass spectrometry (Supporting Information Figures S2–S9) and were consistent with the structures provided by the vendor. Purities were >95% by HPLC.

■ RESULTS AND DISCUSSION

In previous work, we obtained moderate correlations between enzyme inhibition activity and docking scores within a congeneric series of UPPS inhibitors (lipophilic bisphosphonates)^{8,24} using docking methods, so we first examined whether we could obtain similar correlations between docking scores and experimentally determined IC₅₀ values for the 112 known actives. There was no significant correlation between docking scores and pIC₅₀ values (pIC₅₀ = -logIC₅₀), Figure S1 in the Supporting Information. The wide variety of potential binding modes (sites 1, 2, 3, and 4^{8,12}) and protein conformations would be expected to make it difficult to achieve a good correlation between a scoring function and the experimentally determined pIC₅₀ values, in addition to the assumptions made in scoring functions that cause inaccuracy when compared to experimental affinities. Nevertheless, docking studies can provide enrichment of active compounds from large libraries, even though docking scores rarely correlate well with activity when structurally diverse compounds are involved. We thus

next employed an area-under-the-curve (AUC) analysis, also known as the receiver-operating-characteristic (ROC), a method that has been shown to be useful in validating structure-based virtual screening protocols²⁵ and is a standard method for evaluating such protocols.²⁶

We therefore tested 12 EcUPPS X-ray structures for their ability to separate actives (IC₅₀ <100 μM) from decoys (presumed inactive compounds in the decoy library). Several EcUPPS X-ray structures showed a good separation of active from decoy compounds, with AUC values of ~0.8. These structures also demonstrated early enrichment, as evidenced by the steep initial slope of the curve. This means that the best scores were given primarily to active compounds and suggests that, in screening a large compound library, the best scoring compounds would be enriched in UPPS inhibitors. We thus picked the two X-ray structures (PDB codes 2E98 and 4H3A) that provided significant early enrichment and a high AUC in the validation studies, for predictive studies. Using these two X-ray structures, we screened the ChemBridge EXPRESS-pick compound library (after filtering) and determined ~400 hits with GlideXP scores less than -7 kcal/mol (lower energy is better). Since many highly ranked compounds were chemically very similar, we clustered the top scoring compounds and selected representatives from each cluster to ensure chemical diversity among the compounds to be tested.

Discovery of Novel UPPS Inhibitor Cores. The screening of the ChemBridge EXPRESS-pick compound library using the validated docking protocol resulted in the discovery of three new UPPS inhibitor classes: the 4-oxo-2-thioxo-1,3-thiazolidines, also known as rhodanines (e.g., compound 1), dihydroxyphenyls (the resorcinol, compound 2), and pyrimidinetrione (a barbiturate analogue, compound 3). None of these have been previously reported to be UPPS inhibitors. All three compounds are predicted to bind in either site 1 or 3 of the 2E98 crystal structure (Figure 3), although X-ray crystallographic studies will be required to confirm this binding mode (and our attempts to obtain crystal structures of these systems have not been successful). In any case, the three new inhibitors discovered represent UPPS inhibitors with “drug-

Table 1. 4-Oxo-2-thioxo-1,3-thiazolidines Investigated in UPPS and Bacterial Cell Growth Inhibition Assays^a

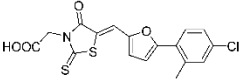
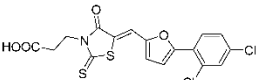
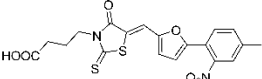
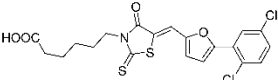
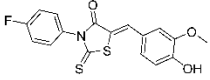
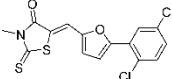
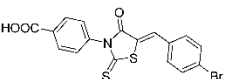
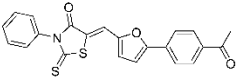
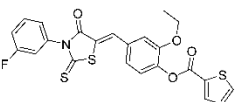
ID	Structure	Vendor #	EcUPPS	SaUPPS	<i>B. subtilis</i>	<i>E. coli</i>	<i>S. cerevisiae</i>
			IC ₅₀	IC ₅₀	IC ₅₀	IC ₅₀	IC ₅₀
1		CB 7471392	260	2.7	2.9	>200	>200
4		CB 5523169	2.1	2.4	0.43	>200	180
5		L339644	>100	5.7	3.2	>200	>200
6		L339822	>100	3.8	5.6	>200	>200
7		CB 5674456	41	150	>200	>200	>200
8		CB 5143717	>200	>200	>200	>200	>200
9		CB 5280379	>200	170	>200	>200	>200
10		CB 5377413	85	32	>200	>200	>200
11		CB 6824270	>200	>200	>200	>200	>200

Table 1. continued

ID	Structure	Vendor #	EeUPPS IC ₅₀	SaUPPS IC ₅₀	<i>B. subtilis</i>	<i>E. coli</i>	<i>S. cerevisiae</i>
					IC ₅₀	IC ₅₀	IC ₅₀
12		NCI 660017	>100	>100	>200	>200	>200
13		NCI 343985	>100	5.3	>200	>200	>200
14		NCI 337736	>100	>100	>200	>200	>200
15		NCI 320208	>100	>100	>200	>200	>200
16		NCI 90950	>100	>100	>200	>200	>200
17		NCI 36005	>100	>100	>200	>200	>200
18		NCI 320207	>100	>100	>200	>200	>200
19		NCI 318219	>100	>100	>200	>200	>200

^aAll concentrations are in μM .

like” physicochemical properties, passing the common drug-like filters as described in the Methods and Materials section. The most potent of the 3 compounds was the 4-oxo-2-thioxo-1,3-thiazolidine 1 (IC₅₀ ~2.6 μM against *S. aureus* UPPS), which in an initial screen for bioactivity was also found to be active against *B. subtilis*, MIC (minimal inhibitory concentration) ~3 $\mu\text{g}/\text{mL}$ (Table 1). For this reason, we chose to next investigate analogues based on the 4-oxo-2-thioxo-1,3-thiazolidine core.

Novel Core SAR. We next obtained 16 additional compounds from ChemBridge, from Sigma-Aldrich, and from the Drug Synthesis and Chemistry Branch, Developmental Therapeutics Program, Division of Cancer Treatment and Diagnosis, National Cancer Institute (4–19, Table 1), containing the 4-oxo-2-thioxo-1,3-thiazolidine core and tested them for activity against SaUPPS and EeUPPS, as well as a preliminary activity screen against *B. subtilis*, *E. coli*, and *S. cerevisiae* (the latter as a general cytotoxicity control, since it

Table 2. MIC Values for Two 4-Oxo-2-thioxo-1,3-thiazolidine Analogues, Compounds 1 and 4, Tested in Diverse Bacterial Cell Growth Inhibition Assays^a

	Compound 1	Compound 4
	 MIC (μg/mL)	 MIC (μg/mL)
<i>Bacillus subtilis</i>	1	0.125
<i>Bacillus anthracis</i>	1-2	0.25
<i>Staphylococcus aureus</i> (MRSA)	8-64	4
<i>Enterococcus faecalis</i> (VRE)	1-8	4
<i>Listeria monocytogenes</i>	4	0.125
<i>Pseudomonas putida</i>	>64	>64
<i>Escherichia coli</i>	>64	>64

^aThe compounds were tested against a panel of both Gram-positive (top five) and Gram-negative (lower two) bacteria.

lacks UPPS). The alkyl carboxylic acid containing compounds with the 4-oxo-2-thioxo-1,3-thiazolidine core were active in assays against *B. subtilis*, and the most potent compound was 4 (an analogue of 1). 4 was roughly equipotent against SaUPPS and EcUPPS with an IC₅₀ of ~2 μM. Additionally, 4 was active against *B. subtilis* with a MIC ~0.43 μg/mL and was very weakly active (~200 μg/mL) against *S. cerevisiae*, indicating that 4 was not generally cytotoxic. Since 1 and 4 showed significant activity in enzymatic and bacterial growth assays, we subsequently tested them against several pathogens. Both 1 and 4 gave MIC values in the high ng/mL to low μg/mL range against *B. anthracis* Sterne, MRSA, VRE, and *L. monocytogenes*, Table 2. This promising antibacterial activity suggested the potential utility of these UPPS inhibitors in synergizing with other cell wall agents but where significant resistance has emerged, such as with methicillin (MRSA) and vancomycin (VRE).

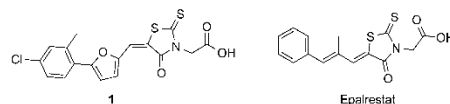
Synergistic Interactions. To investigate the possibility of synergistic interactions with known cell wall biosynthesis inhibitors, we determined the fractional inhibitory concentration index (FICI) values for three systems: MRSA, using 1 + methicillin; VRE, using 1 + vancomycin; and *B. anthracis*, using 1 + ampicillin. The FICI is defined as

$$\begin{aligned} \text{FICI} &= \text{FIC(A)} + \text{FIC(B)} \\ &= \text{MIC(AB)}/\text{MIC(A)} + \text{MIC(BA)}/\text{MIC(B)} \end{aligned}$$

where FIC(A) and FIC(B) are the fractional inhibitory concentrations of drugs A and B, MIC(A) and MIC(B) are the MIC values of drugs A and B acting alone, and MIC(AB) and MIC(BA) are the MIC values of the most effective combination of drug A or B in the presence of drug B or A. Using this method, FICI values of <0.5 represent synergism, >0.5 and <1.0 represent additivity, >1 and <2 represent an indifferent effect, and >2 represents drug antagonism. Isobolo-

grams are shown in Figure 4. As can be seen in Figure 4B, the FICI for 1 + methicillin in MRSA is 0.11, which indicates strong synergism during late stage growth. However, with both VRE (1 + vancomycin) and *B. anthracis* (1 + ampicillin) the FICI values are in the 1–2 range, which indicates an indifferent effect.

What is particularly interesting about the most active species investigated here (1) is that it has a structure that is very similar to that found in the drug epalrestat, an aldolase reductase



inhibitor²⁷ that is used to treat diabetic neuropathy, and is approved for clinical use in Japan, China, and India. This is encouraging because rhodanines as a class are known to often have activity in widely different assays, and indeed computer programs such as PAINS²⁸ categorize, e.g., 1–4 (as well as epalrestat) as possible “pan assay interference compounds”. This can mean that the compounds cause false positives in assays, or that they may be multitarget inhibitors. In some cases multitargeting may be undesirable; however, in the context of anti-infective development, multitargeting is expected to increase efficacy as well as decrease the possibility of resistance development,²⁹ both very desirable features.

CONCLUSIONS

The results described herein are of interest for several reasons. First, we carried out an in silico screen of ~100 known UPPS inhibitors and 1000 decoys using 12 reported UPPS X-ray structures. The two X-ray structures providing the best enrichment in an AUC-ROC analysis were then used to screen

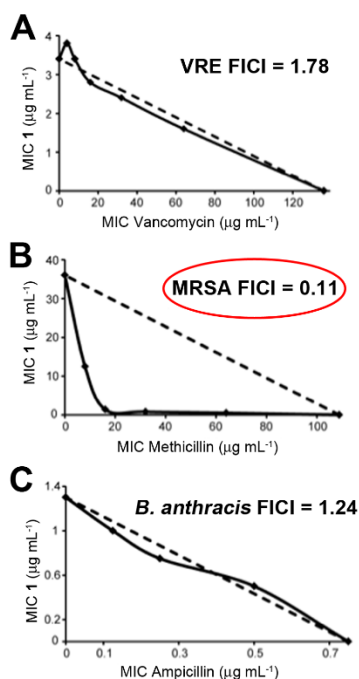


Figure 4. In vitro synergy assays. Isobolograms for growth inhibition of VRE, MRSA, and *B. anthracis* strain Sterne. (A) **1** and vancomycin inhibition of *E. faecalis* U503 (vancomycin-resistant, VRE). FICI = 1.78. (B) **1** and methicillin inhibition of *S. aureus* (USA300). FICI = 0.11. (C) **1** and ampicillin inhibition of *B. anthracis* strain Sterne. FICI = 1.24.

a subset of ~100,000 compounds selected for drug-like activity from an initial ChemBridge library of ~450,000 compounds. We then tested the ~100 in silico hits in vitro against SaUPPS and EcUPPS, leading to several μM UPPS inhibitors (as deduced from both PPi release and radioactive assays). The most potent lead was **1**, which is structurally quite similar to epalrestat, in clinical use to treat diabetic neuropathy. **1** (and its analogue **4**) inhibited the growth of Gram positives; they did not inhibit the growth of Gram negatives (important with *E. coli* in the context of maintaining commensal microflora), and they had no activity against *S. cerevisiae*. Activity against *B. anthracis*, a vancomycin-resistant *Enterococcus* spp., and *Listeria monocytogenes* was good, in the 0.125–4 $\mu\text{g}/\text{mL}$ range, and there was very strong synergy (FICI = 0.11) with methicillin and **1** in a MRSA strain of *S. aureus*, suggesting that **1** could be a promising lead (in combination therapies) for treating staph infections.

■ ASSOCIATED CONTENT

Supporting Information

Correlation of pIC₅₀ with docking scores analysis, as well as ¹H NMR, HPLC analysis, mass spectra, and high-resolution mass

spectra of compounds **1** and **4**. This material is available free of charge via the Internet at <http://pubs.acs.org>.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: wsinko@gmail.com. Phone: 858-234-2905.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported by the United States Public Health Service (National Institutes of Health Grants GM065307, CA158191, GM08326, GM31749, and HD071600); the National Science Foundation (Grant MCB-1020765); a Packard Fellowship for Science and Engineering (D.A.M.), the National Biomedical Computation Resource, the UCSD Center for Theoretical Biological Physics, the Howard Hughes Medical Institute, and the NSF Supercomputer Centers. F.F. acknowledges financial support of the Beatrice de Pinós program from AGAUR for a postdoctoral grant (2010 BP_A_00339). We thank the Drug Synthesis and Chemistry Branch, Developmental Therapeutics Program, Division of Cancer Treatment and Diagnosis, National Cancer Institute, for providing chemicals, and Professors James Wells and Paul Hergenrother for providing bacteria.

■ ABBREVIATIONS USED

MRSA, methicillin-resistant *Staphylococcus aureus*; VRE, vancomycin-resistant *Enterococci*; UPPS, undecaprenyl diphosphate synthase; MIC, minimal inhibitory concentration; FICI, fractional inhibitory concentration index; UPP, undecaprenyl diphosphate; IPP, isopentyl diphosphate; FPP, farnesyl diphosphate; BPH, bisphosphonate; ROC, receiver-operating-characteristic; AUC, area under the curve; LB, Luria–Bertani; PAINS, pan assay interference compounds

■ REFERENCES

- (1) Fischbach, M. A.; Walsh, C. T. Antibiotics for emerging pathogens. *Science* **2009**, *325*, 1089–1093.
- (2) Butler, M. S.; Cooper, M. A. Antibiotics in the clinical pipeline in 2011. *J. Antibiot.* **2011**, *64*, 413–425.
- (3) van Heijenoort, J. Lipid intermediates in the biosynthesis of bacterial peptidoglycan. *Microbiol. Mol. Biol. Rev.* **2007**, *71*, 620–635.
- (4) Apfel, C. M.; Takacs, B.; Fountoulakis, M.; Stieger, M.; Keck, W. Use of genomics to identify bacterial undecaprenyl pyrophosphate synthetase: cloning, expression, and characterization of the essential uppS gene. *J. Bacteriol.* **1999**, *181*, 483–492.
- (5) Payne, D. J.; Gwynn, M. N.; Holmes, D. J.; Pompliano, D. L. Drugs for bad bugs: confronting the challenges of antibacterial discovery. *Nat. Rev. Drug Discovery* **2007**, *6*, 29–40.
- (6) Peukert, S.; Sun, Y.; Zhang, R.; Hurley, B.; Sabio, M.; Shen, X.; Gray, C.; Dzink-Fox, J.; Tao, J.; Cebula, R.; Wattanasin, S. Design and structure-activity relationships of potent and selective inhibitors of undecaprenyl pyrophosphate synthase (UPPS): tetramic, tetronic acids and dihydropyridin-2-ones. *Bioorg. Med. Chem. Lett.* **2008**, *18*, 1840–1844.
- (7) Lee, L. V.; Granda, B.; Dean, K.; Tao, J.; Liu, E.; Zhang, R.; Peukert, S.; Wattanasin, S.; Xie, X.; Ryder, N. S.; Tommasi, R.; Deng, G. Biophysical investigation of the mode of inhibition of tetramic acids, the allosteric inhibitors of undecaprenyl pyrophosphate synthase. *Biochemistry* **2010**, *49*, 5366–5376.
- (8) Guo, R. T.; Cao, R.; Liang, P. H.; Ko, T. P.; Chang, T. H.; Hudock, M. P.; Jeng, W. Y.; Chen, C. K.; Zhang, Y.; Song, Y.; Kuo, C. J.; Yin, F.; Oldfield, E.; Wang, A. H. Bisphosphonates target multiple

- sites in both cis- and trans-prenyltransferases. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 10022–10027.
- (9) Durrant, J. D.; Cao, R.; Gorfe, A. A.; Zhu, W.; Li, J.; Sankovsky, A.; Oldfield, E.; McCammon, J. A. Non-bisphosphonate inhibitors of isoprenoid biosynthesis identified via computer-aided drug design. *Chem. Biol. Drug Des.* **2011**, *78*, 323–332.
- (10) Zhang, Y.; Lin, F.-Y.; Li, K.; Zhu, W.; Liu, Y.-L.; Cao, R.; Pang, R.; Lee, E.; Axelson, J.; Hensler, M.; Wang, K.; Molohon, K. J.; Wang, Y.; Mitchell, D. A.; Nizet, V.; Oldfield, E. HIV-1 integrase inhibitor-inspired antibacterials targeting isoprenoid biosynthesis. *ACS Med. Chem. Lett.* **2012**, *3*, 402–406.
- (11) Lindert, S.; Zhu, W.; Liu, Y.-L.; Pang, R.; Oldfield, E.; McCammon, J. A. Farnesyl diphosphate synthase inhibitors from in silico screening. *Chem. Biol. Drug Des.* **2013**, *81*, 742–748.
- (12) Zhu, W.; Zhang, Y.; Sinko, W.; Hensler, M. E.; Olson, J.; Molohon, K. J.; Lindert, S.; Cao, R.; Li, K.; Wang, K.; Wang, Y.; Liu, Y.-L.; Sankovsky, A.; de Oliveira, C. A. F.; Mitchell, D. A.; Nizet, V.; McCammon, J. A.; Oldfield, E. Antibacterial drug leads targeting isoprenoid biosynthesis. *Proc. Natl. Acad. Sci. U.S.A.* **2013**, *110*, 123–128.
- (13) Halgren, T. A.; Murphy, R. B.; Friesner, R. A.; Beard, H. S.; Frye, L. L.; Pollard, W. T.; Banks, J. L. Glide: A new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening. *J. Med. Chem.* **2004**, *47*, 1750–1759.
- (14) Friesner, R. A.; Murphy, R. B.; Repasky, M. P.; Frye, L. L.; Greenwood, J. R.; Halgren, T. A.; Sanschagrin, P. C.; Mainz, D. T. Extra precision glide: docking and scoring incorporating a model of hydrophobic enclosure for protein-ligand complexes. *J. Med. Chem.* **2006**, *49*, 6177–6196.
- (15) Schrödinger, L. L. C. *Suite 2011: Glide, version 5.7*; Schrödinger LLC: New York, 2011.
- (16) Sastry, G. M.; Adzhigirey, M.; Day, T.; Annabhimoju, R.; Sherman, W. Protein and ligand preparation: parameters, protocols, and influence on virtual screening enrichments. *J. Comput.-Aided Mol. Des.* **2013**, *27*, 221–234.
- (17) *Suite 2012: Camvas version 1.5*; 2012.
- (18) Duan, J.; Dixon, S. L.; Lowrie, J. F.; Sherman, W. Analysis and comparison of 2D fingerprints: insights into database screening performance using eight fingerprint methods. *J. Mol. Graphics Modell.* **2010**, *29*, 157–170.
- (19) Sastry, M.; Lowrie, J. F.; Dixon, S. L.; Sherman, W. Large-scale systematic analysis of 2D fingerprint methods and parameters to improve virtual screening enrichments. *J. Chem. Inf. Model.* **2010**, *50*, 771–784.
- (20) Webb, M. R. A continuous spectrophotometric assay for inorganic phosphate and for measuring phosphate release kinetics in biological systems. *Proc. Natl. Acad. Sci. U.S.A.* **1992**, *89*, 4884–4887.
- (21) Li, H.; Huang, J.; Jiang, X.; Seefeld, M.; McQueney, M.; Macarron, R. The effect of triton concentration on the activity of undecaprenyl pyrophosphate synthase inhibitors. *J. Biomol. Screening* **2003**, *8*, 712–715.
- (22) Molohon, K. J.; Melby, J. O.; Lee, J.; Evans, B. S.; Dunbar, K. L.; Bumpus, S. B.; Kelleher, N. L.; Mitchell, D. A. Structure determination and interception of biosynthetic intermediates for the plantazolicin class of highly discriminating antibiotics. *ACS Chem. Biol.* **2011**, *6*, 1307–1313.
- (23) Leon, A.; Liu, L.; Yang, Y.; Hudock, M. P.; Hall, P.; Yin, F.; Studer, D.; Puan, K.-J.; Morita, C. T.; Oldfield, E. Isoprenoid biosynthesis as a drug target: bisphosphonate inhibition of *Escherichia coli* K12 growth and synergistic effects of fosmidomycin. *J. Med. Chem.* **2006**, *49*, 7331–7341.
- (24) Sinko, W.; de Oliveira, C.; Williams, S.; Van Wynsberghe, A.; Durrant, J. D.; Cao, R.; Oldfield, E.; Andrew McCammon, J. Applying molecular dynamics simulations to identify rarely sampled ligand bound conformational states of undecaprenyl pyrophosphate synthase, an antibacterial target. *Chem. Biol. Drug Des.* **2011**, *77*, 412–420.
- (25) Triballeau, N.; Acher, F.; Brabet, L.; Pin, J.-P.; Bertrand, H.-O. Virtual screening workflow development guided by the “receiver operating characteristic” curve approach. Application to high-throughput docking on metabotropic glutamate receptor subtype 4. *J. Med. Chem.* **2005**, *48*, 2534–2547.
- (26) Christofferson, A. J.; Huang, N. How to benchmark methods for structure-based virtual screening of large compound libraries. *Methods Mol. Biol.* **2012**, *819*, 187–195.
- (27) Ramirez, M. A.; Borja, N. L. Epalrestat: an aldose reductase inhibitor for the treatment of diabetic neuropathy. *Pharmacotherapy* **2008**, *28*, 646–645.
- (28) Baell, J. B.; Holloway, G. A. New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays. *J. Med. Chem.* **2010**, *53*, 2719–2740.
- (29) Li, K.; Schurig-Briccio, L. A.; Feng, X.; Upadhyay, A.; Pujari, V.; Lechartier, B.; Fontes, F. L.; Yang, H.; Rao, G.; Zhu, W.; Gulati, A.; No, J. H.; Cintra, G.; Bogue, S.; Liu, Y. L.; Molohon, K. J.; Orlean, P.; Mitchell, D. A.; Freitas-Junior, L. H.; Ren, F.; Sun, H.; Jiang, T.; Li, Y.; Guo, R. T.; Cole, S. T.; Gennis, R. B.; Crick, D. C.; Oldfield, E. Multi-target drug discovery for tuberculosis and other infectious diseases. *J. Med. Chem.* **2014**, *57*, 3126–3139.

A.4 References

1. Dunbar KL, Chekan JR, Cox CL, Burkhart BJ, Nair SK, and Mitchell DA. (2014) Discovery of a new ATP-binding motif involved in peptidic azoline biosynthesis. *Nature chemical biology* 10: 823-829.
2. Maxson T, Deane CD, Molloy EM, Cox CL, Markley AL, Lee SW, and Mitchell DA. (2015) HIV Protease Inhibitors Block Streptolysin S Production. *ACS chemical biology* 10(5): 1217-1226.
3. Sinko W, Wang Y, Zhu W, Zhang Y, Feixas F, Cox CL, Mitchell DA, Oldfield E, and McCammon JA. (2014) Undecaprenyl diphosphate synthase inhibitors: antibacterial drug leads. *Journal of medicinal chemistry* 57: 5693-5701.