UNIVERSITY OF CALIFORNIA, SAN DIEGO

From Molecules to Genes and Back Again: Tales of Marine Microbes and their Specialized Chemistry

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

in

Marine Biology

by

Michelle Antoinette Schorn

Committee in charge:

Professor Bradley S. Moore, Chair
Professor Eric E. Allen
Professor Lihini Aluwihare
Professor Paul R. Jensen
Professor Kit Pogliano

2018

The dissertation of Michelle Antoinette Schorn is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Chair

University of California, San Diego

2018

For my grandparents and parents, who instilled in me a love for the sea and gave me the foundations to chase my curiosity and follow my dreams.

Table of Contents

List of Figures

List of Tables

Acknowledgements

I would like to sincerely thank my advisor, Bradley S. Moore, who has encouraged, supported, and inspired me throughout my graduate career and has facilitated a collaborative and exploratory laboratory environment that I have greatly enjoyed. I would also like to thank Leonard Kaysser, who served as my first mentor, equipped me with necessary molecular biology tools, and provided me the foundation for becoming an independent researcher. I would like to thank William Gerwick, Lena Gerwick for hosting me in their lab and continuing to be supportive mentors throughout my graduate career. I would like to thank Anton Korobeynikov for allowing access to his assembly algorithms before they were publicly available. I would also like to thank Nadine Ziemert and Evi Stegmann, who hosted me at the University of Tübingen, and their students, Mohammad Alanjary, Kyra Geyer, and Paul Schwarz, who helped me navigate the German laboratory system and aided in experiments. I would like to thank Paul Jensen and Jason Biggs for immersing me in the techniques of marine field collection and serving as excellent dive buddies. I would also like to thank Eric Allen, Sheila Podell, and Jessica Blanton for teaching me invaluable bioinformatic tools and metagenomic analysis. I would like to thank Tommie Lincecum and Kristen Aguinaldo for their help in multiple sequencing projects and providing access to cutting-edge technology. I would also like to thank Vinayak Agarwal and Peter Jordan for helping me become a competent chemist and for being great mentors. Finally, I'd like to thank my family for their unyielding support and constant encouragement, especially my sister, Julia Schorn, who, as a volunteer intern, helped me carry out experiments and has been a constant source of moral support.

Chapter 2, in full, is a reprint of materials as it appears in “Genetic Basis for the Biosynthesis of the Pharmaceutically Important Class of Epoxyketone Proteasome Inhibitors” in *ACS Chemical Biology*, 2014, Schorn, M., Zettler, J., Noel, J. P., Dorrestein, P. C., Moore, B. S. and Kaysser, L. The dissertation author was one of two equally contributing primary investigators and authors of this manuscript.

Chapter 3, in full, is a reprint of materials as it appears in “Sequencing rare marine actinomycete genomes reveals high density of unique natural product biosynthetic gene clusters” in *Microbiology*, 2016. Schorn, M. A., Alanjary, M. M., Aguinaldo, K., Korobeynikov, A., Podell, S., Patin, N., Lincecum, T., Jensen, P. R., Ziemert, N. and Moore, B. S. The dissertation author was the primary investigator and author of this manuscript.

Chapter 4, in part, is currently being prepared for submission of the material. Schorn, M. A., Jordan, P., Podell, S., Blanton, J. M., Agarwal, V., Biggs, J. S., Allen, E. E., and Moore, B. S. The dissertation author is the primary investigator and author of this paper.

Vita

2009 Bachelor of Arts, History of Science and Medicine, Yale University

2010-2012 Scientist I-II, Ion Torrent by Life Technologies, Guilford, CT

2012-2018 Graduate Student Researcher, Scripps Institution of Oceanography, University of California, San Diego

2013 Edna Bailey Sussman Fellow, Scripps Institution of Oceanography, University of California San Diego

2014 Master of Science, Marine Biology, Scripps Institution of Oceanography, University of California, San Diego

2015 Teaching Assistant, Microbial Life in Extreme Environments, University of California, San Diego

2015-2016 Teach at Tübingen Fellow, University of Tübingen, Germany

2016 Claude E. Zobell fellowship awardee, Scripps Institution of Oceanography, University of California, San Diego

2017-2018 Business Development Intern, Scripps Institution of Oceanography, University of California, San Diego

2018 Doctor of Philosophy, Marine Biology, Scripps Institution of Oceanography, University of California, San Diego

Publications

2008 Smith, S. A., Tank, D. C., Boulanger, L. A., Bascom-Slack, C. A., Eisenman, K., Kingery, D., Babbs, B., Fenn, K., Greene, J. S., Hann, B. D., Keehner, J., Kelley-Swift, E. G., Kembaiyan, V., Lee, S. J., Li, P., Light, D. Y., Lin, E. H., Ma, C., Moore, E., Schorn, M. A., Vekhter, D., Nunez, P. V., Strobel, G. A., Donoghue, M. J. & Strobel, S. A. Bioactive endophytes warrant intensified exploration and conservation. *PLoS One* **3**, e3052, 2008. PMID: 18725962.

2009 Bascom-Slack, C. A., Ma, C., Moore, E., Babbs, B., Fenn, K., Greene, J. S., Hann, B. D., Keehner, J., Kelley-Swift, E. G., Kembaiyan, V., Lee, S. J., Li, P., Light, D. Y., Lin, E. H., Schorn, M. A., Vekhter, D., Boulanger, L. A., Hess, W. M., Vargas, P. N., Strobel, G. A. & Strobel, S. A. Multiple, novel biologically active endophytic actinomycetes isolated from upper Amazonian rainforests. *Microb Ecol* **58**, 374-383, 2009. PMID: 19252940.

2011 Rothberg, J. M., Hinz, W., Rearick, T. M., Schultz, J., Mileski, W., Davey, M., Leamon, J. H., Johnson, K., Milgrew, M. J., Edwards, M., Hoon, J., Simons, J. F., Marran, D., Myers, J. W., Davidson, J. F., Branting, A., Nobile, J. R., Puc, B. P., Light, D., Clark, T. A., Huber, M., Branciforte, J. T., Stoner, I. B., Cawley, S. E., Lyons, M., Fu, Y., Homer, N., Sedova, M., Miao, X., Reed, B., Sabina, J., Feierstein, E., Schorn, M., Alanjary, M., Dimalanta, E., Dressman, D., Kasinskas, R., Sokolsky, T., Fidanza, J. A., Namsaraev, E., McKernan, K. J., Williams, A., Roth, G. T. & Bustillo, J. An integrated semiconductor device enabling non-optical genome sequencing. *Nature* **475**, 348-352, 2011.

2014 Schorn, M., Zettler, J., Noel, J. P., Dorrestein, P. C., Moore, B. S. & Kaysser, L. Genetic basis for the biosynthesis of the pharmaceutically important class of epoxyketone proteasome inhibitors. *ACS Chem Biol* **9**, 301-309, 2014. PMID: 24168704.

Agarwal, V., El Gamal, A. A., Yamanaka, K., Poth, D., Kersten, R. D., Schorn, M., Allen, E. E. & Moore, B. S. Biosynthesis of polybrominated aromatic organic compounds by marine bacteria. *Nat Chem Biol* **10**, 640-647, 2014. PMID: 24974229.

2016 Schorn, M. A., Alanjary, M. M., Aguinaldo, K., Korobeynikov, A., Podell, S., Patin, N., Lincecum, T., Jensen, P. R., Ziemert, N. & Moore, B. S. Sequencing rare marine actinomycete genomes reveals high density of unique natural product biosynthetic gene clusters. *Microbiology* **162**, 2075-2086, 2016. PMID: 27902408.

El Gamal, A., Agarwal, V., Diethelm, S., Rahman, I., Schorn, M. A., Sneed, J. M., Louie, G. V., Whalen, K. E., Mincer, T. J., Noel, J. P., Paul, V. J. & Moore, B. S. Biosynthesis of coral settlement cue tetrabromopyrrole in marine bacteria by a uniquely adapted brominase-thioesterase enzyme pair. _Proc Natl Acad Sci U S A_**113**, 3797-3802, 2016. PMID: 27001835.

Harvey, E. L., Deering, R. W., Rowley, D. C., El Gamal, A., Schorn, M., Moore, B. S., Johnson, M. D., Mincer, T. J. & Whalen, K. E. A Bacterial Quorum-Sensing Precursor Induces Mortality in the Marine Coccolithophore, Emiliania huxleyi. *Front Microbiol* **7**, 59, 2016. PMID: 26870019.

Nguyen, D. D., Melnik, A. V., Koyama, N., Lu, X., Schorn, M., Fang, J., Aguinaldo, K., Lincecum, T. L., Jr., Ghequire, M. G., Carrion, V. J., Cheng, T. L., Duggan, B. M., Malone, J. G., Mauchline, T. H., Sanchez, L. M., Kilpatrick, A. M., Raaijmakers, J. M., Mot, R., Moore, B. S., Medema, M. H. & Dorrestein, P. C. Indexing the Pseudomonas specialized metabolome enabled the discovery of poaeamide B and the bananamides. *Nat Microbiol* **2**, 16197, 2016. PMID: 27798598.

2017 Patin, N. V., Schorn, M., Aguinaldo, K., Lincecum, T., Moore, B. S. & Jensen, P. R. Effects of Actinomycete Secondary Metabolites on Sediment Microbial Communities. *Appl Environ Microbiol* **83**, 2017. PMID: 27986719.

Agarwal, V., Blanton, J. M., Podell, S., Taton, A., Schorn, M. A., Busch, J., Lin, Z., Schmidt, E. W., Jensen, P. R., Paul, V. J., Biggs, J. S., Golden, J. W., Allen, E. E. & Moore, B. S. Metagenomic discovery of polybrominated diphenyl ether biosynthesis by marine sponges. *Nat Chem Biol* **13**, 537-543, 2017. PMID: 28319100.

ABSTRACT OF THE DISSERTATION

From Molecules to Genes and Back Again: Tales of Marine Microbes and their Specialized Chemistry

by

Michelle Antoinette Schorn

Doctor of Philosophy in Marine Biology

University of California, San Diego, 2018

Professor Bradley S. Moore, Chair

Nature has created elegant and efficient ways of assembling a wide variety of diverse chemical scaffolds. Microbes are prolific producers of these secondary metabolites, which can have profound bioactivities and be harnessed for use in medicine. Within their genomes, microbes possess the blueprints for making natural products, and recent advances in high-throughput DNA sequencing have revealed a much larger repertoire of specialized chemistry than what can be observed in the lab. This bacterial 'dark matter' has the potential to contain the instructions for making countless new chemical scaffolds and represents an untapped source for drug discovery. Connecting secondary metabolite compounds with their biosynthetic gene clusters can inform novel biosynthetic transformations, provide a renewable source of promising bioactive compounds, and inspire synthetic biology approaches to assembling new molecules. Chapter 2 of this dissertation connects two pharmaceutically relevant natural products with their corresponding biosynthetic gene clusters. The epoxyketone proteasome inhibitors epoxomicin and eponemycin are naturally produced by actinomycete bacteria, and whole genome sequencing revealed their genetic underpinnings and native resistance mechanism. This work represents the first elucidation of biosynthetic gene clusters for epoxyketone proteasome inhibitors. Chapter 3 takes a wider lens to examine an understudied group of bacteria: the rare marine actinomycetes. Whole genome sequencing of a group of rare marine actinomycetes revealed an incredible wealth of biosynthetic potential and diversity not yet represented in current sequencing databases. This study establishes rare marine actinomycetes as a group worthy of further exploration for genome mining and drug discovery. Chapter 4 investigates a more complex system with a focus on yet to be cultured cyanobacterial sponge symbionts. Previous work showed that the genomes of these symbionts contained the genes necessary to produce environmentally relevant poly-brominated diphenyl ethers (PBDEs). In assembling high quality draft genomes of two symbionts of distinct sponges, their full secondary metabolite potential was revealed. Additionally, genome mining led to the identification of two novel dysinosin molecules, representing the first example of mining for gene clusters in a metagenome assembled genome leading to new chemistry.

Chapter 1: Introduction to the Dissertation

1.1 Natural Product Discovery Before the Genomic Age

1.1.1 Twentieth Century Advances in Understanding DNA

As the blueprint of life, deoxyribonucleic acid, or DNA, is the defining molecule that connects biological functions and chemical entities. Understanding this simple and elegant molecule has spurred innumerable scientific advances. The more we learn about DNA, the more we uncover about the biological underpinnings of how life thrives and the chemical complexity that accompanies all life. The discovery of the structure of DNA, the technologies developed to sequence it, and the computational advances to understand these sequences pave the way for discoveries in all fields of science. In the field of microbial natural products, the coupling of DNA sequence and chemical elucidation has greatly expanded scientists' capacity to discover and predict bioactive small molecules. The sequencing revolution has come at a particularly opportune time for the natural products field. As microbial resistance to antibiotics accelerates, and discovery of novel compounds from nature slows due to re-isolation of known compounds, new approaches are necessary for continued fruitful discovery of bioactive molecules from the natural environment. Integrating genomic methods into traditional chemical isolation of natural products has already opened up new chemical and biological spaces to explore and exploit in the search for the next generation of anti-infective molecules. It is clear that the evolution of DNA sequencing, from understanding the structure and composition of DNA, to sequencing one gene at a time, to sequencing full genomes and complex metagenomes, has mirrored the evolution of natural product chemical discovery in the age of DNA.

While the significance of DNA in all disciplines of science seems to be relatively recent, scientists have long been seeking the truth about life's building blocks. The second half of the nineteenth century is when some would say the field of cellular biology was established. Biological thought was shifting from studying whole organisms to studying the component cells that make up all living things. Charles Darwin and Alfred Wallace's theory of evolution, published in 1858, introduced the idea of favorable traits being passed down from one generation to the next by an unknown hereditary substance[1]. Shortly after, in 1865, Gregor Mendel demonstrated the laws of hereditary transmission of traits with pea plants[2]. In 1869, Swiss physician Friedrich Miescher at the University of Tübingen would accidentally precipitate nuclein, and describe, for the first time, the basic physical and chemical properties of DNA[3]. Miescher's astonishing discovery would go largely forgotten and overlooked as the hereditary substance for over seventy years. Proteins were considered the likely culprit as the hereditary substance, as DNA was considered too simple to code for the complexity of all life. In 1928, bacteriologist Frederick Griffith postulated that the "transforming principal" was responsible for converting *Streptococcus pneumoniae* from a nonvirulent form with a rough appearance to a virulent form with a smooth appearance. In his classic experiment, Griffith injected live, nonvirulent, rough cells and dead, virulent, smooth cells into the same mouse. Soon after, the mouse died and Griffith recovered live, smooth, virulent cells from the mouse's blood. Griffith concluded that there must be a "transforming principle" that would transform the nonvirulent strain into a virulent strain, and again all bets were on protein[4]. Immunologist Oswald Avery was intrigued by Griffith's experiments, and set out to ultimately identify the transforming material. Along with Colin MacLeod and Maclyn McCarty, Avery published their results in 1944 which repeated Griffith's experiments, reported methods to isolate and purify the active transforming material, and produced data to show that the active fraction was not made up of protein, unbound lipid, or polysaccharides, but consisted solely of DNA[5]. This momentous discovery was muted at the time, and many scientists still believed that more complex proteins must be the transforming substance, not simple DNA. Alfred Hershey and Martha Chase were able to definitively show that DNA was indeed the inherited material using viral DNA labeled with radioactive phosphorous and viral protein labeled with radioactive sulfur. The resulting progeny, made by infecting bacterial cells, were either phosphorus labeled or not labeled, showing that only the DNA was passed down through the generations of viruses[6]. The elusive hereditary material that Miescher isolated almost a century earlier finally had a proven identity: DNA. Although the study of DNA had a slow start since its discovery, Avery's experiments thrusted the simple molecule into the forefront of scientific exploration. Soon, the ultimate determination of the structure of DNA, the molecule once thought too simple to encode all of life's nuances, would launch the age of molecular biology in the second half of the 20th century[7-9].

The evolution of human understanding of DNA, from an uncharacterized, yet clearly novel substance, to the realization that four nucleotides can code for all of life, to finally solving the double helix structure set the stage for a new scientific era: the sequencing revolution. This progression from an accidental discovery to gradually uncovering attributes of DNA to having a full picture of what DNA and its purpose is, parallels the evolution and acceleration of natural product chemical discovery and understanding over the same time period. While natural products have been purposefully used for millennia, only in the last century have researchers begun to understand where natural products come from, how to discover new ones, and why nature makes such useful molecules, again setting the stage for a renaissance in natural product discovery in the age of DNA sequencing.

1.1.2 Natural Products as Medicines in the Twentieth Century

While cellular biology enjoyed a golden age in the nineteenth century, the chemistry of natural products has long been exploited by humans. The earliest records of natural substances being used as medicinal aids come from Mesopotamia circa 2600 B.C. on cuneiform tablets that documented oils from Cypress and myrrh species. The ancient Egyptians and Chinese also documented pharmacological use of plants as drugs[10]. Until the advent of modern medicine, however, bioactive natural products remained locked up inside plant tissues, and so traditional healers would often make a paste or concoction containing the parts of the plant found to have healing properties. Perhaps the first natural product extracted from a traditional source of medicine is morphine. In 1803, Pharmacist Freidrich Serturner isolated morphine crystals from the long used poppy flower for the treatment of pain[11]. Traditional medicines are almost singularly from plants and while they were fairly well documented by many disparate cultures, knowledge about the true origin of the healing powers of these plants only began to be uncovered in the late nineteenth to early twentieth century, concurrent with the unraveling of the truth about the structure and purpose of DNA.

Natural product discovery is normally done by collecting samples from nature, extracting them and testing the extracts in bioactivity assays. Further fractionation of the extract, followed by more bioactivity assays ultimately leads to a compound or suite of compounds responsible for the bioactivity. Chemical structure elucidation then follows utilizing a host of analytical chemistry technologies, such as tandem mass spectrometry

(MS/MS), high pressure liquid chromatography (HPLC), and nuclear magnetic resonance (NMR). This traditional method is often lengthy and expensive, sometimes taking years to chase a bioactivity, and even longer to elucidate the molecule's structure. It is also limited by the amount of source material collected from nature. This is especially a problem for marine natural products, as many of the sources from the marine environment are not able to be grown or kept in the lab, and are often collected from distant places with little re-accessibility. Yet despite these challenges, natural product discovery thrived even prior to the discovery of the structure of DNA. Natural product discovery flourished during the Golden Age of Antibiotics, spurred by the discovery of the first natural products extracted directly from microbes in the mid-20th century. However, by the late 20th century and extending into the 21st century, many pharmaceutical companies have significantly cut down or eliminated their natural product programs[12]. This lull in the focus on natural products in the search for new drug leads is in part due to a high rate of re-discovery of already known natural products, supply problems that often extend the time required to develop a natural product into a pharmaceutical, and advances in combinatorial chemistry making synthetic approaches attractive[10]. However, natural products remain the best leads for drugs. Nature produces more structural and chemical diversity than any synthetic library of molecules and has provided the basis for over 70% of approved drugs from 1981 – 2010[13]. The DNA sequencing revolution, and especially the advent of whole genome and metagenome sequencing, along with improved molecular biology and chemistry techniques, is ushering natural product discovery into a new Golden Age.

The Golden Age of Antibiotics was built on a strong foundation of the burgeoning field of microbiology. In the late 1800s, scientists like Louis Pasteur and Robert Koch pioneered the science of studying microorganisms, and began to realize their prevalence and importance in disease. Before the turn of the century, bacteriologists had isolated the microbial agents that cause a multitude of diseases, including leprosy, the first human disease to be linked to a bacterium; anthrax; tuberculosis; cholera, and many more[14,15]. Emile Roux and Alexandre Yersin are the first to show that a disease, namely diphtheria, is caused by a toxin released by the bacterium *Cornyebacterium diphtheria* by using a filtrate of the cells that kills laboratory animals[16]. Once the connection between pathogen and disease was well established, many physicians and scientists sought to cure these diseases caused by microscopic organisms. Vaccination had been established at the turn of the 18th century with Edward Jenner's vaccine for smallpox, and now armed with the knowledge that some diseases caused by microorganisms could be cured with vaccines, 19th century bacteriologists made vaccines using attenuated cultures of the disease causing bacteria[17]. This method of using a weakened bacterium as prevention for bacterial infection flourished in the early 20th century, but none thought that live bacteria would ever be weaponized against disease causing bacteria.

In 1929, Alexander Fleming reports his serendipitous observation of *Penicillium* mold lysing the *Staphylococcus* variants he had been studying[18]. He ascertained that the mold produced a bacteriolytic substance, or 'mold juice' that diffused into solid and liquid media. With this happenstance observation, Fleming had observed one of the first beneficial uses of microbes. However, like with many groundbreaking discoveries in the history of science, his discovery was greeted with little enthusiasm. That is, until World War II necessitated the use of Fleming's 'mold juice,' which he named penicillin, to treat scores of infected soldiers[19]. Howard Florey and Ernst Chain followed up on Fleming's discovery and were able to finally isolate appreciable amounts of penicillin from the fermentation broth of *Penicillium*, just in time for its widespread use in World War II[20]. This heroic effort represents the first isolation and production scale up of a microbial chemical for use in human medicine, setting the stage for decades of use of natural products as medicine.

While penicillin fought in the trenches, the urgent pursuance of more antibiotics was at the forefront of medicine. A puzzling genus of bacteria would answer the call and become the most prolific genus of bioactive molecule producers. First described in 1875 by Ferdinand Cohn, *Streptothrix fosteri* appeared to be different from bacillus structures known at the time and reminiscent of branching fungal hyphae. Cohn could not isolate and grow the elusive microbe, but the related causative agent for tuberculosis, *Mycobacterium tuberculosis*, was isolated and extensively studied by Koch in 1882. Decades later, in an attempt to classify these strange microbes, or Actinomycetes, Selman Waksman and Arthur Henrici invented the name *Streptomyces* (meaning "twisted fungus") to apply to one group of branched microbes with spores produced in chains[21]. Despite classifying these microorganisms within the group of Actinomycetes, Waksman and Henrici still could not decipher if Actinomycetes were bacteria, fungi, or something in between. Waksman continued to study these indiscernible microbes in the search for an agent produced by a soil microbe that could be used to kill other pathogenic bacteria. Waksman's use of soil microbes was by design: he had established the field of soil microbiology as a discipline and in 1923, before Fleming's *Penicillium* observation, recorded that "Certain actinomycetes produce substances toxic to bacteria… around an actinomycete colony, upon a plate, zone is found free from bacterial growth"[22]. Waksman's interest in antibiotics was re-awakened with the importance of penicillin during wartime and the emergence of the first bacterial natural product, tyrothricin, discovered by his former student, René Dubos[23]. Waksman decided to target gram-negative organisms, as they were not killed by penicillin or tyrothricin, and chose *Escherichia coli* as the target cell. After months of adding *E. coli* to soil pots and recording decreased numbers of living *E. coli* cells until there were none left living, the soil was plated, and about half of the cultures obtained were able to inhibit the growth of *E. coli*. One such actinomycete was selected, grown in culture, extracted with solvent, and in collaboration with the leader of chemistry at Merck, crystals of actinomycin revealed a potent antibiotic[24,25]. Waksman's team rapidly discovered streptothricin, fumigacin, and clavacin from actinomycetes, but it was soon discovered that these first antibiotics were toxic to animals[23]. The breakthrough came in 1944, when Albert Schatz, another graduate student of Waksman, discovered an antibiotic produced by a strain of *Actinomyces griseus* (later reclassified as *Streptomyces griseus*), the first type strain of the genus. Schatz named the antibiotic streptomycin and found that it killed gram-negative pathogens, while retaining little toxicity in animals[26]. Physicians lost no time in testing streptomycin as a treatment for tuberculosis, and within a year they were able to show that streptomycin was an active and promising drug in the treatment of tuberculosis[27,28]. This rapid progression from discovery to clinic was astonishing and highlighted the dire need for antibiotics at the time. Waskman's astute realization that soil bacteria were constantly at war, and would surely make chemical means of defense, precipitated the fruitful field of actinomycete natural products. The following years saw the discovery of a multitude of antibiotics from actinomycetes: tetracycline from *Streptomyces aureofaciens* in 1945, chloramphenicol from *Streptomyces venezualae* in 1947, erythromycin from *Streptomyces erythraea* (later renamed *Saccharopolyspra erythraea*) in 1949, vancomycin from

Amycolatopsis orientalis in 1953, and rifamycin from *Streptomyces mediterranei* in 1957 to name a few[29].

The Golden Age of Antibiotics launched microbial natural products as a productive field, with scientists isolating and characterizing hundreds of molecules for human use in medicine. The natural products field matured from one of folklore to one of directed chemical discovery and application. Fortuitously, advances in understanding and sequencing DNA, molecular biology techniques, and medicinal chemistry coincided with the rise of natural product discovery and development, resulting in greater understanding of nature’s molecules. The role of genetic manipulation and genomic information would accelerate this endeavor and reveal unanticipated mysteries encoded in the genomes of microbial life.

1.2 Connecting Genes and Molecules

1.2.1 Early Genetic Tools in *Streptomyces* for Connecting Genes and Molecules

With the discovery of DNA’s importance in shaping all life and the maturation of microbiological techniques, it wasn’t long until scientists began probing the genetics of microbes. Naturally, *Streptomyces*, already known to be a prolific antibiotic producer, became the workhorse for genetic manipulation development in Actinomycetes. Genetic recombination in *Streptomyces* was first reported in 1955 by Sermonti and Spada-Sermonti, and was born of the need to understand these peculiar microbes, still not definitively classified as bacteria or fungi[30]. At the time, these ambiguous microbes caught the attention of at least five other groups that had started to genetically probe *Streptomyces*[31]. The husband and wife Sermonti team developed a fruitful collaboration with the young David Hopwood at the John

Innes Centre, who would become the leading pioneer of *Streptomyces* genetics. Hopwood set about the daunting task of illuminating the genetics of *Streptomyces* and chose *Streptomyces coelicolor A3(2)* because it produced a bright blue pigment that he thought would make for a valuable genetic marker, but he wouldn't know just how valuable for another few decades. Hopwood set out to make a linkage map of *S. coelicolor A3(2)* using his invention of the four-on-four cross method, where two parents each have two markers, one which is selectable and one non-selectable[32]. The recombinant progeny colonies could be plated on four selective media to select for the selectable parental markers, while retaining the non-selected marker. Using this method, Hopwood constructed the first relative linkage map, which had six loci in two linkage groups, in his PhD thesis at Cambridge in 1958, and expanded on it the next year[31,33]. Hopwood would go on to further elaborate this linkage map throughout the 1960s, which would in turn facilitate the first studies to connect genes and molecules in *Streptomyces*.

The 1970s saw perhaps the most important advances in understanding *Streptomyces* genetics that would inform natural product biosynthesis, resistance, and regulation in *Streptomyces*. The first genetic tool for connecting genes and molecules was mutation, and with Hopwood's linkage map, these mutations could be mapped to the *S. coelicolor* genome. Hopwood and colleagues were able to show that the biosynthetic genes responsible for making methylenomycin A in *S. coelicolor A3(2),* and the genes responsible for resistance to the antibiotic were present on an unusual plasmid, SCP1. A series of mutations on this plasmid led to a lack of antibiotic synthesis and transfer of the wild type plasmid into two other *Streptomyces* species induced production of methylenomycin A in these non-producing strains. Furthermore, the mutated plasmids that lost the ability to produce methyenomycin A were transferred to non-producing strains, which retained resistance to the antibiotic even though production was absent[34-36]. This represents the very first genetic localization of antibiotic biosynthesis genes, and many thought that this case, namely the location on a plasmid, would be the norm. Ironically, methylenomycin A remains an outlier in terms of where the gene cluster resides[31]. Hopwood and colleagues concurrently showed that the biosynthetic genes for actinorhodin, the blue pigmented molecule that drew Hopwood to *S. coelicolor* in the first place, were present as "several closely linked chromosomal genes controlling its synthesis"[37]. This was perhaps one of the first mentions of biosynthetic genes being clustered in a bacterium's genome, which Hopwood later went on to reiterate after further mutation studies in the actinorhodin biosynthetic gene cluster, and thus the idea of the "biosynthetic gene cluster" was born[38]. It would be almost ten years later that mutational cloning analysis would confirm this suspicion, and determine that the resistance, regulatory, and production genes for methylenomycin are present on a contiguous 17kb region of the SCP1 plasmid[39]. These are the first two examples of the connection of biosynthetic genes to the antibiotic molecules that they make. Although they may seem rudimentary now, mutation studies were one of the first genetic tools to link biosynthetic genes to natural products produced by bacteria.

The most significant advances in connecting genes and molecules, however, would come in the 1970s with the advent of gene cloning methods, polymerase chain reaction (PCR) amplification, and DNA sequencing. Recombinant DNA technology was first developed in *E. coli*, with the discovery of restriction enzymes that could cut and re-join DNA at a specific sequence, often a palindromic sequence[40,41]. Soon after, many independent groups isolated DNA ligases, which have the ability to assemble two pieces of DNA together, and the first recombinant DNA molecules were made in 1972[42]. Now that making recombinant DNA had been established, getting bacteria to take it up would be the next challenge. It was originally believed that *E. coli* was recalcitrant to transformation, but in 1970, it was demonstrated that treating *E. coli* with calcium chloride induced uptake of DNA[43]. This phenomenon was exploited to decisively transfer, for the first time, genes conferring antibiotic resistance, residing on a plasmid, to sensitive *E. coli* strains[44]. One year later, in 1973, the very first recombinant DNA had been introduced by transformation into *E. coli*, launching the field of synthetic biology. Foundational work by Okanishi and colleagues in 1974 elucidated the necessary factors for protoplast formation and regeneration in *Streptomyces*, thus allowing for genetic manipulation and transformation of DNA into *Streptomyces* hosts[45]. Others expanded on this protoplast formation protocol to establish an efficient plasmid transformation protocol[46] as well as protoplast fusion protocols[46,47]. These first genetic manipulation tools for use in the most prolific antibiotic producers in nature set the stage for a flurry of genetic experiments to connect antibiotic biosynthetic genes to their molecules.

In 1980, multiple groups reported the first instances of gene cloning in *Streptomyces*. These first cloning experiments purposefully focused on antibiotic resistance genes, which had recently been used to make the first antibiotic selectable cloning plasmid, pBR322, in *E. coli* in 1977[48]. Use of antibiotic resistance genes as selectable markers on plasmids was a foundational invention, facilitating the establishment of molecular cloning systems and protocols that are still modern gold standards. Mervyn Bibb, having completed his PhD in Hopwood's lab, was successful in cloning the gene encoding resistance for methylenomycin A, making the first antibiotic-selectable plasmid vector for use in *Streptomyces*[49]. Shortly after, researchers in Hopwood's lab cloned the resistance genes for neomycin and thiostrepton, resulting in more selectable plasmids for use in cloning[50]. Additionally, *Streptomyces* promotor regions were isolated and for the first time it was recognized that "genus or species-specific factors may present barriers to the expression of bacterial genetic material in certain heterologous cellular environments," an astute observation that researchers are still exploring today[51]. These early advances laid the foundation for exploring antibiotic biosynthesis in *Streptomyces*. In 1983, the first gene directly involved with the biosynthesis of an antibiotic was cloned, allowing for comparison between two related pathways. Undecylprodigiosin, a red pigment produced by the ever fertile *S. coelicolor* A3(2), is structurally similar to the red pigment, prodigiosin, produced by S*erratia marcescens*. Using co-synthesis studies with blocked mutants, already established in *S. marcescens*, and red mutants generated in *S. coelicolor*, the authors were able to ascertain that the *redE* gene is likely an O-methyltransferase. The *redE* gene was cloned into plasmid pIJ702 and re-introduced to a *S. coelicolor* strain with a mutation in the *redE* gene. This restored production of the methylated undecylprodigiosin, and a methyltransferase activity assay showed a level of methylation was restored to the wild type levels[52]. The authors clearly stated that their motivation for cloning the first biosynthetic gene was to lay the groundwork to further understand genetic control of secondary metabolite production, and perhaps begin to manipulate antibiotic biosynthetic pathways. Subsequent studies were published that same year for cloning biosynthetic genes from the methylenomycin A pathway[53] and the candicidin pathway[54], accelerating the connections of genes to molecules, and furthering understanding of the biosynthetic genes responsible for natural products. What these early studies couldn't do, however, was reveal the true nature of antibiotic pathways. While foundational and necessary to establish molecular cloning in actinomycetes, these first single gene studies looked mainly at resistance factors and tailoring genes, missing the meat of a biosynthetic pathway. The paradigm of modular assembly line pathways remained hidden in the genomes of antibiotic producers.

1.2.2 Cloning Entire Pathways and the First Unnatural Natural Product

It didn't take long to go from cloning one or two genes to cloning the first whole antibiotic biosynthetic pathways. Again the famous blue pigment antibiotic, actinorhodin, was back in the spotlight. Labeled acetate feeding studies had previously determined that the molecule derived from the polyketide pathway, as had been previously shown for similar antibiotics, nanaomycins, naphtocyclinon, and granaticin, but no one knew yet what the polyketide pathway looked like genetically and biochemically[55]. A series of mutants, each blocked in the *act* pathway at different points, were mapped and determined to be in the same segment of the *S. coelicolor* chromosome, making this cluster of genes a good candidate for capturing the entire pathway for actinorhodin. Random pieces of DNA 15-30 kb long were cloned into a *S. coelicolor* mutant that lacked production of actinorhodin. Out of about 8,000 transformants, two produced the characteristic blue color. The isolated plasmids were introduced into *S. coelicolor* mutants and each plasmid complemented some of the mutants, but neither restored production in all mutants, suggesting that the whole cluster had not yet been captured. This prompted the authors to create a third plasmid with the two overlapping portions of the *act* biosynthetic gene cluster. The resulting plasmid restored actinorhodin production in all *S. coelicolor* mutants and a non-producing heterologous host, *Streptomyces parvulus*[56]. This seminal work informed and inspired decades of molecular cloning and heterologous expression of biosynthetic pathways to this day.

Further work with the actinorhodin gene cluster resulted in the creation of a hybrid antibiotic, which is still an elusive feat in synthetic biology today. Specific segments of the *act* gene cluster were cloned into a plasmid vector that was then introduced into *Streptomyces* sp. AM-7161, which makes medermycin (a brown pigment). With the appearance of a purple pigment, Hopwood who had teamed up with American chemist Heinz Floss and Japanese microbiologist Satoshi Omura noticed they had generated a new chemical entity in this hybrid strain, representing a "'cooperation' between gene products in the actinorhodin and garaticin pathways."[57] The new polyketide molecule was isolated and named mederrhodin A, a compound with structural features from both medermycin and actionorhodin. Interestingly, when the whole actinorhodin pathway was introduced to the medermycin producer, only medermycin and actinorhodin were detected. This cutting edge application of genetic tools to produce a hybrid compound is astonishing, especially because it was done "blind," without the guide of sequence information. The first *Streptomyces* gene was sequenced in 1983 by Charles Thompson of the John Innes Centre and Gary Gray of the Swiss pharmaceutical company Biogen, but widespread use of this new technology was sparse. The first *Streptomyces* gene sequence, an aminoglycoside phosphotransferase (*aph*) from the neomycin producer *S. fradiae* ATCC 10745 and also an antibiotic-resistance gene, hinted at the power that genomic sequencing would have to influence the understanding of antibiotic biosynthesis across domains of bacteria. It was in the comparison of the *S. fradiae aph* to other sequenced *aph* genes that the promise of DNA sequencing was realized, that evolutionary comparisons could be made between homologous genes to infer structure and function[58]. In 1988, the *actIII* gene, already deduced by mutation studies to play a role in actinorhodin polyketide chain modification but not chain assembly, was sequenced and revealed to be a putative oxidoreductase by comparison to dehydrogenase genes in *Klebsiella aerogenes* and *Drosophila melanogaster*[59].

Concurrent to the explosion in techniques and tools to connect biosynthetic genes to their molecules, namely in the fruitful *Streptomyces* genus, was an expansion of general molecular biology tools, usually born from fundamental scientific research into bacterial enzymes. In studying antibiotic production and native resistance, the natural products community contributed to one of the most widely used techniques still today: molecular cloning. Exploitation of antibiotic resistance genes as selectable markers and the use of restriction enzymes and ligases isolated from *E. coli* led to sophisticated vector design and development. The following basic science developments, namely DNA sequencing and the invention of the PCR, would revolutionize all fields of science and particularly inform antibiotic biosynthesis in its nascent years.

1.2.3 DNA Sequencing Revolutionizes Natural Product Biosynthetic Understanding

Like most impactful discoveries in science, DNA sequencing had a slow start. It progressed from amino acid sequencing of a protein, to simple RNA molecules, and finally provided the first DNA sequences decades later. Frederick Sanger began his storied career in sequencing in 1951 through a series of heroic efforts to systematically hydrolyze and chromatographically separate pieces of the insulin protein to finally piece together the sequence of amino acids[60]. Protein sequencing flourished throughout the 1950s and 1960s, creating the first collection of sequence data, hinting at the databases of sequence information yet to come. Although proteins may seem intuitively more complex than DNA, they proved easier to sequence because the similar, repeating nucleotides that make up long DNA chains made it more difficult to distinguish between one nucleotide and another. RNA was the next logical step in the effort to sequence, as they were short sequences, unencumbered by a complementary strand, and readily abundant[61]. The first whole nucleic acid sequence, an alanine tRNA from *Saccharomyces cerevisiae*, was completed in 1965 by Robert Holley and colleagues, using RNase enzymes, analytical chemistry, and selective ribonuclease treatments[62]. This process took five people working for three years with one gram of material, resulting in 76 nucleotides. Simultaneously, Sanger was developing a "fingerprinting" method based on the detection of radioactively labeled RNA fragments and subsequent visualization in two dimensions[63]. The subtle complexities of DNA would still evade capture by sequence for some years, and would require creative solutions to solve the first sequences.

The first twelve DNA bases were sequenced using DNA polymerase and primer extension in 1968[64], and another method converting DNA to ribonucleic acid (RNA) and sequencing those fragments decided twenty-four bases in two years[65]. These early methods were not scalable, and it wasn't until 1977 that two methods were reported that could decode hundreds of bases, transforming this budding field practically overnight. Both methods used the distance from a radioactive label to the position of a base within a molecule of DNA. Sanger's method, the chain terminator procedure, involved four extensions of a labeled primer, each with a small amount of a known chain-terminating nucleotide, thus producing different length fragments, depending on where the chain-terminating nucleotide was incorporated[66]. Gilbert's method relied on a terminally labeled DNA restriction fragment, chemically fragmented at specific bases[67]. In both techniques polyacrylamide gels were used to determine the lengths of each DNA fragment, down to the single base pair resolution. While these techniques were used in labs around the world, it wouldn't be for another decade that the first commercially available, automated sequencing machine would be introduced. In the meantime, strategies, such as "shotgun sequencing" of random clones, followed by sequence assembly using overlaps began amplifying the uses of DNA sequencing[68]. By 1982, with the establishment of GenBank, a central repository for all sequence information, over half a million bases had been deposited, and a year before the first commercial sequencers were introduced, in 1986, almost ten million bases were in Genbank[69].

1.2.4 DNA Sequencing Informs Polyketide Synthase (PKS) Pathway Understanding

In natural product research, DNA sequencing would fundamentally inform the biosynthesis of different classes of antibiotics and ultimately aid in the discovery of new bioactive molecules from nature. Perhaps the first groundbreaking development in understanding antibiotic biosynthesis using sequencing came from the seemingly simple polyketide pathway that could make an incredibly diverse structural range of products with widespread functions. In 1989, cloning and sequencing the pathways for granaticin and tetracenomycin C, the first bacterial polyketide synthase (PKS) structures studied by molecular genetics, revealed that the PKS genes are indeed similar in sequence and organization to fatty acid synthases (FASs), as had been postulated based on biochemical studies[70,71]. In both studies, the acyl carrier proteins (ACPs) were also identified using sequence homology, and suggested that different ACPs are utilized in fatty acid and polyketide biosynthesis. Overall, both studies concluded that they had sequenced PKS gene clusters, homologous to FAS type II gene clusters from primary metabolism, consisting of separate subunits, which turns out to be the rule for aromatic, or type II polyketide synthases. Subsequent studies on the sequence of the erythromycin biosynthetic gene cluster in

Saccharopolyspora erythraea revealed an unexpected second group of polyketide biosynthetic genes, previously only known in eukaryotes: type I PKSs. Erythromycin was found to be made not by discrete enzymes encoded by different genes, but instead by large, multi-modular proteins consisting of all the necessary domains for constructing the polyketide backbone[72]. Remarkably, sequencing revealed that multiple modules of domains were present, forming what is now commonly referred to as assembly line PKS biosynthesis[73]. For almost a decade after the discovery of type I PKS pathways in bacteria, researchers believed there were only type I and II PKS systems in bacteria, but again, DNA sequencing would reveal otherwise. In 1999, two groups reported sequencing and characterizing bacterial genes that showed homology to chalcone synthase and stilbene synthase genes, previously known only in plants[74,75]. In both cases, the type III PKS molecule structures were known, but only after genome sequencing was the underlying mechanism for their assembly elucidated. A fourth type of PKS pathway emerged from DNA sequencing data in 2002, when two groups concurrently discovered a new type of PKS system. Jörn Piel noticed a lack of committed acyltransferase (AT) domains in each module in the gene cluster for pederin. Instead, two AT domains are present upstream of the cluster that can act in *trans*, servicing all the ketosynthase (KS) domains in the cluster[76]. Simultaneously, Ben Shen cloned the pathway for leinamycin and documented the "AT-less" nature of the pathway[77,78]. These findings represent the first characterization of *trans*-AT PKS pathways, which are now known to be quite widespread in nature, especially in understudied taxa[79].

Only through sequence analysis and comparison could the underlying mechanisms of *trans*-AT, type I, II, and III PKS biosynthesis be observed. In the case of type II PKS systems, gene sequence directly informed the programming of these assembly line gene clusters, effectively allowing researchers to ‘read’ the construction of a PKS through each module, hinting at the predictive powers to come in PKS biosynthesis[80]. These advances in understanding iterative and assembly line PKS systems further opened the door to informed combinatorial biosynthesis with the ambition of making unnatural natural products.

While the first combinatorial antibiotic had been made in 1985, it was done so in a fairly random manner. Armed with gene sequence information, Chaitan Khosla’s lab at Stanford, in close collaboration with David Hopwood, set about performing directed combinatorial studies that would successfully produce unnatural natural products, while providing new insights into the key features of these programmable enzymes. In 1993, they found that ACPs were interchangeable, carbon chain length is dictated by a chain length factor (CLF) particular to type II PKSs, and that a given ketoreductase (KR) can recognize and reduce different lengths of polyketide chains, among other things[81]. Khosla and colleagues soon devised a set of design rules for the directed manipulation of PKS assembly lines by genetic engineering. They used their rational design to produce two new PKS molecules using natural enzymatic subunits from multiple PKS pathways to obtain the desired molecules[82]. These studies also introduced the idea of the ‘minimal PKS,’ consisting of a KS, ACP, and CLF, which was the starting point for designing new PKS molecules. Just before the turn of the century, KOSAN Biosciences manipulated the erythromycin pathway to produce a library of over 50 macrolides, a feat that would be impractical to pursue using synthetic chemistry and one that would prove that informed design based on sequence information was worth pursuing[83]. Soon researchers could pick and choose particular structural features to gain certain functions in their designer molecules, and the advent of whole bacterial genome sequencing would provide even more insight into the biochemical pathways and regulation of these pathways encoded in bacterial genomes[84].

1.2.5 Patterns in NRPS Domain Architecture Unveiled by DNA Sequencing

Gene sequencing momentously advanced our understanding of PKS pathways in bacteria in ways that wouldn't have been possible without DNA sequencing. Parallel to the unravelling of PKS pathways, understanding of non-ribosomal peptide synthetase (NRPS) pathways was being revealed by DNA sequencing as well. Foundational work by Fritz Lipmann had shown that NRPS pathways were already known to encode large, multi-enzyme complexes with a multi-domain structure (as was shown by sequencing for type I PKSs), but gene sequencing still had more to reveal about NRPS pathways[85]. The first nucleotide sequence of a gene in an NRPS pathway was for the *tycA* gene, encoding the tyrocidine synthase from *Bacillus brevis,* and was reported in 1988 by Mohamed Marahiel's group, but it wouldn't be until the next NRPS genes were sequenced that comparative insights could be gleaned[86]. In 1990, the *acvA* gene, initiating penicillin biosynthesis in *Penicillim chyrsogenum* Oli13, was sequenced and showed homology with the *B. brevis tycA* gene. While the authors recognized that the function of these enzymes is unknown, the similarity between discrete domains within genes known to encode enzymes that adenylate amino acids led them to speculate that homologous domains within *acvA* also activate amino acids. In fact, the authors note that the presence of three such domains may adenylate three amino acids in the resulting penicillin structure[87]. Further studies go on to elaborate the role of adenylation (A) domains in peptide synthetases, definitively showing a link between these repetitive domains and the amino acids selected for in the final product[88-91]. Gene sequencing went on to reveal an NRPS gene with an N-methyl-transferase domain, previously uncharacterized[92]. In 1999, the first bioinformatic tool for predicting A domain specificity was developed. Marahiel, Stachelhaus and colleagues used the recently solved crystal structure of the gramicidin synthetase phenylalanine activating A domain, PheA, to compare residues lining the binding pocket with the corresponding moieties in other A domains[93]. They were able to construct general rules for predicting A domain substrate specificity using in-silico studies and structure-function mutagenesis[94]. This so called Stachelhaus code provides one of the greatest predictive tools in NRPS biosynthesis still today, and provides the basis for many A domain bioinformatic prediction tools discussed in section 1.3.4.

Another tool for connecting NRPS genes with molecules exploited the repetitive, highly homologous nature of NRPS genes. Degenerate primers were designed by sequence comparisons of domains from the gramicidin, tyrocidine, and penicillin synthetases. PCR was then utilized to successfully amplify the previously unidentified surfactin peptide synthetases from *Bacillus subtilis* ATCC 21332[95]. The authors went on to establish this method as a useful protocol for reliably finding NRPS genes in known peptide antibiotic producers, and perhaps in unknown producers as well[96]. This method was used in a wide variety of cyanobacteria and showed that many of the strains tested, whether they produced a toxin or not, contained PCR amplifiable regions of NRPS genes, hinting at the silent gene clusters that whole genome sequencing would soon reveal[97]. Early comparative sequence analysis allowed for the development of these important tools that would play an active role in discovering new NRPS molecules.

Peptide synthetase gene sequence analysis first revealed the highly repetitive A domains as a common feature of NRPS genes, and soon other patterns would emerge from the DNA sequences. Distinct domains were consistently observed upstream of A domains, with the exception of the loading module, and were speculated to be condensation (C) domains, playing a role in peptide elongation. This n-1 domain pattern (where n = the number of A domains) corresponded to the number of peptide bonds in a given NRPS molecule, bolstering the identity of these domains as C domains[98]. Another class of genes, the new superfamily of phosphopantetheinyl (P-pant) transferases, determined to be necessary for non-ribosomal peptide synthesis, were discovered through sequence comparisons by Christopher Walsh's lab. The first P-pant transferase to be sequenced was the *E. coli* ACPS, involved in fatty acid biosynthesis, but surprisingly, initial searches of sequence databases did not turn up any proteins with significant homology[99]. However, after refining sequence alignments, two shared motifs emerged in several putative ACPS homologs. This sequence alignment and comparison led to the discovery of a large family of proteins, putatively P-pant transferases. Walsh, Khosla, Marahiel and colleagues went on to characterize these proteins, confirming P-pant transferase activity[100]. This elegant example shows how comparative sequence analysis can reveal biosynthetic mysteries that may not have been uncovered otherwise.

The examples discussed here are by no means comprehensive, but they do show the vast impact that gene sequencing had on understanding, predicting, and manipulating natural product biosynthesis. Up until this point, however, only single genes or gene clusters had been utilized in the quest to understand natural product biosynthesis at the genetic level. Most early gene sequencing studies focused on known secondary metabolites, which informed the sequencing projects to connect them to genes responsible for assembling them. These studies, while foundational for understanding natural product biosynthesis, yielded few new molecular entities, with the exception of the innovative combinatorial biosynthesis studies, but the full promise of DNA sequencing would soon fill this gap. Coming developments in DNA sequencing technologies to allow for economical whole genome sequencing would greatly expand the success in natural product biosynthesis initiated by gene sequencing efforts, and even lead to a new mode of discovery of natural products from environmental microbes.

1.3 Genome Mining in the Twenty-first Century

1.3.1 Sequencing Phylogenetic Tags Exposes the Uncultured Majority of Microbes

The seemingly small progression from sequencing genes to genomes propelled our understanding of natural product producing organisms to new heights. Early whole genome sequencing was a lengthy process, expensive, and accessible to only a few large sequencing centers[101]. With the introduction of the first commercial Sanger sequencing machines in 1987, single gene sequencing flourished, but to sequence a whole genome was a momentous task. Sanger sequencing, as it relied on gel electrophoresis to read each base, was not a high-throughput initiative. It would take eight years after the first commercial sequencing machines were released for a bacterial full genome sequence to be reported. In 1995, the roughly 1.8 Mb genome of *Heomophilus influenza* was sequenced and reported by Robert Fleischmann and colleagues, followed closely by the report of the 0.58 Mb genome of *Mycoplasma genitalium* by Craig Venter's group[102,103]. In the following decade, about 300 sequenced bacterial genomes would be published[101]. Because of the cost and labor, whole genome sequencing languished behind the invention of the first commercial DNA sequencer. Instead, it was much more common to sequence single genes and sometimes stretches of a chromosome, as was seen in the natural product biosynthesis community. Other researchers sought to use sequencing as a way to probe the environment for microbial diversity, revealing an astonishing abundance of unknown, uncultured microbes, which would prove to be a promising avenue in natural product discovery.

Genetic tools to classify and analyze the phylogenetics of microbes were established by the mid-1980s, with the use of ribosomal RNA sequencing, specifically the conserved marker 16S rRNA[104]. Scientists used this method to show a huge abundance and diversity of uncultured microbes in soil communities[105], extremophilic host-spring environments[106,107], and marine environments, including hydrothermal vent-associated symbionts[108] and pelagic bacterioplankton communities[109,110]. These biodiversity profiling studies established the phenomenon of the ‘uncultured microbial majority,’ with uncultured species vastly outnumbering cultured ones. Large numbers of 16S rRNA sequences began populating GenBank in 1996 as the techniques for 16S rRNA profiling became commonplace[111]. These studies brought recognition to the untapped reservoir of diverse environmental microbes and potential for novel natural product chemistry, giving rise to the field of metagenomics. Metagenomics of soil organisms was of particular interest, and the idea to clone environmental DNA from soil into cultivatable microbes and subsequently screen for new chemistry allowed, for the first time, access to previously inaccessible biosynthetic pathways[112]. The search for novel natural products had helped establish an entirely new discipline and just as the unseen majority of microbes was beginning to be appreciated, so too was the hidden majority of biosynthetic gene clusters locked up in microbial genomes.

Actinomycetes, especially the genus *Streptomyces*, had long been lauded for their abilities to produce an abundance and variety of natural products. Gene sequencing allowed for an in depth look into the biosynthetic pathways of known metabolites and hinted at the ability to find new natural products using degenerate primers and gene probes based on known PKS or NRPS sequences. Sequencing of whole biosynthetic pathways allowed for common features and patterns to emerge from the data, even discovering a whole new superfamily of proteins. But what researchers would find in the first genomes of antibiotic producing *Streptomyces* would change the course of natural product discovery.

1.3.2 Whole Genome Sequencing Reveals Abundance of Cryptic Gene Clusters

The workhorse of actinomycetes, *Streptomyces coelicolor* A3(2), with multiple antibiotics already isolated and many biosynthetic pathways characterized, was the first antibiotic producer to be sequenced in its entirety. The behemoth of a bacterial genome, greater that 8.6 Mb and containing the largest number of genes discovered yet in a bacterium, revealed over twenty predicted gene clusters encoding known and predicted secondary metabolites. The gene clusters responsible for actinorhodin, CDA, and the *whiE* grey spore pigment had been analyzed previous to the whole genome sequencing effort, but the presence of at least eighteen uncharacterized gene clusters containing hallmarks of secondary metabolism was unanticipated. Newly developed bioinformatic tools at the time aided in predicting two NRPS gene clusters responsible for making the siderophores coelichelin and coelibactin[113]. The structure of coelichelin was the first natural product to be predicted solely based on genomic sequence information, giving rise to the field of genome mining[114]. The remaining gene clusters in the complete *S. coelicolor* A3(2) genome sequence were able to be putatively identified, some with predicted molecular products, and at least classified into the type of molecule they would make, again establishing the reach of genome mining in identifying the ‘silent majorty’ of cryptic gene clusters.

The next sequenced *Streptomyces* species, was the industrially important *S. avermitilis*, producer of the antiparasitic avermectins. The *S. avermitilis* genome was revealed to be slightly larger at 9.0 Mb, and was shown to contain thirty gene clusters related to secondary metabolite biosynthesis, with this specialized function comprising 6.6% of the genome[115]. Previous reports of genome sequence of this organism identified 25 secondary metabolite gene clusters, but with complete sequencing, 5 more were revealed[116]. The authors predicted the structures for some of the metabolites encoded and experimentally confirmed production of geosmin, pentalenolactone, squaliene, and pentaene, all predicted from genome sequence. A comparative genomics approach revealed a higher similarity between *S. coelicolor* and *S. avermitilis* than with other whole genome sequenced organisms, including two strains of *M. tuberculosis*, *E. coli*, and *B. subtilis*. The two *Streptomyces* genomes were shown to have 69% similarity in their genes, with unique ORFs largely coding for proteins implicated in secondary metabolism, transcriptional regulation, transposition, and degradation of xenobiotics. Despite high similarity in the two genomes, interesting differences arose. For example, both *Streptomyces* species are resistant to chloramphenicol, but it appeared that *S. coelicolor* gained resistance through efflux pumps shown in other Actinomycetes to confer resistance, but *S. avermitilis* lacked these pumps yet had a chloramphenicol phosphotransferase, which was shown to confer resistance in the chloramphenicol producer, *S. venezuelae*. Furthermore, it was noted that many secondary metabolite biosynthesis genes, including those for producing antibiotics, were not conserved between the two species[115]. This observation was particularly salient and more strains of the same genus sequenced would begin to show this to be true between strains of the same species.

S. coelicolor and *S. avermitilis* had been extensively studied, their genetics probed for decades, and yet the startling discovery that their genomes had so much more to offer prompted a renewed search for natural products using genome mining. With new ways to access unculturable bacteria and the observed abundance of silent, or cryptic, gene clusters the biological and chemical spaces to discover new natural products greatly expanded. The invention of massively parallel, high throughput genome sequencing gave scientists the tools to explore the previously inaccessible spaces. Genomic data trickled in during the age of Sanger sequencing, even the first full genomes were few and far between, but soon, the invention of next-generation sequencing technologies would create a deluge of genomic data.

1.3.3 Next-Generation Sequencing Leads to Explosion in Sequence Data

Sanger sequencing proved to be a reliable method for getting reasonably long sequence reads, and continues to enjoy widespread use today. However, Sanger sequencing would reach a limit in terms of output, as it relies on gel electrophoresis to read each base, using, at most, 384 well capillary machines[101]. An old invention in DNA sequencing would solve this limit, creating a new wave of next-generation DNA sequencers. The basis for the method, called pyrosequencing, took advantage of one of the byproducts of DNA synthesis: the release of a pyrophosphate molecule as each nucleotide is incorporated. The release of pyrophosphate could be detected using a two-enzyme process, which uses ATP sulfurylase to convert pyrophosphate into ATP, which is in turn used as the substrate for luciferase[117]. Exploitation of this natural chemical transformation in DNA synthesis led to a sequencing approached where a template DNA is affixed to a surface, nucleotides are washed over the template, and if a base is added, light can be detected[118]. Although both Sanger sequencing and pyrophosphate sequencing are sequencing by synthesis models and they were developed around the same time, some advantages of pyrosequencing allowed it to be developed into the first major successful commercial sequencing technology. First, expensive, modified dNTPs used in chain-termination methods were not required, and natural bases could be used. Second, the process can be observed in real-time, unlike the lengthy gel electrophoresis readout of Sanger sequencing[119,120]. Pyrosequencing was licensed to 454 Life Sciences where it was developed into a machine with the capability to sequence highly parallelized samples, massively increasing the amount of DNA that could be sequenced in one fell swoop[121]. This paradigm shift re-defined how DNA sequencing is used in experiments, leading to an abundance of whole genome sequences and launching multiple initiatives to sequence DNA cheaper and faster[122-126]. For a complete review of next-generation sequencing technologies, see “Coming of age: ten years of next-generation sequencing technologies” by Goodman, McPherson, and McCombie[127].

The cost for DNA sequencing continues to drop throughout the 21st century with faster, easier, and cheaper methods for sequencing available to almost all labs all over the world. In fact, the capability of DNA sequencers has surpassed the rate observed in the ‘computing revolution’ described by Moore’s law: the number of transistors per unit cost doubles about every two years. DNA sequencing capacity, between 2004-2010, doubled every five months[128]. This democratization of sequencing has driven a ‘genomics revolution’ that the natural products community has enthusiastically taken part in, with the promise of genome mining delivering multiple new chemical entities from predicting sequence data.

1.3.4 Computational Tools Help Predict Natural Products

With new access to unprecedented amounts of DNA sequence data, the natural product community set out to characterize the ‘bacterial dark matter’ hidden within the genomes of bacteria. Bioinformatic tools have been instrumental in making the abundance of data manageable and meaningful. One of the most impactful tools for deducing gene function by homology is the National Center for Biotechnology Information’s (NCBI’s) Basic Local Alignment Search Tool (BLAST)[129]. Genbank, introduced in 1982, was the first public, central repository for genomic sequence data, and eventually, with the amount of sequence data building up, a method to search it and compare sequences was necessary. Additionally, researchers were beginning to notice patterns in the sequences of biosynthetic gene clusters of secondary metabolites, which could be used to make sets of rules for predicting the class of secondary metabolite and even structural features of the resulting molecule. One of the first tools developed examined 102 glycosyltransferases from 52 biosynthetic gene clusters to correlate sequences of glycosyltransferases to their structures and corresponding substrates[130]. SEARCHGTr would highlight the power of comparative genomics of many genes carrying out similar reactions. Another early tool, NRPSPredictor, was made into a freely available online webserver, designed to predict A domain specificity in NRPS gene clusters[131]. NRPSPredictor was first made available in 2005, when widespread full genome sequencing was still in its infancy. The tool was updated and improved in 2011 after more bacterial NRPS sequences became available and was expanded to predict fungal A domain specificity[132]. Since NRPSPredictor, a host of other online tools have been developed to predict substrate specificity of NRPS and PKS pathways, including: NRPSsp[133,134], NRPS/PKS substrate predictor[135], PKS/NRPS Web Server[136], LSI based A-domain function predictor[137], SEQL-

NRPS[138], and PKSIIIexplorer[139]. All of these tools relied on the comparison of thousands of sequences to ultimately achieve high confidence prediction, a feat unfeasible before the introduction of next-generation sequencing.

While DNA sequencing showed early on that NRPS and PKS pathways contained patterns that made substrates straightforward to predict, ribosomally synthesized and postranslationally modified peptides (RiPPs) had been notoriously difficult to computationally identify[140]. RiPP gene clusters can be very small and have no universally shared signature genes, making them difficult to find[141]. With these caveats in mind, the Kuipers group set about making a tool to locate putative bacteriocin, or RiPP, pathways. Their resulting program, BAGEL, uses RiPP databases to identify motifs and takes into account the genomic context so that genes accessory to the core RiPP assembling genes can inform the location of such gene clusters[142,143]. Another early effort manually searched, using PSI-BLAST for homologues of the LanM lanthipeptide dehydratase/cyclase using the sequence of *ltnM1* from the lacticin cluster[144]. This strategy revealed 89 strains containing homologues, 61 of which were not known lanthipeptide producers. These efforts resulted in the isolation of lichenicidin, a novel antibiotic[145].

Perhaps the most widely used computational tool by natural product researchers today is antiSMASH[146-148]. First developed in 2011, antiSMASH is a user-friendly web tool that identifies over 40 types of known secondary metabolite classes. It identifies and annotates secondary metabolite gene clusters and gives rudimentary structural predictions. This holistic software utilizes many standalone tools for an integrated analysis, including MultiGeneBlast (MGB)[149], Natural Product Domain Seeker (NaPDoS)[150], multiple A domain predictor software mentioned previously, and ClusterFinder[151]. MGB is a standalone program that allows similarity searches for cassettes of multiple genes and is useful for determining if a gene cluster is similar to others in sequenced organisms. antiSMASH also uses MGB to compare identified gene clusters to characterized gene clusters in the minimum information about a biosynthetic gene cluster (MIBiG) database[152]. MIBiG is a curated database of experimentally validated biosynthetic gene clusters that is especially useful for dereplicating clusters that make known molecules and identifying clusters similar to characterized ones that might make a structural variant of a known molecule. NaPDoS also utilizes characterized KS and C domains to phylogenetically clade input KS and C domains, classifying them and predicting the kind of chemical transformations they can perform. ClusterFinder can address one of the shortcomings of antiSMASH: predicting noncanonical biosynthetic pathways. While antiSMASH is very good at identifying known classes of biosynthetic gene clusters (ie PKS, NRPS, RiPPs, terpenes, etc), its algorithms rely on the genomic underpinnings of these cluster types. ClusterBlast, however, can pick out putative clusters by identifying genes that are usually involved in secondary metabolism biosynthesis, but may not belong to a class of biosynthetic gene cluster. Another shortcoming of antiSMASH is the crude structural predictions provided. PRISM (Prediction informatics for secondary metabolomes) was introduced in 2015 to create combinatorial structure predictions for NRPS and PKS gene clusters[153]. The structural predictions take into account the permutation of monomers utilized in the structural skeleton and the variability in the action of tailoring enzymes to provide a library of possible structure predictions from a given NRPS or PKS biosynthetic gene cluster. A recent rebuild of the software expanded the cluster types it can predict, widening its scope to twenty-two distinct natural product cluster types[154].

Software prediction and identification of secondary metabolite biosynthetic gene clusters has greatly improved over the last few years, and with the deluge of genomic data, prioritizing gene clusters for further study has become an urgent endeavor. One method of prioritization is to look for clusters that might make molecules with new molecular targets. Often, bacteria will make a natural product that has a cellular target and must avoid suicide while making such a toxic molecule. One way to do this is to have a duplicated copy of the target that is mutated in such a way that confers resistance to its own inhibitor. An elegant example of this is seen in the producer of the potent proteasome inhibitor, salinosporamide A. *Salinispora tropica*, a marine-obligate actinomycete, makes salinosporamide A, a proteasome inhibitor, and possesses proteasome machinery. The biosynthetic gene cluster encoding salinosporamide A also contains a duplicate copy of the gene coding for the beta-proteasome subunit that salinosporamide A attacks. This accessory beta-proteasome subunit has a mutation that confers resistance not only to salinosporamide A, but other proteasome inhibitors as well[155,156]. Co-localizing the biosynthesis of the resistant target with the biosynthesis of the inhibitor is an efficient way to ensure self-resistance. These duplicated, often mutated cellular targets provide the opportunity for target-directed genome mining[157]. A recently developed software tool, the antibiotic resistant target seeker (ARTS) utilizes this concept to identify duplicated, phylogenetically incongruent housekeeping genes localized to secondary metabolite biosynthetic gene clusters[158]. This tool is one of the first to provide a logical basis for gene cluster prioritization based purely on bioinformatics, and more tools like this will be necessary as the surge in genomic information shows no signs of slowing.

1.3.5 Integrating Genomic and Metabolomic Data

A powerful way to utilize genomic information for predicting secondary metabolite structures and dereplicating, or removing redundant, known structures is to couple it with metabolomics data. These methods were first developed for peptide natural products. One of the first developments in this field was the introduction of natural product peptidogenomics (NPP), a mass spectrometry (MS) guided genome mining method connecting peptides detected in tandem MS experiments to their biosynthetic gene clusters[159]. NPP was able to characterize ten diverse peptides using this method of matching observed amino acid fragment masses with expected amino acids from A domain predictions, and served as an early proof-of-concept for automating genome mining by integrating metabolomic data. Multiple tools followed, including Pep2Path[160], which succeeded in automating peptide genome mining pioneered by NPP; cycloquest[161], a tool for matching cyclopeptides to their gene clusters; iSNAP[162], an informatics search strategy developed for NRPS dereplication; and glycogenomics[163], for quick characterization of glycosylated natural products and relation to their corresponding gene clusters. These early specialized tools gave way to more comprehensive tools, such as the genomes to natural products (GNP) platform, which utilizes high-throughput MS/MS data and genomic information to connect NRPS and PKS molecules to their gene clusters[164]. Recently, the first retro-biosynthetic program was introduced as a method to connect known molecules to their unknown biosynthetic gene clusters. The generalized retro-biosynthetic assembly prediction engine (GRAPE) and the global alignment for natural products cheminformatics (GARLIC) endeavor to connect known chemical entities to their gene clusters and aid in the discovery of new molecules from orphan gene clusters[165].

Bioinformatic tools (for a more comprehensive list see the Secondary Metabolite Bioinformatics Portal at http://www.secondarymetabolites.org/) have greatly increased the productivity of genome mining the vast amounts of genomic data available today. Although exact structural predictions solely from genomic information are still not feasible in most cases, integration of easily acquired MS/MS data have helped in connecting observed molecules to sequenced gene clusters. Bioinformatic tools have also aided in the dereplication process that is necessary to deprioritize known molecules and prioritize novel chemical scaffolds. One overall deficit in bioinformatic predictions of natural products is the caveat that we can only find what we're looking for; that is, completely new classes of molecules with unknown underlying genetic basis will be missed by these bioinformatic tools. While ClusterFinder seeks to address this issue, it often identifies so many putative clusters that sifting through and prioritizing them can remain a daunting task. Biochemical characterization of enzymes performing new types of chemistries will surely aid in the discovery of new classes of bioinformatic gene clusters. Additionally, wide-scale genomic comparisons of gene clusters from thousands of bacterial genomes can aid in uncovering more secrets from bacterial genomic dark matter.

1.3.6 Large Scale Analyses of Bacterial Genomes Reveal a Wealth of Undiscovered Biosynthetic Gene Clusters

With bacterial whole genome sequencing relatively commonplace, the secondary metabolite potential, diversity, and distribution in cultured microbes are being revealed on a large scale. Just as the need to prioritize gene clusters has become paramount, the task of cataloging and surveying the biosynthetic capacity of the natural world has become a focus of attention. As previously mentioned, the MIBiG database[152] filled an important void in the cataloging of characterized gene clusters and their corresponding molecules. Likewise, the

Joint Genome Institute (JGI) has curated the Integrated Microbial Genomes—Atlas of Biosynthetic gene Clusters (IMG-ABC)[166] which contains over a million putative biosynthetic gene clusters identified using ClusterFinder. Such large repositories dedicated to secondary metabolism biosynthetic gene clusters are helping to dereplicate characterized gene clusters and better explore the vast majority of uncharacterized gene clusters. One of the first large scale analyses of genomes from talented secondary metabolite producers came from over 75 sequenced genomes of *Salinispora* spp. strains[167]. Ziemert and co-workers found a high level of pathway diversity among these strains with 99% sequence identity for the 16S rRNA gene. Of the 124 distinct "operational biosynthetic units" (OBUs), only nine had characterized natural products associated with them. Furthermore, species specific OBUs were found for the three defined *Salinispora* species, but the majority of OBUs appeared as singletons, residing in only one or two strains. This clustering of similar pathways into OBUs represents one of the first methods for grouping similar gene clusters that likely make similar products.

New methods have emerged for relating similar gene clusters through gene cluster family networks. Multiple groups have now developed algorithms to score gene cluster similarity and group them into gene cluster families based on these scores. The first such large scale analysis was published in 2014 and examined 1,154 genomes spanning the prokaryotic tree[151]. This wide-ranging analysis unveiled a surprising finding that the largest gene cluster family in current sequence databases is responsible for producing aryl polyene carboxylic acids. Additionally, this gene cluster family network utilized ClusterFinder to identify gene clusters of both known and unknown classes. By using a non-restrictive algorithm for gene cluster designation, they were able to identify the presence of large network families that contained widely distributed gene clusters without any characterized members, the largest of which was found to encode aryl polyene carboxylic acid biosynthesis. This kind of global analysis is foundational not only for identifying and defining the genetic underpinnings of new classes of secondary metabolites, but also for prioritizing biosynthetic gene clusters for experimental characterization.

A more focused approach was taken in a gene cluster family network analysis of 830 actinomycete genomes, including 344 newly sequenced genomes for this study[168]. In addition to focusing only on actinomycetes, the study also used a narrowed down list of gene cluster types concentrating on NRPS, type I and II PKS, RiPPs, and thiazole-oxazole modified microcins. Another layer of validation was added to this gene cluster family network by using MS detection of known molecules and correlating them to their gene clusters. Their method was able to link previously unassigned gene cluster families to known molecules, aiding in dereplication and identification of structurally similar congeners. Networking gene cluster families with a focus on a particular class of molecules has revealed that the capacity for making such molecules is much more diverse than what has been presently characterized. For example, a gene cluster family network of lanthipeptide-like biosynthetic pathways in actinomycetes found that lanthionine synthetases are genetically much more diverse that was previously thought[169]. Another study examined more than 10,000 actinomycete genomes for their ability to make phosphonic acid natural products[170]. This large scale analysis found that there are 64 distinct gene cluster families, of which 55 are likely to produce unknown compounds. This study went further to elucidate eleven previously undescribed phosphonic acid natural products.

Another recent gene cluster family algorithm was developed using the prolific *Moorea* genus of cyanobacteria. Leao and coworkers developed BioCompass, a new method that allows for similarity scoring of subclusters, which recognizes and takes advantage of the modular nature of many natural product gene clusters[171]. *Moorea* proved to have an astonishing amount of secondary metabolite biosynthetic potential, with up to a quarter of its genome dedicated to secondary metabolism. The BioCompass gene cluster family network revealed that 59% of the pathways found across four *Moorea* sp. strains have homology only to other *Moorea* clusters, highlighting the uniqueness of these strains. While there is no standard method for genome mining or producing gene cluster family networks, a recently published chapter on such protocols can help guide the identification, classification, and creation of gene cluster family networks for exploring large volumes of genomic information[172]. One advantage of gene cluster family networks is incorporating information from databases, such as MIBiG, to determine if previously characterized clusters exist in such a network[171,173]. This kind of dereplication can aid in prioritizing unknown clusters. Furthermore, gene cluster family networks can reveal the true scope of the biosynthetic potential represented in our sequenced databases. In most cases, the diversity sequenced has not begun to approach saturation, suggesting that much more novelty will be seen as we sequence more rare taxa from underexplored habitats. Some of the most promising sources of biological and chemical diversity are hidden in environmental metagenomes and uncultivatable bacteria that have been widely overlooked in favor of their easily cultured counterparts. Metagenomic sequencing is touted as a powerful tool for examining the vast majority of uncultured organisms and discovering biosynthetic potential previously unknown.

1.3.7 Metagenomic Discovery of Natural Products

Metagenomics and its role in natural product discovery is a newly emerging field, created by the sequencing revolution. It allows for sequencing environmental samples to give a glimpse into the full capacity in any given environment. Previous reliance on culturing organisms for study in the lab is no longer necessary to examine their genomes. Additionally, metagenomic sequencing has brought about the burgeoning field of microbiome studies, which are increasingly seen as productive systems for natural product discovery[29]. In particular, studying the human microbiome has accelerated the recognition of bacterial communities and their specialized chemistry[174,175]. Natural products have been and continue to be isolated from the human microbiome[176]. Furthermore, microbiome communities can give insight into chemical bacterial interactions, including elicitation of specialized chemical signals. Metagenomic sequencing and the study of diverse microbiomes has opened previously inaccessible biological spaces to explore new and functionalized chemistry.

Just as whole genome sequencing of the well-studied *Streptomyces* spp. revealed a wealth of orphan biosynthetic gene clusters, metagenomic sequencing of environmental samples exposes the vast majority of unculturable microbes in the environment and their potential for secondary metabolite production. Soils were the first complex metagenomes realized for their potential in the search for new bioactive molecules. Before metagenomic sequencing of environmental DNA became commonplace, clone libraries made directly from environmental samples, soil for example, were tested for bioactivity and then the individual clone could be sequenced[112]. This allowed for the first metagenomic exploration of environmental samples before the advent of high-throughput sequencing and yielded new natural products[177-179]. Not long after this approach was applied to complex soil samples, researchers drew comparisons to the complex microbiomes seen in symbiotic systems, such as the diverse microbial communities seen in sponges[180]. The depth of sequencing coverage needed for uncultivatable microbes was recognized as a caveat early on, and enriching for a dominant symbiont prior to DNA extraction was suggested as a partial solution to this problem. Methods for combining short insert libraries used for shotgun sequencing to identifying secondary metabolism genes and large insert clone libraries used to screen with the identified sequences were developed prior to economical high-throughput sequencing[181,182]. These early clone library based methods were used to identify and express the first cryptic gene clusters from metagenomic samples[76,183-185].

With the advent of affordable, widespread high-throughput sequencing, whole metagenome or microbiome sequencing has recently become feasible. Metagenomic sequencing is especially powerful in exploring symbiotic animal systems, which have provided an enormity of natural products, but sample size and supply limitations have largely limited them to academic study rather than pharmaceutical applications[186,187]. Directly sequencing metagenomic DNA from a lichen assemblage revealed that the cyanobacteria member of the symbiosis contained the first *trans*-AT PKS cluster seen in cyanobacteria[188]. Direct metagenome sequencing of a marine tunicate and its microbiome yielded the gene cluster responsible for making the important anti-cancer drug, Yondelis[189]. The discovery of the biosynthetic basis for this pharmaceutically important molecule paved the way for direct production of the drug and opened the possibility for metabolic engineering through pathway manipulation. Metagenomic sequencing of three Dysideidae sponges, known both for their abundance of poly-brominated diphenyl ether (PBDE) molecules and their signature cyanobacterial symbiont, revealed the cyanobacterial biosynthesis of PBDEs[190]. The pathway for producing the structurally related pentabromopseudilin had just been described and was used, in part, as a query for the PBDE pathway[191]. Using a biosynthetic hook to search for similar aspects of biosynthesis can be especially useful in examining complex metagenomes. This study also highlights the increasingly feasible practice of *de novo* assembly of microbial genomes from metagenomes[192-195].

Often, metagenomes are so complex that assembly of any one genome from the metagenome is virtually impossible without ample sequencing depth, and the result will inevitably be a population genome. The invention of Multiple Displacement Amplification (MDA) and improvement of the process allowed DNA from single cells to be amplified and sequence for the first time[196,197]. This kind of selective amplification provides a nice complement to whole metagenome sequencing. A filamentous cyanobacteria assemblage was queried using MDA on a single cyanobacteria cell and metagenomics library screening to locate the gene cluster responsible for apratoxin[198]. This kind of hybrid approach was also used to assemble the genomes of the uncultivated *Entotheonella* spp. inhabiting *Theonella swinhoei* sponges[199]. MDA from isolated cells and metagenome sequencing of the whole sponge resulted in a genome for the novel endosymbiont, shown to have immense capacity for secondary metabolite production. Equipped with advanced tools like MDA, genome assembly from metagenomes, large insert clone libraries, heterologous expression and whole genome metagenome sequencing, researchers are finally beginning to dive past the tip of the iceberg when it comes to microbial natural product potential in the environment.

As the cost of sequencing continues to fall, the abundance of genomic information from previously intractable systems will open access to immense new biosynthetic potential. Symbiotic assemblages harboring yet uncultured symbionts are especially promising avenues for new natural product potential. Underexplored environments, now available to exploration through metagenomic sequencing, will also surely yield novel biochemical transformations. We are truly at the cusp of another Golden Age for natural product discovery, facilitated by the explosion in genomic sequencing information and exposing the dark matter hidden inside microbial genomes, one base at a time.

1.4 In This Dissertation

1.4.1 Chapter 2: Genetic Basis for the Biosynthesis of the Pharmaceutically Important Class of Epoxyketone Proteasome Inhibitors

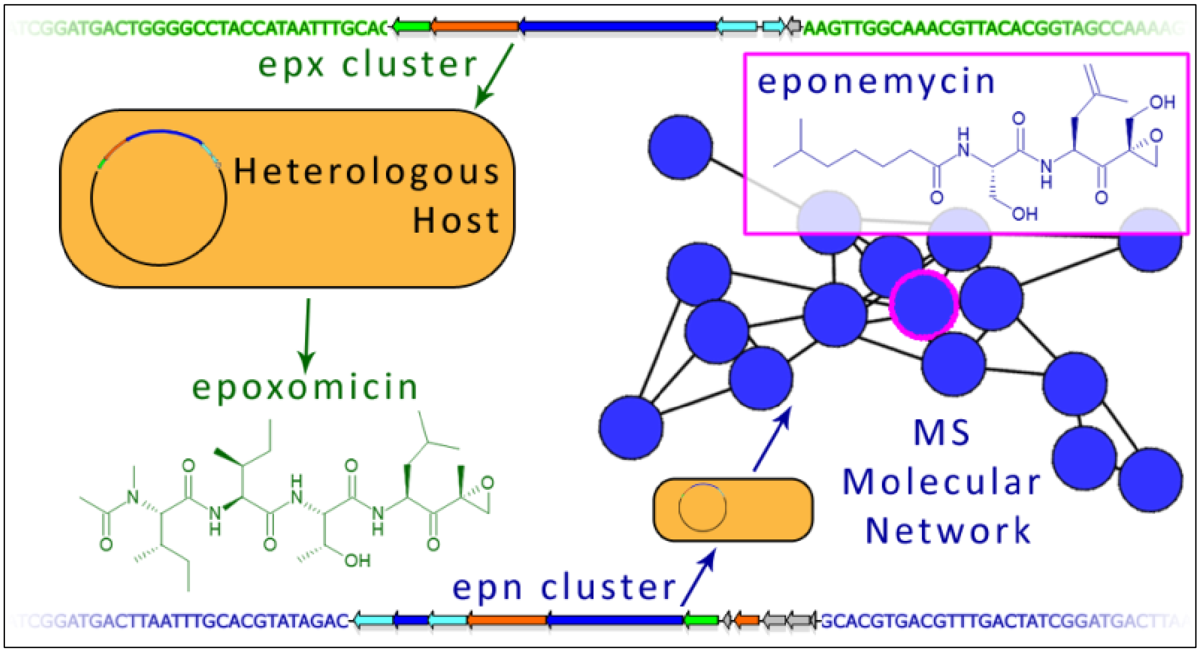

TCGGATGACTGGGGCCTACCATAATTTGCAC
AAGTTGGCAAACGTTACACGGTAGCCAAAAG
epx cluster
Heterologous Host
epoxomicin
eponemycin
MS Molecular Network
epn cluster
TCGGATGACTTAATTTGCACGTATAGAC
GCACGTGACGTTTGACTATCGGATGACTTA

Figure 1. Chapter 2 Overview

The gene clusters for epoxomicin and eponemycin were located in their respective producer's genomes and heterologously expressed. Molecular networking revealed multiple structural congeners produced by the eponemycin gene cluster.

This published chapter (Schorn, M *et al. ACS Chem Biol*, 2013)[200] consists of the first report of biosynthetic gene clusters for epoxyketone proteasome inhibitors, as determined through whole genome sequencing and heterologous expression. In this study, the genomes of two expoxyketone proteasome inhibitors were sequenced using the Ion Torrent PGM with a modified protocol for improvement of high G+C content DNA sequences. The genomes were interrogated to elucidate the putative biosynthetic gene clusters for epoxomicin (*epx*) and eponemycin (*epn*). Two hybrid NRPS/PKS gene clusters with A domain specificities matching the amino acids contained within epoxomicin and eponemycin were chosen for subsequent heterologous expression studies. Additionally, the eponemycin gene cluster contained a second copy of the gene encoding the beta-proteasome subunit, a known resistance factor for proteasome inhibitors in actinomycetes[155]. The epoxomicin producer also contained a second copy of the beta-proteasome subunit in a different region of the genome. These two gene clusters were selected for heterologous expression. The pathways were cloned using a Fosmid Library approach. The libraries were screened using primers designed to probe for regions of the gene cluster. Two clones were identified, each containing the *epx* and *epn* gene clusters in their entirety. The pathways were then integrated into a host strain, *Streptomyces albus* J1074, using triparental intergenic conjugation. The host strains with integrated *epx* and *epn* clusters were fermented and extracted, and the resulting extracts analyzed using HPLC. New peaks corresponding to epoxomicin and eponemycin were identified while comparing to the empty host strain. Additional molecular networking of the empty *S. albus* and *S. albus* + *epn* revealed a number of related analogues present in small amounts. This initial report of the biosynthetic gene clusters for the pharmaceutically important epoxyketone proteasome inhibitors laid the foundation for interrogating the mechanism of formation for the epoxyketone warhead. This study also introduced the use of betaine in Ion Torrent sequencing sample preparation for improved high G+C sequences characteristic of actinomycetes.

1.4.2 Chapter 3: Sequencing Rare Marine Actinomycete Genomes Reveals High Density of Unique Natural Product Biosynthetic Gene Clusters

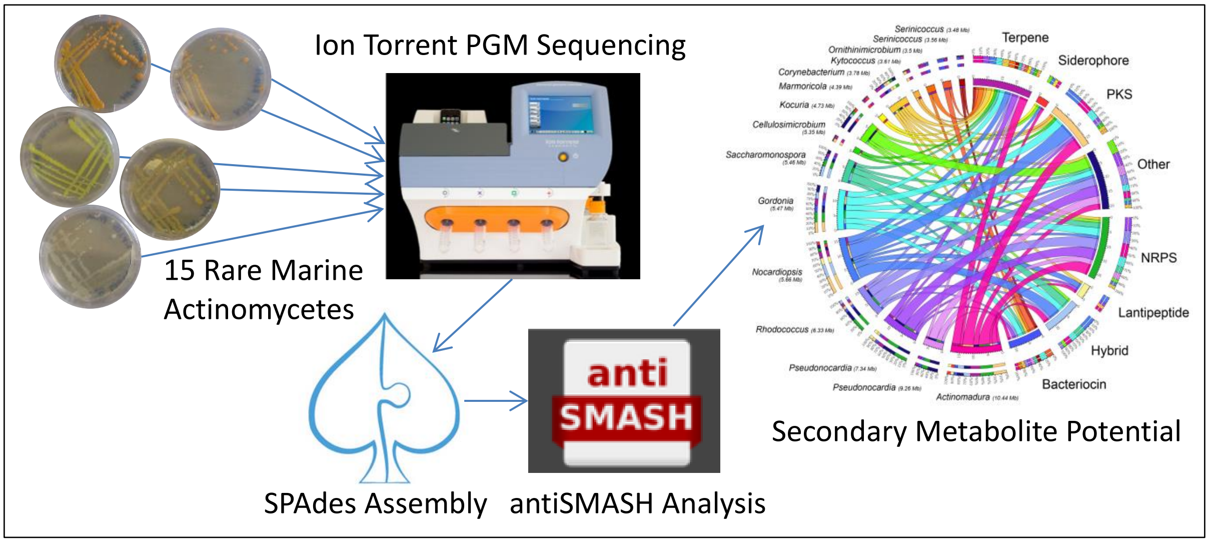


Figure 2. Chapter 3 Workflow

Workflow shown for cultivating rare marine actinomycetes, subsequent sequencing and genome assembly and secondary metabolite gene cluster identification followed by analysis of secondary metabolite potential.

This published chapter (Schorn, M *et al. Microbiology*, 2016)[173] surveys twenty-two rare marine actinomycetes for their biosynthetic potential and compares them with a contemporary database of sequenced actinomycete genomes. The decline in antibiotic discovery, concurrent with the rise of antibiotic resistance necessitates the search for natural products from new sources. Marine *Streptomyces* strains and the marine obligate genus, *Salinispora*, have proven a wealth of biosynthetic potential in marine actinomycetes, but strains from rare, understudied genera have gone largely unexplored[167,201-206]. Twenty two non-*Streptomyces*, or so called rare actinomycete[207], strains from the Fenical/Jensen marine actinomycete collection at the Scripps Institution of Oceanography were selected for whole genome sequencing using the Ion Torrent PGM. Ten strains had been previously screened using degenerate primers for NRPS and PKS biosynthetic genes[208]. Five additional, unscreened strains were chosen to widen the variety of the genera involved in the study. The genomes were assembled using a pre-release version of SPAdesIT from the Pevzner lab[209,210]. The rare marine actinomycete genomes were analyzed for the secondary metabolite biosynthetic potential and a variety of number and types of pathways were observed. To place these rare marine actinomycete genomes in the context of other sequenced actinomycete genomes, biosynthetic gene cluster similarity network was built. This analysis revealed that 87% of the biosynthetic gene clusters in the twenty two genomes were not similar to any sequenced actinomycete gene cluster. Additionally, when compared with marine *Streptomyces*, both showed high levels of pathway uniqueness, suggesting an underrepresentation of marine strains in current whole genome databases. This study employs large-scale bioinformatic analysis to show that marine-derived genera warrant further study and sequencing in natural product discovery. Rare marine actinomycetes, in particular, represent a promising avenue for biosynthetic pathways uncaptured by current sequencing databases.

1.4.3 Chapter 4: Uncultured Cyanobactierial Symbionts of Marine Sponges and Their Natural Products

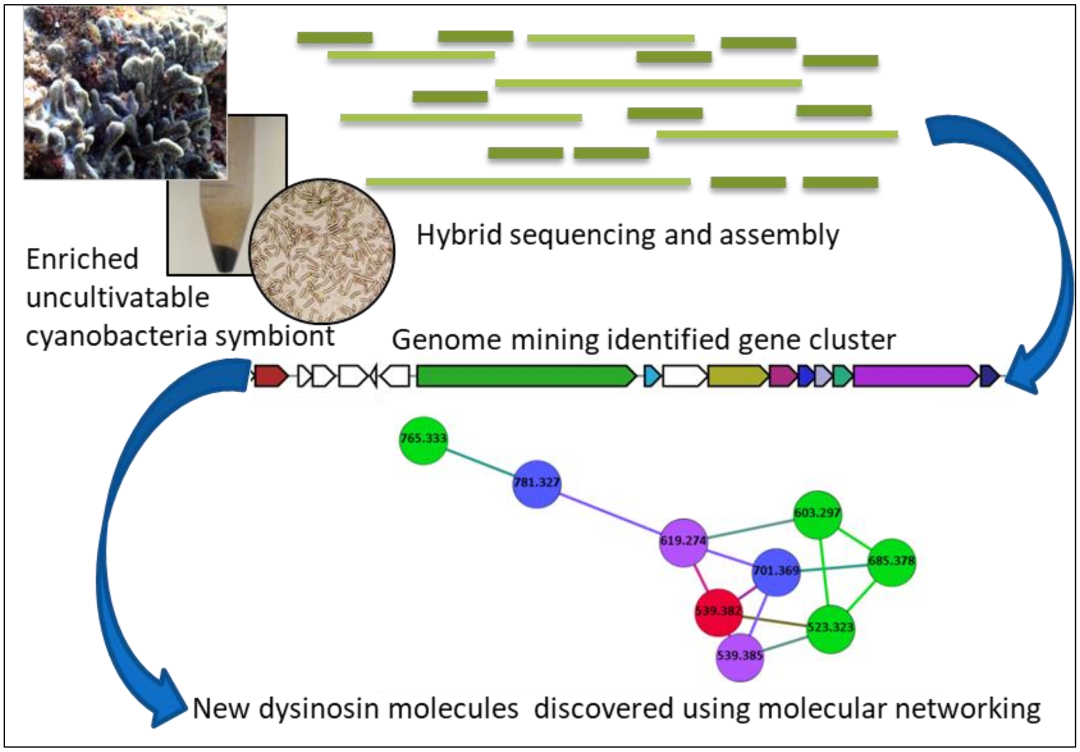

Enriched uncultivatable cyanobacteria symbiont
Hybrid sequencing and assembly
Genome mining identified gene cluster
765.333
781.327
619.274
603.297
685.378
701.369
539.382
523.323
539.385
New dysinosin molecules discovered using molecular networking

Figure 3. Chapter 4 Workflow

Overview of the workflow undertaken in Chapter 4, consisting of hybrid sequencing and assembly of uncultivated sponge cyanobacteria symbionts and subsequent genome mining and molecular networking leading to the identification of novel structures.

This final chapter describes the biosynthetic potential of and the characterized molecules from two Guamanian Dysideidae sponges. Previous studies of the biosynthesis of brominated natural products from marine bacteria[191] and cyanobacterial symbionts in sponges[190] prompted further investigation of the uncultured cyanobacterial symbionts of related specimens of Dysideidae sponges. The persistent cyanobacterial symbionts, well documented in sponges of the family Dysideidae, have been recalcitrant to isolation for decades[211-215]. The cyanobacterial trichomes can, however, be enriched by squeezing the sponge and centrifuging the exudate[212]. These cell enrichments allow for more targeted metagenomic sequencing using both Illumina HiSeq and PacBio RS II platforms to assemble genomes with greater than 90% completion. Two symbionts, one (GUM202_hs) from a sponge harboring poly-brominated diphenyl ethers (PBDEs) and one (GUM007_hs) from a sponge that doesn't contain PBDEs, were selected for this cell enriched sequencing and genome assembly procedure and their genomes interrogated for secondary metabolite biosynthetic potential. The *hs_bmp* pathway, previously characterized as responsible for the production of PBDEs[190], was not present in the GUM007_hs symbiont, but was present with an extra putative halogenase gene in the GUM202_hs symbiont. Also observed and characterized were thirteen PBDEs, with higher degrees of bromination than previously described from the Guamanian sponges collected[190,216]. The non-PBDE producing symbiont from GUM007 contains an NRPS gene cluster closely related to the characterized aeruginosin pathway. Aeruginosin-like compounds, dysinosins A-D, have been isolated from Dysideidae sponges in Australia[217,218]. Bioinformatic comparison leads us to believe that the NRPS pathway in GUM007_hs is responsible for making a dysinosin-like compound. Comparison with dysinosin standards, provided by Ron Quinn at Griffith University, reveals two masses for putative new dysinosins. Fragmentation patterns confirm their structures, and efforts for isolating enough of the compounds for NMR characterization are underway. In an effort to understand why these microbes have yet to be cultured in the lab, the primary metabolism pathways were comparatively analyzed using the *Synecchococcus elongates* PCC 7942 metabolic model to uncover missing genes[219]. This study demonstrates the feasibility of assembling uncultured symbiont genomes from metagenomes, providing a sizeable fraction of cells can be obtained from the host. It also exposes the biosynthetic capacity of these symbionts and possible reasons for their resistance to cultivation. Finally, this study provides the first example, to our knowledge, of genome mining in a metagenome assembled genome from an uncultivatable symbiont resulting in new chemical structures.

1.5 References

1 Darwin, C. & Wallace, A. On the Tendency of Species to form Varieties; and on the Perpetuation of Varieties and Species by Natural Means of Selection. *Journal of the Proceedings of the Linnean Society of London. Zoology* **3**, 45-62, 1858.

2 Education, N. *Gregor Mendel: A Private Scientist | Learn Science at Scitable*, <http://www.nature.com/scitable/topicpage/gregor-mendel-a-private-scientist-6618227> (2013).

3 Dahm, R. Friedrich Miescher and the discovery of DNA. *Dev Biol* **278**, 274-288, 2005. PMID: 15680349.

4 Griffith, F. The Significance of Pneumococcal Types. *J Hyg (Lond)* **27**, 113-159, 1928. PMID: 20474956.

5 Avery, O. T., MacLeod, C. M. & McCarty, M. Studies on the Chemical Nature of the Substance Inducing Transformation of Pneumococcal Types: Induction of Transformation by a Desoxyribonucleic Acid Fraction Isolated from Pneumococcus Type III. *J Exp Med* **79**, 137-158, 1944. PMID: 19871359.

6 Hershey, A. D. & Chase, M. Independent Functions of Viral Protein and Nucleic Acid in Growth of Bacteriophage. *J Gen Physiol* **36**, 39-56, 1952. PMID: 12981234.

7 Watson, J. D. & Crick, F. H. C. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature* **171**, 737, 1953.

8 Franklin, R. E. & Gosling, R. G. Molecular Configuration in Sodium Thymonucleate. *Nature* **171**, 740, 1953.

9 Wilkins, M. H. F., Stokes, A. R. & Wilson, H. R. Molecular Structure of Nucleic Acids: Molecular Structure of Deoxypentose Nucleic Acids. *Nature* **171**, 738, 1953.

10 Dias, D. A., Urban, S. & Roessner, U. A Historical Overview of Natural Products in Drug Discovery. *Metabolites* **2**, 303-336, 2012.

11 Krishnamurti, C. & Rao, S. C. The isolation of morphine by Serturner. *Indian J Anaesth* **60**, 861-862, 2016. PMID: 27942064.

12 Shen, B. A New Golden Age of Natural Products Drug Discovery. *Cell* **163**, 1297-1300, 2015. PMID: 26638061.

13 Newman, D. J. & Cragg, G. M. Natural Products As Sources of New Drugs over the 30 Years from 1981 to 2010. *Journal of Natural Products* **3**, 311-335, 2012.

14 Harboe, M. Armauer Hansen- The Man and His Work. *Int J Leprosy* **41**, 417-424, 1973.

15 Blevins, S. M. & Bronze, M. S. Robert Koch and the 'golden age' of bacteriology. *International Journal of Infectious Diseases* **14**, e744-e751, 2010.

16 Hawgood, B. J. Alexandre Yersin (1863-1943): discoverer of the plague bacillus, explorer and agronomist. *J Med Biogr* **16**, 167-172, 2008. PMID: 18653838.

17 Riedel, S. Edward Jenner and the history of smallpox and vaccination. *Proc (Bayl Univ Med Cent)* **18**, 21-25, 2005. PMID: 16200144.

18 Fleming, A. On the Antibacterial Action of Cultures of a Penicillium, with Special Reference to their Use in the Isolation of B. influenzæ. *Br J Exp Pathol* **10**, 226-236, 1929.

19 Tan, S. Y. & Tatsumura, Y. Alexander Fleming (1881–1955): Discoverer of penicillin. *Singapore Med J* **56**, 366-367, 2015. PMID: 26243971.

20 Ligon, B. L. Sir Howard Walter Florey--the force behind the development of penicillin. *Semin Pediatr Infect Dis* **15**, 109-114, 2004. PMID: 15185195.

21 Hopwood, D. A. *Streptomyces in Nature and Medicine*. (Oxford University Press, Inc., 2007).

22 Waksman, S. A. & Starkey, R. L. Partial Sterilization of Soil, Microbiological Activities and Soil Fertility:III. *Soil Science* **16**, 343-358, 1923.

23 Woodruff, H. B. Selman A. Waksman, Winner of the 1952 Nobel Prize for Physiology or Medicine. *Appl Environ Microbiol* **80**, 2-8, 2014. PMID: 24162573.

24 Waksman, S. A. & Woodruff, H. B. Actinomyces antibioticus, a New Soil Organism Antagonistic to Pathogenic and Non-pathogenic Bacteria. *J Bacteriol* **42**, 231-249, 1941. PMID: 16560451.

25 SA, W. & M., T. The chemical nature of actinomycin, an antimicrobial substance produced by Actinomyces antibioticus. *J. Biol. Chem.* **142**, 277–286, 1942.

26 Schatz, A., Bugle, E. & Waksman, S. A. Streptomycin, a Substance Exhibiting Antibiotic Activity Against Gram-Positive and Gram-Negative Bacteria.∗†. *Experimental Biology and Medicine* **55**, 66-69, 1944.

27 Hinshaw, H. C., Feldman, W. H. & Pfuetze, K. H. Streptomycin in treatment of clinical tuberculosis. *Am Rev Tuberc* **54**, 191-203, 1946. PMID: 20274624.

28 Youmans, G. P. & Mc, C. J. Streptomycin in experimental tuberculosis; its effect on tuberculous infections in mice produced by M. tuberculosis var. hominis. *Am Rev Tuberc* **52**, 432-439, 1945. PMID: 21005387.

29 Milshteyn, A., Schneider, J. S. & Brady, S. F. Mining the metabiome: identifying novel natural products from microbial communities. *Chem Biol* **21**, 1211-1223, 2014. PMID: 25237864.

30 Sermonti, G. & Spada-Sermonti, I. Genetic recombination in Streptomyces. *Nature* **176**, 121, 1955. PMID: 13244631.

31 Hopwood, D. A. Forty years of genetics with Streptomyces: from in vivo through in vitro to in silico. *Microbiology* **145 (Pt 9)**, 2183-2202, 1999. PMID: 10517572.

32 Chater, K. David Hopwood and the emergence of Streptomyces genetics. *Int Microbiol* **2**, 61-68, 1999. PMID: 10943394.

33 Hopwood, D. A. & Botany School, U. o. C., Cambridge, England. Linkage and the Mechanism of Recombination in Streptomyces Coelicolor. *Annals of the New York Academy of Sciences* **81**, 887-898, 1959.

34 Kirby, R., Wright, L. F. & Hopwood, D. A. Plasmid-determined antibiotic synthesis and resistance in Streptomyces coelicolor. *Nature* **254**, 265-267, 1975. PMID: 1113895.

35 Wright, L. F. & Hopwood, D. A. Identification of the antibiotic determined by the SCP1 plasmid of Streptomyces coelicolor A3(2). *J Gen Microbiol* **95**, 96-106, 1976. PMID: 822125.

36 Kirby, R. & Hopwood, D. A. Genetic determination of methylenomycin synthesis by the SCP1 plasmid of Streptomyces coelicolor A3(2). *J Gen Microbiol* **98**, 239-252, 1977. PMID: 833570.

37 Wright, L. F. & Hopwood, D. A. Actinorhodin is a chromosomally-determined antibiotic in Streptomyces coelicolar A3(2). *J Gen Microbiol* **96**, 289-297, 1976. PMID: 993778.

38 Rudd, B. A. & Hopwood, D. A. Genetics of actinorhodin biosynthesis by Streptomyces coelicolor A3(2). *J Gen Microbiol* **114**, 35-43, 1979. PMID: 521794.

39 Chater, K. F. & Bruton, C. J. Resistance, regulatory and production genes for the antibiotic methylenomycin are clustered. *Embo j* **4**, 1893-1897, 1985. PMID: 2992952.

40 Linn, S. & Arber, W. Host specificity of DNA produced by Escherichia coli, X. In vitro restriction of phage fd replicative form. *Proc Natl Acad Sci U S A* **59**, 1300-1306, 1968. PMID: 4870862.

41 Smith, H. O. & Wilcox, K. W. A restriction enzyme from Hemophilus influenzae. I. Purification and general properties. *J Mol Biol* **51**, 379-391, 1970. PMID: 5312500.

42 Jackson, D. A., Symons, R. H. & Berg, P. Biochemical method for inserting new genetic information into DNA of Simian Virus 40: circular SV40 DNA molecules containing lambda phage genes and the galactose operon of Escherichia coli. *Proc Natl Acad Sci U S A* **69**, 2904-2909, 1972. PMID: 4342968.

43 Mandel, M. & Higa, A. Calcium-dependent bacteriophage DNA infection. *J Mol Biol* **53**, 159-162, 1970. PMID: 4922220.

44 Cohen, S. N., Chang, A. C. Y. & Hsu, L. Nonchromosomal Antibiotic Resistance in Bacteria: Genetic Transformation of Escherichia coli by R-Factor DNA*. *Proc Natl Acad Sci U S A* **69**, 2110-2114, 1972. PMID: 4559594.

45 Okanishi, M., Suzuki, K. & Umezawa, H. Formation and reversion of Streptomycete protoplasts: cultural condition and morphological study. *J Gen Microbiol* **80**, 389-400, 1974. PMID: 4207870.

46 Bibb, M. J., Ward, J. M. & Hopwood, D. A. Transformation of plasmid DNA into Streptomyces at high frequency. *Nature* **274**, 398-400, 1978. PMID: 672966.

47 Baltz, R. H. Genetic recombination in Streptomyces fradiae by protoplast fusion and cell regeneration. *J Gen Microbiol* **107**, 93-102, 1978. PMID: 731205.

48 Bolivar, F., Rodriguez, R. L., Greene, P. J., Betlach, M. C., Heyneker, H. L., Boyer, H. W., Crosa, J. H. & Falkow, S. Construction and characterization of new cloning vehicles. II. A multipurpose cloning system. *Gene* **2**, 95-113, 1977. PMID: 344137.

49 Bibb, M., Schottel, J. L. & Cohen, S. N. A DNA cloning system for interspecies gene transfer in antibiotic-producing Streptomyces. *Nature* **284**, 526, 1980.

50 Thompson, C. J., Ward, J. M. & Hopwood, D. A. DNA cloning in Streptomyces: resistance genes from antibiotic-producing species. *Nature* **286**, 525, 1980.

51 Bibb, M. J. & Cohen, S. N. Gene expression in Streptomyces: construction and application of promoter-probe plasmid vectors in Streptomyces lividans. *Mol Gen Genet* **187**, 265-277, 1982. PMID: 6294463.

52 Feitelson, J. S. & Hopwood, D. A. Cloning of a Streptomyces gene for an O-methyltransferase involved in antibiotic biosynthesis. *Mol Gen Genet* **190**, 394-398, 1983. PMID: 6576223.

53 Chater, K. F. & Bruton, C. J. Mutational cloning in Streptomyces and the isolation of antibiotic production genes. *Gene* **26**, 67-78, 1983. PMID: 6323253.

54 Gil, J. A. & Hopwood, D. A. Cloning and expression of a p-aminobenzoic acid synthetase gene of the candicidin-producing Streptomyces griseus. *Gene* **25**, 119-132, 1983. PMID: 6420235.

55 Gorst-Allman, C. P., Rudd, B. A. M., Chang, C.-J. & Floss, H. G. Biosynthesis of actinorhodin. Determination of the point of dimerization. *Journal of Organic Chemistry* **46**, 455-456, 1981.

56 Malpartida, F. & Hopwood, D. A. Molecular cloning of the whole biosynthetic pathway of a Streptomyces antibiotic and its expression in a heterologous host. *Nature* **309**, 462-464, 1984. PMID: 6328317.

57 Hopwood, D. A., Malpartida, F., Kieser, H. M., Ikeda, H., Duncan, J., Fujii, I., Rudd, B. A. M., Floss, H. G. & Ōmura, S. Production of 'hybrid' antibiotics by genetic engineering. *Nature* **314**, 642, 1985.

58 Thompson, C. J. & Gray, G. S. Nucleotide sequence of a streptomycete aminoglycoside phosphotransferase gene and its relationship to phosphotransferases encoded by resistance plasmids. *Proc Natl Acad Sci U S A* **80**, 5190-5194, 1983. PMID: 6310563.

59 Hallam, S. E., Malpartida, F. & Hopwood, D. A. Nucleotide sequence, transcription and deduced function of a gene involved in polyketide antibiotic synthesis in Streptomyces coelicolor. *Gene* **74**, 305-320, 1988. PMID: 2469622.

60 Sanger, F. & Tuppy, H. The amino-acid sequence in the phenylalanyl chain of insulin. I. The identification of lower peptides from partial hydrolysates. *Biochem J* **49**, 463-481, 1951. PMID: 14886310.

61 Heather, J. M. & Chain, B. The sequence of sequencers: The history of sequencing DNA. *Genomics* **107**, 1-8, 2016. PMID: 26554401.

62 Holley, R. W., Apgar, J., Everett, G. A., Madison, J. T., Marquisee, M., Merrill, S. H., Penswick, J. R. & Zamir, A. Structure of a ribonucleic acid. *Science* **147**, 1462-1465, 1965. PMID: 14263761.

63 Sanger, F., Brownlee, G. G. & Barrell, B. G. A two-dimensional fractionation procedure for radioactive nucleotides. *J Mol Biol* **13**, 373-398, 1965. PMID: 5325727.

64 Wu, R. & Kaiser, A. D. Structure and base sequence in the cohesive ends of bacteriophage lambda DNA. *J Mol Biol* **35**, 523-537, 1968. PMID: 4299833.

65 Gilbert, W. & Maxam, A. The nucleotide sequence of the lac operator. *Proc Natl Acad Sci U S A* **70**, 3581-3584, 1973. PMID: 4587255.

66 Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A* **74**, 5463-5467, 1977. PMID: 271968.

67 Maxam, A. M. & Gilbert, W. A new method for sequencing DNA. *Proc Natl Acad Sci U S A* **74**, 560-564, 1977. PMID: 265521.

68 Staden, R. A strategy of DNA sequencing employing computer programs. *Nucleic Acids Res* **6**, 2601-2610, 1979. PMID: 461197.

69 Shendure, J., Balasubramanian, S., Church, G. M., Gilbert, W., Rogers, J., Schloss, J. A. & Waterston, R. H. DNA sequencing at 40: past, present and future. *Nature* **550**, 345, 2017.

70 Bibb, M. J., Biro, S., Motamedi, H., Collins, J. F. & Hutchinson, C. R. Analysis of the nucleotide sequence of the Streptomyces glaucescens tcmI genes provides key information about the enzymology of polyketide antibiotic biosynthesis. *Embo j* **8**, 2727-2736, 1989. PMID: 2684656.

71 Sherman, D. H., Malpartida, F., Bibb, M. J., Kieser, H. M. & Hopwood, D. A. Structure and deduced function of the granaticin-producing polyketide synthase gene cluster of Streptomyces violaceoruber Tu22. *Embo j* **8**, 2717-2725, 1989. PMID: 2583128.

72 Cortes, J., Haydock, S. F., Roberts, G. A., Bevitt, D. J. & Leadlay, P. F. An unusually large multifunctional polypeptide in the erythromycin-producing polyketide synthase of Saccharopolyspora erythraea. *Nature* **348**, 176-178, 1990. PMID: 2234082.

73 Donadio, S., Staver, M. J., McAlpine, J. B., Swanson, S. J. & Katz, L. Modular organization of genes required for complex polyketide biosynthesis. *Science* **252**, 675-679, 1991. PMID: 2024119.

74 Bangera, M. G. & Thomashow, L. S. Identification and characterization of a gene cluster for synthesis of the polyketide antibiotic 2,4-diacetylphloroglucinol from Pseudomonas fluorescens Q2-87. *J Bacteriol* **181**, 3155-3163, 1999. PMID: 10322017.

75 Funa, N., Ohnishi, Y., Fujii, I., Shibuya, M., Ebizuka, Y. & Horinouchi, S. A new pathway for polyketide synthesis in microorganisms. *Nature* **400**, 897-899, 1999. PMID: 10476972.

76 Piel, J. A polyketide synthase-peptide synthetase gene cluster from an uncultured bacterial symbiont of Paederus beetles. *Proc Natl Acad Sci U S A* **99**, 14002-14007, 2002. PMID: 12381784.

77 Cheng, Y. Q., Tang, G. L. & Shen, B. Identification and localization of the gene cluster encoding biosynthesis of the antitumor macrolactam leinamycin in Streptomyces atroolivaceus S-140. *J Bacteriol* **184**, 7013-7024, 2002. PMID: 12446651.

78 Cheng, Y. Q., Tang, G. L. & Shen, B. Type I polyketide synthase requiring a discrete acyltransferase for polyketide biosynthesis. *Proc Natl Acad Sci U S A* **100**, 3149-3154, 2003. PMID: 12598647.

79 Helfrich, E. J. & Piel, J. Biosynthesis of polyketides by trans-AT polyketide synthases. *Nat Prod Rep* **33**, 231-316, 2016. PMID: 26689670.

80 Hopwood, D. A. Genetic Contributions to Understanding Polyketide Synthases. *Chem Rev* **97**, 2465-2498, 1997. PMID: 11851466.

81 McDaniel, R., Ebert-Khosla, S., Hopwood, D. A. & Khosla, C. Engineered biosynthesis of novel polyketides. *Science* **262**, 1546-1550, 1993. PMID: 8248802.

82 McDaniel, R., Ebert-Khosla, S., Hopwood, D. A. & Khosla, C. Rational design of aromatic polyketide natural products by recombinant assembly of enzymatic subunits. *Nature* **375**, 549-554, 1995. PMID: 7791871.

83 McDaniel, R., Thamchaipenet, A., Gustafsson, C., Fu, H., Betlach, M. & Ashley, G. Multiple genetic modifications of the erythromycin polyketide synthase to produce a library of novel "unnatural" natural products. *Proc Natl Acad Sci U S A* **96**, 1846-1851, 1999. PMID: 10051557.

84 Weber, T., Welzel, K., Pelzer, S., Vente, A. & Wohlleben, W. Exploiting the genetic potential of polyketide producing streptomycetes. *J Biotechnol* **106**, 221-232, 2003. PMID: 14651864.

85 Lipmann, F. Bacterial production of antibiotic polypeptides by thiol-linked synthesis on protein templates. *Adv Microb Physiol* **21**, 227-266, 1980. PMID: 6160738.

86 Weckermann, R., Furbass, R. & Marahiel, M. A. Complete nucleotide sequence of the tycA gene coding the tyrocidine synthetase 1 from Bacillus brevis. *Nucleic Acids Res* **16**, 11841, 1988. PMID: 3267240.

87 Smith, D. J., Earl, A. J. & Turner, G. The multifunctional peptide synthetase performing the first step of penicillin biosynthesis in Penicillium chrysogenum is a 421,073 dalton protein similar to Bacillus brevis peptide antibiotic synthetases. *Embo j* **9**, 2743-2750, 1990. PMID: 2118102.

88 Coque, J. J., Martin, J. F., Calzada, J. G. & Liras, P. The cephamycin biosynthetic genes pcbAB, encoding a large multidomain peptide synthetase, and pcbC of Nocardia lactamdurans are clustered together in an organization different from the same genes in Acremonium chrysogenum and Penicillium chrysogenum. *Mol Microbiol* **5**, 1125-1133, 1991. PMID: 1956290.

89 Gutierrez, S., Diez, B., Montenegro, E. & Martin, J. F. Characterization of the Cephalosporium acremonium pcbAB gene encoding alpha-aminoadipyl-cysteinyl-valine synthetase, a large multidomain peptide synthetase: linkage to the pcbC gene as

a cluster of early cephalosporin biosynthetic genes and evidence of multiple functional domains. *J Bacteriol* **173**, 2354-2365, 1991. PMID: 1706706.

90 Turgay, K., Krause, M. & Marahiel, M. A. Four homologous domains in the primary structure of GrsB are related to domains in a superfamily of adenylate-forming enzymes. *Mol Microbiol* **6**, 529-546, 1992. PMID: 1560782.

91 Cosmina, P., Rodriguez, F., de Ferra, F., Grandi, G., Perego, M., Venema, G. & van Sinderen, D. Sequence and analysis of the genetic locus responsible for surfactin synthesis in Bacillus subtilis. *Mol Microbiol* **8**, 821-831, 1993. PMID: 8355609.

92 Haese, A., Schubert, M., Herrmann, M. & Zocher, R. Molecular characterization of the enniatin synthetase gene encoding a multifunctional enzyme catalysing N-methyldepsipeptide formation in Fusarium scirpi. *Mol Microbiol* **7**, 905-914, 1993. PMID: 8483420.

93 Conti, E., Stachelhaus, T., Marahiel, M. A. & Brick, P. Structural basis for the activation of phenylalanine in the non-ribosomal biosynthesis of gramicidin S. *EMBO J* **16**, 4174-4183, 1997. PMID: 9250661.

94 Stachelhaus, T., Mootz, H. D. & Marahiel, M. A. The specificity-conferring code of adenylation domains in nonribosomal peptide synthetases. *Chem Biol* **6**, 493-505, 1999. PMID: 10421756.

95 Borchert, S., Patil, S. S. & Marahiel, M. A. Identification of putative multifunctional peptide synthetase genes using highly conserved oligonucleotide sequences derived from known synthetases. *FEMS Microbiology Letters* **92**, 175-180, 1992.

96 Turgay, K. & Marahiel, M. A. A general approach for identifying and cloning peptide synthetase genes. *Pept Res* **7**, 238-241, 1994. PMID: 7849417.

97 Neilan, B. A., Dittmann, E., Rouhiainen, L., Bass, R. A., Schaub, V., Sivonen, K. & Börner, T. Nonribosomal Peptide Synthesis and Toxigenicity of Cyanobacteria. *J Bacteriol* **181**, 4089-4097, 1999. PMID: 10383979.

98 Marahiel, M. A., Stachelhaus, T. & Mootz, H. D. Modular Peptide Synthetases Involved in Nonribosomal Peptide Synthesis. *Chem Rev* **97**, 2651-2674, 1997. PMID: 11851476.

99 Lambalot, R. H. & Walsh, C. T. Cloning, overproduction, and characterization of the Escherichia coli holo-acyl carrier protein synthase. *J Biol Chem* **270**, 24658-24661, 1995. PMID: 7559576.

100 Lambalot, R. H., Gehring, A. M., Flugel, R. S., Zuber, P., LaCelle, M., Marahiel, M. A., Reid, R., Khosla, C. & Walsh, C. T. A new enzyme superfamily - the phosphopantetheinyl transferases. *Chem Biol* **3**, 923-936, 1996. PMID: 8939709.

101 Land, M., Hauser, L., Jun, S. R., Nookaew, I., Leuze, M. R., Ahn, T. H., Karpinets, T., Lund, O., Kora, G., Wassenaar, T., Poudel, S. & Ussery, D. W. Insights from 20 years of bacterial genome sequencing. *Funct Integr Genomics* **15**, 141-161, 2015. PMID: 25722247.

102 Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J. F., Dougherty, B. A., Merrick, J. M. & et al. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science* **269**, 496-512, 1995. PMID: 7542800.

103 Fraser, C. M., Gocayne, J. D., White, O., Adams, M. D., Clayton, R. A., Fleischmann, R. D., Bult, C. J., Kerlavage, A. R., Sutton, G., Kelley, J. M., Fritchman, R. D., Weidman, J. F., Small, K. V., Sandusky, M., Fuhrmann, J., Nguyen, D., Utterback, T. R., Saudek, D. M., Phillips, C. A., Merrick, J. M., Tomb, J. F., Dougherty, B. A., Bott, K. F., Hu, P. C., Lucier, T. S., Peterson, S. N., Smith, H. O., Hutchison, C. A., 3rd & Venter, J. C. The minimal gene complement of Mycoplasma genitalium. *Science* **270**, 397-403, 1995. PMID: 7569993.

104 Olsen, G. J., Lane, D. J., Giovannoni, S. J., Pace, N. R. & Stahl, D. A. Microbial ecology and evolution: a ribosomal RNA approach. *Annu Rev Microbiol* **40**, 337-365, 1986. PMID: 2430518.

105 Torsvik, V., Goksøyr, J. & Daae, F. L. High diversity in DNA of soil bacteria. *Appl Environ Microbiol* **56**, 782-787, 1990. PMID: 2317046.

106 Ward, D. M., Weller, R. & Bateson, M. M. 16S rRNA sequences reveal numerous uncultured microorganisms in a natural community. *Nature* **345**, 63, 1990.

107 Weller, R., Weller, J. W. & Ward, D. M. 16S rRNA sequences of uncultivated hot spring cyanobacterial mat inhabitants retrieved as randomly primed cDNA. *Appl Environ Microbiol* **57**, 1146-1151, 1991. PMID: 1711832.

108 Stahl, D. A., Lane, D. J., Olsen, G. J. & Pace, N. R. Analysis of hydrothermal vent-associated symbionts by ribosomal RNA sequences. *Science* **224**, 409-411, 1984. PMID: 17741220.

109 Britschgi, T. B. & Giovannoni, S. J. Phylogenetic analysis of a natural marine bacterioplankton population by rRNA gene cloning and sequencing. *Appl Environ Microbiol* **57**, 1707-1713, 1991. PMID: 1714704.

110 Giovannoni, S. J., Britschgi, T. B., Moyer, C. L. & Field, K. G. Genetic diversity in Sargasso Sea bacterioplankton. *Nature* **345**, 60-63, 1990. PMID: 2330053.

111 Rappe, M. S. & Giovannoni, S. J. The uncultured microbial majority. *Annu Rev Microbiol* **57**, 369-394, 2003. PMID: 14527284.

112 Handelsman, J., Rondon, M. R., Brady, S. F., Clardy, J. & Goodman, R. M. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chem Biol* **5**, R245-249, 1998. PMID: 9818143.

113 Bentley, S. D., Chater, K. F., Cerdeno-Tarraga, A. M., Challis, G. L., Thomson, N. R., James, K. D., Harris, D. E., Quail, M. A., Kieser, H., Harper, D., Bateman, A., Brown, S., Chandra, G., Chen, C. W., Collins, M., Cronin, A., Fraser, A., Goble, A., Hidalgo, J., Hornsby, T., Howarth, S., Huang, C. H., Kieser, T., Larke, L., Murphy, L., Oliver, K., O'Neil, S., Rabbinowitsch, E., Rajandream, M. A., Rutherford, K., Rutter, S., Seeger, K., Saunders, D., Sharp, S., Squares, R., Squares, S., Taylor, K., Warren, T., Wietzorrek, A., Woodward, J., Barrell, B. G., Parkhill, J. & Hopwood, D. A. Complete genome sequence of the model actinomycete Streptomyces coelicolor A3(2). *Nature* **417**, 141-147, 2002. PMID: 12000953.

114 Challis, G. L. & Ravel, J. Coelichelin, a new peptide siderophore encoded by the Streptomyces coelicolor genome: structure prediction from the sequence of its non-ribosomal peptide synthetase. *FEMS Microbiol Lett* **187**, 111-114, 2000. PMID: 10856642.

115 Ikeda, H., Ishikawa, J., Hanamoto, A., Shinose, M., Kikuchi, H., Shiba, T., Sakaki, Y., Hattori, M. & Omura, S. Complete genome sequence and comparative analysis of the industrial microorganism Streptomyces avermitilis. *Nat Biotechnol* **21**, 526-531, 2003. PMID: 12692562.

116 Omura, S., Ikeda, H., Ishikawa, J., Hanamoto, A., Takahashi, C., Shinose, M., Takahashi, Y., Horikawa, H., Nakazawa, H., Osonoe, T., Kikuchi, H., Shiba, T., Sakaki, Y. & Hattori, M. Genome sequence of an industrial microorganism Streptomyces avermitilis: deducing the ability of producing secondary metabolites. *Proc Natl Acad Sci U S A* **98**, 12215-12220, 2001. PMID: 11572948.

117 Nyren, P. & Lundin, A. Enzymatic method for continuous monitoring of inorganic pyrophosphate synthesis. *Anal Biochem* **151**, 504-509, 1985. PMID: 3006540.

118 Hyman, E. D. A new method of sequencing DNA. *Anal Biochem* **174**, 423-436, 1988. PMID: 2853582.

119 Ronaghi, M., Karamohamed, S., Pettersson, B., Uhlen, M. & Nyren, P. Real-time DNA sequencing using detection of pyrophosphate release. *Anal Biochem* **242**, 84-89, 1996. PMID: 8923969.

120 Ronaghi, M., Uhlen, M. & Nyren, P. A sequencing method based on real-time pyrophosphate. *Science* **281**, 363, 365, 1998. PMID: 9705713.

121 Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., Braverman, M. S., Chen, Y. J., Chen, Z., Dewell, S. B., Du, L., Fierro, J. M., Gomes, X. V., Godwin, B. C., He, W., Helgesen, S., Ho, C. H., Irzyk, G. P., Jando, S. C., Alenquer, M. L., Jarvie, T. P., Jirage, K. B., Kim, J. B., Knight, J. R., Lanza, J. R.,

Leamon, J. H., Lefkowitz, S. M., Lei, M., Li, J., Lohman, K. L., Lu, H., Makhijani, V. B., McDade, K. E., McKenna, M. P., Myers, E. W., Nickerson, E., Nobile, J. R., Plant, R., Puc, B. P., Ronan, M. T., Roth, G. T., Sarkis, G. J., Simons, J. F., Simpson, J. W., Srinivasan, M., Tartaro, K. R., Tomasz, A., Vogt, K. A., Volkmer, G. A., Wang, S. H., Wang, Y., Weiner, M. P., Yu, P., Begley, R. F. & Rothberg, J. M. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376-380, 2005. PMID: 16056220.

122 Rothberg, J. M., Hinz, W., Rearick, T. M., Schultz, J., Mileski, W., Davey, M., Leamon, J. H., Johnson, K., Milgrew, M. J., Edwards, M., Hoon, J., Simons, J. F., Marran, D., Myers, J. W., Davidson, J. F., Branting, A., Nobile, J. R., Puc, B. P., Light, D., Clark, T. A., Huber, M., Branciforte, J. T., Stoner, I. B., Cawley, S. E., Lyons, M., Fu, Y., Homer, N., Sedova, M., Miao, X., Reed, B., Sabina, J., Feierstein, E., Schorn, M., Alanjary, M., Dimalanta, E., Dressman, D., Kasinskas, R., Sokolsky, T., Fidanza, J. A., Namsaraev, E., McKernan, K. J., Williams, A., Roth, G. T. & Bustillo, J. An integrated semiconductor device enabling non-optical genome sequencing. *Nature* **475**, 348-352, 2011.

123 Valouev, A., Ichikawa, J., Tonthat, T., Stuart, J., Ranade, S., Peckham, H., Zeng, K., Malek, J. A., Costa, G., McKernan, K., Sidow, A., Fire, A. & Johnson, S. M. A high-resolution, nucleosome position map of C. elegans reveals a lack of universal sequence-dictated positioning. *Genome Res* **18**, 1051-1063, 2008. PMID: 18477713.

124 Drmanac, R., Sparks, A. B., Callow, M. J., Halpern, A. L., Burns, N. L., Kermani, B. G., Carnevali, P., Nazarenko, I., Nilsen, G. B., Yeung, G., Dahl, F., Fernandez, A., Staker, B., Pant, K. P., Baccash, J., Borcherding, A. P., Brownley, A., Cedeno, R., Chen, L., Chernikoff, D., Cheung, A., Chirita, R., Curson, B., Ebert, J. C., Hacker, C. R., Hartlage, R., Hauser, B., Huang, S., Jiang, Y., Karpinchyk, V., Koenig, M., Kong, C., Landers, T., Le, C., Liu, J., McBride, C. E., Morenzoni, M., Morey, R. E., Mutch, K., Perazich, H., Perry, K., Peters, B. A., Peterson, J., Pethiyagoda, C. L., Pothuraju, K., Richter, C., Rosenbaum, A. M., Roy, S., Shafto, J., Sharanhovich, U., Shannon, K. W., Sheppy, C. G., Sun, M., Thakuria, J. V., Tran, A., Vu, D., Zaranek, A. W., Wu, X., Drmanac, S., Oliphant, A. R., Banyai, W. C., Martin, B., Ballinger, D. G., Church, G. M. & Reid, C. A. Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* **327**, 78-81, 2010. PMID: 19892942.

125 Bennett, S. Solexa Ltd. *Pharmacogenomics* **5**, 433-438, 2004. PMID: 15165179.

126 Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., Hall, K. P., Evers, D. J., Barnes, C. L., Bignell, H. R., Boutell, J. M., Bryant, J., Carter, R. J., Keira Cheetham, R., Cox, A. J., Ellis, D. J., Flatbush, M. R., Gormley, N. A., Humphray, S. J., Irving, L. J., Karbelashvili, M. S., Kirk, S. M., Li, H., Liu, X., Maisinger, K. S., Murray, L. J., Obradovic, B., Ost, T., Parkinson, M. L., Pratt, M. R., Rasolonjatovo, I. M., Reed, M. T., Rigatti, R., Rodighiero, C., Ross, M. T., Sabot, A., Sankar, S. V., Scally, A., Schroth, G. P., Smith, M. E., Smith, V. P., Spiridou, A., Torrance, P. E., Tzonev, S. S., Vermaas, E. H., Walter, K., Wu, X.,

Zhang, L., Alam, M. D., Anastasi, C., Aniebo, I. C., Bailey, D. M., Bancarz, I. R., Banerjee, S., Barbour, S. G., Baybayan, P. A., Benoit, V. A., Benson, K. F., Bevis, C., Black, P. J., Boodhun, A., Brennan, J. S., Bridgham, J. A., Brown, R. C., Brown, A. A., Buermann, D. H., Bundu, A. A., Burrows, J. C., Carter, N. P., Castillo, N., Chiara, E. C. M., Chang, S., Neil Cooley, R., Crake, N. R., Dada, O. O., Diakoumakos, K. D., Dominguez-Fernandez, B., Earnshaw, D. J., Egbujor, U. C., Elmore, D. W., Etchin, S. S., Ewan, M. R., Fedurco, M., Fraser, L. J., Fuentes Fajardo, K. V., Scott Furey, W., George, D., Gietzen, K. J., Goddard, C. P., Golda, G. S., Granieri, P. A., Green, D. E., Gustafson, D. L., Hansen, N. F., Harnish, K., Haudenschild, C. D., Heyer, N. I., Hims, M. M., Ho, J. T., Horgan, A. M., Hoschler, K., Hurwitz, S., Ivanov, D. V., Johnson, M. Q., James, T., Huw Jones, T. A., Kang, G. D., Kerelska, T. H., Kersey, A. D., Khrebtukova, I., Kindwall, A. P., Kingsbury, Z., Kokko-Gonzales, P. I., Kumar, A., Laurent, M. A., Lawley, C. T., Lee, S. E., Lee, X., Liao, A. K., Loch, J. A., Lok, M., Luo, S., Mammen, R. M., Martin, J. W., McCauley, P. G., McNitt, P., Mehta, P., Moon, K. W., Mullens, J. W., Newington, T., Ning, Z., Ling Ng, B., Novo, S. M., O'Neill, M. J., Osborne, M. A., Osnowski, A., Ostadan, O., Paraschos, L. L., Pickering, L., Pike, A. C., Chris Pinkard, D., Pliskin, D. P., Podhasky, J., Quijano, V. J., Raczy, C., Rae, V. H., Rawlings, S. R., Chiva Rodriguez, A., Roe, P. M., Rogers, J., Rogert Bacigalupo, M. C., Romanov, N., Romieu, A., Roth, R. K., Rourke, N. J., Ruediger, S. T., Rusman, E., Sanches-Kuiper, R. M., Schenker, M. R., Seoane, J. M., Shaw, R. J., Shiver, M. K., Short, S. W., Sizto, N. L., Sluis, J. P., Smith, M. A., Ernest Sohna Sohna, J., Spence, E. J., Stevens, K., Sutton, N., Szajkowski, L., Tregidgo, C. L., Turcatti, G., Vandevondele, S., Verhovsky, Y., Virk, S. M., Wakelin, S., Walcott, G. C., Wang, J., Worsley, G. J., Yan, J., Yau, L., Zuerlein, M., Mullikin, J. C., Hurles, M. E., McCooke, N. J., West, J. S., Oaks, F. L., Lundberg, P. L., Klenerman, D., Durbin, R. & Smith, A. J. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53-59, 2008. PMID: 18987734.

127 Goodwin, S., McPherson, J. D. & McCombie, W. R. Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics* **17**, 333, 2016.

128 Stein, L. D. The case for cloud computing in genome informatics. *Genome Biol* **11**, 207, 2010. PMID: 20441614.

129 Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J Mol Biol* **215**, 403-410, 1990. PMID: 2231712.

130 Kamra, P., Gokhale, R. S. & Mohanty, D. SEARCHGTr: a program for analysis of glycosyltransferases involved in glycosylation of secondary metabolites. *Nucleic Acids Res* **33**, W220-225, 2005. PMID: 15980457.

131 Rausch, C., Weber, T., Kohlbacher, O., Wohlleben, W. & Huson, D. H. Specificity prediction of adenylation domains in nonribosomal peptide synthetases (NRPS) using transductive support vector machines (TSVMs). *Nucleic Acids Res* **33**, 5799-5808, 2005. PMID: 16221976.

132 Rottig, M., Medema, M. H., Blin, K., Weber, T., Rausch, C. & Kohlbacher, O. NRPSpredictor2--a web server for predicting NRPS adenylation domain specificity. *Nucleic Acids Res* **39**, W362-367, 2011. PMID: 21558170.

133 Prieto, C., Garcia-Estrada, C., Lorenzana, D. & Martin, J. F. NRPSsp: non-ribosomal peptide synthase substrate predictor. *Bioinformatics* **28**, 426-427, 2012. PMID: 22130593.

134 Prieto, C. Characterization of Nonribosomal Peptide Synthetases with NRPSsp. *Methods Mol Biol* **1401**, 273-278, 2016. PMID: 26831714.

135 Khayatt, B. I., Overmars, L., Siezen, R. J. & Francke, C. Classification of the adenylation and acyl-transferase activity of NRPS and PKS systems using ensembles of substrate specific hidden Markov models. *PLoS One* **8**, e62136, 2013. PMID: 23637983.

136 Bachmann, B. O. & Ravel, J. Chapter 8. Methods for in silico prediction of microbial polyketide and nonribosomal peptide biosynthetic pathways from DNA sequence data. *Methods Enzymol* **458**, 181-217, 2009. PMID: 19374984.

137 Baranasic, D., Zucko, J., Diminic, J., Gacesa, R., Long, P. F., Cullum, J., Hranueli, D. & Starcevic, A. Predicting substrate specificity of adenylation domains of nonribosomal peptide synthetases and other protein properties by latent semantic indexing. *J Ind Microbiol Biotechnol* **41**, 461-467, 2014. PMID: 24104398.

138 Knudsen, M., Sondergaard, D., Tofting-Olesen, C., Hansen, F. T., Brodersen, D. E. & Pedersen, C. N. Computational discovery of specificity-conferring sites in non-ribosomal peptide synthetases. *Bioinformatics* **32**, 325-329, 2016. PMID: 26471456.

139 Vijayan, M., Chandrika, S. K. & Vasudevan, S. E. PKSIIIexplorer: TSVM approach for predicting Type III polyketide synthase proteins. *Bioinformation* **6**, 125-127, 2011. PMID: 21584189.

140 Arnison, P. G., Bibb, M. J., Bierbaum, G., Bowers, A. A., Bugni, T. S., Bulaj, G., Camarero, J. A., Campopiano, D. J., Challis, G. L., Clardy, J., Cotter, P. D., Craik, D. J., Dawson, M., Dittmann, E., Donadio, S., Dorrestein, P. C., Entian, K. D., Fischbach, M. A., Garavelli, J. S., 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. & van der Donk, W. A. Ribosomally synthesized and post-translationally modified peptide natural products: overview and

recommendations for a universal nomenclature. *Nat Prod Rep* **30**, 108-160, 2013. PMID: 23165928.

141 Medema, M. H. & Fischbach, M. A. Computational approaches to natural product discovery. *Nat Chem Biol* **11**, 639-648, 2015. PMID: 26284671.

142 de Jong, A., van Hijum, S. A., Bijlsma, J. J., Kok, J. & Kuipers, O. P. BAGEL: a web-based bacteriocin genome mining tool. *Nucleic Acids Res* **34**, W273-279, 2006. PMID: 16845009.

143 de Jong, A., van Heel, A. J., Kok, J. & Kuipers, O. P. BAGEL2: mining for bacteriocins in genomic data. *Nucleic Acids Res* **38**, W647-651, 2010. PMID: 20462861.

144 Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389-3402, 1997. PMID: 9254694.

145 Begley, M., Cotter, P. D., Hill, C. & Ross, R. P. Identification of a novel two-peptide lantibiotic, lichenicidin, following rational genome mining for LanM proteins. *Appl Environ Microbiol* **75**, 5451-5460, 2009. PMID: 19561184.

146 Weber, T., Blin, K., Duddela, S., Krug, D., Kim, H. U., Bruccoleri, R., Lee, S. Y., Fischbach, M. A., Muller, R., Wohlleben, W., Breitling, R., Takano, E. & Medema, M. H. antiSMASH 3.0-a comprehensive resource for the genome mining of biosynthetic gene clusters. *Nucleic Acids Res* **43**, W237-243, 2015. PMID: 25948579.

147 Medema, M. H., Blin, K., Cimermancic, P., Jager, V. d., Zakrzewski, P., Fischbach, M. A., Weber, T., Takano, E. & Breitling, R. antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. 2011.

148 Blin, K., Wolf, T., Chevrette, M. G., Lu, X., Schwalen, C. J., Kautsar, S. A., Suarez Duran, H. G., de los Santos, E., Kim, H. U., Nave, M., Dickschat, J. S., Mitchell, D. A., Shelest, E., Breitling, R., Takano, E., Lee, S. Y., Weber, T. & Medema, M. H. antiSMASH 4.0—improvements in chemistry prediction and gene cluster boundary identification. *Nucleic Acids Res* **45**, W36-41, 2017. PMID: 28460038.

149 Medema, M. H., Takano, E. & Breitling, R. Detecting sequence homology at the gene cluster level with MultiGeneBlast. *Mol Biol Evol* **30**, 1218-1223, 2013. PMID: 23412913.

150 Ziemert, N., Podell, S., Penn, K., Badger, J. H., Allen, E. & Jensen, P. R. The natural product domain seeker NaPDoS: a phylogeny based bioinformatic tool to classify secondary metabolite gene diversity. *PLOS ONE* **7**, 2012.

151 Cimermancic, P., Medema, M. H., Claesen, J., Kurita, K., Wieland Brown, L. C., Mavrommatis, K., Pati, A., Godfrey, P. A., Koehrsen, M., Clardy, J., Birren, B. W., Takano, E., Sali, A., Linington, R. G. & Fischbach, M. A. Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters. *Cell* **158**, 412-421, 2014. PMID: 25036635.

152 Medema, M. H., Kottmann, R., Yilmaz, P., Cummings, M., Biggins, J. B., Blin, K., de Bruijn, I., Chooi, Y. H., Claesen, J., Coates, R. C., Cruz-Morales, P., Duddela, S., Dusterhus, S., Edwards, D. J., Fewer, D. P., Garg, N., Geiger, C., Gomez-Escribano, J. P., Greule, A., Hadjithomas, M., Haines, A. S., Helfrich, E. J., Hillwig, M. L., Ishida, K., Jones, A. C., Jones, C. S., Jungmann, K., Kegler, C., Kim, H. U., Kotter, P., Krug, D., Masschelein, J., Melnik, A. V., Mantovani, S. M., Monroe, E. A., Moore, M., Moss, N., Nutzmann, H. W., Pan, G., Pati, A., Petras, D., Reen, F. J., Rosconi, F., Rui, Z., Tian, Z., Tobias, N. J., Tsunematsu, Y., Wiemann, P., Wyckoff, E., Yan, X., Yim, G., Yu, F., Xie, Y., Aigle, B., Apel, A. K., Balibar, C. J., Balskus, E. P., Barona-Gomez, F., Bechthold, A., Bode, H. B., Borriss, R., Brady, S. F., Brakhage, A. A., Caffrey, P., Cheng, Y. Q., Clardy, J., Cox, R. J., De Mot, R., Donadio, S., Donia, M. S., van der Donk, W. A., Dorrestein, P. C., Doyle, S., Driessen, A. J., Ehling-Schulz, M., Entian, K. D., Fischbach, M. A., Gerwick, L., Gerwick, W. H., Gross, H., Gust, B., Hertweck, C., Hofte, M., Jensen, S. E., Ju, J., Katz, L., Kaysser, L., Klassen, J. L., Keller, N. P., Kormanec, J., Kuipers, O. P., Kuzuyama, T., Kyrpides, N. C., Kwon, H. J., Lautru, S., Lavigne, R., Lee, C. Y., Linquan, B., Liu, X., Liu, W., Luzhetskyy, A., Mahmud, T., Mast, Y., Mendez, C., Metsa-Ketela, M., Micklefield, J., Mitchell, D. A., Moore, B. S., Moreira, L. M., Muller, R., Neilan, B. A., Nett, M., Nielsen, J., O'Gara, F., Oikawa, H., Osbourn, A., Osburne, M. S., Ostash, B., Payne, S. M., Pernodet, J. L., Petricek, M., Piel, J., Ploux, O., Raaijmakers, J. M., Salas, J. A., Schmitt, E. K., Scott, B., Seipke, R. F., Shen, B., Sherman, D. H., Sivonen, K., Smanski, M. J., Sosio, M., Stegmann, E., Sussmuth, R. D., Tahlan, K., Thomas, C. M., Tang, Y., Truman, A. W., Viaud, M., Walton, J. D., Walsh, C. T., Weber, T., van Wezel, G. P., Wilkinson, B., Willey, J. M., Wohlleben, W., Wright, G. D., Ziemert, N., Zhang, C., Zotchev, S. B., Breitling, R., Takano, E. & Glockner, F. O. Minimum Information about a Biosynthetic Gene cluster. *Nat Chem Biol* **11**, 625-631, 2015. PMID: 26284661.

153 Skinnider, M. A., Dejong, C. A., Rees, P. N., Johnston, C. W., Li, H., Webster, A. L., Wyatt, M. A. & Magarvey, N. A. Genomes to natural products PRediction Informatics for Secondary Metabolomes (PRISM). *Nucleic Acids Res* **43**, 9645-9662, 2015. PMID: 26442528.

154 Skinnider, M. A., Merwin, N. J., Johnston, C. W. & Magarvey, N. A. PRISM 3: expanded prediction of natural product chemical structures from microbial genomes. *Nucleic Acids Res* **45**, W49-w54, 2017. PMID: 28460067.

155 Kale, A. J., McGlinchey, R. P., Lechner, A. & Moore, B. S. Bacterial self-resistance to the natural proteasome inhibitor salinosporamide A. *ACS Chem Biol* **6**, 1257-1264, 2011. PMID: 21882868.

156 Kale, A. J. & Moore, B. S. Molecular mechanisms of acquired proteasome inhibitor resistance. *J Med Chem* **55**, 10317-10327, 2012. PMID: 22978849.

157 Tang, X., Li, J., Millan-Aguinaga, N., Zhang, J. J., O'Neill, E. C., Ugalde, J. A., Jensen, P. R., Mantovani, S. M. & Moore, B. S. Identification of thiotetronic acid antibiotic biosynthetic pathways by target-directed genome mining. *ACS Chem Biol* **10**, 2841-2849, 2015. PMID: 26458099.

158 Alanjary, M., Kronmiller, B., Adamek, M., Blin, K., Weber, T., Huson, D., Philmus, B. & Ziemert, N. The Antibiotic Resistant Target Seeker (ARTS), an exploration engine for antibiotic cluster prioritization and novel drug target discovery. *Nucleic Acids Res* **45**, W42-48, 2017. PMID: 28472505.

159 Kersten, R. D., Yang, Y. L., Xu, Y., Cimermancic, P., Nam, S. J., Fenical, W., Fischbach, M. A., Moore, B. S. & Dorrestein, P. C. A mass spectrometry-guided genome mining approach for natural product peptidogenomics. *Nat Chem Biol* **7**, 794-802, 2011. PMID: 21983601.

160 Medema, M. H., Paalvast, Y., Nguyen, D. D., Melnik, A., Dorrestein, P. C., Takano, E. & Breitling, R. Pep2Path: automated mass spectrometry-guided genome mining of peptidic natural products. *PLoS Comput Biol* **10**, e1003822, 2014. PMID: 25188327.

161 Mohimani, H., Liu, W. T., Mylne, J. S., Poth, A. G., Colgrave, M. L., Tran, D., Selsted, M. E., Dorrestein, P. C. & Pevzner, P. A. Cycloquest: identification of cyclopeptides via database search of their mass spectra against genome databases. *J Proteome Res* **10**, 4505-4512, 2011. PMID: 21851130.

162 Ibrahim, A., Yang, L., Johnston, C., Liu, X., Ma, B. & Magarvey, N. A. Dereplicating nonribosomal peptides using an informatic search algorithm for natural products (iSNAP) discovery. *Proc Natl Acad Sci U S A* **109**, 19196-19201, 2012. PMID: 23132949.

163 Kersten, R. D., Ziemert, N., Gonzalez, D. J., Duggan, B. M., Nizet, V., Dorrestein, P. C. & Moore, B. S. Glycogenomics as a mass spectrometry-guided genome-mining method for microbial glycosylated molecules. *Proc Natl Acad Sci U S A* **110**, E4407-4416, 2013. PMID: 24191063.

164 Johnston, C. W., Skinnider, M. A., Wyatt, M. A., Li, X., Ranieri, M. R., Yang, L., Zechel, D. L., Ma, B. & Magarvey, N. A. An automated Genomes-to-Natural Products platform (GNP) for the discovery of modular natural products. *Nat Commun* **6**, 8421, 2015. PMID: 26412281.

165 Dejong, C. A., Chen, G. M., Li, H., Johnston, C. W., Edwards, M. R., Rees, P. N., Skinnider, M. A., Webster, A. L. & Magarvey, N. A. Polyketide and nonribosomal peptide retro-biosynthesis and global gene cluster matching. *Nat Chem Biol* **12**, 1007-1014, 2016. PMID: 27694801.

166 Hadjithomas, M., Chen, I. M., Chu, K., Ratner, A., Palaniappan, K., Szeto, E., Huang, J., Reddy, T. B., Cimermancic, P., Fischbach, M. A., Ivanova, N. N., Markowitz, V. M., Kyrpides, N. C. & Pati, A. IMG-ABC: a knowledge base to fuel discovery of biosynthetic gene clusters and novel secondary metabolites. *MBio* **6**, e00932, 2015. PMID: 26173699.

167 Ziemert, N., Lechner, A., Wietz, M., Millan-Aguinaga, N., Chavarria, K. L. & Jensen, P. R. Diversity and evolution of secondary metabolism in the marine actinomycete genus Salinispora. *Proc Natl Acad Sci U S A* **111**, E1130-1139, 2014. PMID: 24616526.

168 Doroghazi, J. R., Albright, J. C., Goering, A. W., Ju, K. S., Haines, R. R., Tchalukov, K. A., Labeda, D. P., Kelleher, N. L. & Metcalf, W. W. A roadmap for natural product discovery based on large-scale genomics and metabolomics. *Nat Chem Biol* **10**, 963-968, 2014. PMID: 25262415.

169 Zhang, Q., Doroghazi, J. R., Zhao, X., Walker, M. C. & van der Donk, W. A. Expanded natural product diversity revealed by analysis of lanthipeptide-like gene clusters in actinobacteria. *Appl Environ Microbiol* **81**, 4339-4350, 2015. PMID: 25888176.

170 Ju, K. S., Gao, J., Doroghazi, J. R., Wang, K. K., Thibodeaux, C. J., Li, S., Metzger, E., Fudala, J., Su, J., Zhang, J. K., Lee, J., Cioni, J. P., Evans, B. S., Hirota, R., Labeda, D. P., van der Donk, W. A. & Metcalf, W. W. Discovery of phosphonic acid natural products by mining the genomes of 10,000 actinomycetes. *Proc Natl Acad Sci U S A* **112**, 12175-12180, 2015. PMID: 26324907.

171 Leao, T., Castelao, G., Korobeynikov, A., Monroe, E. A., Podell, S., Glukhov, E., Allen, E. E., Gerwick, W. H. & Gerwick, L. Comparative genomics uncovers the prolific and distinctive metabolic potential of the cyanobacterial genus Moorea. *Proc Natl Acad Sci U S A* **114**, 3198-3203, 2017. PMID: 28265051.

172 Adamek, M., Spohn, M., Stegmann, E. & Ziemert, N. Mining Bacterial Genomes for Secondary Metabolite Gene Clusters. *Methods Mol Biol* **1520**, 23-47, 2017. PMID: 27873244.

173 Schorn, M. A., Alanjary, M. M., Aguinaldo, K., Korobeynikov, A., Podell, S., Patin, N., Lincecum, T., Jensen, P. R., Ziemert, N. & Moore, B. S. Sequencing rare marine actinomycete genomes reveals high density of unique natural product biosynthetic gene clusters. *Microbiology* **162**, 2075-2086, 2016. PMID: 27902408.

174 Turnbaugh, P. J., Ley, R. E., Hamady, M., Fraser-Liggett, C. M., Knight, R. & Gordon, J. I. The Human Microbiome Project. *Nature* **449**, 804, 2007.

175 Ursell, L. K., Metcalf, J. L., Parfrey, L. W. & Knight, R. Defining the Human Microbiome. *Nutr Rev* **70**, S38-44, 2012. PMID: 22861806.

176 Wilson, M. R., Zha, L. & Balskus, E. P. Natural product discovery from the human microbiome. *J Biol Chem* **292**, 8546-8552, 2017. PMID: 28389564.

177 Courtois, S., Cappellano, C. M., Ball, M., Francou, F. X., Normand, P., Helynck, G., Martinez, A., Kolvek, S. J., Hopke, J., Osburne, M. S., August, P. R., Nalin, R., Guerineau, M., Jeannin, P., Simonet, P. & Pernodet, J. L. Recombinant environmental libraries provide access to microbial diversity for drug discovery from natural products. *Appl Environ Microbiol* **69**, 49-55, 2003. PMID: 12513976.

178 Wang, G. Y., Graziani, E., Waters, B., Pan, W., Li, X., McDermott, J., Meurer, G., Saxena, G., Andersen, R. J. & Davies, J. Novel natural products from soil DNA libraries in a streptomycete host. *Org Lett* **2**, 2401-2404, 2000. PMID: 10956506.

179 Daniel, R. The soil metagenome--a rich resource for the discovery of novel natural products. *Curr Opin Biotechnol* **15**, 199-204, 2004. PMID: 15193327.

180 Hildebrand, M., Waggoner, L. E., Lim, G. E., Sharp, K. H., Ridley, C. P. & Haygood, M. G. Approaches to identify, clone, and express symbiont bioactive metabolite genes. *Nat Prod Rep* **21**, 122-142, 2004. PMID: 15039839.

181 Zazopoulos, E., Huang, K., Staffa, A., Liu, W., Bachmann, B. O., Nonaka, K., Ahlert, J., Thorson, J. S., Shen, B. & Farnet, C. M. A genomics-guided approach for discovering and expressing cryptic metabolic pathways. *Nat Biotechnol* **21**, 187-190, 2003. PMID: 12536216.

182 Brady, S. F., Simmons, L., Kim, J. H. & Schmidt, E. W. Metagenomic approaches to natural products from free-living and symbiotic organisms. *Nat Prod Rep* **26**, 1488-1503, 2009. PMID: 19844642.

183 Schmidt, E. W., Nelson, J. T., Rasko, D. A., Sudek, S., Eisen, J. A., Haygood, M. G. & Ravel, J. 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, 2005.

184 Piel, J., Hui, D., Wen, G., Butzke, D., Platzer, M., Fusetani, N. & Matsunaga, S. Antitumor polyketide biosynthesis by an uncultivated bacterial symbiont of the marine sponge Theonella swinhoei. *PNAS*, 2004.

185 Brady, S. F., Chao, C. J., Handelsman, J. & Clardy, J. Cloning and heterologous expression of a natural product biosynthetic gene cluster from eDNA. *Org Lett* **3**, 1981-1984, 2001. PMID: 11418029.

186 Molinski, T. F., Dalisay, D. S., Lievens, S. L. & Saludes, J. P. Drug development from marine natural products. *Nat Rev Drug Discov* **8**, 69-85, 2009. PMID: 19096380.

187 Donia, M. S., Ruffner, D. E., Cao, S. & Schmidt, E. W. Accessing the Hidden Majority of Marine Natural Products through Metagenomics. *ChemBioChem* **12**, 1230-1236, 2011.

188 Kampa, A., Gagunashvili, A. N., Gulder, T. A., Morinaka, B. I., Daolio, C., Godejohann, M., Miao, V. P., Piel, J. & Andresson, O. Metagenomic natural product discovery in lichen provides evidence for a family of biosynthetic pathways in diverse symbioses. *Proc Natl Acad Sci U S A* **110**, E3129-3137, 2013. PMID: 23898213.

189 Rath, C. M., Janto, B., Earl, J., Ahmed, A., Hu, F. Z., Hiller, L., Dahlgren, M., Kreft, R., Yu, F., Wolff, J. J., Kweon, H. K., Christiansen, M. A., Hakansson, K., Williams, R. M., Ehrlich, G. D. & Sherman, D. H. Meta-omic characterization of the marine invertebrate microbial consortium that produces the chemotherapeutic natural product ET-743. *ACS Chem Biol* **6**, 1244-1256, 2011. PMID: 21875091.

190 Agarwal, V., Blanton, J. M., Podell, S., Taton, A., Schorn, M. A., Busch, J., Lin, Z., Schmidt, E. W., Jensen, P. R., Paul, V. J., Biggs, J. S., Golden, J. W., Allen, E. E. & Moore, B. S. Metagenomic discovery of polybrominated diphenyl ether biosynthesis by marine sponges. *Nat Chem Biol* **13**, 537-543, 2017. PMID: 28319100.

191 Agarwal, V., El Gamal, A. A., Yamanaka, K., Poth, D., Kersten, R. D., Schorn, M., Allen, E. E. & Moore, B. S. Biosynthesis of polybrominated aromatic organic compounds by marine bacteria. *Nat Chem Biol* **10**, 640-647, 2014. PMID: 24974229.

192 Nielsen, H. B., Almeida, M., Juncker, A. S., Rasmussen, S., Li, J., Sunagawa, S., Plichta, D. R., Gautier, L., Pedersen, A. G., Le Chatelier, E., Pelletier, E., Bonde, I., Nielsen, T., Manichanh, C., Arumugam, M., Batto, J. M., Quintanilha Dos Santos, M. B., Blom, N., Borruel, N., Burgdorf, K. S., Boumezbeur, F., Casellas, F., Dore, J., Dworzynski, P., Guarner, F., Hansen, T., Hildebrand, F., Kaas, R. S., Kennedy, S., Kristiansen, K., Kultima, J. R., Leonard, P., Levenez, F., Lund, O., Moumen, B., Le Paslier, D., Pons, N., Pedersen, O., Prifti, E., Qin, J., Raes, J., Sorensen, S., Tap, J., Tims, S., Ussery, D. W., Yamada, T., Renault, P., Sicheritz-Ponten, T., Bork, P., Wang, J., Brunak, S. & Ehrlich, S. D. Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nat Biotechnol* **32**, 822-828, 2014. PMID: 24997787.

193 Lin, H. H. & Liao, Y. C. Accurate binning of metagenomic contigs via automated clustering sequences using information of genomic signatures and marker genes. *Sci Rep* **6**, 24175, 2016. PMID: 27067514.

194 Papudeshi, B., Haggerty, J. M., Doane, M., Morris, M. M., Walsh, K., Beattie, D. T., Pande, D., Zaeri, P., Silva, G. G. Z., Thompson, F., Edwards, R. A. & Dinsdale, E. A. Optimizing and evaluating the reconstruction of Metagenome-assembled microbial genomes. *BMC Genomics* **18**, 915, 2017. PMID: 29183281.

195 Sangwan, N., Xia, F. & Gilbert, J. A. Recovering complete and draft population genomes from metagenome datasets. *Microbiome* **4**, 8, 2016. PMID: 26951112.

196 Lizardi, P. M., Huang, X., Zhu, Z., Bray-Ward, P., Thomas, D. C. & Ward, D. C. Mutation detection and single-molecule counting using isothermal rolling-circle amplification. *Nat Genet* **19**, 225-232, 1998. PMID: 9662393.

197 Dean, F. B., Nelson, J. R., Giesler, T. L. & Lasken, R. S. Rapid amplification of plasmid and phage DNA using Phi 29 DNA polymerase and multiply-primed rolling circle amplification. *Genome Res* **11**, 1095-1099, 2001. PMID: 11381035.

198 Grindberg, R. V., Ishoey, T., Brinza, D., Esquenazi, E., Coates, R. C., Liu, W. T., Gerwick, L., Dorrestein, P. C., Pevzner, P., Lasken, R. & Gerwick, W. H. Single cell genome amplification accelerates identification of the apratoxin biosynthetic pathway from a complex microbial assemblage. *PLoS One* **6**, e18565, 2011. PMID: 21533272.

199 Wilson, M. C., Mori, T., Ruckert, C., Uria, A. R., Helf, M. J., Takada, K., Gernert, C., Steffens, U. A., Heycke, N., Schmitt, S., Rinke, C., Helfrich, E. J., Brachmann, A. O., Gurgui, C., Wakimoto, T., Kracht, M., Crusemann, M., Hentschel, U., Abe, I., Matsunaga, S., Kalinowski, J., Takeyama, H. & Piel, J. An environmental bacterial taxon with a large and distinct metabolic repertoire. *Nature* **506**, 58-62, 2014. PMID: 24476823.

200 Schorn, M., Zettler, J., Noel, J. P., Dorrestein, P. C., Moore, B. S. & Kaysser, L. Genetic basis for the biosynthesis of the pharmaceutically important class of epoxyketone proteasome inhibitors. *ACS Chem Biol* **9**, 301-309, 2014. PMID: 24168704.

201 Jensen, P. R., Williams, P. G., Oh, D. C., Zeigler, L. & Fenical, W. Species-Specific Secondary Metabolite Production in Marine Actinomycetes of the Genus Salinispora▿. *Appl Environ Microbiol* **73**, 1146-1152, 2007. PMID: 17158611.

202 Udwary, D. W., Zeigler, L., Asolkar, R. N., Singan, V., Lapidus, A., Fenical, W., Jensen, P. R. & Moore, B. S. Genome sequencing reveals complex secondary metabolome in the marine actinomycete Salinispora tropica. *Proc Natl Acad Sci U S A* **104**, 10376-10381, 2007. PMID: 17563368.

203 Moore, B. S., Kalaitzis, J. A. & Xiang, L. Exploiting marine actinomycete biosynthetic pathways for drug discovery. *Antonie Van Leeuwenhoek* **87**, 49-57, 2005. PMID: 15726291.

204 Dharmaraj, S. Marine Streptomyces as a novel source of bioactive substances. *World Journal of Microbiology and Biotechnology* **26**, 2123-2139, 2010.

205 Zhao, X. & Yang, T. Draft genome sequence of the marine sediment-derived actinomycete Streptomyces xinghaiensis NRRL B24674T. *Journal of Bacteriology* **193**, 2011.

206 Dalisay, D. S., Williams, D. E., Wang, X. L., Centko, R., Chen, J. & Andersen, R. J. Marine sediment-derived Streptomyces bacteria from British Columbia, Canada are a promising microbiota resource for the discovery of antimicrobial natural products. *PLoS One* **8**, e77078, 2013. PMID: 24130838.

207 Berdy, J. Bioactive microbial metabolites. *J Antibiot (Tokyo)* **58**, 1-26, 2005. PMID: 15813176.

208 Gontang, E., Gaudencio, S., Fenical, W. & Jensen, P. Sequence-based analysis of secondary metabolite biosynthesis in marine Actinobacteria. *Appl Environ Microbiol* **76**, 2487-2499, 2010.

209 Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko, S. I., Pham, S., Prjibelski, A. D., Pyshkin, A. V., Sirotkin, A. V., Vyahhi, N., Tesler, G., Alekseyev, M. A. & Pevzner, P. A. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *J Comput Biol* **19**, 455-477, 2012. PMID: 22506599.

210 Nurk, S., Bankevich, A., Antipov, D., Gurevich, A. A., Korobeynikov, A., Lapidus, A., Prjibelski, A. D., Pyshkin, A., Sirotkin, A., Sirotkin, Y., Stepanauskas, R., Clingenpeel, S. R., Woyke, T., McLean, J. S., Lasken, R., Tesler, G., Alekseyev, M. A. & Pevzner, P. A. Assembling single-cell genomes and mini-metagenomes from chimeric MDA products. *J Comput Biol* **20**, 714-737, 2013. PMID: 24093227.

211 Berthold, R. J., Borowitzka, M. A. & Mackay, M. A. The ultrastructure of Oscillatoria spongeliae, the blue-green algal endosymbiont of the sponge Dysidea herbacea. *Phycologia* **21**, 327-335, 1982.

212 Hinde, R., Pironet, F. & Borowitzka, M. A. Isolation of *Oscillatoria spongeliae*, the filamentous cyanobacterial symbiont of the marine sponge *Dysidea herbacea*. *Marine Biology* **119**, 99-104, 1994.

213 Thacker, R. W. & Starnes, S. Host specificity of the symbiotic cyanobacterium Oscillatoria spongeliae in marine sponges, Dysidea spp. *Marine biology* **142**, 2003.

214 Ridley, C. P., John Faulkner, D. & Haygood, M. G. Investigation of Oscillatoria spongeliae-Dominated Bacterial Communities in Four Dictyoceratid Sponges. *Appl Environ Microbiol* **71**, 7366-7375, 2005. PMID: 16269779.

215 Ridley, C. P., Bergquist, P. R., Harper, M. K., Faulkner, D. J., Hooper, J. N. A. & Haygood, M. G. Speciation and Biosynthetic Variation in Four Dictyoceratid Sponges and Their Cyanobacterial Symbiont, Oscillatoria spongeliae. *Chemistry & Biology* **12**, 397-406, 2005.

216 Agarwal, V., Li, J., Rahman, I., Borgen, M., Aluwihare, L. I., Biggs, J. S., Paul, V. J. & Moore, B. S. Complexity of naturally produced polybrominated diphenyl ethers

revealed via mass spectrometry. *Environ Sci Technol* **49**, 1339-1346, 2015. PMID: 25559102.

217 Carroll, A. R., Pierens, G. K., Fechner, G., De Almeida Leone, P., Ngo, A., Simpson, M., Hyde, E., Hooper, J. N., Bostrom, S. L., Musil, D. & Quinn, R. J. Dysinosin A: a novel inhibitor of Factor VIIa and thrombin from a new genus and species of Australian sponge of the family Dysideidae. *J Am Chem Soc* **124**, 13340-13341, 2002. PMID: 12418859.

218 Carroll, A. R., Buchanan, M. S., Edser, A., Hyde, E., Simpson, M. & Quinn, R. J. Dysinosins B-D, inhibitors of factor VIIa and thrombin from the Australian sponge Lamellodysidea chlorea. *J Nat Prod* **67**, 1291-1294, 2004. PMID: 15332844.

219 Broddrick, J. T., Rubin, B. E., Welkie, D. G., Du, N., Mih, N., Diamond, S., Lee, J. J., Golden, S. S. & Palsson, B. O. Unique attributes of cyanobacterial metabolism revealed by improved genome-scale metabolic modeling and essential gene analysis. *Proc Natl Acad Sci U S A* **113**, E8344-e8353, 2016. PMID: 27911809.

Chapter 2: Genetic Basis for the Biosynthesis of the Pharmaceutically Important Class of Epoxyketone Proteasome Inhibitors

2.1 Introduction to Chapter 2

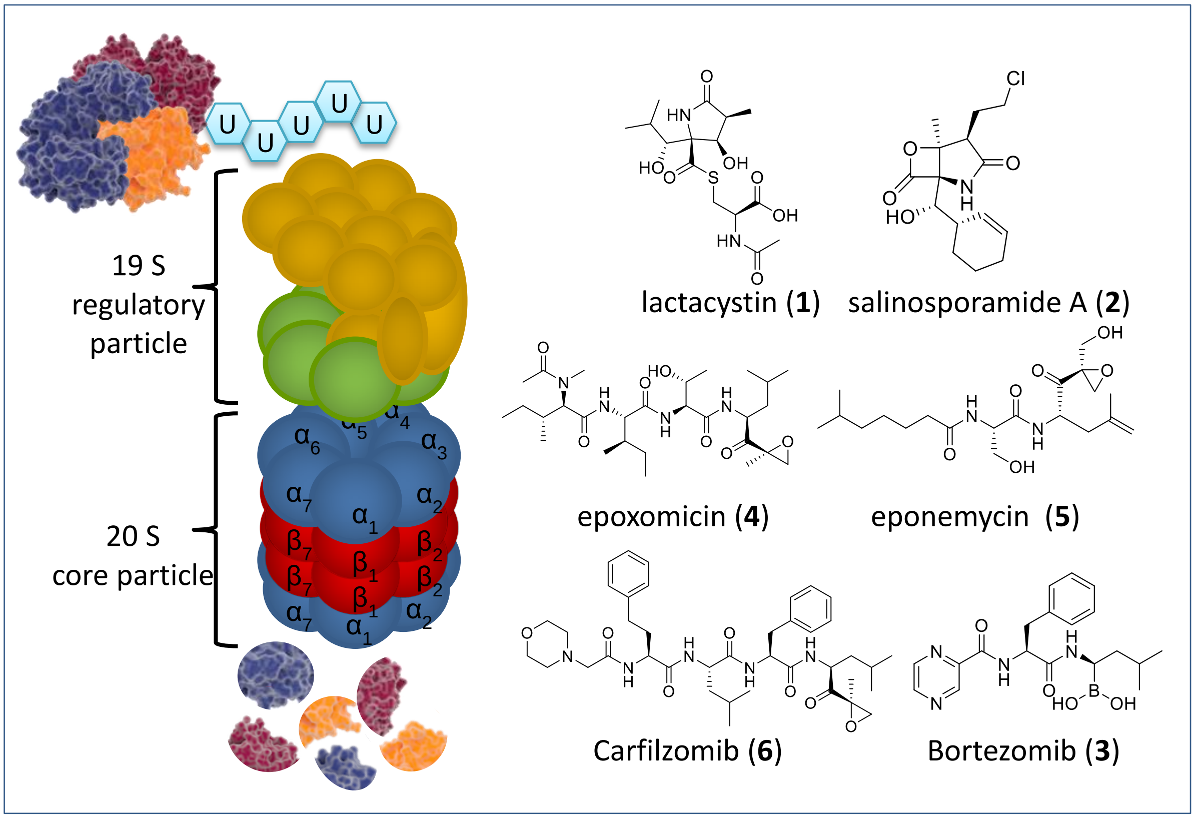

U U U U U
19 S regulatory particle
20 S core particle
α_6 α_5 α_4 α_3
α_7 α_1 α_2
β_7 β_1 β_2
β_7 β_1 β_2
α_7 α_1 α_2
lactacystin (1)
salinosporamide A (2)
epoxomicin (4)
eponemycin (5)
Carfilzomib (6)
Bortezomib (3)

Figure 4. The 20S Proteasome and Proteasome Inhibitors

The 20S proteasome is comprised of a core made up of two β-rings between two α-rings, each consisting of seven subunits, and a 19S regulatory particle that recognizes poly-ubiquinated proteins. The proteins enter the catalytic chamber where they are degraded into small peptides. Four natural product proteasome inhibitors (**1, 2, 4, 5**) and two synthetic proteasome inhibitors currently on the market as anti-cancer agents (**3, 6**) are shown.

Cancer is a tenacious and devastating disease affecting an increasing amount of the human population. According to the World Health Organization, cancer is the second leading cause of death globally, with almost 1 in 6 deaths attributed to cancer worldwide in 2015[1]. Many current cancer treatments have harsh side effects and tumor cells have been developing resistance to clinical drugs, so the search for new anticancer agents is pressing. One mechanism by which cancerous cells can be halted is by the inactivation of the proteasome, an essential protein-degradation macromolecular complex[2].

In eukaryotic cells, the 26S proteasome works in tandem with the ubiquitin system, which tags proteins with polyubiquitin chains that are recognized by the proteasome which proceeds with nonlysosomal protein degradation (Figure 1). The proteasome is responsible for 90% of protein degradation in the cell. The 26S proteasome is involved in a variety of cell functions from DNA repair to cell cycle control to apoptosis, through its most fundamental purpose: irreversibly inactivating targeted proteins. This targeted degradation can be used to regulate concentrations of proteins within the cell. Inhibiting the proteasome has many medicinally relevant consequences, including cell apoptosis, anti-viral activity, anti-tuberculosis activity, and anti-cancer therapy. Arresting the function of the proteasome specifically suppresses angiogenesis, which can go unregulated in cancerous cells causing tumor growth, and ultimately leads to apoptosis in oncogenic cells while retaining limited toxicity to normal cells[3]. This large protein complex is composed of the core 20S proteasome and the 19S regulatory particle, both made up of multiple subunits. The 20S proteasome core is formed by stacking two outer α-rings and two inner β-rings, each ring made up of seven subunits. Of the 14 β subunits, β1, β2, and β5 have specific hydrolytic activity. The β1-subunit has caspase-like activity, the β2-subunit has trypsin-like activity, and the β5-subunit has chymotrypsin-like activity, all with active sites on the inner surface of the β-rings, forming a proteolytic chamber[4].

Proteasome inhibitors (PIs) bind, via a variety of mechanisms, to the hydrolytic β-subunits either irreversibly or reversibly. The first natural product PI, lactacystin (**1**), was isolated from a *Streptomyces* sp. in 1991[5]. Subsequent studies showed that at low concentrations of lactacystin, the eukaryotic cell cycle would be halted and at higher concentrations, would causes apoptosis[6]. Further studies showed that lactacystin was selectively causing apoptosis in malignant cells, rather than healthy cells. This fueled the speculation that cancerous cells rely more heavily on proteasomal degradation than do normal cells[7]. There are eight structural classes of PIs, five of which have natural products among them. One example is salinosporamide A (**2**) of the β-lactone class, produced by the marine actinomycete *Salinispora tropica*[8]. It is currently in clinical trials, known as Marizomib, for treatment of multiple myeloma and glioma. Bortezomib (**3**) was the first PI approved by the FDA for treatment of multiple myeloma and mantle cell lymphoma. However, bortezomib binds reversibly, has significant side effects, and has developed resistance in some patients. Resistance to bortezomib was first seen in cancer cell lines and subsequently in patients. It was found that repeated exposure to bortezomib caused some cell lines to become resistant. Further studies showed an up-regulation of β-subunits and/or a mutation in the gene encoding for the β_5-subunit found in these cell lines[9].

Insights into PI resistance can be gleaned by studying the biosynthetic pathways of actinomycetes that produce PIs, as they have functioning proteasomes and must have some resistance mechanism to survive. In the salinosporamide A pathway there is an extra gene, SalI, which encodes a mutated proteasome β-subunit. This gene is accessory to the normal proteasome machinery in *Salinispora tropica*, and when expressed will complex with the primary α-rings to form a proteasome resistant to salinosporamide A. It was also found to be resistant to bortezomib, suggesting a mutation in the substrate binding pocket. When comparing the β-subunit S1 binding pocket protein sequence residues of multiple actinomycetes, *Saccharomyces cerevisiae*, and *Homo sapiens*, a mutation at position 49 is apparent. All “non-resistant” sequences contain an alanine at this residue, while SalI has a valine in this position. Site directed mutagenesis at the 49 position confirmed that a mutation to valine causes a loss in hydrolytic activity[10].

The most specific and potent class of PIs is the α'β'-epoxyketones, which covalently and irreversibly bind to the proteolytically active subunits only in the proteasome, not other proteases in the cell. The first two PIs with this unique epoxyketone moiety were epoxomicin (**4**) and eponemycin (**5**), both naturally produced by actinomycetes discovered in the early 1990s[11,12]. An analogue of epoxomicin, carfilzomib (**6**), was recently approved by the FDA in July 2012 as a third line treatment against multiple myeloma[13]. The mechanism by which the epoxyketone functional group selectively binds to and halts the function of the proteasome was investigated using co-crystallization of epoxomicin bound to a yeast 20S proteasome. The exquisite specificity with which epoxyketone PIs inhibit the proteasome is explained by a six-membered morpholine ring formed by the opening of the epoxyketone warhead by the N-terminal threonine of the catalytic subunits[14]. This morpholino adduct reveals the exploitation of the 26S proteasome's unusual catalytic mechanism; the free α-amino group of the threonine opens the epoxide to form the ring, and thus serine and cysteine proteases will not be inhibited as these residues do not have a free α-amino group to open the epoxide[15].

While the mechanism by which the epoxyketone binds to and arrests the function of the proteasome has been described, the biosynthesis of these natural products is unknown. Unveiling the biosynthetic pathways to assemble these molecules could enable production of large quantities of starting material for synthetic derivatives and precursors. Of particular interest to us is the assembly of the unique epoxyketone moiety and the presence of proteasome subunit resistance genes in the biosynthetic gene clusters encoding epoxyketone PIs. This research represents the identification of the first biosynthetic gene clusters for epoxyketone PIs.

This work has been published in ACS Chemical Biology in January 2014, Volume 9, Issue 1, Pages 301- 309[16]. This publication resulted from a collaboration with Dr. Leonard Kaysser who wrote the majority of the paper, planned the experiments with Dr. Bradley Moore, and carried out some of the lab work. I wrote portions of the paper and performed the majority of the lab work. Judith Zettler curated and submitted the genomic information to NCBI. Dr. Joseph Noel provided the Ion Torrent PGM machine I used to sequence the genomes. Dr. Pieter Dorrestein provided the LC/MS instruments used for molecular networking.

We obtained the strains *Streptomyces* sp. ATCC 53904, the epoxomicin producer, and *Streptomyces* sp. ATCC 53709, the eponemycin producer and isolated genomic DNA. I then sequenced the two strains using the Ion Torrent^{TM} PGM with a modified protocol that I devised for improved sequencing with high G + C content DNA. I assembled the genomes using the CLC Genomic Workbench software suite. The assembled genomes were submitted to antiSMASH, and we identified the hybrid non-ribosomal peptide synthetase (NRPS)/polyketide synthase (PKS) putative gene clusters responsible for the production of epoxomicin and eponemycin. I then constructed two fosmid libraries using the CopyControl^{TM} Fosmid Library Production Kit. I screened each library using primers specific for genes within each cluster and identified a positive clone with the entire gene cluster intact for each fosmid library. We then proceeded to create integration cassettes for stable chromosomal integration into heterologous hosts, using triparental intergenic conjugation. Dr. Kaysser then analyzed the heterologous hosts using HPLC to separate the culture extracts and find the peaks representing epoxomicin and eponemycin. He then went on to do molecular networking of the eponemycin host to reveal various structural analogues produced in the heterologous host containing the eponemycin gene cluster. We found that the gene cluster for eponemycin contained a mutated, secondary β-subunit, and while the gene cluster for epoxomicin did not contain the accessory β-subunit, there was a second, mutated copy elsewhere in the genome. This accessory subunit acts as the producer's defense mechanism against its own toxic molecule.

As a whole, this project provided the perfect foundation for my continuing studies on connecting genes and molecules in antibiotic producing bacteria. The skills I learned in this project lead me to develop a larger scale sequencing effort of rare marine actinomycetes to genome mine these underutilized taxa for the secondary metabolite potential.

Section 2.3 is a reprint of material as it appears in "Genetic Basis for the Biosynthesis of the Pharmaceutically Important Class of Epoxyketone Proteasome Inhibitors" in *ACS Chemical Biology*, 2014, Schorn, M., Zettler, J., Noel, J. P., Dorrestein, P. C., Moore, B. S. & Kaysser, L.

2.2 Chapter 2 Introduction References

1 *WHO | Cancer*, <http://www.who.int/cancer/en/> (2017).

2 Crawford, L. J., Walker, B. & Irvine, A. E. Proteasome inhibitors in cancer therapy. *J Cell Commun Signal* **5**, 101-110, 2011. PMID: 21484190.

3 Borissenko, L. & Groll, M. 20S proteasome and its inhibitors: crystallographic knowledge for drug. *Chem Rev* **107**, 687-717, 2007. PMID: 17316053.

4 Murata, S., Yashiroda, H. & Tanaka, K. Molecular mechanisms of proteasome assembly. *Nat Rev Mol Cell Biol* **10**, 104-115, 2009. PMID: 19165213.

5 Omura, S., Fujimoto, T., Otoguro, K., Matsuzaki, K., Moriguchi, R., Tanaka, H. & Sasaki, Y. Lactacystin, a novel microbial metabolite, induces neuritogenesis of neuroblastoma cells. *J Antibiot (Tokyo)* **44**, 113-116, 1991. PMID: 1848215.

6 Imajoh-Ohmi, S., Kawaguchi, T., Sugiyama, S., Tanaka, K., Omura, S. & Kikuchi, H. Lactacystin, a specific inhibitor of the proteasome, induces apoptosis in human monoblast U937 cells. *Biochem Biophys Res Commun* **217**, 1070-1077, 1995. PMID: 8554559.

7 Delic, J., Masdehors, P., Omura, S., Cosset, J. M., Dumont, J., Binet, J. L. & Magdelenat, H. The proteasome inhibitor lactacystin induces apoptosis and sensitizes chemo- and radioresistant human chronic lymphocytic leukaemia lymphocytes to TNF-alpha-initiated apoptosis. *Br J Cancer* **77**, 1103-1107, 1998. PMID: 9569046.

8 Feling, R. H., Buchanan, G. O., Mincer, T. J., Kauffman, C. A., Jensen, P. R. & Fenical, W. Salinosporamide A: a highly cytotoxic proteasome inhibitor from a novel microbial source, a marine bacterium of the new genus salinospora. *Angew Chem Int Ed Engl* **42**, 355-357, 2003. PMID: 12548698.

9 Kale, A. J. & Moore, B. S. Molecular mechanisms of acquired proteasome inhibitor resistance. *J Med Chem* **55**, 10317-10327, 2012. PMID: 22978849.

10 Kale, A. J., McGlinchey, R. P., Lechner, A. & Moore, B. S. Bacterial self-resistance to the natural proteasome inhibitor salinosporamide A. *ACS Chem Biol* **6**, 1257-1264, 2011. PMID: 21882868.

11 Hanada, M., Sugawara, K., Kaneta, K., Toda, S., Nishiyama, Y., Tomita, K., Yamamoto, H., Konishi, M. & Oki, T. Epoxomicin, a new antitumor agent of microbial origin. *J Antibiot (Tokyo)* **45**, 1746-1752, 1992. PMID: 1468981.

12 Sugawara, K., Hatori, M., Nishiyama, Y., Tomita, K., Kamei, H., Konishi, M. & Oki, T. Eponemycin, a new antibiotic active against B16 melanoma. I. Production. *J Antibiot (Tokyo)* **43**, 8-18, 1990. PMID: 2106503.

13 Kortuem, K. M. & Stewart, A. K. Carfilzomib. *American Society of Hematology*, 2012.

14 Groll, M., Kim, K., Kairies, N. & Crews, C. Crystal Structure of Epoxomicin:20s Proteasome Reveals a Molecular Basis for Selectivity of Alpha,beta-epoxyketone Proteasome Inhibitors. *J Am Chem Soc*, 2000.

15 Kisselev, A. F., van der Linden, W. A. & Overkleeft, H. S. Proteasome Inhibitors: An Expanding Army Attacking a Unique Target. *Chem Biol* **19**, 99-115, 2012. PMID: 22284358.

16 Schorn, M., Zettler, J., Noel, J. P., Dorrestein, P. C., Moore, B. S. & Kaysser, L. Genetic basis for the biosynthesis of the pharmaceutically important class of epoxyketone proteasome inhibitors. *ACS Chem Biol* **9**, 301-309, 2014. PMID: 24168704.

2.3 Reprint of: “Genetic Basis for the Biosynthesis of the Pharmaceutically Important Class of Epoxyketone Proteasome Inhibitors”

ACS chemical biology

Articles

pubs.acs.org/acschemicalbiology

Genetic Basis for the Biosynthesis of the Pharmaceutically Important Class of Epoxyketone Proteasome Inhibitors

Michelle Schorn,[†] Judith Zettler,[‖,⊥] Joseph P. Noel,[‡] Pieter C. Dorrestein,[§] Bradley S. Moore,*[,†,§] and Leonard Kaysser*[,†,‖,⊥]

[†]Scripps Institution of Oceanography, University of California, San Diego, California 92093, United States of America

[‡]Jack H. Skirball Center for Chemical Biology and Proteomics, Salk Institute for Biological Studies, La Jolla, California 92037, United States of America

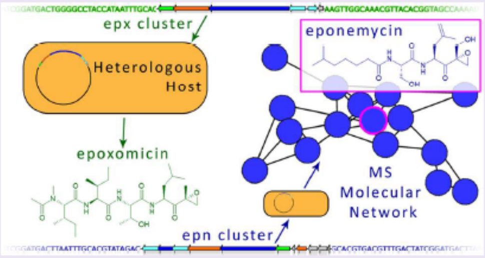
[§]Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California, San Diego, California 92093, United States of America

[‖]Pharmaceutical Biology, Eberhard Karls University Tübingen, 72076 Tübingen, Germany

[⊥]German Center for Infection Research (DZIF), partner site Tübingen, 72076 Tübingen, Germany

S *Supporting Information*

ABSTRACT: The epoxyketone proteasome inhibitors are an established class of therapeutic agents for the treatment of cancer. Their unique α',β'-epoxyketone pharmacophore allows binding to the catalytic β-subunits of the proteasome with extraordinary specificity. Here, we report the characterization of the first gene clusters for the biosynthesis of natural peptidyl-epoxyketones. The clusters for epoxomicin, the lead compound for the anticancer drug Kyprolis, and for eponemycin were identified in the actinobacterial producer strains ATCC 53904 and *Streptomyces hygroscopicus* ATCC 53709, respectively, using a modified protocol for Ion Torrent PGM genome sequencing. Both gene clusters code for a hybrid nonribosomal peptide synthetase/polyketide synthase multifunctional enzyme complex and homologous redox enzymes. Epoxomicin and eponemycin were heterologously produced in *Streptomyces albus* J1046 via whole pathway expression. Moreover, we employed mass spectral molecular networking for a new comparative metabolomics approach in a heterologous system and discovered a number of putative epoxyketone derivatives. With this study, we have definitively linked epoxyketone proteasome inhibitors and their biosynthesis genes for the first time in any organism, which will now allow for their detailed biochemical investigation.



The 26S proteasome is the essential enzymatic complex for nonlysosomal proteolytic degradation in eukaryotes.[1] It mediates levels of key factors in a variety of essential cellular processes that are deregulated in cancer cells and pivotal elements in carcinogenesis and tumorigenesis. Thus, the inhibition of the proteasome specifically targets heavily proliferating cells over quiescent cells.[2,3] The first proteasome inhibitor bortezomib (Figure 1, **1**) (marketed as Velcade by Millenium Pharmaceuticals) was approved by the U.S. Food and Drug Administration (FDA) in 2003. It is currently applied as a first-line treatment for multiple myeloma and mantle cell lymphoma. However, intravenous administration of the drug is associated with significant side effects. The development of proteasome inhibitors with improved properties is therefore an ongoing effort.

A number of potent proteasome inhibitors have been isolated from nature, predominantly from microorganisms.[4] The most prominent class are the peptide epoxyketones, which comprise epoxomicin (**2**),[5] eponemycin (**3**),[6] and several related compounds (**4**–**8**)[7–9] (Figure 1). All these molecules consist of a short peptidic core structure with a terminal C_3-extended leucine derivative. Compound **2** is particularly potent with IC_{50} values against the proteasome as low as 2.5 nM.[7] The compound has been used as a lead for the development of carfilzomib (**9**, Kyprolis, Onyx Pharmaceuticals), which was granted accelerated approval by the FDA in July 2012 for the treatment of refractory and relapsed multiple myeloma.[10] The drug appears to be better tolerated by patients than **1** and can therefore be applied in higher and more effective doses.[11,12] Beside their usage as anticancer drugs, epoxyketones have shown excellent activity against parasites.[13] In particular, *Plasmodium falciparum*, the deadly malaria pathogen, is vitally affected by this class of proteasome inhibitors at different stages of its life cycle.[14] Co-crystallization experiments with **2** and the

Received: September 11, 2013
Accepted: October 29, 2013

dx.doi.org/10.1021/cb400699p | *ACS Chem. Biol.* XXXX, XXX, XXX–XXX

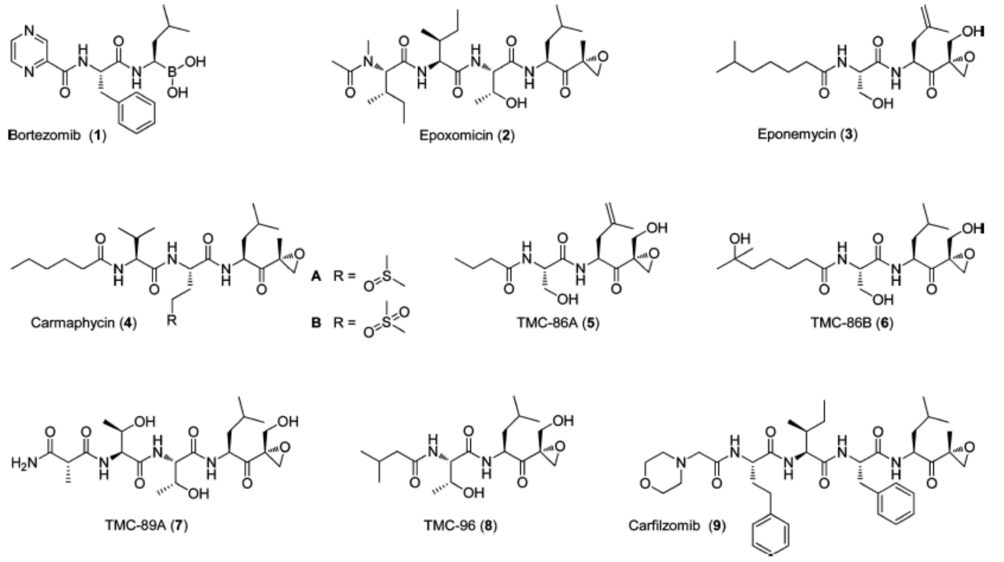


Figure 1. Chemical structures of proteasome inhibitors.

yeast proteasome revealed an irreversible two-step binding mechanism and the essential role of the α',β'-epoxyketone warhead.[15] The unprecedented terminal α',β'-position of the C3-carbonyl and the ring-strained epoxide constitutes two strongly electrophilic groups in the immediate proximity to each other, which are very accessible for nucleophilic attack. We were thus highly interested in the biosynthesis of these compounds and their unique pharmacophore. The biotransformation of the carboxy-terminus of a polyketide intermediate to an epoxide is biochemically not trivial and may involve either new polyketide biochemistry and/or unique enzymatic redox reactions.

In this study, we present the epoxomicin and the eponemycin gene cluster from an unspecified actinomycete strain ATCC 53904 and *Streptomyces hygroscopicus* strain ATCC 53709, respectively. Both compounds were produced by heterologous pathway expression in *S. albus* J1074 to definitively link epoxyketone proteasome inhibitors and their biosynthesis genes for the first time in any organism. This genetic linkage allowed us to locate homologous orphan gene clusters in various bacteria that promise the discovery of new bioactive derivative molecules.

■ RESULTS AND DISCUSSION

Identification of the Epoxomicin and Eponemycin Gene Clusters by Ion PGM Genome Sequencing. To investigate the biosynthetic pathways of epoxyketone proteasome inhibitors, we attempted to isolate the genes responsible for the formation of the prototypes **2** and **3**. To this end, we subjected genomic DNA of the producer strains ATCC 53904,[5] an unspecified actinomycete, and *S. hygroscopicus* ATCC 53709[6] to semiconductor sequencing from Ion Torrent.[16] Recently, we employed Ion Torrent technology in the *de novo* sequencing of the draft genome of *Thalassospira* sp. CNJ-328, which has a GC-content at around 50%.[17] However, we were unsuccessful in the sequencing of DNA with high GC-content such as from actinobacteria using standard protocols provided by the manufacturer. To address the sequencing problems, we slightly modified the procedure for manual template preparation as described in the Ion PGM 200 Xpress Template Kit. Betaine has been shown previously to substantially improve the amplification of difficult GC-rich DNA sequences.[18] As the Ion PGM template preparation is PCR based, we thus added betaine to a final concentration of 1 M to the amplification mix. After template preparation, enrichment, and sequencing with the Ion PGM system, the assembly of the obtained sequence data resulted in the generation of two draft genomes. The assembled sequence of the epoxomicin producer ATCC 53904 genome contains 8.9 Mb with a GC-content of 71.8% and was presented on 426 contigs with a 66-fold coverage. Similarly, the 9.8 Mb assembled genome sequence of the eponemycin producer ATCC 53709 has a GC-content of 71.1% and was presented on 490 contigs with a 79-fold coverage. Our modified protocol proved efficient and might thus facilitate the future application of semiconductor technology for the genome sequencing of other high-GC bacteria.

To assess the secondary metabolomic potential of both strains, we submitted the draft genome sequences to *in silico* analysis via the software antiSMASH.[19] Among 52 and 70 preliminary putative biosynthetic gene clusters, we identified potential clusters for the formation of **2** and **3** at 27.9 kb and 23.8 kb, respectively (Figure 2A). The homologous clusters encode hybrid nonribosomal peptide synthetase (NRPS)/polyketide synthase (PKS) multifunctional enzymes consistent with the putative formation of the core structure of the compounds. In addition, we identified analogous genes for a putative P450 monooxygenase and a conserved acyl-CoA dehydrogenase. A detailed summary of the proposed function of the genes can be found in the Supporting Information, Table S1.

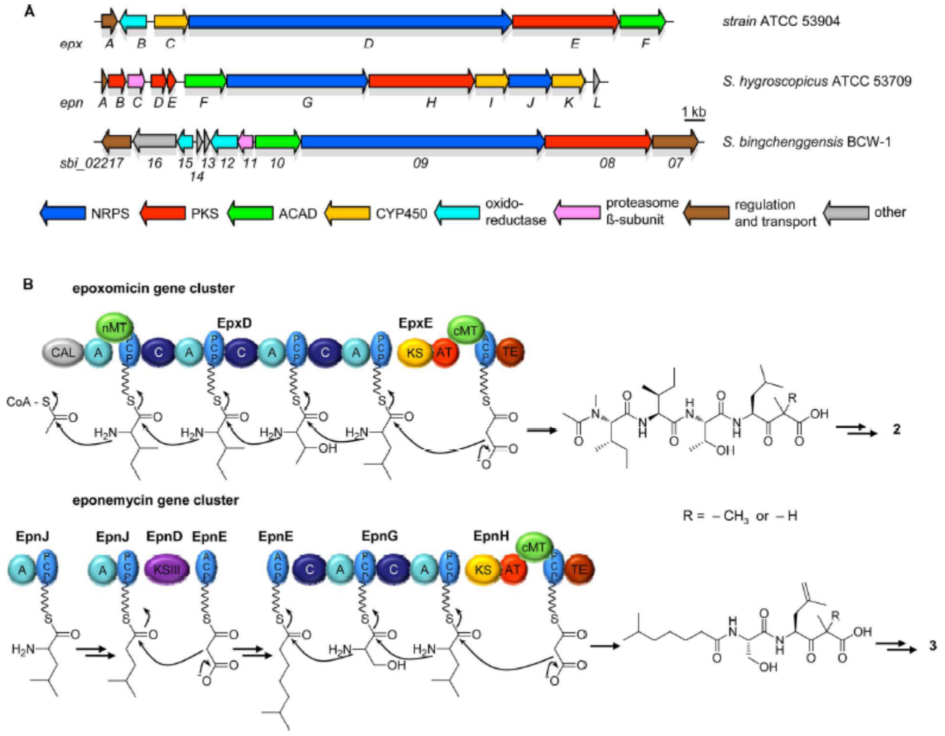


Figure 2. Biosynthesis of natural epoxyketones. (A) Relative organization of the epoxomicin (*epx*), eponemycin (*epn*) and orphan *S. bingchenggensis* gene clusters. (B) Predicted NRPS/PKS (nonribosomal peptide synthetase/polyketide synthase) assembly line synthesis of epoxomicin (**2**) and eponemycin (**3**). Abbreviations: ACAD, acyl-CoA dehydrogenase; CYP450, cytochrome P450; A, adenylation domain; ACP, acyl carrier protein; PCP, peptidyl carrier protein; C, condensation domain; MT, methyltransferase; TE, thioesterase; KS ketosynthase; AT acyltransferase domain.

These observations and the presence of a gene putatively encoding a resistant β-proteasome subunit homologous to the salinosporamide resistance enzyme[20] further suggested that the identified gene clusters code for epoxomicin (*epx*) and eponemycin (*epn*) production.

Heterologous Production of Epoxomicin and Eponemycin in *S. albus*. To confirm the suspected functions of the *epx* and *epn* gene clusters, we designed experiments to produce **2** and **3** in a surrogate host organism. To this end, we generated two fosmid libraries from the genomic DNA of both producer strains, ATCC 53904 and ATCC 53709, to isolate the gene clusters. The genomic libraries comprised ∼1800 individual clones each and were screened by PCR. Both clusters were found intact on single fosmids, the epoxomicin gene cluster on fosmid 15C3 and the eponemycin gene cluster on fosmid 2H4. A heterologous expression approach was used to confirm the identity of the clusters.

For this purpose, we replaced the chloramphenicol resistance gene in the fosmid backbone by λ-Red-mediated recombination with an integration cassette we generated previously.[21] The cassette int_neo contains the attP attachment site and the integrase gene (*int*) of phage ΦC31, a kanamycin resistance gene (*neo*) and an origin of transfer (*oriT*), and allows site-specific integration in most *Streptomyces* chromosomes.[22] The resulting fosmids were named epxMS01 and epnLK01. In order to express the introduced pathway, the transcriptional and translational machinery of the host strain must recognize the promoter, ribosomal binding sites (RBS) and the regulatory system of the gene cluster. Consequently, the transfer of a biosynthetic gene cluster into a phylogenetically related strain is preferable for heterologous expression. While the eponemycin producer ATCC 53709 belongs to the genus *Streptomyces*, the taxonomic specification of the epoxomicin producer ATCC 53904 was not defined. We thus analyzed two phylogenetic markers, the 16S rRNA and the *rpoB* gene, for their relatedness to genes from other bacteria. Both markers classify the strain as a member of the Actinobacteria and the taxonomic family Pseudonocardiaceae. The *rpoB* gene shows highest homology (89% identity) to *Actinosynnema mirum* DSM 43827. The 16S rRNA confirms the close relatedness to *A. mirum* (96% identity) but indicates an even greater similarity to *Goodfellowia coeruleoviolacea* NRRL-B 24058 at 99% sequence identity.[23] Hence, we strongly consider ATCC 53904 to be a *Goodfellowia* species.

Both the epoxomicin and the eponemycin gene cluster were transferred into *S. albus* J1046 by intergeneric conjugation,[24] as this strain is one of the few in which non-*Streptomyces*-derived gene clusters have before been successfully expressed.[25] Three

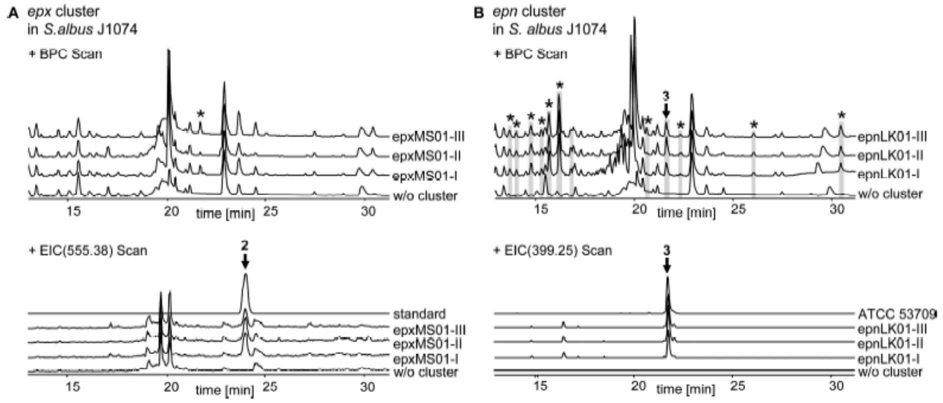


Figure 3. HPLC-MS analysis of (A) *S. albus* epxMS01-(1−3) expressing the epoxomicin gene cluster and (B) *S. albus* epnLK01-(1−3) expressing the eponemycin gene cluster. LC-MS base peak chromatograms (BPC, top) and extracted ion chromatograms (EIC, bottom) are depicted. Mass peaks of epoxomicin (**2**) and eponemycin (**3**) are indicated as well as unique peaks (*) in the *S. albus* heterologous mutants correlating to derivatives of **2** and **3** (see Supporting Information (SI) Table S2).

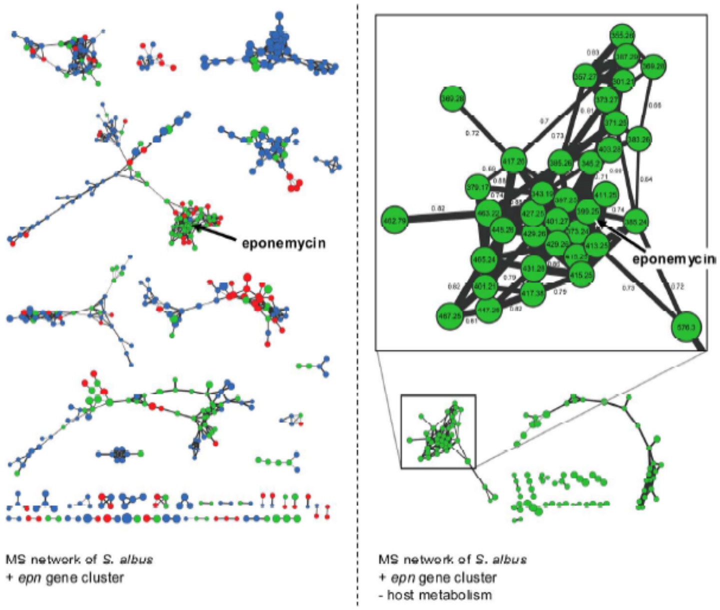


Figure 4. Molecular networks of mass spectra from *S. albus* containing the eponemycin gene cluster. Node (circle) colors indicate the source of the ions: blue (found in *S. albus* J1046), red (found only in *S. albus* epnLK01−1, −2, or −3), and green (found only in all three *S. albus* epnLK01-1, -2, and -3 strains). Node size indicates mass range of parent ions (m/z 200.17 − 1329.51). Edge line width indicates relatedness of MS/MS spectra represented by two connected nodes (cos 0.59−0.99). Selected masses and cosine values are noted in the eponemycin clade.

individual kanamycin resistant clones were selected and named *S. albus* epxMS01-(1−3) and *S. albus* epnLK01-(1−3), respectively. Ethyl acetate culture extracts of the mutant strains were analyzed by HPLC-MS and compared to the metabolic profiles of native *S. albus* J1064 and the producer strains ATCC 53904 and ATCC 53709 (Figure 3).

In the case of *S. albus* epnLK01, we discovered a new chromatographic peak not observed in *S. albus* J1064 (**3**, Figure 3B). Based on its chromatographic properties, its high-resolution mass, and its mass spectral fragmentation fingerprint (SI Figure S1), the product of the *S. albus*-expressed *epn* gene cluster is eponemycin (m/z calcd. 399.2495 $[M+H]^+$; obsvd. 399.2500). Moreover, we clearly identified at least 11 additional *epn*-based products suggestive of the production of a series of eponemycin derivatives (Figure 3B and SI Table S2).

Analysis of the *S. albus* epxMS01 extracts (Figure 3A) confirmed the accumulation of **2** (m/z calcd. 555.3758 [M+H]$^+$; obsvd. 555.3758). However, heterologous production of **2** was low and detected only by HPLC retention time and MS/MS comparisons with authentic material. A possible explanation for its low production is the phylogenetic distance of *S. albus* and the epoxomicin producer ATCC 53904 and the resulting incompatibility of their expression system and regulatory networks. However, a distinct mass peak present in the ion chromatograms of the mutant strains clearly indicates the accumulation of an epoxomicin congener (Figure 3B and SI Table S2). Our genetic experiments successfully linked the biosynthesis of **2** and **3** to the *epx* and *epn* gene sets, respectively. This experimental outcome represents the first genes-to-molecules confirmation of the pharmaceutically important family of proteasome epoxyketone inhibitors.

MS Molecular Networking Reveals New Eponemycin Derivatives. While examining the base peak chromatograms of *S. albus* containing the epoxyketone gene clusters, we observed a number of additional unique mass peaks (Figure 3B), particularly in the *epn* mutant. Motivated by the prospect of new eponemycin derivatives, we subjected *S. albus* epnLK01 to MS molecular networking for comparative metabolic profiling. The spectral networks paradigm was originally developed for application in proteomics[26] but has recently been adapted as a general MS/MS-data analysis tool.[27,28] A molecular network is created based on the relationships of MS/MS spectra for any molecule that is captured by mass spectrometry, even across multiple experiments. The network we generated from the combined data of *S. albus* J1046 with and without the *epn* gene cluster displayed ten major MS/MS clusters comprising six or more nodes (circles) of distinct mass fragmentation (Figure 4). These clusters likely represent structurally related molecules and are also referred to as molecular families (MFs).[29] Upon subtracting mass ions consistent with the *S. albus* J1046 wildtype or with only a subset of the three heterologous mutants, eight of the ten major MFs were eliminated. One of the two remaining clusters forms a tight network centered on a node with m/z 399.246 representing **3** (Figure 4). Notably, at least twelve individual mass ions are directly related to this node.

The analysis of the MS/MS spectra incorporated in the network revealed a number of structural derivatives of **3** that are produced by the heterologous mutants (SI Figure S2). This is especially interesting as **3** has been reported as a single compound from the wildtype producer ATCC 53709. Based on the fragmentation patterns, we postulate that most of the variation applies to the length and oxidation status of the fatty acid side chain. The heterologously produced eponemycin analogues contain shorter (C_4) or longer (C_9) acyl moieties with double bonds, hydroxyl and/or keto groups consistent with distinct HPLC retention times (SI Table S2 and Figure S2). This observation suggests that the enzyme responsible for the attachment of the fatty acid group to the peptide is promiscuous and therefore may facilitate future bioengineering efforts. In addition, some of the mass spectra suggest the production of congeners of **3** with an altered epoxyketone pharmacophore. Notably, di- and tetrahydro derivatives are so far only known as synthetic compounds[6] but may be prominent metabolites in the *S. albus* epnLK01 extracts (SI Table S2). We plan to report the structures and biological properties of new eponemycin analogues separately.

Biosynthetic Pathways for Epoxomicin and Eponemycin. The peptidic backbone of the two epoxyketone compounds is assembled by hybrid nonribosomal peptide synthetase/polyketide synthase multifunctional enzymes (Figure 2B). EpxD consists of a tetra-modular NRPS and analysis of the adenylation (A) domain substrate specificities strongly correlated A_1 and A_3 to the activation of isoleucine and threonine, respectively.[30,31] The specificities for the A_2 and A_4 domains were less evident (SI Table S3). The megasynthetase begins with a putative primer fatty acyl condensation (C) domain[32] known to transfer fatty acids onto the initial amino acid residue.[33] We thus postulate that this domain is responsible for the installation of the acetyl moiety in **2**. The methyltransferase (MT) domain in the first module of EpxD probably further modifies the Ile1 residue by introducing the *N*-methyl group.

Interestingly, the corresponding eponemycin NRPS, EpnG, does not contain a specific primer C-domain. This observation rather suggests that the eponemycin fatty acyl moiety is not constructed in the same way as in **2**. Branched odd-chain fatty acids in the C_6-C_{10} range, such as 6-methyl heptanoic acid (6-MHA), can be found in natural products[34,35] but are not common in bacterial primary metabolism. Hence, a dedicated pathway for the generation and incorporation of 6-MHA might be encoded in the *epn* cluster. Normally, the biosynthesis of branched chain fatty acids involves FabH (KAS III), which accepts small CoA-activated acyl groups derived from leucine, valine, or isoleucine.[36] FabH catalyzes the Claisen condensation of these substrates with one unit of malonyl-ACP to initiate type II fatty acid synthesis (FAS). For the formation of 6-MHA, one would typically expect that valine-derived isobutyrate undergoes two rounds of malonate extension, one round with the help of FabH and the other carried out by common type II FAS enzymes. Notably, EpnD is a FabH-homologue similar to other enzymes that are occasionally encoded in secondary metabolite gene clusters and participate in the priming of type II PKS systems with unusual starter units[37] or the generation of acyl side chains in lipopeptide biosynthesis.[35] However, the presence of the discrete NRPS tridomain EpnJ, consisting of an mbtH-like protein, a leu-specific A-domain, and a peptidyl carrier protein (PCP), might indicate a more unusual mechanism for the biosynthesis of 6-MHA (Figure 2B). Here, EpnJ-bound leucine may be subjected to deamination and reduction prior to C_2-extension by the FabH homologue EpnD using malonyl-EpnE. Further reduction to afford 6-MHA would likely be performed by primary fatty acid synthase enzymes.

The terminal PKS-modules of the epoxomicin and the eponemycin assembly lines, EpxE and EpnH, respectively, are identically organized. Both comprise a malonyl-CoA-specific acyltransferase (AT)-domain and a putative C-methyltransferase (cMT) domain, which suggests that the substituted epoxy moiety does not derive from methylmalonyl-CoA but rather from malonyl-CoA and S-adenosylmethionine (Figure 2B). A C-terminal thioesterase (TE) domain in EpxE and EpnH intimates that the peptide-polyketide hybrid product is released from the enzyme as the carboxylic acid. In this case, the construction of the rare α',β'-epoxyketone unit would have to be mediated by auxiliary enzymes likely acting subsequent or *in trans* to the assembly of the core structure. The transformation from the free acid to the epoxide may include two reductions, a dehydration and an epoxidation. Consequently, we analyzed the conserved biosynthesis genes in the epoxomicin and eponemycin loci, and gratifyingly, we identified two homologous genes common to both clusters—the putative acyl-CoA dehydrogenases (ACADs) *epxF*/*epnF* and the cytochrome P450 (CYP) monooxygenases *epxC*/*epnI*.

Cytochrome P450 enzymes are known to catalyze epoxidation reactions.[38] EpxC and EpnI are therefore plausible candidates to introduce an epoxy group into an unsaturated precursor molecule. Consequently, the reduction of the acid to the alcohol or the olefin may be performed by the ACADs EpxF/EpnF. Bacterial 4e$^-$ reductases that catalyze such reactions have been studied in the reductive off-loading of NRPSs[39] and in wax formation.[40] However, these enzymes rather belong to the short-chain dehydrogenase/reductase (SDR) superfamily, which is clearly distinct from the flavin adenine dinucleotide (FAD)-dependent ACADs. An alternative biosynthetic route to the epoxyketone moiety may involve the EpxE/EpnH cMT-domain if it introduces two methyl groups at C2. This scenario has been proposed in various polyketides with gem-dimethyl groups.[41] Subsequently, decarboxylation of the terminal carboxylic acid group to the dimethylketone may initiate a concerted reaction involving EpxF/EpnF and EpxC/EpnI to install the epoxide. On the basis of the biosynthetic features of the *epx* and *epn* gene clusters, the construction of the α',β'-epoxyketone pharmacophore is anticipated to involve new biochemical reactions. Detailed investigations of the pathway are therefore now underway. The distinct C8-OH group in eponemycin is likely implemented by the second CYP EpnK, which is unique to the *epn* cluster.

Genome Mining Identifies Homologous Gene Clusters in Various Bacteria. We next explored other microorganisms with the prospect to identify analogous pathways to new proteasome inhibitors. EpxF, the ACAD we predict to be essential for epoxyketone formation, was employed as a probe for a BLAST sequence homology search in the National Center for Biotechnology Information (NCBI) database. Strikingly, the genes with highest similarity to *epxF* (Expect (*E*) value < 1 e^{-70}) are all colocated with other genes from secondary metabolism. Most of these orphan biosynthetic pathways involve small hybrid NRPS/PKS systems and are present in a variety of different bacterial families (SI Figure S3). Notably, such homologous gene clusters can be found in a rhizosphere-associated *S. canus* 299MFChir4.1 strain as well as the human pathogen *Nocardia cyriacigeorgica* GUH-2. It might therefore be interesting to investigate if the compounds produced by these clusters have similar functions in symbiosis, virulence, and predation as the proteasome inhibitor family of syrbactins.[42,43] The cluster from *S. bingchenggensis* BCW-1 is particularly intriguing because it is highly analogous to the epoxomicin gene cluster (Figure 2A). We previously identified this set of genes and suggested it encodes a proteasome inhibitor, as it contains a secondary proteasome ß-subunit.[20] However, the *S. bingchenggensis* gene cluster lacks CYP homologues such as EpxC/EpnC that we propose are involved in the epoxidation reaction. We thus predict that this cluster may alternatively direct the formation of an acylated tripeptide in which the terminal amino acid moiety is modified not to an epoxyketone but rather as a vinylketone (enone) functional group. We recently showed in a separate study that synthetic carmaphycin enone derivatives are potent and irreversible proteasome inhibitors (B.S. Moore, unpublished data). Thus, we postulate that the orphan pathway in *S. bingchenggensis* BCW-1 encodes an unprecedented peptidic proteasome inhibitor.

In conclusion, we identified the epoxomicin and the eponemycin gene clusters, the first clusters for the biosynthesis of natural peptidyl epoxyketones. The genetic information suggests that their powerful pharmacophore is generated by a series of unprecedented biotransformations. With this study, we set the genetic basis to study the formation of the natural epoxyketone proteasome inhibitors in detail. Unexpectedly, we found a set of eponemycin congeners in extracts of a surrogate host organism with the *epn* cluster by molecular networking, demonstrating the benefits of this new technique for comparative metabolomics. Moreover, the presence of homologous gene clusters in other bacteria may facilitate the discovery of more new bioactive derivatives in the near future.

■ METHODS

Bacterial Strains and General Methods. Chemicals, microbiological, and molecular biological agents were purchased from standard commercial sources. Actinomycete strain ATCC 53904, *Streptomyces hygroscopicus* ATCC 53709, *Streptomyces albus* J1046, and their respective derivatives were maintained and grown on either MS agar (2% (w/v) soy flour, 2% (w/v) mannitol, 2% (w/v) agar; components purchased from Becton Dickinson) or TSB medium (Becton Dickinson). *Escherichia coli* strains were cultivated in LB medium (components purchased from Becton Dickinson) supplemented with appropriate antibiotics. DNA isolation and manipulations were carried out according to standard methods for *E. coli*[44] and *Streptomyces*.[45]

MS data were collected with an Agilent 6530 Accurate-Mass Quadrupole Time-of-flight (QTOF) LC-MS instrument (Agilent Technologies), and the analytes were separated with a reversed-phase C_{18} column (Phenomenex Luna 5 μ C18(2), 4.6 mm ×150 mm) on a 1260 Infinity LC-System (Agilent Technologies) using a flow rate of 0.1 mL/min.

Production of Epoxyketone Proteasome Inhibitors. TSB broth (10 mL) was inoculated with a spore suspension of *Streptomyces albus* J1046/epxMS01 or *Streptomyces albus* J1046/epnLK01 and incubated for 2 days at 30 °C and 200 rpm. Then 1% of the culture was transferred to 50 mL of R5 medium (103 g L^{-1} sucrose, 0.25 g L^{-1} K_2SO_4, 10.12 g L^{-1} $MgCl_2{\cdot}6H_2O$, 10 g L^{-1} glucose, 0.1 g L^{-1} casaminoacids, 5 g L^{-1} yeast extract, 5.73 g L^{-1} TES (*N*-tris(hydroxymethyl)methyl-2-aminoethanesulfonic acid), 80 μg L^{-1} $ZnCl_2$, 400 μg L^{-1} $FeCl_3{\cdot}6H_2O$, 20 μg L^{-1} $CuCl_2{\cdot}2H_2O$, 20 μg L^{-1} $MnCl_2{\cdot}4H_2O$, 20 μg L^{-1} $Na_2B_4O_7{\cdot}10H_2O$, 20 μg L^{-1} $(NH_4)_6Mo_7O_{24}{\cdot}4H_2O$, 50 mg L^{-1} KH_2PO_4, 3 g L^{-1} L-proline, 2.94 g L^{-1} $CaCl_2$, and 280 μg L^{-1} NaOH). After 6 days of incubation at 28 °C, 50 mL of EtOAc was added and incubated for 1 h at 200 rpm, and the EtOAc layer was recovered. The solvent was evaporated under reduced pressure. The residue was dissolved in 1 mL of MeCN, and the solution was filtered through a C_{18} sorbent (Spice C18 Sample Preparation Cartridges, Analtech). The filtrate was evaporated under reduced pressure in a 14-mL scintillation vial, and the residue was stored at −20 °C until LC/MS analysis.

Analysis of Culture Extracts and MS Molecular Networking. The residue was dissolved in 1000 μL of MeCN, and 5 μL of the dissolved residue was injected onto a reversed-phase HPLC column coupled to an mass spectrometer with an electrospray ionization interface (ESI) interface (heated capillary temperature 320 °C; sheath and collision gas nitrogen). The following solvent composition was used to separate the analytes: 10% (v/v) MeCN in H_2O for 4 min, 10−100% (v/v) MeCN in H_2O for 36 min, 100% (v/v) MeCN in H_2O for 3 min, 100−10% (v/v) MeCN in H_2O for 2 min. HR-MS data were acquired in positive mode ((+)-ESI). MS and MS/MS spectra were recorded with a scan rate of one and four spectra/second, respectively. Collision energy was 10 eV. Molecular formulas were calculated from monoisotopic masses using ChemCalc.[46]

To construct molecular networks MS/MS spectra recorded from extracts of *S. albus* J1046/epnLK01-(1−3) and *S. albus* J1046 wildtype were clustered using MS-Cluster.[47] Cluster-consensus spectra were further processed as described by Watrous et al.[28] Each spectrum comprised the 10 highest-cosine alignments in both directions. To define the MS/MS network pair wise alignments were considered with cosine ≥0.55 and ≥6 matched peaks. Custom scripts and the attributes to the molecular network were added as described.[28] The MS networks were visualized with Cytoscape (2.8.3.).[48]

DNA Sequencing and Bioinformatic Analysis. Ion Torrent libraries were prepared from genomic DNA (gDNA) of strains ATCC 53904 (epoxomicin producer) and ATCC 53709 (eponemycin producer), respectively, using the Ion Plus Fragment Library kit (Life Technologies) with a gDNA input of 1 μg. An S220 Focused-ultrasonicator (Covaris Inc.) was used to shear gDNA to obtain fragments of 100−250 bp size. Separation and extraction of DNA fragments was performed on the electrophoresis platform Pippin Prep (Sage Science Inc.). The Ion Library Quantitation Kit (Life Technologies) was used to quantify the libraries by quantitative real-time PCR on a LightCycler 480 (Roche Applied Science). No additional library amplification was performed. Samples were prepared manually using the Ion PGM 200 Xpress Template Kit (Life Technologies). To facilitate the amplification of the high GC content DNA, betaine was added to the amplification mix to a final concentration of 1M. The thermo profile was modified (95 °C for 10 min; 15 cycles, 95 °C for 30 s and 68 °C for 4 min; 30 cycles, 95 °C for 30 s and 68 °C for 6 min). Enrichment was performed on the Ion OneTouch ES (Life Technologies). Two sequencing runs per library were conducted on an Ion Personal Genome Machine (PGM) System using Ion 316 chips and the Ion PGM Sequencing 200 Kit v2 (Life Technologies). Sequencing data was assembled with the CLC Genomics Workbench software version 5.01 (CLC Bio). The draft genomes were subjected to the online tool antibiotics and Secondary Metabolite Analysis Shell (antiSMASH).[19] The *epx* and *epn* gene clusters were both found split on two contigs. The sequence gaps were closed by Sanger sequencing using primer walking and the shotgun method (GenoTech, Baejeon, Korea). Manual in silico sequence analysis was performed using GC frame-plot[49] and BLAST.[50] The Geneious software package (Biomatters Ltd.) and Artemis (Wellcome Trust Genome Campus) were used for sequence analysis and annotation.

Generation and Screening of Fosmid Libraries. The genomic fosmid libraries were constructed for strains ATCC 53904 and *Streptomyces hygroscopicus* ATCC 53709. High-molecular weight chromosomal DNA was randomly sheared to obtain fragments of ~40 kb size and cloned into pCC1FOS (Epicentre Biotechnologies). Fosmid libraries with ~1800 clones each were generated in *E. coli* EPI300 according to the manufacturers' instructions. For identification of the biosynthetic gene clusters the fosmid libraries were screened by PCR with primer pairs epxA_f GAATCTCAAGCGCGAGGGG/ epxA_r GGTGTCGCGGAAGTAGTCC and epxF_f GCGCACCATGTCGCTGTTG/epxF_r GTAGTCGGGTGTCTCCTCC (library ATCC 53904) as well as epnA_f GTGTGGCCGTGAGCGGATTC/epnA_r GCGGCCACGTTCCGATCTTG and epnK_f CAGCATGCTGCTGCAAGCCC/epnK_r CCCGGATGAAGTTCGACCGC (library ATCC 53709). The primers were designed to amplify small specific fragments (0.3−0.5 kb) from the borders of the respective gene clusters.

Heterologous Expression of the *epx* and the *epn* Gene Clusters. An *Xba*I restriction fragment from merLK01 was generated representing an integration cassette (int_neo) for stable chromosomal integration. int_neo was used to replace *cat* in fosmids 15C3 (*epx* cluster) and 2H4 (*epn* cluster), as described previously.[21] The resulting fosmids epxMS01 and epnLK01 were verified by restriction analysis. The fosmids were transferred into *E. coli* ET12567 and introduced into *S. albus* J1046 by triparental intergeneric conjugation with the help of *E. coli* ET12567/pUB307. Kanamycin resistance mutants were selected and designated as *S. albus* J1046/epxMS01-(1−3) and *S. albus* J1046/epnLK01-(1−3), respectively.

■ ASSOCIATED CONTENT

❺ Supporting Information

Supporting tables and graphics. This material is available free of charge via the Internet at http://pubs.acs.org.

Accession Codes

The complete sequences of the epoxomicin (27908 bp) and the eponemycin (23794 bp) biosynthetic gene clusters were deposited at NCBI under the accession numbers KF647219 and KF647220, respectively.

■ AUTHOR INFORMATION

Corresponding Authors

*E-mail: leonard.kaysser@pharm.uni-tuebingen.de.
*E-mail: bsmoore@ucsd.edu.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We thank Dr. R. Kersten for assistance with mass spectrometry analysis. We are grateful to Dr. A. Lechner and H. Sun for help with molecular networking and semiconductor sequencing, respectively. This work was supported by the National Institutes of Health (NIH) (CA127622 to B.S.M. and GM097509 to P.C.D. and B.S.M.), a Feodor Lynen postdoctoral fellowship from the Alexander von Humboldt Foundation to L.K., and an NIH instrument grant (S10-OD010640).

■ REFERENCES

(1) Ciechanover, A. (1994) The ubiquitin-proteasome proteolytic pathway. *Cell 79*, 13−21.

(2) Lopes, U. G., Erhardt, P., Yao, R., and Cooper, G. M. (1997) p53-dependent induction of apoptosis by proteasome inhibitors. *J. Biol. Chem. 272*, 12893−12896.

(3) Drexler, H. C. (1997) Activation of the cell death program by inhibition of proteasome function. *Proc. Natl. Acad. Sci. U.S.A. 94*, 855−860.

(4) Grawert, M. A., and Groll, M. (2012) Exploiting nature's rich source of proteasome inhibitors as starting points in drug development. *Chem. Commun. (Camb.) 48*, 1364−1378.

(5) Hanada, M., Sugawara, K., Kaneta, K., Toda, S., Nishiyama, Y., Tomita, K., Yamamoto, H., Konishi, M., and Oki, T. (1992) Epoxomicin, a new antitumor agent of microbial origin. *J. Antibiot. (Tokyo) 45*, 1746−1752.

(6) Sugawara, K., Hatori, M., Nishiyama, Y., Tomita, K., Kamei, H., Konishi, M., and Oki, T. (1990) Eponemycin, a new antibiotic active against B16 melanoma. I. Production, isolation, structure, and biological activity. *J. Antibiot. (Tokyo) 43*, 8−18.

(7) Pereira, A. R., Kale, A. J., Fenley, A. T., Byrum, T., Debonsi, H. M., Gilson, M. K., Valeriote, F. A., Moore, B. S., and Gerwick, W. H. (2012) The carmaphycins: New proteasome inhibitors exhibiting an α,β-epoxyketone warhead from a marine cyanobacterium. *Chembiochem 13*, 810−817.

(8) Koguchi, Y., Kohno, J., Suzuki, S., Nishio, M., Takahashi, K., Ohnuki, T., and Komatsubara, S. (2000) TMC-86A, B and TMC-96, new proteasome inhibitors from *Streptomyces sp.* TC 1084 and *Saccharothrix sp.* TC 1094. II. Physico-chemical properties and structure determination. *J. Antibiot. (Tokyo) 53*, 63−65.

(9) Koguchi, Y., Nishio, M., Suzuki, S., Takahashi, K., Ohnuki, T., and Komatsubara, S. (2000) TMC-89A and B, new proteasome inhibitors from *Streptomyces sp.* TC 1087. *J. Antibiot. (Tokyo) 53*, 967−972.

(10) Kuhn, D. J., Chen, Q., Voorhees, P. M., Strader, J. S., Shenk, K. D., Sun, C. M., Demo, S. D., Bennett, M. K., van Leeuwen, F. W., Chanan-Khan, A. A., and Orlowski, R. Z. (2007) Potent activity of carfilzomib, a novel, irreversible inhibitor of the ubiquitin-proteasome

pathway, against preclinical models of multiple myeloma. *Blood 110*, 3281−3290.

(11) McCormack, P. L. (2012) Carfilzomib: In relapsed, or relapsed and refractory, multiple myeloma. *Drugs 72*, 2023−2032.

(12) Curran, M. P., and McKeage, K. (2009) Bortezomib: A review of its use in patients with multiple myeloma. *Drugs 69*, 859−888.

(13) Glenn, R. J., Pemberton, A. J., Royle, H. J., Spackman, R. W., Smith, E., Jennifer Rivett, A., and Steverding, D. (2004) Trypanocidal effect of α',β'-epoxyketones indicates that trypanosomes are particularly sensitive to inhibitors of proteasome trypsin-like activity. *Int. J. Antimicrob. Agents 24*, 286−289.

(14) Czesny, B., Goshu, S., Cook, J. L., and Williamson, K. C. (2009) The proteasome inhibitor epoxomicin has potent *Plasmodium falciparum* gametocytocidal activity. *Antimicrob. Agents Chemother. 53*, 4080−4085.

(15) Groll, M., Kim, K. B., Kairies, N., Huber, R., and Crews, C. M. (2000) Crystal structure of epoxomicin: 20S proteasome reveals a molecular basis for selectivity of α',β'-epoxyketone proteasome inhibitors. *J. Am. Chem. Soc. 122*, 1237−1238.

(16) Rothberg, J. M., Hinz, W., Rearick, T. M., Schultz, J., Mileski, W., Davey, M., Leamon, J. H., Johnson, K., Milgrew, M. J., Edwards, M., Hoon, J., Simons, J. F., Marran, D., Myers, J. W., Davidson, J. F., Branting, A., Nobile, J. R., Puc, B. P., Light, D., Clark, T. A., Huber, M., Branciforte, J. T., Stoner, I. B., Cawley, S. E., Lyons, M., Fu, Y., Homer, N., Sedova, M., Miao, X., Reed, B., Sabina, J., Feierstein, E., Schorn, M., Alanjary, M., Dimalanta, E., Dressman, D., Kasinskas, R., Sokolsky, T., Fidanza, J. A., Namsaraev, E., McKernan, K. J., Williams, A., Roth, G. T., and Bustillo, J. (2011) An integrated semiconductor device enabling non-optical genome sequencing. *Nature 475*, 348−352.

(17) Ross, A. C., Xu, Y., Lu, L., Kersten, R. D., Shao, Z., Al-Suwailem, A. M., Dorrestein, P. C., Qian, P. Y., and Moore, B. S. (2013) Biosynthetic multitasking facilitates thalassospiramide structural diversity in marine bacteria. *J. Am. Chem. Soc. 135*, 1155−1162.

(18) Henke, W., Herdel, K., Jung, K., Schnorr, D., and Loening, S. A. (1997) Betaine improves the PCR amplification of GC-rich DNA sequences. *Nucleic Acids Res. 25*, 3957−3958.

(19) Medema, M. H., Blin, K., Cimermancic, P., de Jager, V., Zakrzewski, P., Fischbach, M. A., 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 Res. 39*, 339−346.

(20) Kale, A. J., McGlinchey, R. P., Lechner, A., and Moore, B. S. (2011) Bacterial self-resistance to the natural proteasome inhibitor salinosporamide A. *ACS Chem. Biol. 6*, 1257−1264.

(21) Kaysser, L., Bernhardt, P., Nam, S. J., Loesgen, S., Ruby, J. G., Skewes-Cox, P., Jensen, P. R., Fenical, W., and Moore, B. S. (2012) Merochlorins A−D, cyclic meroterpenoid antibiotics biosynthesized in divergent pathways with vanadium-dependent chloroperoxidases. *J. Am. Chem. Soc. 134*, 11988−11991.

(22) Bierman, M., Logan, R., O'Brien, K., Seno, E. T., Rao, R. N., and Schoner, B. E. (1992) Plasmid cloning vectors for the conjugal transfer of DNA from *Escherichia coli* to *Streptomyces spp. Gene 116*, 43−49.

(23) Labeda, D. P., and Kroppenstedt, R. M. (2006) *Goodfellowia gen. nov.*, a new genus of the Pseudonocardineae related to Actinoalloteichus, containing *Goodfellowia coeruleoviolacea gen. nov.*, comb. nov. *Int. J. Syst. Evol. Microbiol. 56*, 1203−1207.

(24) Flett, F., Mersinias, V., and Smith, C. P. (1997) High efficiency intergeneric conjugal transfer of plasmid DNA from *Escherichia coli* to methyl DNA-restricting streptomycetes. *FEMS Microbiol. Lett. 155*, 223−229.

(25) Lombo, F., Velasco, A., Castro, A., de la Calle, F., Brana, A. F., Sanchez-Puelles, J. M., Mendez, C., and Salas, J. A. (2006) Deciphering the biosynthesis pathway of the antitumor thiocoraline from a marine actinomycete and its expression in two streptomyces species. *ChemBioChem 7*, 366−376.

(26) Bandeira, N., Tsur, D., Frank, A., and Pevzner, P. A. (2007) Protein identification by spectral networks analysis. *Proc. Natl. Acad. Sci. U.S.A. 104*, 6140−6145.

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

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

(29) Nguyen, D. D., Wu, C. H., Moree, W. J., Lamsa, A., Medema, M. H., Zhao, X., Gavilan, R. G., Aparicio, M., Atencio, L., Jackson, C., Ballesteros, J., Sanchez, J., Watrous, J. D., Phelan, V. V., van de Wiel, C., Kersten, R. D., Mehnaz, S., De Mot, R., Shank, E. A., Charusanti, P., Nagarajan, H., Duggan, B. M., Moore, B. S., Bandeira, N., Palsson, B. O., Pogliano, K., Gutierrez, M., and Dorrestein, P. C. (2013) MS/MS networking guided analysis of molecule and gene cluster families. *Proc. Natl. Acad. Sci. U.S.A. 110*, 2611−2620.

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

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

(32) Rausch, C., Hoof, I., Weber, T., Wohlleben, W., and Huson, D. H. (2007) Phylogenetic analysis of condensation domains in NRPS sheds light on their functional evolution. *BMC Evol. Biol. 7*, 78.

(33) Kraas, F. I., Giessen, T. W., and Marahiel, M. A. (2012) Exploring the mechanism of lipid transfer during biosynthesis of the acidic lipopeptide antibiotic CDA. *FEBS Lett. 586*, 283−288.

(34) Wilkinson, S., and Lowe, L. A. (1964) Structure of polymyxin B1. *Nature 202*, 1211.

(35) Powell, A., Borg, M., Amir-Heidari, B., Neary, J. M., Thirlway, J., Wilkinson, B., Smith, C. P., and Micklefield, J. (2007) Engineered biosynthesis of nonribosomal lipopeptides with modified fatty acid side chains. *J. Am. Chem. Soc. 129*, 15182−15191.

(36) Revill, W. P., Bibb, M. J., Scheu, A. K., Kieser, H. J., and Hopwood, D. A. (2001) β-ketoacyl acyl carrier protein synthase III (FabH) is essential for fatty acid biosynthesis in *Streptomyces coelicolor* A3(2). *J. Bacteriol. 183*, 3526−3530.

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

(38) Thibodeaux, C. J., Chang, W. C., and Liu, H. W. (2012) Enzymatic chemistry of cyclopropane, epoxide, and aziridine biosynthesis. *Chem. Rev. 112*, 1681−1709.

(39) Read, J. A., and Walsh, C. T. (2007) The lyngbyatoxin biosynthetic assembly line: Chain release by four-electron reduction of a dipeptidyl thioester to the corresponding alcohol. *J. Am. Chem. Soc. 129*, 15762−15763.

(40) Hofvander, P., Doan, T. T., and Hamberg, M. (2011) A prokaryotic acyl-CoA reductase performing reduction of fatty acyl-CoA to fatty alcohol. *FEBS Lett. 585*, 3538−3543.

(41) Gulder, T. A., Freeman, M. F., and Piel, J. (2011) The catalytic diversity of multimodular polyketide synthases: Natural product biosynthesis beyond textbook assembly rules. *Top. Curr. Chem.*, DOI: 10.1007/128_2010_113.

(42) Stein, M. L., Beck, P., Kaiser, M., Dudler, R., Becker, C. F., and Groll, M. (2012) One-shot NMR analysis of microbial secretions identifies highly potent proteasome inhibitor. *Proc. Natl. Acad. Sci. U.S.A. 109*, 18367−18371.

(43) Krahn, D., Ottmann, C., and Kaiser, M. (2011) The chemistry and biology of syringolins, glidobactins, and cepafungins (syrbactins). *Nat. Prod. Rep. 28*, 1854−1867.

(44) Sambrook, J., and Russell, D. W. (2001) *Molecular Cloning. A Laboratory Manual*, Cold Spring Harbor Laboratory Press, New York.

(45) Kieser, T., Bibb, M., Buttner, M., Chater, K., and Hopwood, D. (2000) *Practical Streptomyces Genetics*, The John Innes Foundation, Norwich, U.K.

(46) Patiny, L., and Borel, A. (2013) ChemCalc: A building block for tomorrow's chemical infrastructure. *J. Chem. Inf. Model* 53, 1223–1228.

(47) Frank, A. M., Bandeira, N., Shen, Z., Tanner, S., Briggs, S. P., Smith, R. D., and Pevzner, P. A. (2008) Clustering millions of tandem mass spectra. *J. Proteome Res.* 7, 113–122.

(48) Smoot, M. E., Ono, K., Ruscheinski, J., Wang, P. L., and Ideker, T. (2011) Cytoscape 2.8: New features for data integration and network visualization. *Bioinformatics* 27, 431–432.

(49) Bibb, M. J., Findlay, P. R., and Johnson, M. W. (1984) The relationship between base composition and codon usage in bacterial genes and its use for the simple and reliable identification of protein-coding sequences. *Gene* 30, 157–166.

(50) Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402.

Supporting Information

The genetic basis for the biosynthesis of the pharmaceutically important class of epoxyketone proteasome inhibitors

Michelle Schorn,[†] Judith Zettler[§,⊥], Joseph P. Noel,[‡] Pieter C. Dorrestein,[¶] Bradley S. Moore[*,†,¶] and Leonard Kaysser[*,†,§,⊥]

[†] Scripps Institution of Oceanography, University of California, San Diego, CA 92093, [‡] Jack H. Skirball Center for Chemical Biology and Proteomics, Salk Institute for Biological Studies, La Jolla, CA 92037, [¶] Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California, San Diego, CA 92093. [§]Pharmaceutical Biology, Eberhard Karls University Tübingen, 72076 Germany. [⊥]German Center for Infection Research (DZIF), partner site Tübingen, 72076 Tübingen, Germany.

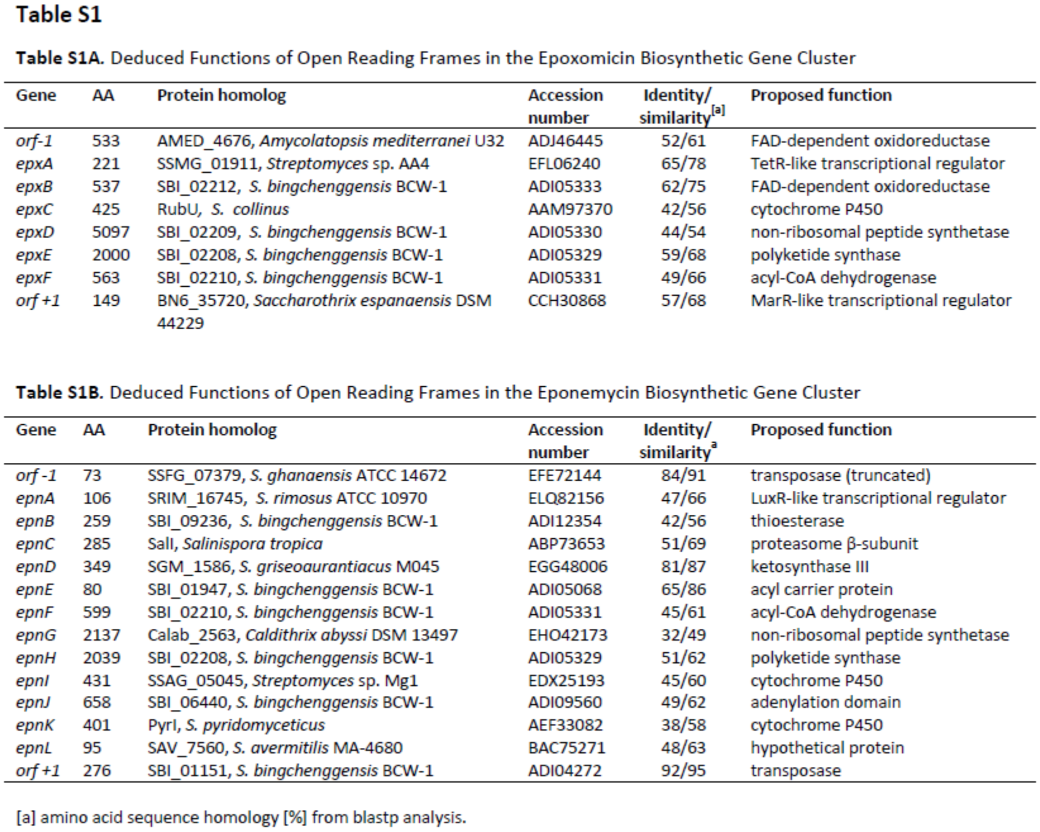


Table S1

Table S1A. Deduced Functions of Open Reading Frames in the Epoxomicin Biosynthetic Gene Cluster

Gene	AA	Protein homolog	Accession number	Identity/ similarity[a]	Proposed function
orf-1	533	AMED_4676, *Amycolatopsis mediterranei* U32	ADJ46445	52/61	FAD-dependent oxidoreductase
epxA	221	SSMG_01911, *Streptomyces* sp. AA4	EFL06240	65/78	TetR-like transcriptional regulator
epxB	537	SBI_02212, *S. bingchenggensis* BCW-1	ADI05333	62/75	FAD-dependent oxidoreductase
epxC	425	RubU, *S. collinus*	AAM97370	42/56	cytochrome P450
epxD	5097	SBI_02209, *S. bingchenggensis* BCW-1	ADI05330	44/54	non-ribosomal peptide synthetase
epxE	2000	SBI_02208, *S. bingchenggensis* BCW-1	ADI05329	59/68	polyketide synthase
epxF	563	SBI_02210, *S. bingchenggensis* BCW-1	ADI05331	49/66	acyl-CoA dehydrogenase
orf +1	149	BN6_35720, *Saccharothrix espanaensis* DSM 44229	CCH30868	57/68	MarR-like transcriptional regulator

Table S1B. Deduced Functions of Open Reading Frames in the Eponemycin Biosynthetic Gene Cluster

Gene	AA	Protein homolog	Accession number	Identity/ similarity[a]	Proposed function
orf -1	73	SSFG_07379, *S. ghanaensis* ATCC 14672	EFE72144	84/91	transposase (truncated)
epnA	106	SRIM_16745, *S. rimosus* ATCC 10970	ELQ82156	47/66	LuxR-like transcriptional regulator
epnB	259	SBI_09236, *S. bingchenggensis* BCW-1	ADI12354	42/56	thioesterase
epnC	285	SalI, *Salinispora tropica*	ABP73653	51/69	proteasome β-subunit
epnD	349	SGM_1586, *S. griseoaurantiacus* M045	EGG48006	81/87	ketosynthase III
epnE	80	SBI_01947, *S. bingchenggensis* BCW-1	ADI05068	65/86	acyl carrier protein
epnF	599	SBI_02210, *S. bingchenggensis* BCW-1	ADI05331	45/61	acyl-CoA dehydrogenase
epnG	2137	Calab_2563, *Caldithrix abyssi* DSM 13497	EHO42173	32/49	non-ribosomal peptide synthetase
epnH	2039	SBI_02208, *S. bingchenggensis* BCW-1	ADI05329	51/62	polyketide synthase
epnI	431	SSAG_05045, *Streptomyces* sp. Mg1	EDX25193	45/60	cytochrome P450
epnJ	658	SBI_06440, *S. bingchenggensis* BCW-1	ADI09560	49/62	adenylation domain
epnK	401	PyrI, *S. pyridomyceticus*	AEF33082	38/58	cytochrome P450
epnL	95	SAV_7560, *S. avermitilis* MA-4680	BAC75271	48/63	hypothetical protein
orf +1	276	SBI_01151, *S. bingchenggensis* BCW-1	ADI04272	92/95	transposase

[a] amino acid sequence homology [%] from blastp analysis.

Table S2

Table S2. Putative compounds corresponding to unique mass peaks in the ion chromatograms of *S. albus* epxMS01 and epnLK01 mutant strains.

unique mass peaks at Rt[a]	corresponding ions	putative properties (predicted structure)
S. albus epxMS01		
25.0 min	*m/z* 541.3976 [M+H]$^+$ *m/z* 563.3798 [M+Na]$^+$ *m/z* 1103.7680 [2M+Na]$^+$	epoxomicin derivative
S. albus epnLK01		
13.6 min	*m/z* 343.1874 [M+H]$^+$ *m/z* 325.1777 [M-H$_2$0]$^+$ *m/z* 365.1701 [M+Na]$^+$ *m/z* 707.3503 [2M+Na]$^+$	eponemycin derivative (see Figure S2, A)
14.0 min	*m/z* 415.2449 [M+H]$^+$ *m/z* 397.2344 [M-H$_2$0]$^+$ *m/z* 437.2283 [M+Na]$^+$ *m/z* 851.4651 [2M+Na]$^+$	eponemycin derivative (see Figure S2, C)
14.8 min	*m/z* 413.2297 [M+H]$^+$ *m/z* 435.2124 [M+Na]$^+$	eponemycin derivative (see Figure S2, G)
15.2 min	*m/z* 411.2505 [M-H$_2$0]$^+$ *m/z* 451.2431 [M+Na]$^+$	eponemycin derivative (see Figure S2, F)
15.6 min	*m/z* 429.2604 [M+H]$^+$ *m/z* 451.2433 [M+Na]$^+$	eponemycin derivative (see Figure S2, L)
16.3 min	*m/z* 427.2446 [M+H]$^+$ *m/z* 449.2273 [M+Na]$^+$	eponemycin derivative (see Figure S2, K)
16.9 min	*m/z* 429.2602 [M+H]$^+$ *m/z* 451.2437 [M+Na]$^+$	eponemycin derivative (see Figure S2, M)
20.6 min	*m/z* 403.2818 [M+H]$^+$	tetrahydroeponemycin (see Figure S2, E)
	m/z 419.2533 [M+H]$^+$	unknown compound
	m/z 491.8402 [M+H]$^+$	unknown compound
22.3 min	*m/z* 401.2658 [M+H]$^+$ *m/z* 423.2486 [M+Na]$^+$	dihydroeponemycin (see Figure S2, D)
26.0 min	*m/z* 357.2760 [M+H]$^+$ *m/z* 379.2587 [M+Na]$^+$ *m/z* 735.5269 [2M+Na]$^+$	eponemycin derivative
27.2 min	*m/z* 515.0984 [M+H]$^+$	unknown compound
30.5 min	*m/z* 479.2039 [M+H]$^+$	unknown compound
	m/z 501.1868 [M+H]$^+$	unknown compound

[a] retention time

Table S3

Table S3. Analysis of NRPS A-domains from the epoxomicin and eponemycin gene clusters and the cluster of the unknown proteasome inhibitor from *S. bingchenggensis* BCW-1: Predicted substrate specificities*(1)* and postulated functions.

Protein	A-domains	Specificity signature	Predicted substrate [a]
epx cluster			
EpxD	A 1	D A Y F W G V T F K	valine/leucine/isoleucine
	A 2	D F W S T G V I L K	threonine/valine [b]
	A 3	D F W N I G M V H K	threonine
	A 4	D L L H L G L I L K	glycine/hydroxyphenylglycine/valine [b]
epn cluster			
EpnG	A 1	D V W S I A M V H K	serine/threonine
	A 2	D A W Y L A N V V K	valine/leucine
EpnJ	A 1	D A W L A G L T A K	leucine
sbi cluster			
SBI_02209	A 1	D A Y W W G G T F K	valine/leucine
	A 2	D F W N I G M V H K	threonine
	A 3	D A S V V G C V T K	lysine [b]

[a] substrate specificities as predicted by comparison of the 10 amino acid code defined by Stachelhaus et al.[1]

[b] predictions with low confidence (≤ 70% identity)

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

Figure S1

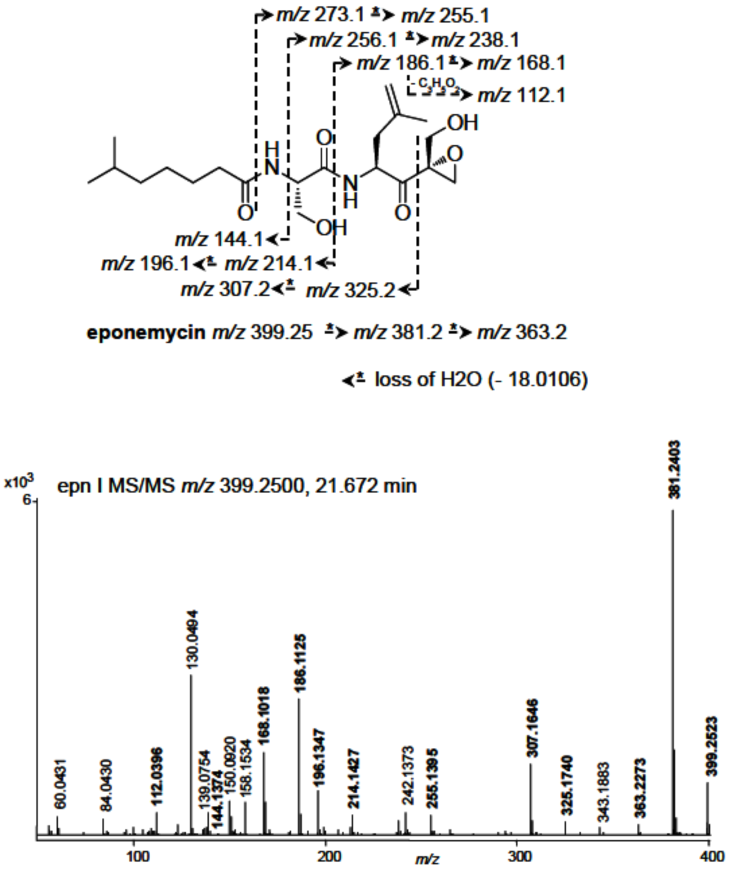
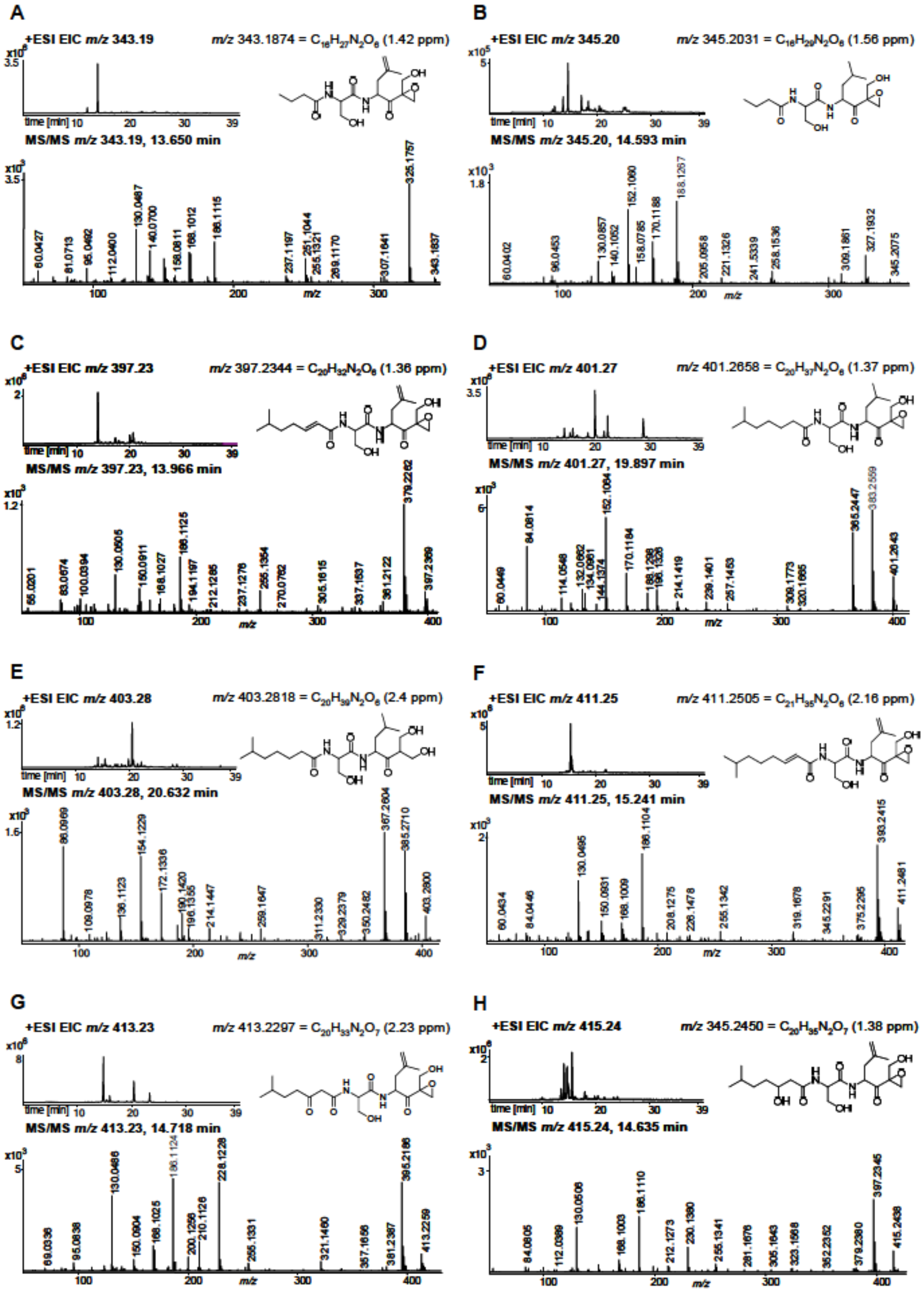


Figure S1. Proposed fragmentation scheme and MS/MS spectrum of eponemycin as produced in *S. albus* J0146/epnLK01. Assigned fragments are indicated as bold.

Figure S2


A
+ESI EIC m/z 343.19
m/z 343.1874 = $C_{16}H_{27}N_2O_6$ (1.42 ppm)
x10^6 3.5
time [min] 10 20 30 39
MS/MS m/z 343.19, 13.650 min
x10^3 3.5
60.0427 81.0713 95.0492 112.0400 130.0487 140.0700 158.0811 168.1012 186.1115 237.1197 251.1044 255.1321 269.1170 307.1641 325.1757 343.1837
100 200 m/z 300
OH O H N
B
+ESI EIC m/z 345.20
m/z 345.2031 = $C_{16}H_{29}N_2O_6$ (1.56 ppm)
x10^5 5
time [min] 10 20 30 39
MS/MS m/z 345.20, 14.593 min
x10^3 1.8
60.0402 96.0453 130.0857 140.1052 152.1060 158.0785 170.1188 188.1267 205.0958 221.1326 241.5339 258.1536 309.1861 327.1932 345.2075
100 200 m/z 300
C
+ESI EIC m/z 397.23
m/z 397.2344 = $C_{20}H_{32}N_2O_6$ (1.36 ppm)
x10^6 2
time [min] 10 20 30 39
MS/MS m/z 397.23, 13.966 min
x10^3 1.2
55.0201 83.0674 100.0394 130.0505 150.0911 168.1027 186.1125 194.1197 212.1285 237.1276 255.1354 270.0762 305.1615 337.1537 361.2122 379.2262 397.2369
100 200 m/z 300 400
D
+ESI EIC m/z 401.27
m/z 401.2658 = $C_{20}H_{37}N_2O_6$ (1.37 ppm)
x10^6 3.5
time [min] 10 20 30 39
MS/MS m/z 401.27, 19.897 min
x10^3 6
60.0449 84.0814 114.0548 132.0662 134.0961 144.1374 152.1064 170.1184 188.1288 198.1326 214.1419 239.1401 257.1453 309.1773 320.1665 365.2447 383.2559 401.2643
100 200 m/z 300 400
E
+ESI EIC m/z 403.28
m/z 403.2818 = $C_{20}H_{39}N_2O_6$ (2.4 ppm)
x10^6 1.2
time [min] 10 20 30 39
MS/MS m/z 403.28, 20.632 min
x10^3 1.6
86.0969 109.0978 136.1123 154.1229 172.1336 190.1420 196.1355 214.1447 259.1647 311.2330 329.2379 350.2482 367.2604 385.2710 403.2800
100 200 m/z 300 400
F
+ESI EIC m/z 411.25
m/z 411.2505 = $C_{21}H_{35}N_2O_6$ (2.16 ppm)
x10^6 5
time [min] 10 20 30 39
MS/MS m/z 411.25, 15.241 min
x10^3 2
60.0434 84.0446 130.0495 150.0931 168.1009 186.1104 208.1275 226.1478 255.1342 319.1678 345.2291 375.2295 393.2415 411.2481
100 200 m/z 300 400
G
+ESI EIC m/z 413.23
m/z 413.2297 = $C_{20}H_{33}N_2O_7$ (2.23 ppm)
x10^6 8
time [min] 10 20 30 39
MS/MS m/z 413.23, 14.718 min
x10^3 5
69.0336 95.0838 130.0486 150.0904 168.1025 186.1124 200.1256 210.1126 228.1228 255.1331 321.1460 357.1656 381.2387 395.2186 413.2259
100 200 m/z 300 400
H
+ESI EIC m/z 415.24
m/z 345.2450 = $C_{20}H_{35}N_2O_7$ (1.38 ppm)
x10^6 2
time [min] 10 20 30 39
MS/MS m/z 415.24, 14.635 min
x10^3 3
84.0805 112.0389 130.0506 168.1003 186.1110 212.1273 230.1380 255.1341 281.1676 305.1643 323.1568 352.2352 379.2380 397.2345 415.2438
100 200 m/z 300 400

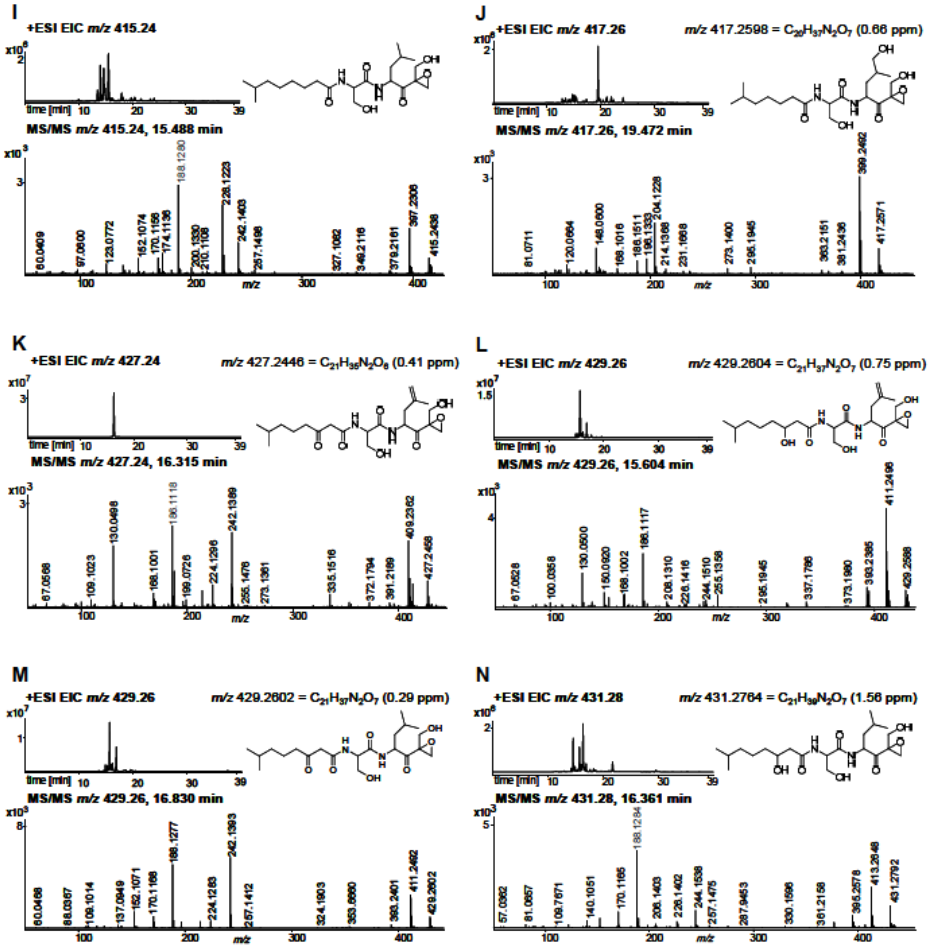


Figure S2. MS analysis of putative eponemycin derivatives produced by *S. albus* J1046 as found by molecular networking. MS/MS spectra, extracted ion chromatograms (EICs) and monoisotopic masses are shown. Plausible structures for the eponemycin analogues deduced from the MS data are depicted.

Figure S3

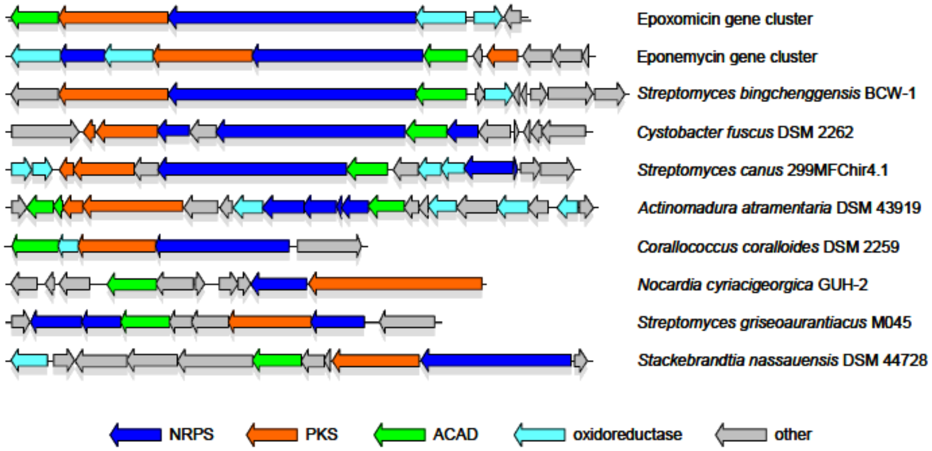

Epoxomicin gene cluster
Eponemycin gene cluster
Streptomyces bingchenggensis BCW-1
Cystobacter fuscus DSM 2262
Streptomyces canus 299MFChir4.1
Actinomadura atramentaria DSM 43919
Corallococcus coralloides DSM 2259
Nocardia cyriacigeorgica GUH-2
Streptomyces griseoaurantiacus M045
Stackebrandtia nassauensis DSM 44728
NRPS
PKS
ACAD
oxidoreductase
other

Figure S3. NRPS/PKS hybrid gene clusters from various bacteria containing an EpxF (ACAD) homolog.

2.4 Acknowledgements

This publication resulted from a collaboration with Dr. Leonard Kaysser who wrote the majority of the paper, planned the experiments with Dr. Bradley Moore, and carried out some of the lab work. Michelle A. Schorn wrote portions of the paper and performed the majority of the lab work. Judith Zettler curated and submitted the genomic information to NCBI. Dr. Joseph Noel provided the Ion Torrent PGM machine used to sequence the genomes. Dr. Pieter Dorrestein provided the LC/MS instruments used for molecular networking. We would also like to acknowledge Roland Kersten for assistance with mass spectrometry analysis, Anna Lechner for help with molecular networking, and Helena Sun for help sequencing. This work was supported by the National Institutes of Health (NIH) (CA127622 to B.S.M. and GM097509 to P.C.D. and B.S.M.), a Feodor Lynen postdoctoral fellowship from the Alexander von Humboldt Foundation to L.K., and an NIH instrument grant (S10-OD010640).

Chapter 2, in full, is a reprint of materials as it appears in “Genetic Basis for the Biosynthesis of the Pharmaceutically Important Class of Epoxyketone Proteasome Inhibitors” in *ACS Chemical Biology*, 2014, Schorn, M., Zettler, J., Noel, J. P., Dorrestein, P. C., Moore, B. S. and Kaysser, L. The dissertation author was one of two equally contributing primary investigators and authors of this manuscript.

Chapter 3: Sequencing Rare Marine Actinomycete Genomes Reveals High Density of Unique Natural Product Biosynthetic Gene Clusters

3.1 Introduction to Chapter 3

The scientific community is currently poised at an exciting inflection point with the escalating “democratization of sequencing.” Just a few years ago, whole genome sequencing was used sparingly as it was extremely expensive and labor intensive, and thus outsourced. Now, genome sequencing, especially for microbial genomes, is rapid, affordable and can take place in the lab[1]. Scientists are finding new ways to utilize the growing abundance of genomic DNA sequences. Bacterial whole genome sequencing is being developed as a new tool to discover promising novel chemicals from marine microbes by identifying specific secondary metabolic gene clusters using genome mining and expressing them in heterologous hosts[2].

The conventional path to discovering new compounds from microbes is lengthy and work intensive. Culturing the microbes can take weeks and a large percentage of the microbes from environmental samples cannot be cultured. The bioassays to detect bioactivity may be incomplete in the scope of natural product activities and can be time intensive. The downstream extraction, fractionation, and elucidation may also miss molecules not produced in a laboratory setting. This new era of fast, easy sequencing and synthetic biology can open doors to exploring microbial communities, even those uncultivatable, harboring unprecedented natural chemistry.

The actinobacteria (referred to here as actinomycetes) represents a diverse phylum of Gram-positive bacteria with a high GC DNA content, capable of immense secondary metabolic capacity. They are renowned for producing biomedically important molecules, from antibiotics to anti-cancer compounds[3]. Actinomycetes produce about two-thirds of known antibiotics, with the majority discovered from the genus *Streptomyces*[4]. Many of the more widely known prolific genera, such as *Streptomyces*, *Mycobacterium*, and *Salinispora* have been studied in depth with hundreds genomes sequenced[5]. *Salinispora* is arguably the most prolific marine actinomycete genus discovered to date, producing a suite of potent bioactive chemicals[6]. Lesser known marine actinobacteria have been discovered, however, their genomes remain largely unexplored. In their 2010 paper, Gontang and coworkers probed a variety of marine sediment actinomycetes, from the popular *Salinispora* and *Streptomyces* genera to unexploited genera such as *Marmoricola* and *Serinicoccus*, for biosynthetic pathways indicative of secondary metabolite production[7]. The majority of the microbes assayed contained either or both polyketide synthase (PKS) and non-ribosomal peptide synthetase (NRPS) genes, suggesting the presence of biosynthetic machinery for making yet unknown novel chemicals. The so called rare actinomycetes, defined by Bérdy as non-*Streptomyces* actinomycetes, represent a group worthy of further exploration through genome sequencing and genome mining[8].

At the time of this project, Ion Torrent boasted the fastest and most economical sequencer on the market, as well as long read lengths of over 400 base pairs[9,10]. With access to such technology, it is possible to generate a draft genome of a bacterium within a few days. Actinobacteria are characterized as having high G + C content in their DNA, which is often hard to amplify without bias, and thus provide a problem for next-generation sequencing. In this project, I used the optimized sample preparation for high G + C DNA that I developed in the previous chapter, sequencing the proteasome inhibitor producers. A TAR cloning vector has been developed in the Moore lab specifically for expression of large actinomycete gene clusters[11]. This method is faster and more targeted than the traditional fosmid library construction method, which relies on random DNA shearing and probability of an intact cluster being located within one fosmid. The intersection of benchtop sequencing and genetic engineering provides one method to go from genomic sequence to chemical product, bypassing traditional selective processes.

The prospect of leaping from genomic code to molecule is fast approaching, with technological advances in sequencing, bioinformatics and heterologous gene expression paving the way to discover and manufacture unknown molecules from proposed biosynthetic pathways. Rare marine actinobacteria are the perfect candidates for exploration, as their relatives are historically rich in secondary metabolites and are relatively unstudied.

This project has been published in the journal *Microbiology* in December 2016 Issue 12, Volume 162 on pages 2075-2086[12]. It was chosen as the Editor's Choice for the December 2016 issue. This work was planned and conceptualized by myself, Dr. Brad Moore, and Dr. Tommie Lincecum. I first selected 22 strains of rare marine actinomycetes from the Fenical/Jensen collection at the Scripps Institution of Oceanography. Five of the strains were isolated, described, and provided by Nastassia Patin, a member of Dr. Paul Jensen's lab[13]. I grew each strain and isolated high molecular weight gDNA by using the Qiagen Genomic Tip Kit. I then prepared Ion Torrent 400bp sequencing libraries. Dr. Lincecum and Kristen Aguinaldo of Ion Torrent helped to optimize a thermoprofile for these difficult to sequence genomes.

I originally assembled the genomes using CLC Genomics Workbench, as was done in Chapter 2. I was able to get basic assembly statistics and biosynthetic gene cluster completeness from these original assemblies, as well as determined which genomes needed re-sequencing. Subsequently, I established a collaboration with Dr. Anton Korobeynikov from

Dr. Pavel Pevzner's lab, who was developing an improved assembly algorithm for Ion Torrent data, SPAdesIT[14,15]. Anton re-assembled all my genomes using SPAdesIT before the assembly program was released to the public. Dr. Sheila Podell performed the genome completeness assessment of the final assemblies to show that these genomes were indeed complete and of high enough quality for subsequent analysis. Mohammad Alanjary, of Dr. Nadine Ziemert's lab, and I planned the bioinformatic analyses of biosynthetic gene clusters in all genomes. I constructed the Circos[16] figure showing the variety and distributions of pathways within the rare marine actinomycetes I sequenced. Mohammad performed the biosynthetic gene cluster networking using a combination of custom scripts and published methods[17]. We then analyzed the networks and summarized the findings. Mohammad and I also built the 16S phylogenetic tree, placing each strain within the larger context of the actinomycetes. Dr. Ziemert provided guidance and suggestions for phylogenetic and bioinformatic analyses and Dr. Jensen provided helpful advice about growing marine actinomycetes and insights into their genomes.

This highly collaborative project allowed us to reveal, for the first time, the secondary metabolite potential of a sliver of the rare marine actinomycetes isolated by the Jensen and Fenical labs. We were able to show that these taxa are underrepresented in our current sequencing databases, and they yield a high density of previously un-sequenced, possibly novel biosynthetic gene clusters. This work provides the genetic foundation and motivation to pursue underexplored marine actinomycetes as a source for novel compounds.

The genome mining I performed in this project led me to pursue two biosynthetic gene clusters from two of the genomes from this study. I chose two gene clusters from two different *Nocardia* isolates: strains CNB044 and CNY236. Both clusters were successfully captured using Transformation Associated Recombination (TAR), but downstream integration and expression of the clusters proved unyielding. Despite attempts to manipulate regulatory elements present in one cluster, no production of new compounds was ever detected. The results of these attempts, and the insights they provided to problems facing activation of silent gene clusters are discussed in the Chapter 3 Appendix.

Section 3.3 is a reprint of material as it appears in “Sequencing Rare Marine Actinomycete Genomes Reveals High Density of Unique Natural Product Biosynthetic Gene Clusters” in *Microbiology*, 2016. Schorn, M. A., Alanjary, M. M., Aguinaldo, K., Korobeynikov, A., Podell, S., Patin, N., Lincecum, T., Jensen, P. R., Ziemert, N. & Moore, B. S.

3.2 Chapter 3 Introduction References

1 Shendure, J., Balasubramanian, S., Church, G. M., Gilbert, W., Rogers, J., Schloss, J. A. & Waterston, R. H. DNA sequencing at 40: past, present and future. *Nature* **550**, 345, 2017.

2 Ziemert, N., Alanjary, M. & Weber, T. The evolution of genome mining in microbes - a review. *Nat Prod Rep* **33**, 988-1005, 2016. PMID: 27272205.

3 Barka, E. A., Vatsa, P., Sanchez, L., Gaveau-Vaillant, N., Jacquard, C., Klenk, H. P., Clement, C., Ouhdouch, Y. & van Wezel, G. P. Taxonomy, Physiology, and Natural Products of Actinobacteria. *Microbiol Mol Biol Rev* **80**, 1-43, 2016. PMID: 26609051.

4 Newman, D. J. & Cragg, G. M. Natural Products as Sources of New Drugs from 1981 to 2014. *J Nat Prod* **79**, 629-661, 2016. PMID: 26852623.

5 Gomez-Escribano, J. P., Alt, S. & Bibb, M. J. Next generation sequencing of Actinobacteria for the discovery of novel natural products. *Mar Drugs* **14**, 2016. PMID: 27089350.

6 Jensen, P. R., Moore, B. S. & Fenical, W. The marine actinomycete genus Salinispora: a model organism for secondary metabolite discovery. *Nat Prod Rep* **32**, 738-751, 2015. PMID: 25730728.

7 Gontang, E., Gaudencio, S., Fenical, W. & Jensen, P. Sequence-based analysis of secondary metabolite biosynthesis in marine Actinobacteria. *Appl Environ Microbiol* **76**, 2487-2499, 2010.

8 Berdy, J. Bioactive microbial metabolites. *J Antibiot (Tokyo)* **58**, 1-26, 2005. PMID: 15813176.

9 Rothberg, J. M., Hinz, W., Rearick, T. M., Schultz, J., Mileski, W., Davey, M., Leamon, J. H., Johnson, K., Milgrew, M. J., Edwards, M., Hoon, J., Simons, J. F., Marran, D., Myers, J. W., Davidson, J. F., Branting, A., Nobile, J. R., Puc, B. P., Light, D., Clark, T. A., Huber, M., Branciforte, J. T., Stoner, I. B., Cawley, S. E., Lyons, M., Fu, Y., Homer, N., Sedova, M., Miao, X., Reed, B., Sabina, J., Feierstein, E., Schorn, M., Alanjary, M., Dimalanta, E., Dressman, D., Kasinskas, R., Sokolsky, T., Fidanza, J. A., Namsaraev, E., McKernan, K. J., Williams, A., Roth, G. T. & Bustillo, J. An integrated semiconductor device enabling non-optical genome sequencing. *Nature* **475**, 348-352, 2011.

10 Merriman, B. & Rothberg, J. M. Progress in ion torrent semiconductor chip based sequencing. *Electrophoresis* **33**, 3397-3417, 2012. PMID: 23208921.

11 Yamanaka, K., Reynolds, K. A., Kersten, R. D., Ryan, K. S., Gonzalez, D. J., Nizet, V., Dorrestein, P. C. & Moore, B. S. Direct cloning and refactoring of a silent lipopeptide biosynthetic gene cluster yields the antibiotic taromycin A. *Proc Natl Acad Sci U S A* **111**, 1957-1962, 2014. PMID: 24449899.

12 Schorn, M. A., Alanjary, M. M., Aguinaldo, K., Korobeynikov, A., Podell, S., Patin, N., Lincecum, T., Jensen, P. R., Ziemert, N. & Moore, B. S. Sequencing rare marine actinomycete genomes reveals high density of unique natural product biosynthetic gene clusters. *Microbiology* **162**, 2075-2086, 2016. PMID: 27902408.

13 Patin, N. V., Duncan, K. R., Dorrestein, P. C. & Jensen, P. R. Competitive strategies differentiate closely related species of marine actinobacteria. *Isme j* **10**, 478-490, 2016. PMID: 26241505.

14 Antipov, D., Korobeynikov, A., McLean, J. S. & Pevzner, P. A. hybridSPAdes: an algorithm for hybrid assembly of short and long reads. *Bioinformatics* **32**, 1009-1015, 2016. PMID: 26589280.

15 Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko, S. I., Pham, S., Prjibelski, A. D., Pyshkin, A. V., Sirotkin, A. V., Vyahhi, N., Tesler, G., Alekseyev, M. A. & Pevzner, P. A. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* **19**, 455-477, 2012. PMID: 22506599.

16 Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S. J. & Marra, M. A. Circos: an information aesthetic for comparative genomics. *Genome Res* **19**, 1639-1645, 2009. PMID: 19541911.

17 Cimermancic, P., Medema, M. H., Claesen, J., Kurita, K., Wieland Brown, L. C., Mavrommatis, K., Pati, A., Godfrey, P. A., Koehrsen, M., Clardy, J., Birren, B. W., Takano, E., Sali, A., Linington, R. G. & Fischbach, M. A. Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters. *Cell* **158**, 412-421, 2014. PMID: 25036635.

3.3 Reprint of: “Sequencing Rare Marine Actinomycete Genomes Reveals High Density of Unique Natural Product Biosynthetic Gene Clusters”

Microbiology (2016), **162**, 2075–2086 **DOI** 10.1099/mic.0.000386

Editor’s Choice

Sequencing rare marine actinomycete genomes reveals high density of unique natural product biosynthetic gene clusters

Michelle A. Schorn,[1] Mohammad M. Alanjary,[2] Kristen Aguinaldo,[3] Anton Korobeynikov,[4,5] Sheila Podell,[1] Nastassia Patin,[1] Tommie Lincecum,[3] Paul R. Jensen,[1,6] Nadine Ziemert[2] and Bradley S. Moore[1,6,7]

Correspondence
Bradley S. Moore
bsmoore@ucsd.edu

[1]Center for Marine Biotechnology and Biomedicine, Scripps Institution of Oceanography, University of California, San Diego, USA

[2]German Centre for Infection Research (DZIF), Interfaculty Institute for Microbiology and Infection Medicine Tuebingen (IMIT), University of Tuebingen, Tuebingen, Germany

[3]Thermo Fisher Scientific, Carlsbad, CA, USA

[4]Center for Algorithmic Biotechnology, St. Petersburg State University, St. Petersburg, Russia

[5]Department of Statistical Modeling, St. Petersburg State University, St. Petersburg, Russia

[6]Center for Microbiome Innovation, University of California, San Diego, USA

[7]Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California, San Diego, USA

Received 16 August 2016
Accepted 25 October 2016

Traditional natural product discovery methods have nearly exhausted the accessible diversity of microbial chemicals, making new sources and techniques paramount in the search for new molecules. Marine actinomycete bacteria have recently come into the spotlight as fruitful producers of structurally diverse secondary metabolites, and remain relatively untapped. In this study, we sequenced 21 marine-derived actinomycete strains, rarely studied for their secondary metabolite potential and under-represented in current genomic databases. We found that genome size and phylogeny were good predictors of biosynthetic gene cluster diversity, with larger genomes rivalling the well-known marine producers in the *Streptomyces* and *Salinispora* genera. Genomes in the *Micrococcineae* suborder, however, had consistently the lowest number of biosynthetic gene clusters. By networking individual gene clusters into gene cluster families, we were able to computationally estimate the degree of novelty each genus contributed to the current sequence databases. Based on the similarity measures between all actinobacteria in the Joint Genome Institute’s Atlas of Biosynthetic gene Clusters database, rare marine genera show a high degree of novelty and diversity, with *Corynebacterium*, *Gordonia*, *Nocardiopsis*, *Saccharomonospora* and *Pseudonocardia* genera representing the highest gene cluster diversity. This research validates that rare marine actinomycetes are important candidates for

Abbreviations: ABC, Atlas of Biosynthetic gene Clusters; BGC, biosynthetic gene cluster; gDNA, genomic DNA; IMG, Integrated Microbial Genomes; JGI, Joint Genome Institute; NRPS, nonribosomal peptide synthetase; PGM, Personal Genome Machine; PKS, polyketide synthase; RMA, rare marine actinomycete.

This Whole Genome Shotgun project has been deposited at DDBJ/ENA/GenBank under the BioProject number PRJNA344658, and individual genome assembly accession numbers are as follows: CNU-125-MKJY01000000, CUA-896-MKKH01000000, CNJ-954-MKKI01000000, CNJ-863-MKKG01000000, CNJ-787-SRX873596*, CNJ-770-MKJW01000000, CUA-901-MKKB01000000, CNB-394-GCA_000374985.1, CNS-044-SRX873600*, CNY-236-GCA_000482385.1, CNR-923-MKKC01000000, CNS-639-GCA_000381685.1, CNT-312-GCA_000515115.1, CNJ-824-MKKA01000000, CNS-004-MKJV01000000, CNS-139-MKJX01000000, CUA-806-MKKD01000000, CUA-673-MKKE01000000, CNQ-490-GCA_000527075.1, CNJ-927-MKKF01000000 and CUA-874-MKJZ01000000. *These accession numbers are for the NCBI Sequence Read Archive (SRA) database.

Six supplementary figures and five supplementary tables are available with the online Supplementary Material.

exploration, as they are relatively unstudied, and their relatives are historically rich in secondary metabolites.

INTRODUCTION

The Actinobacteria represent a diverse phylum of bacteria capable of immense secondary metabolic capacity (Monciardini *et al*., 2014). They are renowned for producing biomedically important molecules such as antibiotics and anticancer compounds and include human and plant pathogens. Many of the better-studied genera in the terrestrial and clinical spheres, such as *Streptomyces* and *Mycobacterium*, have hundreds of genomes sequenced (Doroghazi & Metcalf, 2013; Nett *et al*., 2009). Though these terrestrial strains have been exploited for centuries, the rate of discovery of new chemical entities from terrestrial microbes has slowed in recent years (Bérdy, 2012). The conventional path to natural product discovery relies heavily on the ability to coax cultured strains of bacteria to produce metabolites at detectable levels by varying laboratory conditions. As time goes on, the likelihood of re-discovering a known compound inevitably increases, further reducing the productivity of traditional natural product discovery methods. However, a new era of fast and cheap sequencing has transformed the natural products discovery field by revealing the undetected majority of gene clusters harboured in bacterial genomes (Bentley *et al*., 2002; Ikeda *et al*., 2003; Udwary *et al*., 2007). A bacterium's genomic sequence contains the blueprint of potential molecules the strain is capable of producing. Mining bacterial genomes has shown that their potential for producing secondary metabolites is much higher than what is observed in the laboratory (Bachmann *et al*., 2014). As bioinformatic tools for assessing bacterial secondary metabolite biosynthesis advance, the power in genome mining amplifies, allowing for de-replication of known products, compound class identification, structural predictions and, in some cases, target identification (Jensen *et al*., 2014; Tang *et al*., 2015). Likewise, advances in heterologous expression and regulation manipulation allow increased access to silent natural product biosynthetic pathways (Tang *et al*., 2015; Yamanaka *et al*., 2014).

Marine actinobacteria have recently come into the spotlight as fruitful producers of structurally diverse secondary metabolites and remain relatively untapped (Fenical & Jensen, 2006; Moore *et al*., 2005; Subramani & Aalbersberg, 2013; Zotchev, 2012). Over 70 bioactive compounds have been isolated from marine actinobacteria, most from the genus *Streptomyces* (Manivasagan *et al*., 2014). However, in the marine ecosystem, genera previously underexplored for natural product research are being reported on a regular basis as a source of new metabolites (Tiwari & Gupta, 2012). These so-called rare actinomycetes are defined as strains from actinomycete genera other than *Streptomyces* (Bérdy, 2005) or strains from genera that are isolated less frequently than *Streptomyces* species, although they may not be rare in abundance (Baltz, 2006; Lazzarini *et al*., 2001). Our understanding of the genetic potential of rare marine actinomycetes (RMAs) is incomplete. Although there are many reported compounds from rare actinomycetes (Bérdy, 2005; Choi *et al*., 2015), most rare genera have few genomes published and little is known about their genomic capacity to produce natural products, especially those from the ocean. Insight into RMA genomes may reveal an important untapped resource for unique biosynthetic gene clusters (BGCs) and inform future collection efforts in the search for new bioactive natural products.

Aside from *Streptomyces*, *Salinispora* is arguably the most prolific marine actinomycete genus, in terms of secondary metabolite production, discovered to date, producing a suite of potent bioactive chemicals (Jensen *et al*., 2015; Manivasagan *et al*., 2014). Sequencing of 75 genomes from three *Salinispora* species has revealed that up to 10 % of the genomes are dedicated to secondary metabolite biosynthesis and that many BGCs are uniquely present in only one or two strains (Ziemert *et al*., 2014). These rare BGCs show that sequencing just one strain of a species does not capture the entire repertoire of potential natural products, thus validating further sequencing efforts to discover new BGCs. *Salinispora* is a prime example of how new or poorly explored taxa can lead to the discovery of new molecules through genome mining (Eustáquio *et al*., 2011; Udwary *et al*., 2007). Therefore, further investigation of RMA genomes could give access to a pool of untapped natural product biosynthetic potential. Currently, the Joint Genome Institute (JGI) Integrated Microbial Genomes (IMG) database houses over 4600 actinobacteria genomes. The top four represented orders – *Bifidobacteriales*, *Corynebacteriales*, *Micrococcales* and *Streptomycetales* – account for over 80 % of the genomes in the database.

Advancements in four key areas will aid in overcoming the current lull in novel natural product discovery: studying and learning to cultivate strains from poorly explored habitats, advances in genome sequencing and assembly for unfragmented genome blueprints, improved bioinformatics tools for predicting elusive biosynthetic origins and reliable heterologous expression of cryptic pathways. Development in these areas will unveil unexploited resources, such as RMAs for novel BGC sequence information, and could therefore lead to new chemical entities for use in medicine and biotechnology. While individual metabolites have been reported from RMAs at an increasing rate over the past decade, there has not yet been an analysis of the biosynthetic potential of a large set of RMAs. In this study, we add to the growing pool of sequenced RMA genomes and utilize gene cluster similarity networks to compare RMA gene clusters with gene clusters from the JGI IMG database to assess the likelihood of discovering new natural products. These gene cluster similarity networks allow for the rapid comparison of tens of thousands of gene clusters to quickly determine RMA gene cluster novelty and the diversity of classes they are distributed in.

METHODS

Genomic DNA isolation and genome sequencing. Strains (listed in Table S1, available in the online Supplementary Material) were obtained using previously detailed methods from various locations, grown in pure culture and stored at −80 °C (Gontang *et al.*, 2007; Jensen *et al.*, 2005). Genomic DNA (gDNA) extraction, sequencing and assembly of strains by JGI are previously detailed (Ziemert *et al.*, 2014). Strains sequenced in-house were grown in 5 ml of A1 medium (28 g Instant Ocean distributed by United Pet Group, 10 g starch, 4.0 g yeast extract, 2.0 g peptone and 1 litre deionized water) for 7 days, shaking at 220 r.p.m. at 28 °C on an Innova 2350 platform shaker (New Brunswick Scientific). Five millilitres of culture was then used for gDNA extraction using the Qiagen Genomic-tip 20/G kit (Qiagen). Strains that yielded little or no DNA using this kit and were extracted using a modified protocol developed for *Salinispora* spp. (Gontang *et al.*, 2007). gDNA was checked for quality by running on a 1 % agarose gel at 70–80 V for 1.5–2 h and stained using 1× GelRed (Biotium) in the gel. gDNA was quantified using the Qubit dsDNA HS Assay Kit with the Qubit Fluorometer (Thermo Fisher Scientific).

Ion Torrent 400 bp sequencing libraries were made using the Ion Xpress Plus gDNA Fragment Library kit according to the user guide (Thermo Fisher Scientific). The Covaris S2 (Covaris) was used to shear 1 µg of gDNA to about 500 bp. The samples were then processed according to the user guide for end repair and adapter ligation. The Pippen Prep (Sage Science) instrument was used, according to the provided protocol, to size select using a 2 % agarose gel cassette with DNA marker B, which allows for size selection between 100 and 600 bp, using a narrow size range targeting 475 bp. Libraries were not amplified and were analysed for quality and quantitated using the BioAnalyzer High Sensitivity DNA kit on the BioAnalyzer 2100 System (Agilent). The Ion Personal Genome Machine (PGM) 400 Template OT2 400 bp Kit (Thermo Fisher Scientific) was used for sample preparation with the Ion OneTouch 2 System according to the protocol with a modified 400 bp thermoprofile. The melting temperature was increased to 97 °C with elongated extension times. Sequencing was performed using an Ion Torrent PGM (Thermo Fisher Scientific) with an Ion PGM Hi-Q Sequencing Kit (Thermo Fisher Scientific), according to the standard protocol, on a 318 v2 sequencing chip (Thermo Fisher Scientific).

Genome assembly and annotation. SPAdes version 3.1.1 with Ion Torrent and single cell options was run with each fastq sequencing file (Bankevich *et al.*, 2012; Nurk *et al.*, 2013). K-mer sizes of 21, 33, 55 and 77 and single cell mode are recommended for high G+C genomes and were run for each genome assembly. Scaffolds smaller than 1 kb in length were discarded, unless 16S rRNA gene information was present. Each genome assembly was submitted to the JGI IMG Expert Review pipeline for genome annotation and is publicly available through http://genomeportal.jgi.doe.gov/. Additionally, this Whole Genome Shotgun project has been deposited at DDBJ/ENA/GenBank under the BioProject number PRJNA344658, and individual genome assembly accession numbers can be found in Table S1.

Contigs containing potential contaminants were identified based on a combination of assembly coverage depth, nucleotide composition (percent G+C) and number of predicted amino acid sequences having closest matches in GenBank nr from actinobacterial versus non-actinobacterial sources, as determined using DarkHorse software, version 1.5 (Podell & Gaasterland, 2007). One genome, CUA-806 *Rhodococcus* sp., was contaminated by a low G+C *Staphylococcus* sp., sequenced at a much higher coverage. In this case, we were able to extract the low coverage, high G+C scaffolds such that the final assembly only contains scaffolds attributed to *Rhodococcus*.

Genome quality was assessed as described in *Standards in Genomic Science* in 2014 (Land *et al.*, 2014). Scaffold number; length of continuous, un-gapped nucleotides; length of 5S, 16S and 23S rRNAs; number and amino acid specificity of tRNAs and the presence or absence of genes encoding 102 housekeeping proteins found in nearly all bacteria were combined into a single normalized, composite score, enabling direct comparison to previously published quality data for 32 000 microbial genomes already in public databases.

All genome assemblies were subsequently analysed using antiSMASH v3.0 with and without ClusterFinder enabled (Weber *et al.*, 2015). The gene clusters identified without ClusterFinder were further curated using NaPDoS (Natural Product Domain Seeker) (Ziemert *et al.*, 2012) to determine which polyketide synthase (PKS) and nonribosomal peptide synthetase (NRPS) partial clusters most likely belonged in the same gene cluster (Fig. S1). The final numbers of gene clusters were corrected to incorporate this additional information on inferred connections, consolidating gene clusters likely to have been split across multiple contigs during sequence assembly. This correction step was necessary to avoid potential overestimation of the number of PKS and NRPS clusters in each genome. The resulting table (Table S4) of genome versus gene cluster type and number was visualized using Circos (Krzywinski *et al.*, 2009).

16S phylogenetic tree. The 16S rRNA sequences were extracted from the whole-genome assemblies for all strains. Type strains within the same families and, if possible, from marine sources were selected from the list of prokaryotic names with standing in nomenclature (LPSN; bacterio.net) and 16S gene sequences for these strains were collected from NCBI. All sequences were aligned with MAFFT (Katoh *et al.*, 2005) using the accurate L-INS-I mode. Alignments were manually inspected and trimmed with trimAl (Capella-Gutiérrez *et al.*, 2009) using the automated1 setting resulting in approximately 1500 bp of aligned sequence. The GTR+I+G evolutionary model was selected using MrModeltest (Nylander, 2004) and a maximum-likelihood tree was built using RAxML (Stamatakis, 2006) with 1000 bootstrap replicates. Visual aids were then added using the Interactive Tree of Life (iTOL) (Letunic & Bork, 2016).

Gene cluster family networking. BGC similarity was determined using a distance measure based on Pfam composition as detailed in Cimermancic *et al.* (2014). The method uses an optimized weighted distance that incorporates the Jaccard index and domain duplication similarity measures as employed in a previous method to determine protein similarity (Lin *et al.*, 2006). Pairwise distances between BGCs in RMA strains and Actinobacteria from JGI were generated using a custom python script and were then visualized in a network using Gephi (Bastian *et al.*, 2009). A similarity threshold was set at 0.6 to limit network complexity while retaining meaningful connections and to minimize the number of connected nodes with different cluster type annotations (Fig. S2). Clustering using the OpenORD and the YifanHu force-directed algorithm was performed for the final layout (Hu, 2006). Before final networking, a de-replication step was necessary to account for the number of re-sequencing projects of the same isolate present in the JGI dataset. Replicate gene clusters, ones with 0.99 similarity score and matching organism identification, were condensed into a single node before visualization. This was accomplished using an initial network of nodes over the 0.99 threshold. Attributes were inherited by the new nodes with the addition of a node size attribute, used to visualize the number of de-replicated gene clusters in the final network.

Annotations for secondary metabolite products from the JGI dataset were augmented by including homology results to the MIBiG database

(Medema *et al.*, 2015). A random sampling of clusters in the set were screened using MultiGeneBlast (Medema *et al.*, 2013) against MIBiG and paired with the highest scoring hit. All hits that did not have 80 % of the genes in the query cluster were filtered out. Nodes used as example compounds were also manually screened for accuracy and MIBiG BGC numbers are provided in Fig. 1 caption.

RESULTS AND DISCUSSION

Rare actinomycete sequencing and genome assembly

We chose 21 strains for genome sequencing and secondary metabolite analysis (Table S1). In their paper, Gontang *et al.* (2010) probed a variety of marine sediment actinomycetes for pathways indicative of secondary metabolite production. The majority of the microbes assayed contained either or both PKS and NRPS pathways, suggesting the presence of biosynthetic machinery for making yet to be identified molecules. We selected nine strains from the Gontang study, as well as five newly isolated RMAs, for whole-genome sequencing in-house, and another seven strains for sequencing at the JGI. We sequenced the 14 strains using the Ion Torrent PGM, with 400 bp sequencing libraries and 400 bp sample preparation and sequencing chemistry (Rothberg *et al.*, 2011; Yamanaka *et al.*, 2014). The limitations of using next-generation sequencing, specifically for the discovery of secondary metabolites, are many and varied (Gomez-Escribano *et al.*, 2016). Read length and high G+C content of the DNA are the two biggest hurdles to overcome. At the time of sequencing, 400 bp contiguous sequences were considered quite long. To address the high G+C nature of these actinomycete genomes, we created a modified thermoprofile which incorporated high denaturation temperatures and longer extension times to improve amplification of these long, high G+C sequencing libraries. Each genome was run on one 318 chip, giving 1–1.8 Gb of information per genome.

We used SPAdes (Bankevich *et al.*, 2012; Nurk *et al.*, 2013) for genome assembly for many reasons. First, it is one of the only non-commercial assemblers that can be used with Ion Torrent reads, and has an error correction module, IonHammer, specific to Ion Torrent errors. Additionally, the single cell option, which was developed for improved assembly of multiple displacement amplified genomes from a single cell, aided with assembly of these high G+C genomes because of the non-uniform coverage profiles associated with these genomes. Preliminary assembly using commercial CLC Genomics Workbench (Qiagen) software resulted in very few complete secondary metabolite gene clusters; most clusters were truncated at the beginning or at the end of a contig. However, after using SPAdes, over 50 complete gene clusters, including notoriously difficult to assemble PKS and NRPS clusters, emerged. The single cell option in SPAdes helped us to solve the complications associated with high G+C sequences.

All genome assemblies were confirmed to be of sufficiently high quality for comparative analysis (Table S2) based on recommended guidelines set forth in 2014 in *Standards in Genomic Science* (Land *et al.*, 2014). This definitive monograph states that genome assemblies with quality scores above 0.8 can be safely used for comparative genomic analysis, but those with scores below 0.6 should not be used. Of the 21 RMA genomes, 19 in the current study had quality scores above 0.8. The two assemblies below 0.8 were *Actinomadura* sp. CNU-125 (0.76) and *Kytococcus* sp. CUA-901 (0.79), but both of these were still well above the 0.6 minimum threshold value.

Variety of secondary metabolite biosynthetic gene clusters

Many of the under-exploited genera from this study contain pathways of various classes that warrant further exploration (Table S3). The variety and number of pathways, as found by antiSMASH 3.0 without ClusterFinder, for a representative strain of each genus sequenced are displayed in the Circos diagram (Fig. 2) (Krzywinski *et al.*, 2009). Non-NRPS-PKS hybrid clusters were separated into their component parts so as to more easily visualize the variety of categories present, while the 'Hybrid' category retains only NRPS-PKS hybrid clusters (Table S4). As is expected, the number and variety of pathways generally increases as the size of the genome increases, with few exceptions. Also not surprising are the ubiquitous terpene pathways, present in every genome. Recent genome mining efforts revealed wide distribution of terpene synthases in bacterial genomes and led to the creation of a new hidden Markov model to identify bacterial terpene synthases (Cane & Ikeda, 2012; Yamada *et al.*, 2015).

The second most pervasive and most abundant class of BGCs are PKS pathways, present in all large genomes, and some small genomes as well. The next most represented class of BGCs is NRPS clusters. Identification of siderophores based on bioinformatics criteria alone can be difficult, and many siderophores are made by NRPS pathways that antiSMASH identifies as NRPS and not Siderophore. In fact, all genomes that did not have an explicitly identified siderophore cluster contained at least one NRPS pathway with >50 % gene homology, according to antiSMASH, to known siderophore pathways, such as coelichelin and fuscachelin. Interestingly, one of the larger genomes, *Pseudonocardia* strain CNS-004, contains a relatively low number of secondary metabolite pathways, with no PKS or Hybrid clusters and only two NRPS pathways (Table S2). In contrast, *Pseudonocardia* strain CNS-139 contains a large number of clusters, including Hybrid, PKS and many NRPS clusters. This difference in total pathway abundance between two members of the same genus is also seen in another *Pseudonocardineae* genus, *Saccharomonospora* (Fig. 3).

The actinomycete 16S phylogeny (Fig. 3) reveals some patterns in the number of pathways present in the RMA genomes. Some of the strains sequenced have numbers nd MIBiG BGC numbers are provand varieties of pathway

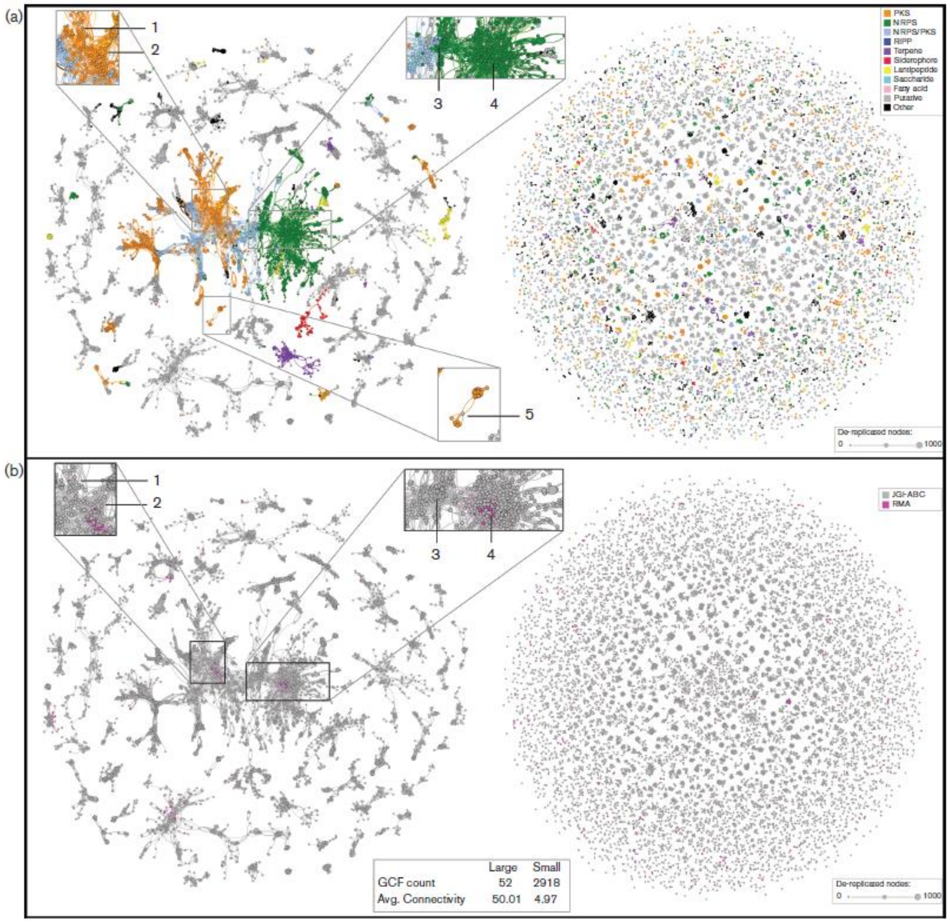


Fig. 1. JGI ABC gene cluster similarity network. Gene cluster similarity networks generated using >3000 Actinobacteria genomes from JGI ABC and the RMA genomes introduced in this study. Gene clusters were identified using antiSMASH3.0 with ClusterFinder (probability >0.8). Each node represents one sequenced gene cluster; any identical gene clusters from multiple sequencings of the same genorme were de-replicated and that information is stored in the size of the node. If cluster category was assigned in JGI ABC, those colours were included and correspond to the antiSMASH colouring scheme in (a). RMA BGCs are highlighted in pink in (b). For better visualization, the network was split into large communities (left) and small communities (right). Selected zoom panels in (a) show examples of BGCs with known products. Type I PKS macrolides such as oligomycin (1) (BGC0000117) and erythromycin (2) (BGC0000055) are contained within the larger PKS GCF. Siderophores, such as mycobactin (3) (BGC0001021), lie within the hybrid NRPS-PKS section, while cyclic depsipeptides, including homologues to pristinamycin (4) (BGC0000952), reside in the NRPS GCF. Rifamycin (5) (BGC0000136) and analogues form their own separate GCF. The two zoom panels in (b) show RMA nodes found in proximity to the identified BGCs for known compounds. General network statistics are shown at the bottom of (b).

classes that rival the well-known *Streptomyces* and *Salinispora* producers, which can contain as many as 30 BGCs. These include strains in the genera *Nocardia*, *Rhodococcus*, *Actinomadura*, *Micromonospora*, *Nocardiopsis* and *Gordonia*. Perhaps more surprising are the genera that lack a large number of secondary metabolite BGCs. These are small genome (<4 Mb) actinomycetes, which also lack NRPS and non-fatty-acid PKS pathways, that belong to the suborder

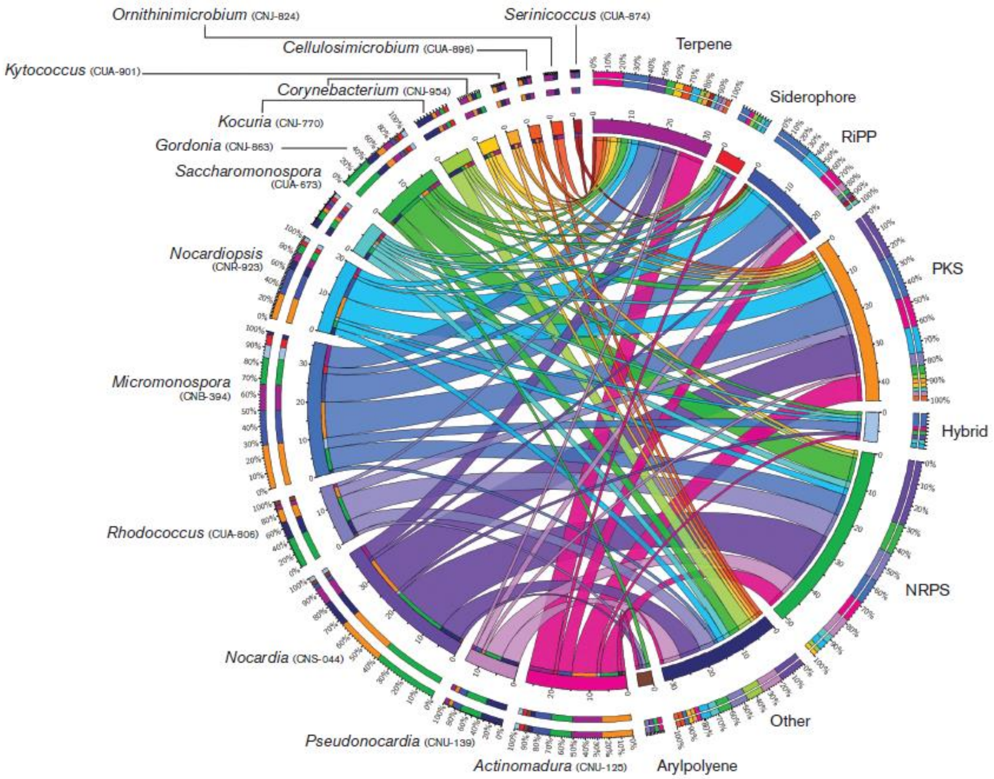


Fig. 2. Circos diagram of RMA pathway diversity. The genomes in this Circos figure include one representative from each genus sequenced and are arranged in ascending genome size, from the top of the circle, counterclockwise to the bottom. Each genome is represented by a different coloured band (left half of the circle) that can be traced from the organism to the types of gene clusters found in that genome (right half of the circle). The width of these bands indicates the number of pathways of that type. The cluster types are also assigned colours based on the colours antiSMASH uses to represent pathway types. The cluster types are designated by their respective colours that make up the outer two rings next to each genome to easily see what portion of the pathways belong to each category. Conversely, the outer two rings next to the gene cluster categories show the proportion of that pathway attributed to each genome represented by the genome colour. The 'Other' category includes clusters that antiSMASH calls Other and Ectoine, as well as uncommon cluster types found in only one or two genomes, such as the following: Butyrolactone, Phenazine, Homoserine lactone, Aminoglycosides, Oligosaccharide and Nucleoside. The Hybrid category includes only Hybrid NRPS-PKS gene clusters. All other hybrid clusters were split into their component parts to get a better overview of gene cluster category diversity (see Tables S3 and S4 for full gene cluster counts). The RiPP category includes clusters identified as Bacteriocin, Lantipeptide or Thiopeptide.

Micrococcineae. Actinomycetes are generally lauded for the production of multiple complex metabolites made by NRPS and PKS pathways; thus, to observe a subset of genomes without this capacity is quite unusual. This is not to say that they do not produce interesting natural products but that these small genome actinomycetes may have different mechanisms of producing secondary metabolites undetectable by current bioinformatic analysis. A taxonomic search in MarinLit for all families in the Fig. S5 suborder returned three recently published groups of compounds: the dermacozines A–J, phenazine compounds isolated from the deep sea *Dermacoccus abyssi* (Abdel-Mageed *et al.*, 2010; Wagner *et al.*, 2014); microluside A, a glycosylated xanthone from a sponge associated *Micrococcus* sp. EG45 (Eltamany *et al.*, 2014); and indole alkaloids from the deep sea *Serinicoccus profundi* (Yang *et al.*, 2013). Additionally, seriniquinone was isolated from *Serinicoccus* strain CNJ-927, a strain sequenced as part of this study (Trzoss *et al.*, 2014). Kocurin, a thiozolyl peptide, has been reported from marine sponge-derived strains *Kocuria marina* F-276310, *Kocuria palustris* F-276345 and *Micrococcus yunnanensis* F-256446 and is hypothesized to be a product of a RiPP (ribosomally synthesized and post-translationally modified peptide) pathway (Palomo *et al.*, 2013). Aside from kocurin, the reported structures from Fig. S5 strains have bioinformatically elusive biosynthetic origins that would likely not be identified by current automated genome mining programs.

Gene cluster similarity network

To address the novelty of the BGCs within the RMA genomes, we classified gene cluster families (GCFs) via a similarity network (Fig. 1). GCF similarity networks have been increasingly utilized to compare gene clusters from large sequencing datasets (Cimermancic *et al.*, 2014; Doroghazi *et al.*, 2014; Ziemert *et al.*, 2014). Grouping gene clusters into larger families allows for quick prioritization of potentially new classes of gene clusters and dereplication of known gene clusters and their associated products (Medema & Fischbach, 2015). In this case, the degree to which the gene clusters we found in the RMAs network with other sequenced gene clusters can give insight into how rare the pathways are in this subset of marine bacteria and whether they are worth pursuing for novel gene cluster discovery.

In order to determine which RMA gene clusters are similar to already sequenced, but not necessarily experimentally confirmed, gene clusters, we used a gene cluster similarity networking approach with the JGI Atlas of Biosynthetic gene Clusters (ABC) database as our comparison set (Hadjithomas *et al.*, 2015). This dataset is composed of all genomes deposited in JGI and run with ClusterFinder and antiSMASH to locate and annotate secondary metabolite gene clusters. We only used actinomycete genomes from the database for comparison to our RMA genomes and dereplicated identical gene clusters from replicate sequencings of the same genome. The resulting network is split into large and small GCFs coloured by BGC type (Fig. 1a), and to show where RMA clusters are incorporated in the network, Fig. 1(b) shows RMA BGCs highlighted in pink. Each BGC is represented by a node, and BGCs that do not have a similarity score over the 0.6 threshold do not appear in the network. Related BGC nodes are connected by edges, and an inclusive set of connected nodes is called a GCF, as was used in Cimermancic *et al.* (2014). In the JGI ABC gene cluster similarity network, 311 of the 1382 (22 %) RMA BGCs appear in the network as nodes, and of those, only 179 nodes (13 % of the total number of predicted gene clusters) are directly linked with JGI ABC nodes. This suggests that 87 % of the RMA gene clusters have similarity scores lower than the threshold used for the network in comparison with any known actinomycete sequence in the JGI ABC database. The dispersal of RMA nodes can be seen in Fig. 1 (b); note that the RMA nodes in the smaller half of the network tend to group into GCFs comprised only of RMA nodes. The distribution of similarity scores (Fig. S3) also shows an average drop for RMA sequences relative to the JGI dataset. Additionally, because a significant fraction of the JGI dataset is represented in the network, we expect that the exclusion of the many RMA gene clusters is not significantly explained by shortcomings with the distance method.

For comparison, we looked at the uniqueness of pathways in the well-characterized genomes of *Streptomyces coelicolor* (non-rare) and *Saccharopolyspora erythraea* (rare). For all *Streptomyces coelicolor* strains included in the JGI ABC set, 54 of the 73 (74 %) gene clusters network with other JGI ABC clusters. For all *Saccharopolyspora erythraea* strains in the JGI ABC set, 25 of the 183 (14 %) gene clusters network with other JGI ABC clusters. This analysis sets a benchmark that, for non-rare strains, the similarity of clusters is quite high while, for rare strains, the connectivity of their clusters is lower, as is seen with the RMA gene clusters.

To assess if the marine environment plays a role in pathway uniqueness, we looked at *Streptomyces xinghaiensis*, the first marine-derived streptomycete to be sequenced (Zhao & Yang, 2011). As a streptomycete, we would expect a high number of pathways shared with other genomes in the database. However, we see that only 9 clusters out of 60 total (15 %) are included in the network, which is on par with the RMA genomes. To further explore the uniqueness of marine streptomycetes, we analysed 24 additional marine isolates (strains listed in the online Supplementary Material). Of the 1925 gene clusters in this group, 412 clusters are in the network (21 %) (Table 1). This low number suggests that marine-derived streptomycetes also harbour unique gene clusters, and we propose that the under-sampling of marine genomes is likely the reason for such novelty.

While the number of unique pathways is important, the diversity of networked gene clusters can also direct which strains to pursue for greater novelty and variety. Diversity indices are used to measure species diversity (Tuomisto, 2010); however, here, we have applied the measure for True diversity (as a function of the Shannon Index) (Jost, 2006) to BGC diversity. To compare the diversity between BGCs in RMA and marine streptomycete genomes, we calculated diversity and normalized it by number of BGCs (Table 1). This measure gives insight into the degree of re-occurring GCFs and, by extension, re-occurring classes of compounds, where a higher value represents a wider range of GCFs and therefore increased likelihood of product diversity. While both RMAs and marine streptomycetes have a low amount of clusters that network, those that do are more diverse in RMAs (0.28) than in marine streptomycetes (0.18). Furthermore, the overlap between RMA and marine streptomycetes is very low, with only four GCFs in common between the two groups (3 %) (Fig. S4). Thus, it is worthwhile to continue sequencing both marine streptomycetes

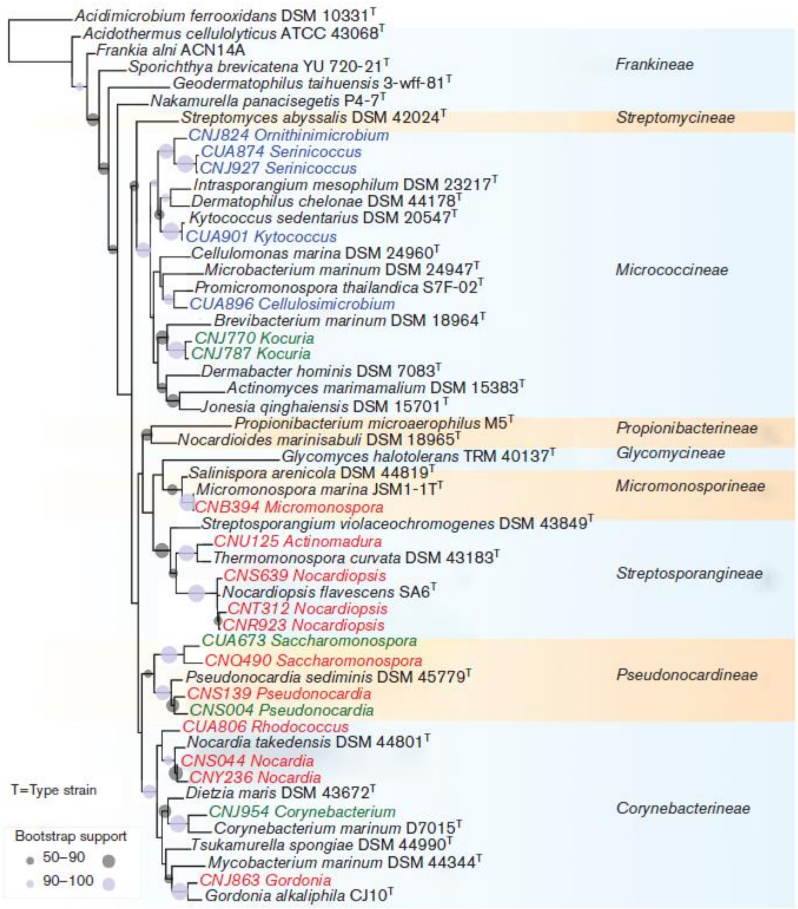

Acidimicrobium ferrooxidans DSM 10331^{T}
Acidothermus cellulolyticus ATCC 43068^{T}
Frankia alni ACN14A
Sporichthya brevicatena YU 720-21^{T}
Geodermatophilus taihuensis 3-wff-81^{T}
Nakamurella panacisegetis P4-7^{T}
Frankineae
Streptomyces abyssalis DSM 42024^{T}
Streptomycineae
CNJ824 Ornithinimicrobium
CUA874 Serinicoccus
CNJ927 Serinicoccus
Intrasporangium mesophilum DSM 23217^{T}
Dermatophilus chelonae DSM 44178^{T}
Kytococcus sedentarius DSM 20547^{T}
CUA901 Kytococcus
Cellulomonas marina DSM 24960^{T}
Microbacterium marinum DSM 24947^{T}
Promicromonospora thailandica S7F-02^{T}
CUA896 Cellulosimicrobium
Micrococcineae
Brevibacterium marinum DSM 18964^{T}
CNJ770 Kocuria
CNJ787 Kocuria
Dermabacter hominis DSM 7083^{T}
Actinomyces marimamalium DSM 15383^{T}
Jonesia qinghaiensis DSM 15701^{T}
Propionibacterium microaerophilus M5^{T}
Propionibacterineae
Nocardioides marinisabuli DSM 18965^{T}
Glycomyces halotolerans TRM 40137^{T}
Glycomycineae
Salinispora arenicola DSM 44819^{T}
Micromonospora marina JSM1-1T^{T}
CNB394 Micromonospora
Micromonosporineae
Streptosporangium violaceochromogenes DSM 43849^{T}
CNU125 Actinomadura
Thermomonospora curvata DSM 43183^{T}
CNS639 Nocardiopsis
Nocardiopsis flavescens SA6^{T}
CNT312 Nocardiopsis
CNR923 Nocardiopsis
Streptosporangineae
CUA673 Saccharomonospora
CNQ490 Saccharomonospora
Pseudonocardia sediminis DSM 45779^{T}
CNS139 Pseudonocardia
CNS004 Pseudonocardia
Pseudonocardineae
CUA806 Rhodococcus
Nocardia takedensis DSM 44801^{T}
CNS044 Nocardia
CNY236 Nocardia
Dietzia maris DSM 43672^{T}
CNJ954 Corynebacterium
Corynebacterium marinum D7015^{T}
Corynebacterineae
Tsukamurella spongiae DSM 44990^{T}
Mycobacterium marinum DSM 44344^{T}
CNJ863 Gordonia
Gordonia alkaliphila CJ10^{T}
T=Type strain
Bootstrap support
50–90
90–100

Fig. 3. 16S rRNA phylogenetic tree of RMA and select type strains. The 21 RMAs used in this study are compared with representative type strains. When possible, marine type strains were chosen. Strains coloured red contain a high number of pathways (>13), those coloured green have a medium number of pathways (5–8) and at least one PKS and/or NRPS pathway (with the exception of CNJ-787) and those coloured blue have a low number of pathways (1–3) and do not contain NRPS or non-fatty-acid PKS pathways. The number of gene clusters in each genome was determined using antiSMASH 3.0 without ClusterFinder. Bootstrap values are indicated by grey circles (50–90) and blue circles (90–100) with increasing size representing increasing confidence.

but, perhaps more importantly, RMAs that have novelty as well as diversity.

Marine-derived genera warranting further study

While we can see that RMAs in general maintain unique gene clusters, it would be beneficial to know which genera contain the least replicated and more diverse gene clusters. We therefore analysed each genus individually and compared the number of new GCFs present from RMAs (Table S5); these represent GCFs not present in the currently sequenced genomes within the same genus. To isolate the effect that the marine environment has, we combined any genome in JGI ABC that had any metadata indicating isolation from a marine environment with those strains sequenced as part of this study. For example, 89 GCFs make

Table 1. RMA versus marine *Streptomyces* true diversity

This table compares 21 RMA genomes from this study and 24 marine streptomycete genomes. While the two groups have similar percentages for in network BGCs, the diversity of RMA BGCs is greater.

	Total BGCs	No. of strains	BGCs in network	No. of GCFs	% BGCs in network	RMA GCFs shared	True diversity	True diversity/ BGC
RMA	1386	21	311	153	22.44	153	86.1595	0.2770
Marine *Streptomyces*	1925	24	412	143	21.40	4	73.6128	0.1787

up the *Nocardiopsis* network (Fig. 4), which is composed of 18 non-marine JGI ABC strains and four RMA strains (three from this study and one from JGI). Of the 38 GCFs that contain RMA gene clusters, 26 (68 %) are exclusively composed of RMA gene clusters. This indicates that the marine *Nocardiopsis* strains include substantial biosynthetic gene diversity not observed in the 18 sequenced *Nocardiopsis* strains that came from sources other than the marine environment. Furthermore, the True diversity, normalized by number of BGCs, is higher for RMA BGCs (0.54) than for non-marine *Nocardiopsis* BGCs (0.17). Because the *Nocardiopsis* strains in the JGI ABC database are from terrestrial, host-associated, aquatic non-marine and other non-marine sources (Fig. S5), the added variety of clusters from the RMA genomes may be due to their marine nature. A recent review details the current natural products from *Nocardiopsis* species and the unique potential of marine *Nocardiopsis* strains (Bennur *et al.*, 2016). This genomic analysis corroborates their assertion that marine *Nocardiopsis* strains are promising in the pursuit of novel bioactive small molecules. In fact, for all genera examined here, all but two (*Actinomadura* and *Kocuria*) have higher normalized True diversity ratios for marine genomes when compared with their non-marine counterparts from JGI ABC (Table S5).

In order to account for how phylogenetic distance between strains affects BGC diversity, we plotted 16S rRNA percent identity against GCF overlap for each pair in the following groups: RMA genomes sequenced as part of this study, JGI marine streptomycetes and non-marine rare actinomycete genomes within the genera examined in this study (Fig. S6). The general trend shows that, with increasing phylogenetic similarity, more GCFs are shared for all three groups.

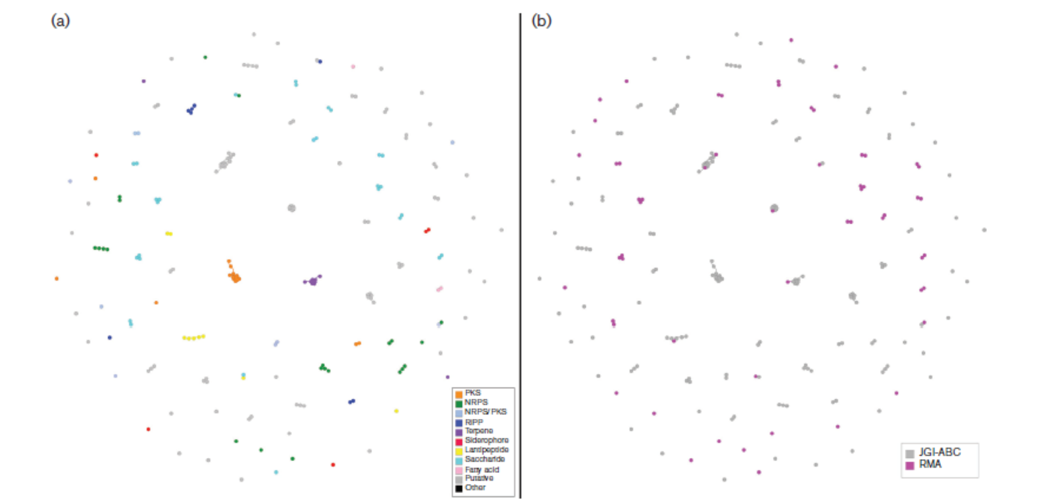


Fig. 4. Gene cluster similarity network of *Nocardiopsis* strains. Gene clusters from all non-marine *Nocardiopsis* genomes in JGI ABC and the three RMA *Nocardiopsis* genomes from this study and one marine *Nocardiopsis* genome from JGI that network with any gene cluster in the large network are retained in this *Nocardiopsis* subset network. Gene clusters coloured by type are shown in (a), while (b) highlights in pink those clusters from the four marine *Nocardiopsis* genomes. RMA clusters tend to form their own GCFs, with only seven RMA nodes connecting with other *Nocardiopsis* nodes.

However, only the non-marine group has pairs that share greater than 20 % GCFs at low phylogenetic distances (below 94 %). This could be due to the lower number of marine genomes in current sequence databases, but it could also indicate more diversity in marine genomes at lower phylogenetic distances. This observed increase in diversity for marine BGCs indicates a hidden biosynthetic potential that warrants more sequencing of RMAs. Although the number of RMA strains included in these analyses is low, with more sequencing, a more complete picture of their full potential will emerge.

The prospect of leaping from genomic code to molecule is fast approaching, with technological advances in sequencing, bioinformatics and heterologous gene expression paving the way to discover and manipulate unknown molecules from proposed biosynthetic pathways. The conventional path to discovering new compounds from microbes is lengthy, work intensive and does not always capture the full potential of high secondary metabolite producing bacteria. Additionally, a large percentage of the microbes from environmental samples have yet to be obtained in culture, representing an impressive biodiversity that remains largely inaccessible to natural product discovery. This new era of fast, easy sequencing can open doors to exploring microbial communities harbouring unprecedented natural chemistry. Though they have been examined on an individual basis for natural product discovery, RMAs as a group represent a widely untapped resource for new BGCs. Certain suborders of RMAs have very high potential to possess unique gene clusters, which may encode unprecedented chemical scaffolds. Further efforts should focus on culturing RMAs and sequencing their genomes to survey their full biosynthetic potential.

Gene cluster similarity networks provide a powerful tool to quickly assess the uniqueness of a given genome's biosynthetic pathways. However, there are limitations to such an automated method. We encountered gene cluster trimming to be a setback in some cases where antiSMASH would overcall the number of genes in a gene cluster. Automated and accurate gene cluster boundary delineation will only improve the precision of gene cluster similarity networks.

It has been estimated that all Actinobacteria biosynthetic diversity can be reached by sequencing only 15 000 actinomycete genomes (Doroghazi *et al.*, 2014). However, as the authors state, this estimation is based on what is currently in our sequenced databases, which is largely terrestrial streptomycetes and clinical isolates. With the amount of novelty seen in RMA and even marine streptomycete genomes, this estimation can be re-visited. We may be saturating the pool with terrestrial strains, but thus far, the potential from marine genomes has yet to be fully realized. Aside from the marine environment, other unexplored habitats will likely expand the number of genomes we need to sequence to see full secondary metabolite pathway potential. Bacteria in symbioses (such as endophytes in plants and endosymbionts in sponges), microbes living in extreme environments and un-cultured bacteria all represent large potential reservoirs of unknown biosynthetic capacity (Brader *et al.*, 2014; Chávez *et al.*, 2015). Using the RMA genomes as a glimpse into the potential of under-sampled genomes showcases the importance of expanding our sequencing efforts from the mainly terrestrial and clinical isolates that exist today. Improved sampling and culturing practices, along with enhanced molecular biology, sequencing and metagenome assembly techniques, will pave the way for accessing previously inaccessible genomes in the search for new biosynthetic potential.

Availability of data and material

The genome assemblies for each strain in this article are available at DDBJ/ENA/GenBank under the BioProject number PRJNA344658, and individual genome assembly accession numbers can be found in Table S1. Additionally, these strains can be found through the JGI IMG Expert Review portal at https://img.jgi.doe.gov. Genome assemblies can be accessed using the NCBI accession numbers or JGI OIDs provided in Table S1. Custom python scripts written to create the gene cluster similarity network are deposited and annotated at: https://bitbucket.org/malanjary_ut/clustsimscore.

ACKNOWLEDGEMENTS

M. A. S. acknowledges support from the Edna Bailey Sussman Foundation, the Louise Torraco Memorial Doctorate in Science Scholarship from UNICO and the Institutional Strategy of the University of Tübingen (Deutsche Forschungsgemeinschaft, ZUK 63). B. S. M. and P. R. J. acknowledge support from the National Institutes of Health (NIH) grant R01-GM085770. The genome sequencing conducted by the U.S. Department of Energy Joint Genome Institute is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. A. K. was supported by the Russian Science Foundation (grant 14-50-00069). N. Z. and M. M. A. acknowledge support from the German Center for Infection Research (DZIF) (TTU 09.704).

REFERENCES

Abdel-Mageed, W. M., Milne, B. F., Wagner, M., Schumacher, M., Sandor, P., Pathom-aree, W., Goodfellow, M., Bull, A. T., Horikoshi, K. & other authors (2010). Dermacozines, a new phenazine family from deep-sea dermacocci isolated from a Mariana Trench sediment. *Org Biomol Chem* **8**, 2352–2362.

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

Baltz, R. H. (2006). Marcel Faber Roundtable: is our antibiotic pipeline unproductive because of starvation, constipation or lack of inspiration? *J Ind Microbiol Biotechnol* **33**, 507–513.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko, S. I., Pham, S. & other authors (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* **19**, 455–477.

Bastian, M., Heymann, S. & Jacomy, M. (2009). Gephi: an open source software for exploring and manipulating networks. In *Third International*

AAAI Conference on Weblogs and Social Media. San Jose McEnery Convention Center.

Bennur, T., Ravi Kumar, A., Zinjarde, S. S. & Javdekar, V. (2016). *Nocardiopsis* species: a potential source of bioactive compounds. *J Appl Microbiol* **120**, 1–16.

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

Bérdy, J. (2005). Bioactive microbial metabolites. *J Antibiot* **58**, 1–26.

Bérdy, J. (2012). Thoughts and facts about antibiotics: where we are now and where we are heading. *J Antibiot* **65**, 441.

Brader, G., Compant, S., Mitter, B., Trognitz, F. & Sessitsch, A. (2014). Metabolic potential of endophytic bacteria. *Curr Opin Biotechnol* **27**, 30–37.

Cane, D. E. & Ikeda, H. (2012). Exploration and mining of the bacterial terpenome. *Acc Chem Res* **45**, 463–472.

Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973.

Choi, S.-S., Kim, H.-J., Lee, H.-S., Kim, P. & Kim, E.-S. (2015). Genome mining of rare actinomycetes and cryptic pathway awakening. *Process Biochem* **50**, 1184–1193.

Chávez, R., Fierro, F., García-Rico, R. O. & Vaca, I. (2015). Filamentous fungi from extreme environments as a promising source of novel bioactive secondary metabolites. *Front Microbiol* **6**, 903.

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

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

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

Eltamany, E. E., Abdelmohsen, U. R., Ibrahim, A. K., Hassanean, H. A., Hentschel, U. & Ahmed, S. A. (2014). New antibacterial xanthone from the marine sponge-derived *Micrococcus* sp. EG45. *Bioorg Med Chem Lett* **24**, 4939–4942.

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

Fenical, W. & Jensen, P. R. (2006). Developing a new resource for drug discovery: marine actinomycete bacteria. *Nat Chem Biol* **2**, 666–673.

Gomez-Escribano, J., Alt, S. & Bibb, M. (2016). Next generation sequencing of Actinobacteria for the discovery of novel natural products. *Mar Drugs* **14**, 78.

Gontang, E. A., Fenical, W. & Jensen, P. R. (2007). Phylogenetic diversity of Gram-positive bacteria cultured from marine sediments. *Appl Environ Microbiol* **73**, 3272–3282.

Gontang, E. A., Gaudêncio, S. P., Fenical, W. & Jensen, P. R. (2010). Sequence-based analysis of secondary-metabolite biosynthesis in marine Actinobacteria. *Appl Environ Microbiol* **76**, 2487–2499.

Hadjithomas, M., Chen, I. M., Chu, K., Ratner, A., Palaniappan, K., Szeto, E., Huang, J., Reddy, T. B., Cimermančič, P. & other authors (2015). IMG-ABC: a knowledge base to fuel discovery of biosynthetic gene clusters and novel secondary metabolites. *MBio* **6**, e00932-15.

Hu, Y. (2006). Efficient, high-quality force-directed graph drawing. *Mathematica J* **10**, 37–71.

Ikeda, H., Ishikawa, J., Hanamoto, A., Shinose, M., Kikuchi, H., Shiba, T., Sakaki, Y., Hattori, M. & Omura, S. (2003). Complete genome sequence and comparative analysis of the industrial microorganism *Streptomyces avermitilis*. *Nat Biotechnol* **21**, 526–531.

Jensen, P. R., Gontang, E., Mafnas, C., Mincer, T. J. & Fenical, W. (2005). Culturable marine actinomycete diversity from tropical Pacific Ocean sediments. *Environ Microbiol* **7**, 1039–1048.

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

Jensen, P. R., Moore, B. S. & Fenical, W. (2015). The marine actinomycete genus *Salinispora*: a model organism for secondary metabolite discovery. *Nat Prod Rep* **32**, 738–751.

Jost, L. (2006). Entropy and diversity. *Oikos* **113**, 363–375.

Katoh, K., Kuma, K.-i., Toh, H. & Miyata, T. (2005). MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res* **33**, 511–518.

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S. J. & Marra, M. A. (2009). Circos: an information aesthetic for comparative genomics. *Genome Res* **19**, 1639–1645.

Land, M. L., Hyatt, D., Jun, S. R., Kora, G. H., Hauser, L. J., Lukjancenko, O. & Ussery, D. W. (2014). Quality scores for 32,000 genomes. *Stand Genomic Sci* **9**, 20.

Lazzarini, A., Cavaletti, L., Toppo, G. & Marinelli, F. (2001). Rare genera of actinomycetes as potential producers of new antibiotics. *Antonie Van Leeuwenhoek* **79**, 399–405.

Letunic, I. & Bork, P. (2016). Interactive Tree of Life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res* **44**, W242–W245.

Lin, K., Zhu, L. & Zhang, D. Y. (2006). An initial strategy for comparing proteins at the domain architecture level. *Bioinformatics* **22**, 2081–2086.

Manivasagan, P., Venkatesan, J., Sivakumar, K. & Kim, S. K. (2014). Pharmaceutically active secondary metabolites of marine actinobacteria. *Microbiol Res* **169**, 262–278.

Medema, M. H. & Fischbach, M. A. (2015). Computational approaches to natural product discovery. *Nat Chem Biol* **11**, 639–648.

Medema, M. H., Takano, E. & Breitling, R. (2013). Detecting sequence homology at the gene cluster level with MultiGeneBlast. *Mol Biol Evol* **30**, 1218–1223.

Medema, M. H., Kottmann, R., Yilmaz, P., Cummings, M., Biggins, J. B., Blin, K., de Bruijn, I., Chooi, Y. H., Claesen, J. & other authors (2015). Minimum information about a biosynthetic gene cluster. *Nat Chem Biol* **11**, 625–631.

Monciardini, P., Iorio, M., Maffioli, S., Sosio, M. & Donadio, S. (2014). Discovering new bioactive molecules from microbial sources. *Microb Biotechnol* **7**, 209–220.

Moore, B. S., Kalaitzis, J. A. & Xiang, L. (2005). Exploiting marine actinomycete biosynthetic pathways for drug discovery. *Antonie van Leeuwenhoek* **87**, 49–57.

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

Nurk, S., Bankevich, A., Antipov, D., Gurevich, A. A., Korobeynikov, A., Lapidus, A., Prjibelski, A. D., Pyshkin, A., Sirotkin, A. & other authors (2013). Assembling single-cell genomes and mini-metagenomes from chimeric MDA products. *J Comput Biol* **20**, 714–737.

Nylander, J. A. A. (2004). MrModeltest v2. Evolutionary Biology Centre, Uppsala University: Program distributed by the author.

Palomo, S., González, I., de la Cruz, M., Martín, J., Tormo, J. R., Anderson, M., Hill, R. T., Vicente, F., Reyes, F. & other authors (2013). Sponge-derived *Kocuria* and *Micrococcus* spp. as sources of the new thiazolyl peptide antibiotic Kocurin. *Mar Drugs* **11**, 1071–1086.

Podell, S. & Gaasterland, T. (2007). DarkHorse: a method for genome-wide prediction of horizontal gene transfer. *Genome Biol* **8**, R16.

Rothberg, J. M., Hinz, W., Rearick, T. M., Schultz, J., Mileski, W., Davey, M., Leamon, J. H., Johnson, K., Milgrew, M. J. & other authors (2011). An integrated semiconductor device enabling non-optical genome sequencing. *Nature* **475**, 348–352.

Stamatakis, A. (2006). RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**, 2688–2690.

Subramani, R. & Aalbersberg, W. (2013). Culturable rare Actinomycetes: diversity, isolation and marine natural product discovery. *Appl Microbiol Biotechnol* **97**, 9291–9321.

Tang, X., Li, J., Millán-Aguiñaga, N., Zhang, J. J., O'Neill, E. C., Ugalde, J. A., Jensen, P. R., Mantovani, S. M. & Moore, B. S. (2015). Identification of thiotetronic acid antibiotic biosynthetic pathways by target-directed genome mining. *ACS Chem Biol* **10**, 2841–2849.

Tiwari, K. & Gupta, R. K. (2012). Rare actinomycetes: a potential storehouse for novel antibiotics. *Crit Rev Biotechnol* **32**, 108–132.

Trzoss, L., Fukuda, T., Costa-Lotufo, L. V., Jimenez, P., La Clair, J. J. & Fenical, W. (2014). Seriniquinone, a selective anticancer agent, induces cell death by autophagocytosis, targeting the cancer-protective protein dermcidin. *Proc Natl Acad Sci U S A* **111**, 14687–14692.

Tuomisto, H. (2010). A consistent terminology for quantifying species diversity? Yes, it does exist. *Oecologia* **164**, 853–860.

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

Wagner, M., Abdel-Mageed, W. M., Ebel, R., Bull, A. T., Goodfellow, M., Fiedler, H. P. & Jaspars, M. (2014). Dermacozines H-J isolated from a deep-sea strain of *Dermacoccus abyssi* from Mariana Trench sediments. *J Nat Prod* **77**, 416–420.

Weber, T., Blin, K., Duddela, S., Krug, D., Kim, H. U., Bruccoleri, R., Lee, S. Y., Fischbach, M. A., Müller, R. & other authors (2015). anti-SMASH 3.0 — a comprehensive resource for the genome mining of biosynthetic gene clusters. *Nucleic Acids Res* **43**, W237–W243.

Yamada, Y., Kuzuyama, T., Komatsu, M., Shin-ya, K., Omura, S., Cane, D. E. & Ikeda, H. (2015). Terpene synthases are widely distributed in bacteria. *Proc Natl Acad Sci U S A* **112**, 857–862.

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

Yang, X.-W., Zhang, G.-Y., Ying, J.-X., Yang, B., Zhou, X.-F., Steinmetz, A., Liu, Y.-H. & Wang, N. (2013). Isolation, characterization, and bioactivity evaluation of 3-((6-methylpyrazin-2-yl)methyl)-1*H*-indole, a new alkaloid from a deep-sea-derived actinomycete *Serinicoccus profundi* sp. nov. *Mar Drugs* **11**, 33–39.

Zhao, X. & Yang, T. (2011). Draft genome sequence of the marine sediment-derived actinomycete *Streptomyces xinghaiensis* NRRL B24674T. *J Bacteriol* **193**, 5543.

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

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

Zotchev, S. B. (2012). Marine actinomycetes as an emerging resource for the drug development pipelines. *J Biotechnol* **158**, 168–175.

Edited by: P. W. O'Toole and Y. Ohnishi

Sequencing Rare Marine Actinomycete Genomes Reveals High Density of Unique Natural Product Biosynthetic Gene Clusters

Supplementary Information

Michelle A. Schorn[1], Mohammad M. Alanjary[2], Kristen Aguinaldo[3], Anton Korobeynikov[4,5], Sheila Podell[1], Nastassia Patin[1], Tommie Lincecum[3], Paul R. Jensen[1,6] Nadine Ziemert[2] and Bradley S. Moore[1,6,7]*

[1] Center for Marine Biotechnology and Biomedicine, Scripps Institution of Oceanography, University of California, San Diego, USA.

[2] German Centre for Infection Research (DZIF), Interfaculty Institute for Microbiology and Infection Medicine Tuebingen (IMIT), University of Tuebingen, Germany

[3] Ion Torrent by Thermo Fisher Scientific, Carlsbad, California, USA.

[4] Center for Algorithmic Biotechnology, St. Petersburg State University, St. Petersburg, Russia.

[5] Department of Statistical Modeling, St. Petersburg State University, St. Petersburg, Russia.

[6] Center for Microbiome Innovation, University of California, San Diego, USA

[7] Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California, San Diego, USA.

*e-mail: bsmoore@ucsd.edu

Table S1. RMA Strains in this Study

NCBI Accession Number	JGI Taxon OID	Strain	Genus	Sequencing	Collection	Reference(s)
MKJY01000000	2675903202	CNU-125	*Actinomadura*	In-house	Palau / sediment	(Gontang *et al.*, 2010)
MKKH01000000	2675903201	CUA-896	*Cellulosimicrobium*	In-house	Mexico / sediment	(Patin *et al.*, 2016)
MKKI01000000	2675903203	CNJ-954	*Corynebacterium*	In-house	Palau / sediment	(Gontang *et al.*, 2010)
MKKG01000000	2596583509	CNJ-863	*Gordonia*	In-house	Palau / sediment	(Gontang *et al.*, 2010)
SRX873596*	2561511136	CNJ-787	*Kocuria*	JGI	Palau / sediment	(Gontang *et al.*, 2010)
MKJW01000000	2675903205	CNJ-770	*Kocuria*	In-house	Palau / sediment	(Gontang *et al.*, 2010)
MKKB01000000	2675903206	CUA-901	*Kytococcus*	In-house	Mexico / sediment	(Patin *et al.*, 2016)
GCA_000374985.1	2517572149	CNB-394	*Micromonospora*	JGI	Bahamas / sediment	(Mincer *et al.*, 2002)
SRX873600*	2563366738	CNS-044	*Nocardia*	JGI	Palau / sediment	(Gontang *et al.*, 2010)
GCA_000482385.1	2528311129	CNY-236	*Nocardia*	JGI	Fiji / sediment	New to this study
MKKC01000000	2675903207	CNR-923	*Nocardiopsis*	In-house	Palau / sediment	(Gontang *et al.*, 2010)
GCA_000381685.1	2519899670	CNS-639	*Nocardiopsis*	JGI	Fiji / sediment	New to this study
GCA_000515115.1	2515154089	CNT-312	*Nocardiopsis*	JGI	Fiji / sediment	New to this study
MKKA01000000	2675903208	CNJ-824	*Ornithinimicrobium*	In-house	Palau / sediment	(Gontang *et al.*, 2010)
MKJV01000000	2675903200	CNS-004	*Pseudonocardia*	In-house	Palau / sediment	(Gontang *et al.*, 2010)
MKJX01000000	2675903209	CNS-139	*Pseudonocardia*	In-house	Palau / sediment	(Gontang *et al.*, 2010)
MKKD01000000	2675903210	CUA-806	*Rhodococcus*	In-house	Mexico / sediment	(Patin *et al.*, 2016)
MKKE01000000	2675903211	CUA-673	*Saccharomonospora*	In-house	San Diego / sponge	(Patin *et al.*, 2016)
GCA_000527075.1	2515154179	CNQ-490	*Saccharomonospora*	JGI	San Diego / sediment	(Yamanaka *et al.*, 2014)
MKKF01000000	2675903068	CNJ-927	*Serinicoccus*	In-house	Palau / sediment	(Trzoss *et al.*, 2014)
MKJZ01000000	2675903212	CUA-874	*Serinicoccus*	In-house	Mexico / sediment	(Patin *et al.*, 2016)

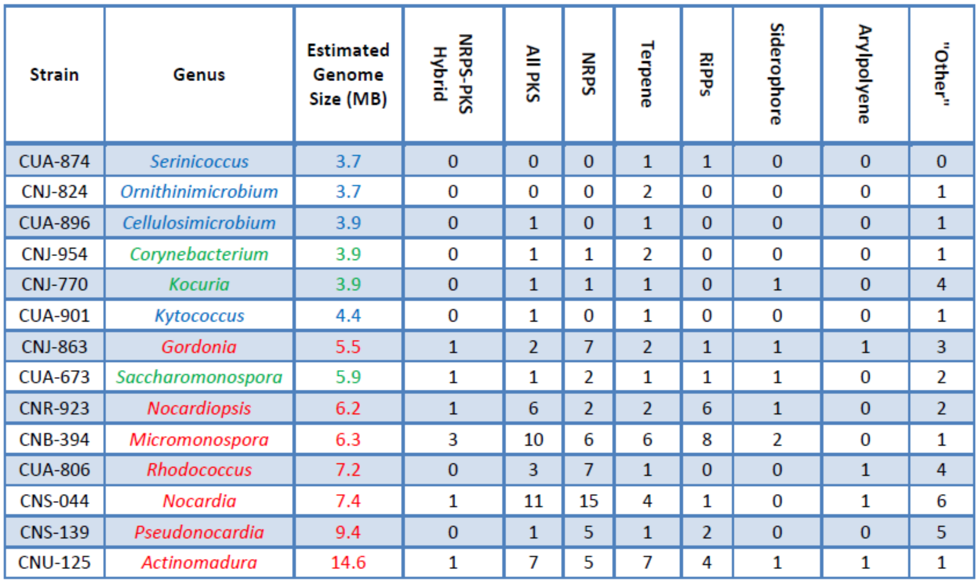
Table S1. NCBI accession numbers, JGI Organism ID (OID) numbers, genera, sequencing center, country and source of collection and previous references for all strains sequenced as part of this study are included in SI Table 1. Each strain is deposited and annotated in JGI and NCBI. *These accession numbers are for the NCBI Sequence Read Archive (SRA) database.

Table S3. antiSMASH 3.0 All Clusters

Strain	Genus	Total Clusters	Hybrid	Type 1 PKS	Type 2 PKS	Type 3 PKS	Other KS	NRPS	Terpene	Bacteriocin	Ectoine	Siderophore	Lantipeptide	Oligo-saccharide	Other	Butyrolactone	Phenazine	Nucleoside	Homoserine lactone	Arylpolyene	Indole	Hybrid Types
CNU-125	*Actinomadura*	44 (25)	2	4 (3)	3 (1)	2 (1)	2	18 (3)	6	1	1	1	3	0	0	0	0	0	0	1	0	OtherKS-T1PKS-NRPS, NRPS-Terpene
CUA-896	*Cellulosimicrobium*	3	0	0	0	1	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	
CNJ-954	*Corynebacterium*	5	0	1	0	0	0	1	2	0	0	0	0	0	1	0	0	0	0	0	0	
CNJ-863	*Gordonia*	20 (17)	2	1	1	0	0	9 (6)	2	1	1	0	0	0	2	0	0	0	0	1	0	NRPS-Siderophore, NRPS-T1PKS
CNJ-787	*Kocuria*	6	1	0	0	0	0	0	0	1	0	1	0	0	3	0	0	0	0	0	0	Terpene-T3PKS
CNJ-770	*Kocuria*	7	1	0	0	0	0	1	0	0	0	1	0	0	4	0	0	0	0	0	0	T3PKS-Terpene
CUA-901	*Kytococcus*	3	0	0	0	1	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	
CNB-394	*Micromonospora*	34 (27)	9	9 (4)	1	1	0	5 (3)	4	0	0	1	4	0	0	0	0	0	0	0	0	TransatPKS-NRPS-OtherKS, Oligosaccharide-NRPS-Terpene, NRPS-T1PKS, Siderophore-NRPS-Lantipeptide-T1PKS-OtherKS, NRPS-Lantipeptide-T1PKS, Lantipeptide-T2PKS, NRPS-T1PKS, Bacteriocin-Terpene
CNS-044	*Nocardia*	35	4	6	0	2	0	14	2	1	1	0	0	0	2	2	0	0	0	1	0	NRPS-Terpene, NRPS-T1PKS, T1PKS-TransatPKS-Oligosaccharide, Terpene-T1PKS-Butyrolactone
CNY-236	*Nocardia*	38 (31)	5	7 (5)	1	0	0	14 (9)	2	0	1	0	0	0	6	1	0	1	0	0	0	NRPS-Terpene, Amglyccycl-OtherKS, NRPS-Bacteriocin, NRPS-T1PKS, T1PKS-NRPS-OtherKS
CNR-923	*Nocardiopsis*	16	5	3	0	0	0	1	2	0	1	1	2	0	0	1	0	0	0	0	0	OtherKS-T2PKS, NRPS_Lantipeptide, T1PKS-NRPS, Thiopeptide-Lantipeptide, Lantipeptide-T1PKS
CNS-639	*Nocardiopsis*	24 (22)	6	1	1	1	0	5 (3)	2	0	1	1	1	1	1	2	1	0	0	0	0	Bacteriocin-Terpene, Lantipeptide-T1PKS, Lantipeptide-Oligosaccharide, Oligosaccharide-OtherKS-Lantipeptide, NRPS-T1PKS, Indole-NRPS
CNT-312	*Nocardiopsis*	22 (20)	0	3	0	0	0	7 (5)	2	3	3	0	3	0	0	0	1	0	0	0	0	
CNJ-824	*Ornithinimicrobium*	3	0	0	0	0	0	0	2	0	1	0	0	0	0	0	0	0	0	0	0	
CNS-004	*Pseudonocardia*	8	0	0	0	0	0	2	1	1	1	0	0	0	3	0	0	0	0	0	0	
CNS-139	*Pseudonocardia*	14 (13)	1	0	1	0	0	6 (5)	1	1	0	0	1	0	2	1	0	0	0	0	0	Arylpolyene-Butyrolactone
CUA-806	*Rhodococcus*	20 (15)	1	0	0	0	2	12 (7)	1	0	1	0	0	0	2	0	0	0	0	1	0	Oligosaccharide-T2PKS
CUA-673	*Saccharomonospora*	7	3	0	0	0	0	2	1	1	0	0	0	0	0	0	0	0	0	0	0	Oligosaccharide-T2PKS, Siderophore-Ectoine, NRPS-T1PKS
CNQ-490	*Saccharomonospora*	25 (21)	2	8 (4)	1	2	0	3	2	3	1	0	0	0	0	0	0	0	1	1	1	NRPS-T1PKS, T1PKS-Siderophore
CNJ-927	*Serinicoccus*	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	
CUA-874	*Serinicoccus*	2	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	

Table S3. This table includes the number of pathways for each category as output by antiSMASH 3.0. Genera are colored by total pathway number, with blue = low, green = medium, red = high amount of clusters. Each hybrid cluster is detailed in the last column. For NRPS and PKS clusters joined by NaPDoS (Fig. S1), the final predicted number of putative clusters is in parentheses.

Table S4: antiSMASH 3.0 Table used for Circos



Strain	Genus	Estimated Genome Size (MB)	NRPS-PKS Hybrid	All PKS	NRPS	Terpene	RiPPs	Siderophore	Arylpolyene	"Other"
CUA-874	*Serinicoccus*	3.7	0	0	0	1	1	0	0	0
CNJ-824	*Ornithinimicrobium*	3.7	0	0	0	2	0	0	0	1
CUA-896	*Cellulosimicrobium*	3.9	0	1	0	1	0	0	0	1
CNJ-954	*Corynebacterium*	3.9	0	1	1	2	0	0	0	1
CNJ-770	*Kocuria*	3.9	0	1	1	1	0	1	0	4
CUA-901	*Kytococcus*	4.4	0	1	0	1	0	0	0	1
CNJ-863	*Gordonia*	5.5	1	2	7	2	1	1	1	3
CUA-673	*Saccharomonospora*	5.9	1	1	2	1	1	1	0	2
CNR-923	*Nocardiopsis*	6.2	1	6	2	2	6	1	0	2
CNB-394	*Micromonospora*	6.3	3	10	6	6	8	2	0	1
CUA-806	*Rhodococcus*	7.2	0	3	7	1	0	0	1	4
CNS-044	*Nocardia*	7.4	1	11	15	4	1	0	1	6
CNS-139	*Pseudonocardia*	9.4	0	1	5	1	2	0	0	5
CNU-125	*Actinomadura*	14.6	1	7	5	7	4	1	1	1

Table S4. This table was used to create the Circos diagram (Fig. 1); it includes a representative genome from each genus sequenced and run through antiSMASH 3.0 without ClusterFinder. PKS and NRPS clusters that could be connected by NaPDoS are included in this table (Fig. S1). Hybrid clusters were separated into their composite categories (i.e. an NRPS-Siderophore hybrid is split into an NRPS cluster and a Siderophore cluster) to better assess the spread of cluster categories across genera. The hybrid category now only contains NRPS-PKS hybrids. All types of PKS clusters were also collapsed into one category. Lantipeptide, Bacteriocin, Thiopeptide, and Lassopeptide are collapsed into the category "RiPPs" (Ribosomally synthesized and Post-translationally modified Peptides). All minor categories, present in less than 5 genomes, were collapsed into the "Other" category, along with clusters designated by antiSMASH as Other. Ectoine was also included in the "Other" category, although ectoine clusters were present in 13/21 genomes. Those categories included in the "Other" category are: Other, Ectoine, Oligosaccharide, Butyrolactone, Phenazine, Nucleoside, Homoserine lactone, Aminoglycoside and Indole.

Figure S1. NaPDoS Cluster Connection

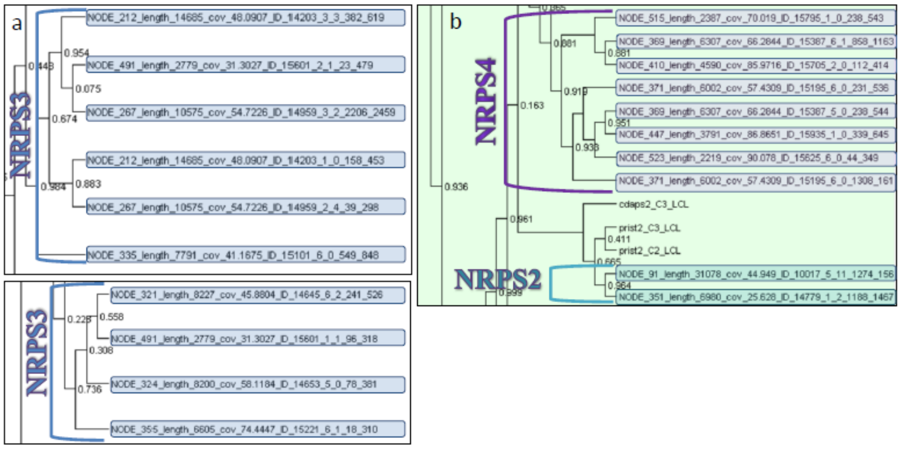
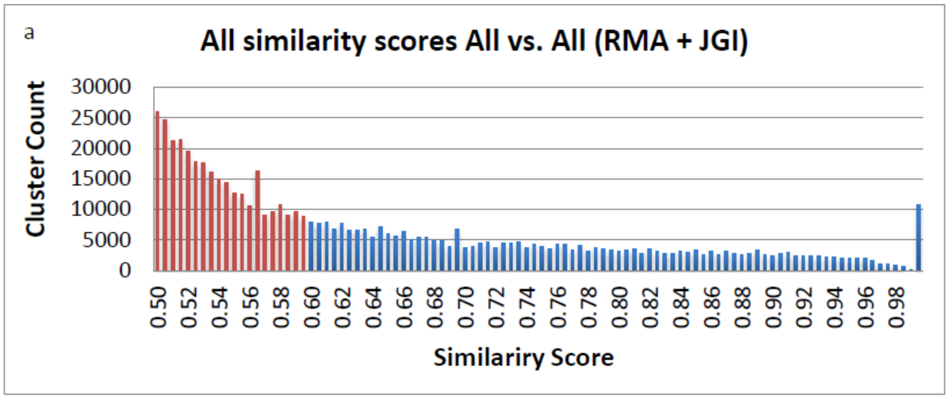

a
NRPS3
NODE_212_length_14685_cov_48.0907_ID_14203_3_3_382_619
0.954
0.448
NODE_491_length_2779_cov_31.3027_ID_15601_2_1_23_479
0.075
0.674
NODE_267_length_10575_cov_54.7226_ID_14959_3_2_2206_2459
NODE_212_length_14685_cov_48.0907_ID_14203_1_0_158_453
0.984
0.883
NODE_267_length_10575_cov_54.7226_ID_14959_2_4_39_298
NODE_335_length_7791_cov_41.1675_ID_15101_6_0_549_848
NRPS3
NODE_321_length_8227_cov_45.8804_ID_14645_6_2_241_526
0.228
0.558
NODE_491_length_2779_cov_31.3027_ID_15601_1_1_96_318
0.308
NODE_324_length_8200_cov_58.1184_ID_14653_5_0_78_381
0.736
NODE_355_length_6605_cov_74.4447_ID_15221_6_1_18_310
b
0.865
NODE_515_length_2387_cov_70.019_ID_15795_1_0_238_543
0.881
NODE_369_length_6307_cov_66.2844_ID_15387_6_1_858_1163
0.881
NODE_410_length_4590_cov_85.9716_ID_15705_2_0_112_414
NRPS4
NODE_371_length_6002_cov_57.4309_ID_15195_6_0_231_536
0.919
0.163
NODE_369_length_6307_cov_66.2844_ID_15387_5_0_238_544
0.951
NODE_447_length_3791_cov_86.8651_ID_15935_1_0_339_645
0.938
NODE_523_length_2219_cov_90.078_ID_15625_6_0_44_349
0.936
NODE_371_length_6002_cov_57.4309_ID_15195_6_0_1308_161
cdaps2_C3_LCL
0.961
prist2_C3_LCL
0.411
prist2_C2_LCL
0.665
NRPS2
NODE_91_length_31078_cov_44.949_ID_10017_5_11_1274_156
0.964
0.999
NODE_351_length_6980_cov_25.628_ID_14779_1_2_1188_1467

Figure S1. NaPDoS was used to connect NRPS and PKS clusters split onto two or more contigs. For example, *Actinomadura* CNU-125 NRPS3 (a), NRPS4 (b) and NRPS2 (b) clusters are made up of multiple sister taxa condensation domain sequences present on separate nodes (contigs). Secondary metabolite gene clusters can be inherited through horizontal gene transfer from other phylogenetically distant bacteria. The transferred gene cluster harbors the genetic signature of its historical relative and thus contigs/scaffolds containing pieces of one gene cluster are likely to phylogenetically clade together. While this is not always the case, it is a good tool to narrow down a more accurate number of NRPS/PKS pathways present in fragmented next-generation sequencing assemblies. Genomes with more than three NRPS or PKS clusters, as identified by antiSMASH 3.0 without ClusterFinder, were submitted to NaPDoS and KS and/or C domains were identified and NaPDoS constructed a tree. If the cluster was in the middle of a contig (i.e. has sequence before and after the region antiSMASH identified), it is considered complete. If domains on different contigs were sister taxa in the NaPDoS outputted tree, the clusters on the two or more contigs were considered part of one cluster. The total length of the prospective gene cluster was also taken into consideration. For each genome, the sum of the lengths of all clusters was divided by the average length of all the complete clusters. The resulting measure is the expected number of clusters based on an average length, specific to each genome. These estimates support the joining of clusters using NaPDoS.

Figure S2. Similarity Scores and Mismatch BGC Type suggest 0.6 cutoff



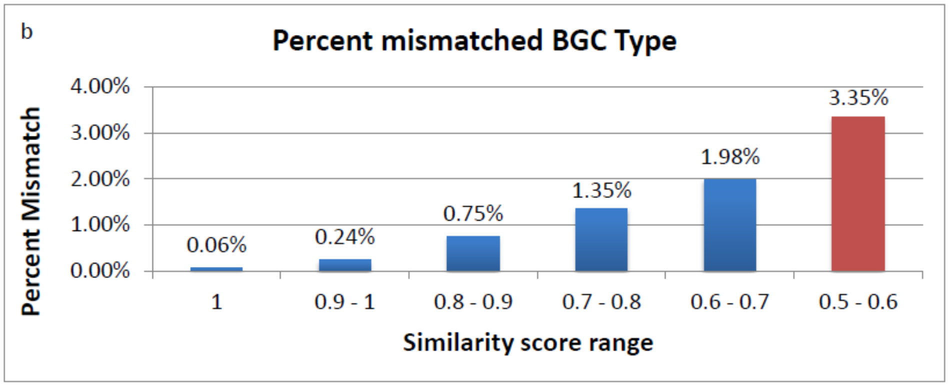
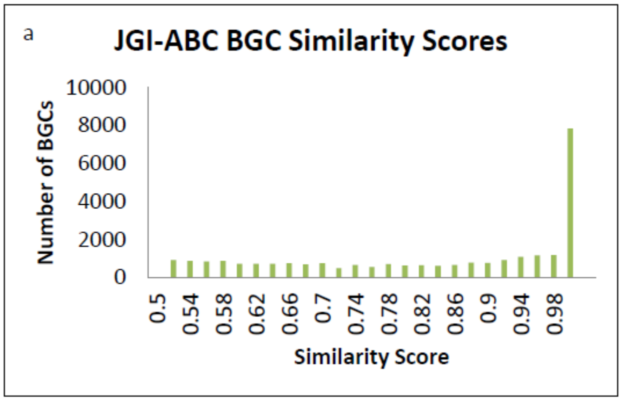


Figure S2. (a) All pairwise similarity scores are shown from 0.5 to 1. A threshold at 0.6 was chosen to significantly reduce the number of edges in the network. (b) Percent of pairwise connections where BGC Type did not match for nodes with BGC Type annotated by JGI. The raise in mismatches between connected nodes for the 0.5 to 0.6 similarity score range corroborates the 0.6 cutoff for clustering.

Figure S3. BGC Similarity Score Distributions



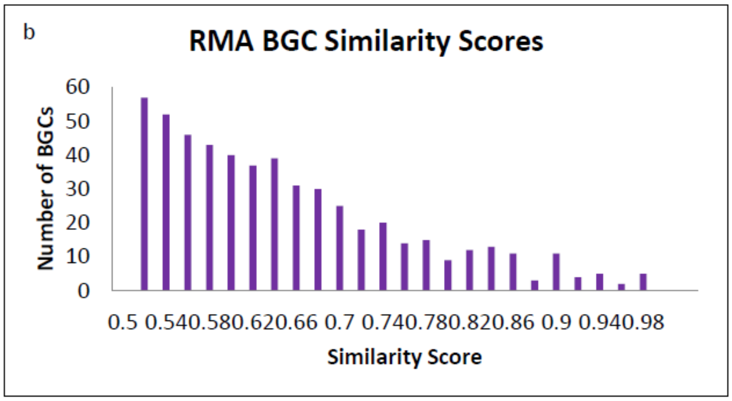


Figure S3. Calculated similarity scores >0.5 are shown for: (a) JGI-ABC clusters versus all clusters. Note the relatively flat distribution of scores. The spike in scores at 1 is due to replicate sequencings of the same strain, and were de-replicated in the similarity network. Similarity scores shown in (b) are from RMA clusters versus all clusters, including self-similarity between RMA clusters. Note that the RMA scores skew more heavily toward lower similarity scores, suggesting that they are more unique than what is currently available in the JGI-ABC database.

The 24 marine *Streptomyces* strains included in the comparison against RMA strains (Table 1) are: *Streptomyces* spp. CNB-091, CNB-632, CNH-099, CNH-189, CNH-287, CNQ-525, CNQ-329, CNQ-766, CNQ-865, CNR-698, CNS-335, CNS-606, CNS-615, CNT-302, CNT-318, CNT-360, CNT-371, CNT-372, CNX-435, CNY-228, CNY-243, TAA-040, TAA-204, and TAA-486.

Table S5. Networking Breakdown by Genus

	RMA							In-Network Cluster Diversity		JGI - Not Marine						In-Network Cluster Diversity	
Genus	Total Clusters	# of Strains	Clusters in network	# of GCFs	% Clusters in Network	# Unique genus GCFs	% New genus GCFs	True Diversity (q=1, D1)	True Diversity / Cluster	Total Clusters**	# of Strains	Clusters in Network	# of GCFs	% Clusters in Network	RMA GCFs shared	True Diversity (q=1, D1)	True Diversity / Cluster
Serinicoccus	99	4	43	21	43%	21	100 %	20.49	**0.4764**	0	0	0	0	0%	0	N/A	**N/A**
Nocardiopsis	292	4	57	38	20%	26	68%	34.93	**0.6128**	1147	18	163	51	14%	12	25.99	**0.1595**
Actinomadura	153	1	27	8	18%	5	63%	4.46	**0.1653**	696	7	81	27	12%	3	20.80	**0.2568**
Saccharomonospora	221	4	40	28	18%	20	71%	24.23	**0.6057**	365	8	77	35	21%	8	26.46	**0.3436**
Pseudonocardia	186	2	11	7	6%	4	57%	6.64	**0.6040**	613	7	35	20	6%	3	16.19	**0.4625**
Micromonospora	209	4	100	52	48%	13	25%	41.97	**0.4197**	1886	40	701	146	37%	39	76.49	**0.1091**
Ornithinimicrobium	35	1	2	2	6%	2	100%	2.00	**1.0000**	27	1	1	1	4%	0	1.00	**1.0000**
Kocuria	123	3	27	17	22%	11	65%	16.16	**0.5984**	131	6	17	13	13%	6	12.27	**0.7217**
Kytococcus	49	2	5	3	10%	3	100%	2.59	**0.5173**	0	0	0	0	0%	0	N/A	**N/A**
Cellulosimicrobium	36	1	2	2	6%	0	0%	2.00	**1.0000**	170	7	39	19	23%	2	16.04	**0.4114**
Nocardia	289	3	57	28	20%	11	39%	18.28	**0.3207**	3718	36	716	176	19%	17	49.23	**0.0688**
Corynebacterium	68	3	19	17	28%	4	24%	17.66	**0.9296**	2165	143	291	84	13%	13	46.52	**0.1598**
Gordonia	66	1	20	17	30%	1	6%	15.16	**0.7579**	1665	31	410	87	25%	16	37.83	**0.0923**
Rhodococcus	244	3	69	46	29%	5	11%	26.72	**0.3872**	3565	46	1094	242	31%	41	95.52	**0.0873**

**After de-replication

Table S5. This table breaks down the number of BGCs and GCFs by genus, comparing the RMA strains (from this study, as well as those labelled as marine in JGI: Figure S5) against the same genera from the JGI-ABC database. Novel contributions, in the form of GCFs not previously present in the JGI-ABC database for that genus, can be seen for each genus sampled in this study. True diversity was calculated according to equation (3) in (Jost & Baños, 2016). This equation is the exponent of the Shannon Index when q = 1.

Figure S4. RMA and Marine-derived *Streptomyces* Network

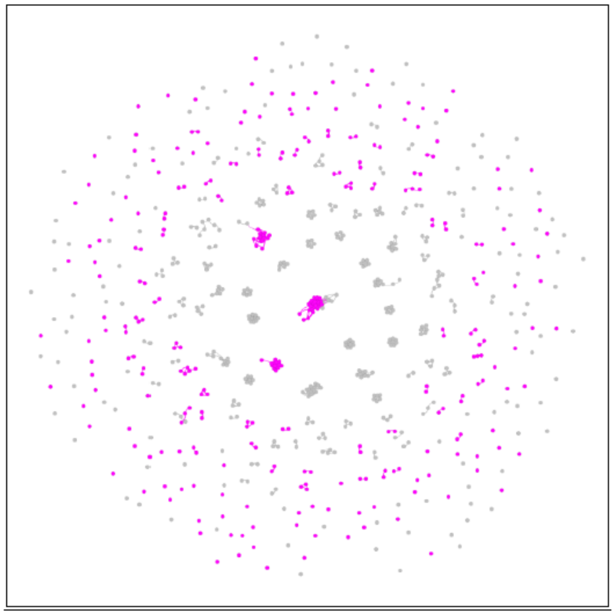


Figure S4. This BGC network includes the 21 RMA strains sequenced as part of this study and 24 marine-derived *Streptomyces* strains from the JGI database. Those nodes colored in pink are BGCs from RMAs and marine-derived *Streptomyces* BGCs are in grey. Notice that there is little overlap between RMA and marine-derived *Streptomyces* BGCs.

Figure S5. Environments/Sources of Genomes in each Genus

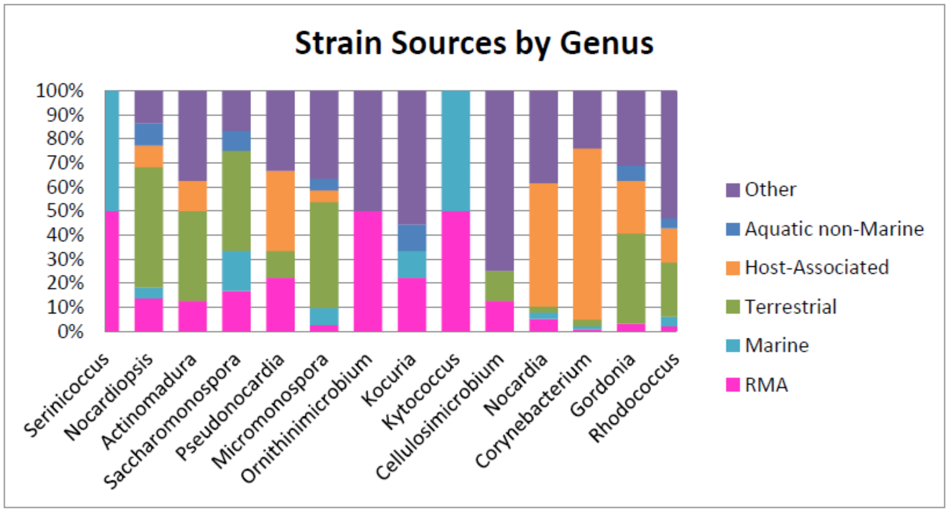


Figure S5. This stacked bar graph shows the sources of genome sequenced strains in JGI for each genus studied. RMA genomes from this study are colored in pink. JGI genomes were categorized by scanning all metadata fields. If no metadata was present, the genome was categorized as Other, so it is possible that marine genomes were included in the "non-marine" JGI-ABC calculations of True Diversity. Strains with species name "marina" (i.e. *Micromonospora marina*) with no metadata were looked up and determined as marine. These designated marine genomes were excluded when calculating Total Diversity in SI Table 6.

Figure S6. Phylogenetic Similarity vs. Shared GCFs

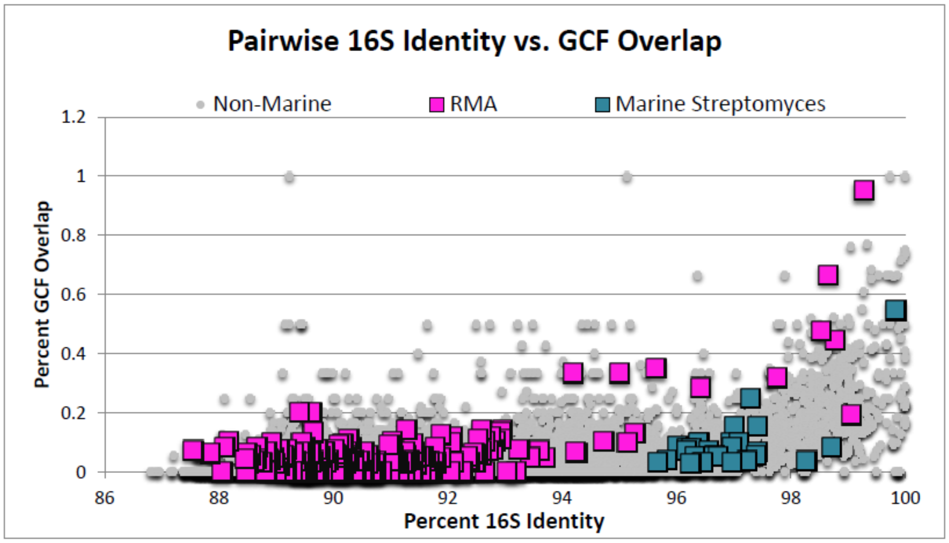


Figure S6. Each point represents a pairwise distance of 16S rRNA percent identity vs the GCF overlap between two genomes. Each pair is part of a larger group: non-marine JGI genomes from the genera examined in this study (grey), RMA genomes from this study (pink), and marine streptomycetes (teal).

<u>Supplementary References</u>

Gontang, E., Gaudencio, S., Fenical, W. & Jensen, P. (2010). Sequence-based analysis of secondary metabolite biosynthesis in marine Actinobacteria. *Appl Environ Microbiol* **76**, 2487-2499.

Jost, L. & Baños, T., Ecuador (loujost@yahoo.com). (2016). Entropy and diversity. *Oikos* **113**, 363-375.

Land, M. L., Hyatt, D., Jun, S. R., Kora, G. H., Hauser, L. J., Lukjancenko, O. & Ussery, D. W. (2014). Quality scores for 32,000 genomes. *Stand Genomic Sci* **9**, 20.

Mincer, T. J., Jensen, P. R., Kauffman, C. A. & Fenical, W. (2002). Widespread and persistent populations of a major new marine actinomycete taxon in ocean sediments. *Appl Environ Microbiol* **68**, 5005-5011.

Patin, N. V., Duncan, K. R., Dorrestein, P. C. & Jensen, P. R. (2016). Competitive strategies differentiate closely related species of marine actinobacteria. *Isme j* **10**, 478-490.

Trzoss, L., Fukuda, T., Costa-Lotufo, L. V., Jimenez, P., La Clair, J. J. & Fenical, W. (2014). Seriniquinone, a selective anticancer agent, induces cell death by autophagocytosis, targeting the cancer-protective protein dermcidin. *Proc Natl Acad Sci U S A* **111**, 14687-14692.

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

3.4 Acknowledgements

This work was planned and conceptualized by Michelle A. Schorn, Dr. Brad Moore, and Dr. Tommie Lincecum. Twenty-two strains of rare marine actinomycetes were obtained from Dr. Paul Jensen and Dr. Bill Fenical. Five of the strains were isolated and provided by Dr. Nastassia Patin. Dr. Lincecum and Kristen Aguinaldo of Ion Torrent helped with genome sequencing. Dr. Anton Korobeynikov from Dr. Pavel Pevzner's lab, assembled the genomes using SPAdesIT before it was released to the public. Dr. Sheila Podell performed the genome completeness assessment of the final assemblies. Mohammad Alanjary, of Dr. Nadine Ziemert's lab, directed the bioinformatic analyses of biosynthetic gene clusters. Dr. Ziemert provided guidance and suggestions for phylogenetic and bioinformatic analyses and Dr. Jensen provided helpful advice about growing marine actinomycetes and insights into their genomes

Chapter 3, in full, is a reprint of materials as it appears in "Sequencing rare marine actinomycete genomes reveals high density of unique natural product biosynthetic gene clusters" in *Microbiology*, 2016. Schorn, M. A., Alanjary, M. M., Aguinaldo, K., Korobeynikov, A., Podell, S., Patin, N., Lincecum, T., Jensen, P. R., Ziemert, N. and Moore, B. S. The dissertation author was the primary investigator and author of this manuscript.

3.5 Chapter 3 Appendix

Genome mining the twenty-two strains described in Chapter 3 led to the discovery of two pathways chosen for pathway capture and expression using transformation associated recombination (TAR)[1-3]. Although both pathways were successfully captured using TAR, subsequent integration and expression in heterologous hosts was unsuccessful. These biosynthetic gene clusters and the attempts to express them are described here.

1. Type 2 PKS (T2PKS10) + glycosylation genes (GLY) from *Nocardia* sp. CNY236

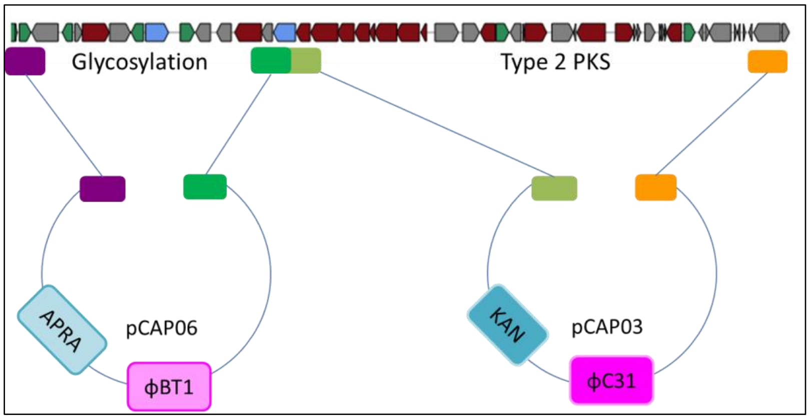


Figure 5. Dual capture scheme for the T2PKS10+GLY gene cluster from *Nocardia* sp. CNY236

A strategy for dual capture and integration is depicted here. The T2PKS portion of the cluster was captured using pCAP03, while the glycosylation and other tailoring genes were captured on pCAP06. The vectors are designed such that they can both be integrated in a host organism simultaneously.

pCAP06 Vector Construction

A T2PKS gene cluster was identified by antiSMASH[4] and upon further inspection, was found to have a set of glycosylation genes putatively involved in making and appending a sugar group to the PKS core. Because the T2PKS portion and the neighboring glycosylation genes were too large for capture together (about 80 kb), we decided to capture and express them separately. This approach however necessitated the design of a new capture vector for a double integration system (Figure 5). This new vector, named pCAP06 (Figure 6), was designed to work in tandem with pCAP03, as they have different antibiotic resistance markers and different integrase sites. pCAP06 was constructed via Gibson Assembly with three pieces. The first two were amplified from pCC02 (pCAP02), a first generation TAR vector, similar to pCC01 (pCAP01), but with apramycin resistance and the φBT1 integrase site. In order to incorporate the counter-selection introduced in the second generation TAR vector, pCAP03, the continuous URA3 and pADH site was amplified from pCAP03 using primers designed to have overlapping regions with the ends of the pCC02 amplified pieces. Because the apramycin cassette has an XhoI cut site in it, this vector cannot be digested with XhoI and NdeI for use in TAR. Instead, it can be amplified in two overlapping pieces and combined with overlapping capture arms using a three piece Gibson Assembly.

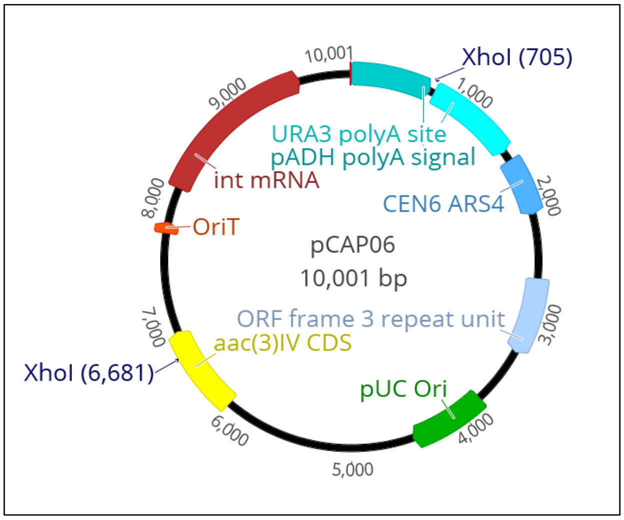


Figure 6. Vector map of pCAP06

pCAP06 replaces the kanamycin resistance cassette in pCAP03 with an apramycin resistance cassette aac(3)IV (in yellow). Additionally, the φC31 integrase is replaced with the φBT1 integrase (in red). All other elements of pCAP03 are retained.

TAR capture of T2PKS and glycosylation genes and integration attempts

The TAR 2.0 protocol[2] was followed to capture both the T2PKS portion of the cluster in pCAP03 and the glycosylation genes in pCAP06. Confirmation of the captured clusters was verified by restriction enzyme digest. Isolated plasmids pCAP03/PKS10 and pCAP06/GLY were taken to the Ziemert lab at the University of Tübingen for integration and heterologous expression in *Amycolatopsis japonica*. This host strain was chosen for its closer phylogenetic relationship to *Nocardia* and it had been recently developed as a suitable heterologous host for *Nocardia* pathways[5,6].

A sequential integration approach was taken with the two parts of the cluster. A first attempt used tri-parental conjugation as described in Flett *et al*[7] and mycelial conjugation as described by Stegmann *et al*[8] for both pCAP03/PKS10 and pCAP06/GLY in *A. japonica*, kindly provided by the Evi Stegmann at the University of Tübingen. Only the pCAP06/GLY integration was successful in this first attempt, verified by sequencing multiple regions of the GLY cluster from DNA isolated from the heterologous host. Multiple subsequent attempts to integrate pCAP03/PKS10 into *A. japonica* with mycelial and spore conjugation failed. A possible reason was that the glycosylation portion of the gene cluster contained some exporters that may be necessary to transport a toxic PKS molecule out of the cell. Therefore, *A. japonica* + pCP06/GLY was prepared as a host strain for pCAP03/PKS10. Integration of pCAP03/PKS10 into *A. japonica* + pCAP06/GLY was seemingly successful by the presence of exconjugants after every attempt at conjugation. However, each time these exconjugants were screened, they contained only the GLY portion of the gene cluster, and never the PKS portion, yet they had resistance to both kanamycin and apramycin. The apramycin resistance cassette could not be amplified from the exconjugants, suggesting possible cross resistance between kanamycin and apramycin.

Replacement of kanamycin resistance with hygromycin resistance in pCAP03/PKS10

After discussing cross resistance of kanamycin and apramycin with Leonard Kaysser and Bertold Gust at the University of Tübingen, it appears that cross resistance to the two antibiotics can occur when apramycin resistance is introduced first and then kanamycin resistance. That is, strains carrying apramycin resistance can also be resistant to kanamycin, but the opposite is not always observed. Because integration of pCAP03/PKS10 (with kanamycin resistance) could never be achieved, possibly due to self-toxicity issues, conjugation with the apramycin resistant plasmid had to occur first. To overcome this, the kanamycin resistance cassette was replaced with hygromycin resistance, using λ red recombination to "knock out" the kanamycin resistance gene, replacing it with hygromycin resistance. The PCR targeting protocol reported by Gust *et al*[9] detailing λ RED recombination in *Streptomyces* was used. The resulting plasmid, pCAP03/hygΔkan-PKS10, was used for conjugation attempts in *A. japonica* + pCAP06/GLY. More than six conjugation attempts were made. Again, exconjugants grew and continued to grow after multiple re-streakings, but they never carried the pCAP03/hygΔkan-PKS10 plasmid. Some exconjugants began producing a dark blue/black pigment, but again were devoid of pCAP03/hygΔkan-PKS10. All plasmids and strains used in these experiments have been stored in replicates and saved at -20°C or -80°C.

Future Directions

It is puzzling that exconjugants appear after conjugation in *A. japonica* + GLY with pCAP03/hygΔkan-PKS10 and grow when re-streaked, but never contain the PKS portion of the gene cluster. It is possible that *A. japonica* + GLY has weak resistance to hygromycin, although tests of *A. japonica* WT showed no resistance to hygromycin. Future conjugation attempts should check for presence of the hygromycin resistance cassette in exconjugants, as it is possible that the plasmid is partially conjugated. Another antibiotic could be substituted for hygromycin, or increased amounts of hygromycin could be used. Alternatively, the PKS cluster is not able to integrate into *A. japonica*. It could be that it is too toxic to the host organism, and necessary resistance genes were not captured in either cluster.

2. NRPS20 gene cluster from *Nocardia* sp. CNS044

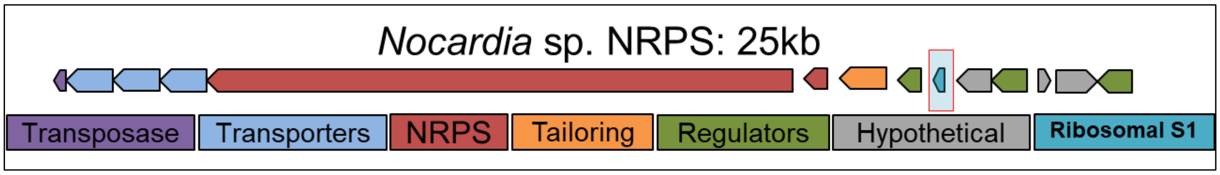


Figure 7. NRPS20 cluster from *Nocardia* sp. CNS044

Figure 7 shows the NRPS20 gene cluster as identified by antiSMASH. It consists of one large NRPS gene with four A domains and no specificities predicted. An additional partial copy of the ribosomal S1 subunit is also present in the cluster (shown in a teal box) and could be a possible resistant target.

Selection of NRPS cluster, TAR capture and initial conjugation

A small NRPS gene cluster (~25kb) was identified by antiSMASH (Figure 7). It contains one large NRPS gene with four A domains with unpredictable specificities. Three regulators precede the cluster and are designated as PadR, TetR1, and TetR2. Upstream of the PadR regulator is an oxidoreductase and a standalone A domain. It also contains a duplicate copy of the ribosomal S1 subunit (RS1). The ribosomal S1 subunit is not currently a target of any antibiotics and could be a putative resistant target if the product of this gene cluster encodes a ribosome inhibitor. Thus, this cluster was chosen for TAR capture and expression. TAR capture was successful in pCAP03 and the resulting plasmid, pCAP03/NRPS20 was integrated by triparental conjugation and spore conjugation as previously described into *Streptomyces coelicolor* M1152. However, HPLC analysis of the host with empty vector and host with pCAP03/NRPS20 did not identify any new peaks.

Conjugation in *A. japonica* and regulator knock outs

Because no production was observed in *S. coelicolor* M1152, integration in the more closely related host, *A. japonica*, was undertaken in the Ziemert lab. Triparental conjugation was successful in *A. japonica*, but still no production was seen. To check for transcription of the cluster, reverse transcriptase (RT) PCR was performed. RT-PCR primers were designed for the NRPS gene, the RS1 gene, the *tetR1* gene and the *padR* gene. *sigB* was used as a positive control. No transcription was seen in the heterologous host. The TetR1 and PadR regulators were individually deleted using λ red recombination. RNA was converted to cDNA with (+) RT and controls without (-) RT were performed. The designed PCR primers were used to amplify from both the + and – RT samples. No transcription was observed for the NRPS, *tetR1* and *padR* genes (Figure 8). Transcription of *sigB* in the + RT samples and not the – RT samples confirms that the cDNA conversion was successful. The transcription of RS1 was not able to be resolved due to a persistent band in the negative control. PadR regulators have been previously shown to be repressors of multi-drug resistance transport systems[10,11]. Three multi-drug resistance transporters are directly upstream of and translationally coupled to the NRPS gene, so repression of the transporters could also cause repression of the NRPS gene. Therefore, an ErmE promotor was used to replace PadR in an attempt to induce expression. RT-PCR of cDNA made from RNA from *A. japonica* WT, the *A. japonica* + pCAP03/NRPS20 knockout strain and the *A. japonica* + pCAP03/NRPS20ErmEΔPadR replacement strain revealed no transcription of the NRPS gene, but transcription of the RS1 and *sigB* genes (Figure 9).

Future Directions

It is unknown why the replacement of a putative repressor with a strong promoter did not successfully activate this silent cluster. The RT-PCR primers in the NRPS gene are far upstream, so it's possible that part of this large gene was being transcribed, but not the region where the RT-PCR primers were designed. Future experiments could check transcription in multiple position of the NRPS gene, the standalone A domain, the oxidoreductase, and in the *Nocardia* sp. CNS044 parent strain. RT-PCR primers were designed for the standalone A domain, the oxidoreductase, and the beginning and middle of the NRPS gene. An initial attempt at RT-PCR was unsuccessful. All negative controls had a band in them, suggesting that the cDNA had been contaminated. All plasmids and heterologous hosts with knockouts are stored in replicates and saved at -20°C or -80°C.

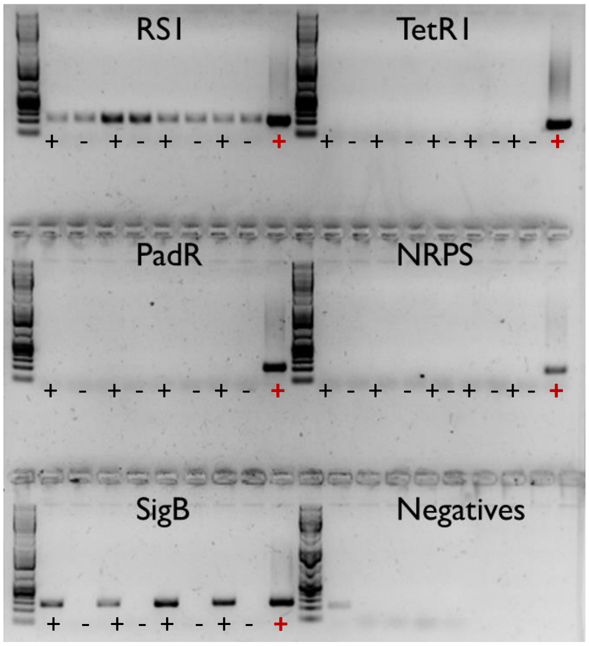


Figure 8. Transcription of five genes in wild type (WT) and regulator knockouts

This gel shows amplification of five genes (RS1, TetR1, PadR, NRPS, and SigB) from cDNA made from RNA extracted from the *A. japonica* WT (+) and without (-) RT (lanes one and two), *A. japonica* + pCAP03/NRPS20 + and - RT (lanes three and four), *A. japonica* + pCAP03/NRPS20ΔTetR1 + and - RT (lanes five and six) and *A. japonica* + pCAP03/NRPS20ΔPadR + and - RT (lanes seven and eight). Lane nine is the positive control for each gene. Negative controls are shown in the last gel section.

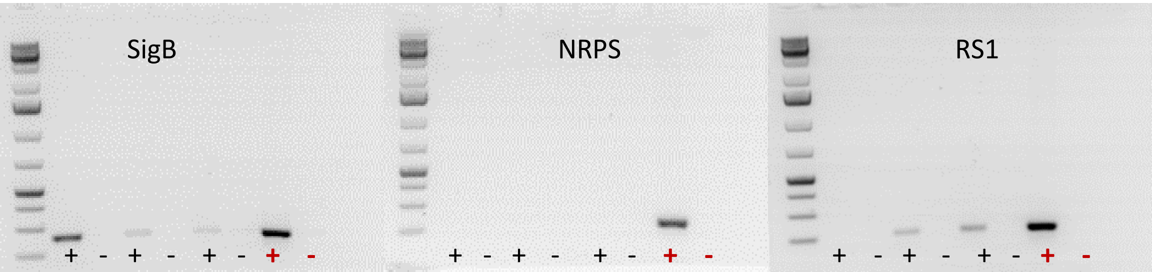


Figure 9. Transcription of X genes in WT and PadRΔErmE replacement

These gels show amplification with SigB, NRPS, and RS1 primers from cDNA made from RNA extracted from *A. japonica* WT + and - RT (lanes one and two), *A. japonica* + pCAP03/NRPS20 + and - RT (lanes three and four), and *A. japonica* + pCAP03/NRPS20ErmEΔPadR + and – RT (lanes five and six). The last two lanes in each set are the positive control and negative control, respectively.

Appendix References

1 Yamanaka, K., Reynolds, K. A., Kersten, R. D., Ryan, K. S., Gonzalez, D. J., Nizet, V., Dorrestein, P. C. & Moore, B. S. Direct cloning and refactoring of a silent lipopeptide biosynthetic gene cluster yields the antibiotic taromycin A. *Proc Natl Acad Sci U S A* **111**, 1957-1962, 2014. PMID: 24449899.

2 Tang, X., Li, J., Millan-Aguinaga, N., Zhang, J. J., O'Neill, E. C., Ugalde, J. A., Jensen, P. R., Mantovani, S. M. & Moore, B. S. Identification of thiotetronic acid antibiotic biosynthetic pathways by target-directed genome mining. *ACS Chem Biol* **10**, 2841-2849, 2015. PMID: 26458099.

3 Kouprina, N. & Larionov, V. Selective isolation of genomic loci from complex genomes by transformation-associated recombination cloning in the yeast Saccharomyces cerevisiae. *Nat Protoc* **3**, 371-377, 2008. PMID: 18323808.

4 Medema, M. H., Blin, K., Cimermancic, P., Jager, V. d., Zakrzewski, P., Fischbach, M. A., Weber, T., Takano, E. & Breitling, R. antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. 2011.

5 Schwarz, P. N., Buchmann, A., Roller, L., Kulik, A., Gross, H., Wohlleben, W. & Stegmann, E. The Immunosuppressant Brasilicardin: Determination of the Biosynthetic Gene Cluster in the Heterologous Host Amycolatopsis japonicum. *Biotechnol J* **13**, 2018. PMID: 29045029.

6 Schwarz, P. N., Roller, L., Kulik, A., Wohlleben, W. & Stegmann, E. Engineering metabolic pathways in Amycolatopsis japonicum for the optimization of the precursor supply for heterologous brasilicardin congeners production. *Synthetic and Systems Biotechnology*, 2018.

7 Flett, F., Mersinias, V. & Smith, C. P. High efficiency intergeneric conjugal transfer of plasmid DNA from Escherichia coli to methyl DNA-restricting streptomycetes. *FEMS Microbiol Lett* **155**, 223-229, 1997. PMID: 9351205.

8 Stegmann, E., Pelzer, S., Wilken, K. & Wohlleben, W. Development of three different gene cloning systems for genetic investigation of the new species Amycolatopsis japonicum MG417-CF17, the ethylenediaminedisuccinic acid producer. *J Biotechnol* **92**, 195-204, 2001. PMID: 11640989.

9 Gust, B., Challis, G. L., Fowler, K., Kieser, T. & Chater, K. F. PCR-targeted Streptomyces gene replacement identifies a protein domain needed for biosynthesis of the sesquiterpene soil odor geosmin. *Proc Natl Acad Sci U S A* **100**, 1541-1546, 2003. PMID: 12563033.

10 Huillet, E., Velge, P., Vallaeys, T. & Pardon, P. LadR, a new PadR-related transcriptional regulator from Listeria monocytogenes, negatively regulates the

expression of the multidrug efflux pump MdrL. *FEMS Microbiol Lett* **254**, 87-94, 2006. PMID: 16451184.

11 Fibriansah, G., Kovacs, A. T., Pool, T. J., Boonstra, M., Kuipers, O. P. & Thunnissen, A. M. Crystal structures of two transcriptional regulators from Bacillus cereus define the conserved structural features of a PadR subfamily. *PLoS One* **7**, e48015, 2012. PMID: 23189126.

12 Kieser, T., Bibb, M., Buttner, M., Chater, K. & Hopwood, D. *Practical Streptomyces Genetics*. (John Innes Foundation, 2000).

.

Chapter 4: Uncultured Cyanobacterial Symbionts of Marine Sponges and Their Natural Products

4.1 Introduction

4.1.1 Diversity of Chemistry from *Lamellodysidea herbacea* Sponges

Marine sponges of the family Dysideidae are widely distributed in tropical and subtropical waters and have been investigated as a source of diverse natural products for over forty years[1]. The *Lamellodysidea* (formerly *Dysidea*) genus within the Dysideidae family is particularly prolific in the variety of bioactive natural products that have been isolated (Figure 10). The most well-known class of compounds, due to the environmental toxicity of their anthropogenic counterparts, are the polybrominated diphenyl ethers (PBDEs), like **1**, first isolated in 1972 from *Dysidea herbacea*, and subsequently from numerous Dysideidae sponges[2-8]. Astonishingly, PBDEs can make up as much as 12% of the sponge's dry weight[9]. Polychlorinated molecules are often found in Dysideidae sponges, including amino acid derivatives, such as dysidin (**2**)[10], dysidenin (**3**)[11], the dysideathiazoles[9], herbaceamide[12], the dysideaprolines[13], herbamide A[14], and several diketopiperazines, like **4**[15-17]. Numerous distinct terpene molecules have been isolated from these sponges, including the sesquiterpenes: spirodysin (**5**) [18], herbadysidolide[19], lamellodysidines[20], hydroxybutenolide[21], furodysin (**6**)[22] and others, including an anti-fouling furano-sesquiterpene[23]. Two other diverse classes of compounds have been found in *Dysidea herbacea* samples: the dysibetaines, three novel betaines, like **7**[24], and multiple novel polyoxygenated sterols, like **8**[25,26]. A novel UV absorbing mycosporine-like amino acid (MAA) pigment (**9**), commonly found in cyanobacteria, was isolated from an Australian *Dysidea herbacea*[27]. Dysinosins A-D (dysinosin B **10**) were also isolated from Australian collections of *Dysidea herbacea* specimens [28,29]. The dysinosins strongly resemble the common cyanobacterial metabolites, the aeruginosides, which have a representative biosynthetic pathway already characterized from free-living cyanobacteria[30].

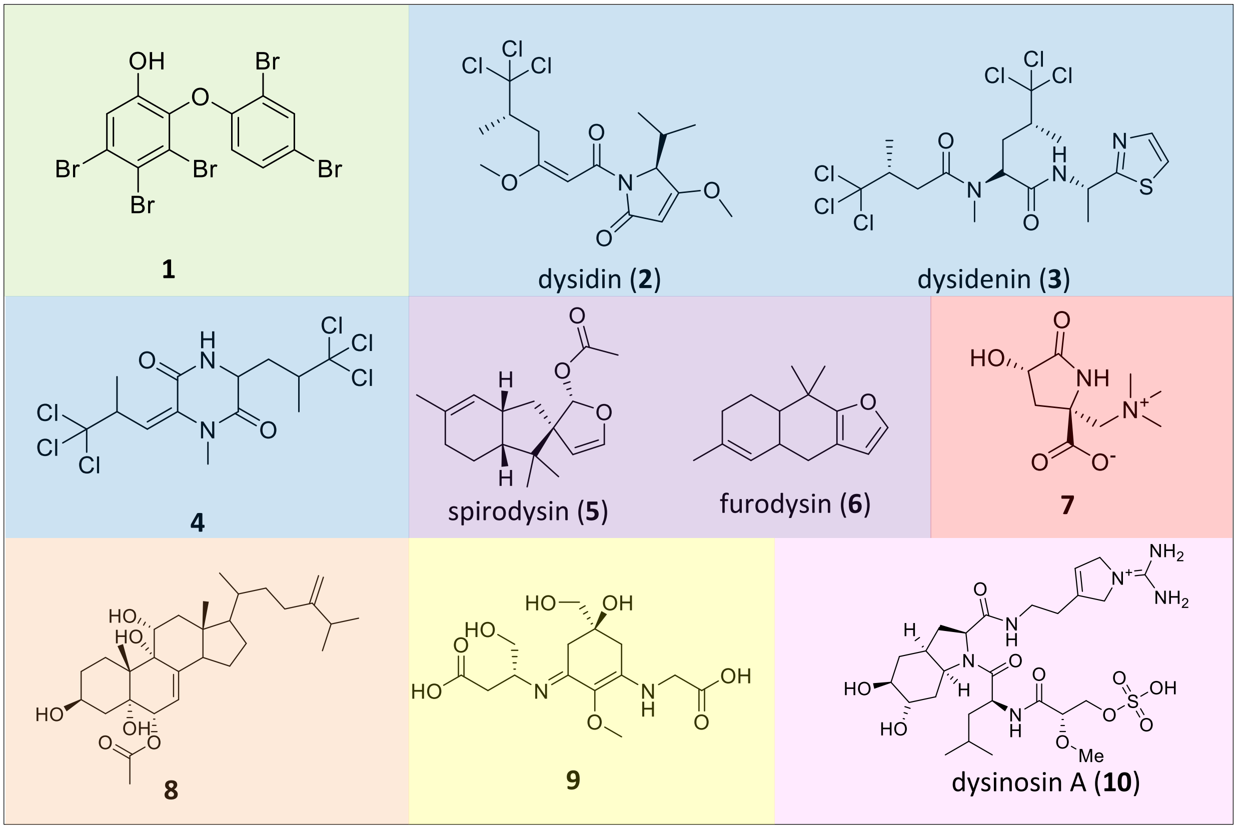


Figure 10. Secondary metabolites previously isolated from *Lamellodysidea herbacea*.

Representative examples of secondary metabolites from seven classes isolated from *L. herbacea* specimens. Brominated and chlorinated molecules are seen in the first two classes (**1-4**). Diverse sesquiterpines and other terpenoids are represented in the third class (**5-6**). Novel betaines, polyoxygenated sterols, and mycosporine-like amino acids make up the next three classes with one representative shown for each (**7-9**). Finally, highly modified NRPS compounds similar to aeruginosides make up the last class, with dysinosin A (**10**) shown here.

Such a wealth of chemistry from one family of sponges, with a focus on one species here, *Lamellodysidea herbacea*, is remarkable, but with the abundance of bacteria residing in sponges, the true source of these diverse compounds is difficult to decipher. In many cases, sequencing the dominant symbiont of an invertebrate assemblage gives indication that the bacterial symbiont is a major producer of natural products[31]. For example, in the marine ascidian, *Lissoclinum patella*, the dominant, uncultured cyanobacterial symbiont, *Prochloron didemni*, was demonstrated to be the producer of patellamides A and C through pathway sequencing and heterologous expression[32]. Genomic sequencing of the unculturable symbiont, *Entotheonella swinhoeii* of the sponge *Theonella swinhoeii*, has exposed the elusive symbiont as the producer of almost all polyketides and peptides isolated from these chemically rich sponges[33]. These and numerous other studies show that access to genomic information can readily inform the true producers of secondary metabolites in symbiotic systems. Recently, we reported the biosynthetic gene clusters responsible for PBDEs within the genomes of three sponge cyanobacterial symbionts, establishing bacterial biosynthesis of at least one major metabolite isolated from *L. herbacea* sponges[34].

4.1.2 *Hormoscilla spongelia* as an Obligate Symbiont and Natural Product Producer

A defining feature of *Lamellodysidea* sponges is the persistent presence of a filamentous cyanobacteria symbiont, *Hormoscilla spongeliae* (formerly *Oscillatoria spongeliae*)[35]. These symbionts have been shown to consistently inhabit all studied *L. herbacea* sponges, with distinct strains inhabiting morphologically discrete hosts[36,37]. Despite repeated attempts, these symbionts have been recalcitrant to culturing efforts and are therefore assumed to be obligate symbionts, unable to live outside their host system[38]. Despite their inability to be cultured, the cyanobacterial trichomes make up a large portion of the tissue volume of the sponge, up to 40%, and are easily isolated by simply squeezing the sponge tissue, which is fortuitous for obtaining enriched, albeit unviable, fractions of the cyanobacterial cells[35].

H. spongeliae symbionts have long been suspected to be the natural producers of some compounds isolated from *L. herbacea* sponges. Cell sorting studies showed that the chlorinated amino acid derivative 13-demethylisodysidenin was localized to the cyanobacterial cells, while the sesquiterpenes spirodysin and dysidenin were found to be co-localized with the sponge cells[39]. A subsequent cell sorting study showed that the presence of a PBDE molecule was limited to the cyanobacterial cells and not the sponge cells[40]. A more recent study used the *barB1* sequence from the tri-chloromethyl containing cyanobacterial secondary metabolite, barbamide, to amplify a homologue from an *L. herbacea* sponge. This homologue was used as a fluorescently labeled DNA probe, which CARD-FISH analysis showed was hybridizing to sequences within the cyanobacterial symbiont[41]. While these localization studies all suggest that *H. spongeliae* is the producer of various halogenated natural products isolated from *L. herbacea* sponges, genomic evidence is the most definitive method to determine the real producers of sponge natural products.

Here we report the >90% complete population genomes of two *H. spongeliae* populations, isolated from chemically and morphologically distinct *L. herbacea* sponges, collected in Guam. We used a hybrid sequencing and assembly approach with short-read Illumina sequencing and long-read PacBio sequencing of enriched cyanobacterial cell fractions. Insights from these genomes of uncultured symbionts shed light on their biosynthetic capacity for secondary metabolites and give clues to their seemingly obligate symbiont lifestyle. We also report the isolation and characterization of fourteen PBDEs (**11-24**), including penta- and hexa-brominated molecules not yet reported from Guamanian *L. herbacea* sponges. A unique halogenase in the new hs_bmp cluster may be responsible for the observed extra bromination. Additionally, two new putative dysinosins (**25-26**) were discovered using genome mining and LCMS comparison with standards.

4.2 Methods

4.2.1 Sponge Collection and Cyanobacteria Enrichment

The sponges used in the study are designated GUM007 and GUM202, with their corresponding symbionts named GUM007_hs and GUM202_hs. GUM007 was collected in July 2015 by snorkel in Pago Bay, located just outside of the University of Guam Marine Lab. GUM202 was collected in December 2016 by SCUBA diving in 20-40 feet of water at Anae Island in Guam. Both sponges were processed in the same way. Samples were taken back to the University of Guam Marine Lab and immediately processed as follows. Fresh sponge tissue was sectioned off using a sterile razor blade into sterile petri dishes. Sponge pieces were washed with 5mL of sterile artificial seawater. 5 mL of sterile artificial seawater or phosphate buffered saline were added to the sponge, and the sponge tissue squeezed, resulting in visible exudation of cyanobacterial trichomes. This liquid was carefully transferred to a 15 mL conical tube on top of 5 mL of RNAlater (25 mM Sodium Citrate, 10 mM EDTA, 70 g ammonium sulfate/100 ml solution, filtered, pH 5.2). These tubes were spun at 100xg for 5 minutes, yielding a dense dark green pellet of cyanobacterial trichomes, which were positively identified by light microscopy (Figure 11).

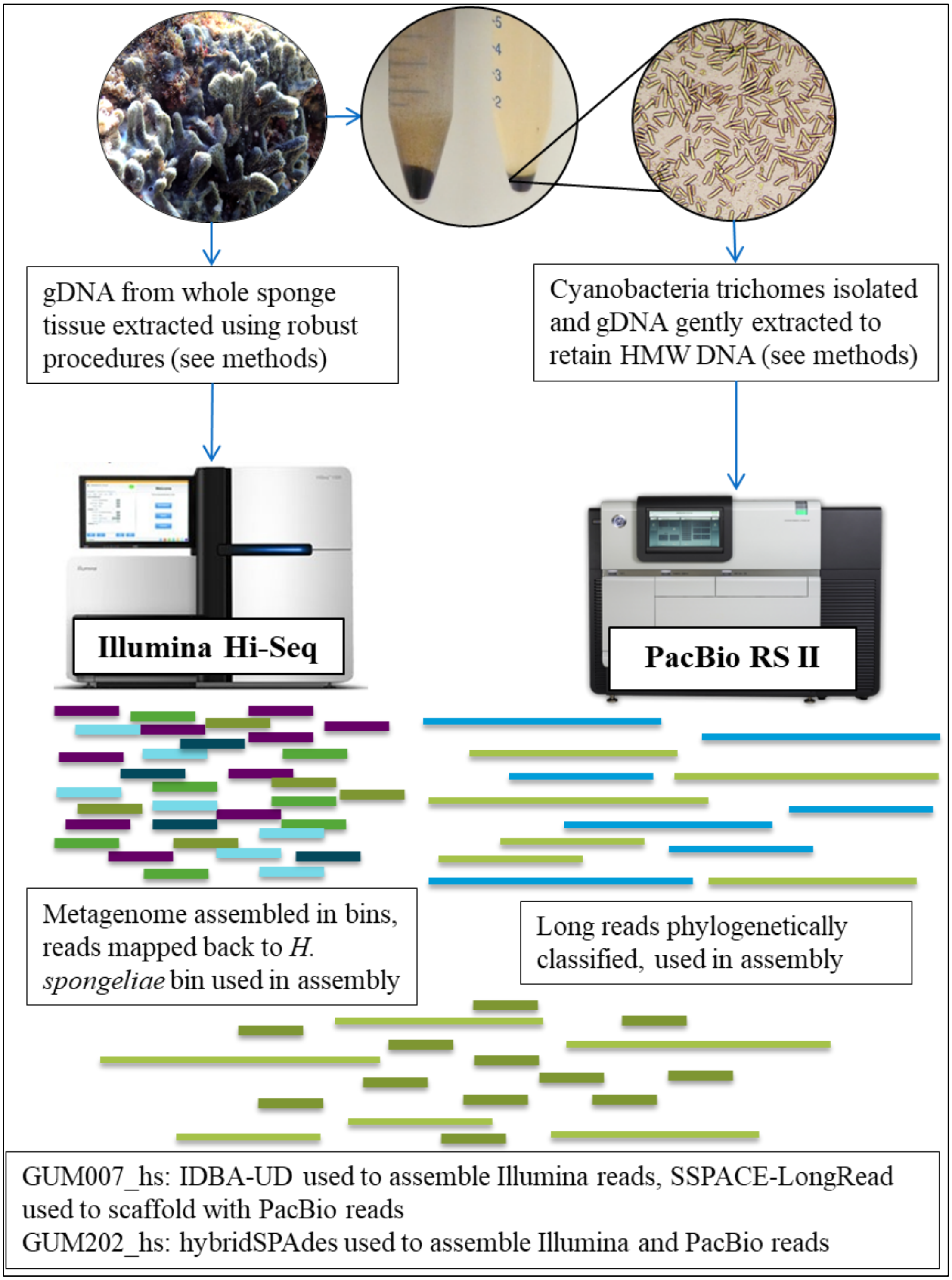

gDNA from whole sponge tissue extracted using robust procedures (see methods)
Cyanobacteria trichomes isolated and gDNA gently extracted to retain HMW DNA (see methods)
Illumina Hi-Seq
PacBio RS II
Metagenome assembled in bins, reads mapped back to *H. spongeliae* bin used in assembly
Long reads phylogenetically classified, used in assembly
GUM007_hs: IDBA-UD used to assemble Illumina reads, SSPACE-LongRead used to scaffold with PacBio reads
GUM202_hs: hybridSPAdes used to assemble Illumina and PacBio reads

Figure 11. Workflow for processing whole sponges and enriched cyanobacteria fractions

Simplified workflow showing parallel processing and sequencing for whole sponge metagenomes using Illumina and enriched cyanobacteria using PacBio, requiring high molecular weight (HMW) DNA for long-read sequencing. Two different hybrid assembly methods were used to obtain genome assemblies.

4.2.2 Enriched Genomic DNA Extraction and Sequencing

The enriched cyanobacterial fractions were subjected to genomic DNA isolation using a previously published protocol[42] to obtain high molecular weight DNA enriched for *H. spongeliae* from both GUM007 and GUM202. Isolated DNA was either left dry or resuspended in Tris-EDTA (TE) and frozen, and shipped back to the Scripps Institution of Oceanography along with frozen sponge tissue stored (1) dry, (2) in RNAlater, and (3) in Calcium-Magnesium-Free Artificial Seawater with EDTA and glycerol (449 mM NaCl, 33 mM Na2SO4, 9 mM KCl, 2.5 mM NaHCO3, 1 mM EDTA, 15% w/v glycerol, pH 8.0). Unenriched 147etagenomics DNA was extracted from GUM007 and GUM202 whole sponge tissue frozen in RNAlater as previously described[34]. Preparation of all libraries and sequencing was performed by the UC San Diego Institute for Genomic Medicine. Illumina libraries were constructed from unenriched 147etagenomics sponge DNA and sequenced as previously described[34]. PacBio libraries were constructed from the enriched, high molecular weight DNA isolated in Guam (Figure 11). PacBio sequencing libraries were generated using SMRTbell® Template Preparation Reagent Kits (Pacific Biosciences) and libraries >6 Kb were selected using a PippinHT® (Sage Science). Libraries were sequenced on a PacBio RS II® sequencer (UCSD IGM Genomics Center, La Jolla, CA) via four hour movies using the DNA/Polymerase Binding Kit Version P6 V2 with C4 sequencing chemistry. Whole sponge metagenomes for both GUM007 and GUM202 were sequenced following previously published methods[34].

4.2.3 Population Genome Assemblies and Quality Assessment

The GUM007_hs and GUM202_hs genomes were assembled using two slightly different methods (Figure 11). GUM007_hs was assembled first by using quality trimmed Illumina reads that had been classified by DarkHorse version 1.5[43] and binned, as previously described[34]. The PacBio reads were then used to scaffold the Illumina assembled genome using SSPACE-LongRead[44] to give the resulting assembly on 64 scaffolds. GUM202_hs was assembled by first assembling the GUM202 *L. herbacea* metagenome using Illumina reads and binning based on phylogenetic and coverage information. Illumina reads were then mapped to the *H. spongeliae* bin to give Illumina reads that should belong only to GUM202_hs. The PacBio reads were also phylogenetically classified by blasting against a custom made database of protein coding sequences from the GUM007_hs genome previously assembled and a database made of protein coding sequences for all *Oscillatoria* genomes in the Joint Genome Institute (JGI) database. Any PacBio reads that contained a hit to one of these databases was retained. The raw, binned Illumina reads and the raw positive hit PacBio reads were then used for assembly with hybridSPAdes[45]. The Illumina reads and the GUM202_hs hybrid assembly were then used as input for GapFiller to obtain the final assembly[46].

To determine the quality and completeness of genome assembly, CheckM was used[47]. In both cases, genome completeness exceeded 90% according to CheckM parameters. Average nucleotide identity (ANI) was calculated by submitting both genome assemblies to the ANI Calculator, hosted by the Kostas lab at http://enve-omics.ce.gatech.edu/ani/. Genome synteny was examined using Mauve version 20150226 build 10[48].

4.2.4 Phylogenetic Analysis

To place *H. spongeliae* in the broader context of sequenced cyanobacteria, we used a Multi-Locus Sequence Analysis (MLSA) to build a phylogenetic tree using twenty-five conserved housekeeping genes (Table 1) from a set of thirty one genes previously defined for bacterial MLSA[49]. We chose to use twenty-five genes in order to include a wider variety of genomes that did not have all thirty one genes. All 442 available cyanobacteria genomes in JGI as of January 22, 2018 were searched for the presence of the twenty-five housekeeping genes, of which 305 genomes contained each gene in one copy. After eliminating duplicate genomes and limiting redundant species to five representatives, a final set of 197 cyanobacteria genomes were used in the MLSA (see Supplementary Information). Because the GUM007_hs genome was submitted to JGI before their "Metagenome Assembled Genome" category was available, GUM007_hs is classified as a metagenome and was searched manually for all housekeeping genes using HMMER3[50], of which it contained all twenty five in single copy. *Chloroflexus aurantiacus* J-10-fl was used as an outgroup to root the tree. Each of the twenty five gene sets were individually aligned using MAFFT v7.310[51] with high accuracy local iterative mode using 100 iterations. Next, each alignment was trimmed using trimAl v1.2rev59[52] and the "automated1" option optimized for maximum-likelihood tree construction. The resulting trimmed, aligned files were concatenated using a custom python script and the resulting supermatix was processed with IQ-TREE v1.6.1[53] with 1000 ultra-bootstrap replicates using UF:Boot2[54] and ModelFinder[55] for each gene partition. The tree was visualized using interactive tree of life (iTOL) v3[56]. Bootstrap values under 95 are displayed in the tree.

Table 1. Cyanobacteria housekeeping genes used in MLSA

Twenty-five genes were chosen to construct a MLSA of cyanobacteria. These genes were chosen from a set of thirty one suggested for building bacterial phylogenies. Because many of the cyanobacteria in the JGI database do not have all thirty one genes in single copy, the data set was narrowed to 25 conserved genes to retain more genomes in the analysis.

Gene Name	Abbreviation	pfam or TIGRFAM
Phosphoglycerate kinase	pgk	pfam00162
Ribosomal protein L5	rplE	pfam00281
Ribosomal protein L3	rplC	pfam00297
Ribosomal protein L6	rplF	pfam00347
Ribosomal protein S9/S16	rpsI	pfam00380
Ribosomal protein S11	rpsK	pfam00411
Ribosomal protein S13/S18	rpsM	pfam00416
SsrA-binding protein	smpB	TIGR00086
translation elongation factor Ts	tsf	TIGR00116
ribosome recycling factor	frr	TIGR00496
ribosomal protein S2, bacterial type	rpsB	TIGR01011
ribosomal protein S5, bacterial/organelle type	rpsE	TIGR01021
ribosomal protein L19, bacterial type	rplS	TIGR01024
ribosomal protein L20	rplT	TIGR01032
ribosomal protein S10, bacterial/organelle	rpsJ	TIGR01049
ribosomal protein S19, bacterial/organelle	rpsS	TIGR01050
ribosomal protein L13, bacterial type	rplM	TIGR01066
ribosomal protein L14, bacterial/organelle	rplN	TIGR01067
ribosomal protein L16, bacterial/organelle	rplP	TIGR01164
ribosomal protein L1, bacterial/chloroplast	rplA	TIGR01169
ribosomal protein L2, bacterial/organellar	rplB	TIGR01171
50S ribosomal protein uL11, bacterial form	rplK	TIGR01632
transcription termination factor NusA	nusA	TIGR01953
DNA-directed RNA polymerase, beta subunit	rpoB	TIGR02013
50S ribosomal protein L4, bacterial/organelle	rplD	TIGR03953

4.2.5 Primary Metabolism Analysis

To determine if any essential genes were missing from the GUM007_hs and GUM202_hs genomes, we used the metabolic model developed experimentally for *Synecchococcus elongatus*

PCC 7942, a free-living cyanobacterium, as a comparison[57]. Protein coding regions of both GUM_hs genomes were BLASTed against the *S. elongatus* protein coding regions, resulting in 155 genes present in *S. elongatus*, but absent in both GUM genomes. The missing genes were converted into their corresponding KO and pfam numbers for analysis in JGI IMG/MER. The missing KO numbers were used as a search query in both GUM genomes. KOs that were absent in both genomes were cross-referenced against gene essentiality in *S. elongatus*, as experimentally determined using hundreds of thousands of transposon mutants[58]. The same process was done using pfams, which often catch genes that may not have been assigned a KO function by JGI. The intersection of the two datasets was a set of seven genes missing in the GUM_hs genomes and essential in *S. elongatus* PCC 7942 (Table 2). In order to examine if the pathways containing missing genes were actually incomplete, we used ec2kegg[59] to generate metabolic maps for pathway comparison in *S. elongatus* PCC 7942 and the GUM_hs genomes. *Moorea producens* was also used as a reference genome for a closer phylogenetic comparison in ec2kegg. The presence of protein secretion systems was assessed using the hmm models developed by Abby et al (2016), utilizing the TXSScan tool in Galaxy[60,61]. Five *Roseofilum* spp. genomes were also downloaded from JGI for comparison in TXSScan and ec2kegg.

4.2.6 Biosynthetic Gene Cluster Identification and Networking

Both genomes were submitted to JGI’s Integrated Microbial Genomes and Microbiomes Expert Review (IMG/MER) for detailed annotation. The two genomes were also submitted to antiSMASH 4.1.0 with standard options and ClusterBlast selected in extra features[62]. The results were examined using the Homologous Gene Clusters output for similarity to publicly available sequences and to characterized clusters in the MIBiG database[63]. When similarity to a known cluster was detected, all genes in the cluster were submitted to BLAST to determine closest matches and assign putative functions; JGI annotations were also taken into account to assign putative functions[64]. NaPDoS[65] was used to further analyze PKS clusters. Biosynthetic gene cluster networking was done using a locally installed version of the BiG-SCAPE software[66]. Each cluster Genbank file was extracted from the antiSMASH output and used as input for BiG-SCAPE with the local option enabled. The resulting pairwise scores were filtered for those above 0.40 and were visualized as a network using Gephi v0.9.1[67].

Table 2. Essential genes from *Synechococcus elongatus* PCC 7942 missing from GUM202_hs and GUM007_hs

The essential gene analysis using the well curated *S. elongatus* PCC 7942 metabolic model as a reference, resulted in eight genes that are essential for *S. elongatus* and missing in both GUM007_hs and GUM202_hs. Subsequent inspection of each pathway showed that only the biosynthetic pathway for histidine was actually incomplete.

***S. elongatus* PCC 7942 Locus tag**	**KO / pfam / EC**	**Name**	**Associated pathway (gene)**	**Pathway Incomplete?**
Synpcc7942_0125	K01693 / pfam00475 / 4.2.1.19	imidazoleglycerol-phosphate dehydratase	Histidine metabolism (hisB)	Yes
Synpcc7942_0475	K03689 / pfam03742	cytochrome b6-f complex subunit 8	Photosynthesis (petN)	No; also missing in *M. producens*
Synpcc7942_0849	K08973 / pfam03653	putative membrane protein	Chlorophyll Biosynthesis (K08973)	No; HemY present
Synpcc7942_1044	K03465 / pfam02511 / 2.1.1.148	thymidylate synthase (FAD)	Pyrimidine metabolism (thyX, thy1)	No; ThyA present
Synpcc7942_1447	K01673 / pfam00484 / 4.2.1.1	carbonic anhydrase	Nitrogen Metabolism (cynT, can)	No; CcmM present
Synpcc7942_2062	K06443 / pfam05834 / 5.5.1.19	lycopene beta-cyclase	Carotenoid Biosynthesis (lcyB, crtL1, crtY)	No; CruA and CruP present
Synpcc7942_2290	K00979 / pfam02348 / 2.7.7.38	3-deoxy-manno-octulosonate cytidylyltransferase (CMP-KDO synthetase)	Lipopolysaccharide biosynthesis (kdsB)	Cannot biosynthesize Kdo

4.2.7 PBDE Structure Elucidation from GUM202

Lyophilized GUM202 sponge tissue (1.5 grams) was ground using a mortar and pestle, and extracted with methanol (MeOH) 3x 20mL for 30-60 minutes on a benchtop nutator. The combined extracts were dried *in vacuo* and resuspended in dichloromethane (CH_2Cl_2) and then washed with water. The resulting CH_2Cl_2 layer was dried using magnesium sulfate, filtered, and dried *in vacuo*. Preparative HPLC solvents used were HPLC grade water with 0.1% trifluoroacetic acid (TFA) and HPLC grade acetonitrile (MeCN) with 0.1% TFA. Preparative HPLC was carried out using an Agilent 218 purification system (ChemStation software, Agilent) equipped with a ProStar 410 automatic injector, Agilent ProStar UV-Vis Dual Wavelength Detector, a 440-LC fraction collector and an Agilent Pursuit XRs 5 C18 100 x 21.2 mm preparative HPLC column (0-5 min 5% MECN isocratic, 5-10 min 10-25% MECN, 10-35 min 65-75% MECN, 35-40 min 75-100% MECN, 40-45 min 100% MECN isocratic). Thirteen fractions were collected and analyzed using LC-MS/MS performed on an Agilent 1260 LC system with diode array detector and Phenomenex Kinetex 5μ C18(2) 100 A, 150 x 4.6 mm column in negative mode. All LCMS solvents used were LCMS grade water with 0.1% formic acid and LCMS grade acetonitrile with 0.1% formic acid. MS were analyzed with Agilent MassHunter Qualitative Analysis version B.05.00. Eleven of the thirteen fractions were pure enough for NMR. Each fraction was dissolved in deuterated methanol and ^{1}H-NMR, COSY, HSQC, and HMBC experiments were run on a MACHINE 600NMR. Spectra were analyzed with SOFTWARE and were matched to previously reported PBDEs[68]. LCMS and NMR spectra used to characterize the eleven PBDEs can be found in the Chapter 4 Appendix.

4.2.8 Identification of Putative Dysinosin Analogues from GUM007

Due to sample limitations, dysinosin analogues were isolated from GUM007 sponge samples stored dry and in RNAlater. Lyophilized GUM007 sponge was ground with a mortar and pestle, extracted with 3x 20mL MeOH for 30-60 minutes on a benchtop nutator. The combined extracts were dried *in vacuo*. The MeOH extract and the RNALater solution used to store the sponges were loaded onto a C18 solid phase extraction column and fractionated (5% MECN, 10% MECN, 15% MECN, 20% MECN, 25% MECN, 30% MECN, 40% MECN, 100% MECN). Each fraction was analyzed by LC-MS/MS (see below) and the 15% and 20% MECN fractions contained the dysinosin analogues. Four standards, dysinosin A (**28**), dysinosin B (**10**), dysinosin C (**27**), and **29**, kindly provided by Professor Ron Quinn from Griffith University, were prepared as 0.1 mg/ml solutions in MeOH and analyzed in parallel using the Agilent 6530 Accurate-Mass Q-TOF MS mentioned above with a Phenomenex Kinetex 5µ C18(2) 100 A, 150 x 4.6 mm column (0-3 min 5% MECN isocratic, 3-23 min 5-100% MECN, 23-26 min 100% MECN isocratic at 0.7 mL/min). Molecular networking, using the Global Natural Products Social (GNPS)[69] molecular networking platform, was used to visualize the relation of the new dysinosins and dysinosin standards.

4.3 Results and Discussion

4.3.1 Enriched Metagenome Assembled Genomes

Obtaining high quality draft genomes from uncultured symbionts has proven a unique challenge in metagenomics. Often, short-read metagenomics sequencing alone does not produce high enough quality complete genomes, especially when examining highly repetitive biosynthetic gene clusters, such as non-ribosomal peptide synthetase (NRPS) and polyketide synthase (PKS) containing clusters[70]. For studies of sponge associated obligate symbionts, which can have reduced/minimized genomes[71,72], genome completeness is an imperative measure. Recently developed sequencing techniques for high-throughput, long-reads, such as PacBio RSII® (Pacific Biosciences) and MinION® (Oxford Nanopore), can be complemented with high-accuracy, short-reads obtained from Illumina sequencing[73,74]. Although *H. spongeliae* has never been cultured from any sponge host, we are fortunate that the denser cyanobacteria trichomes can be easily squeezed out of sponge tissue and separated from the less dense sponge cells by centrifugal partitioning, yielding an enriched fraction containing *H. spongeliae* cells[38], as assessed by light microscopy.

Two sponges, GUM007 and GUM202, yielded large cyanobacterial pellets and also exhibited different chemotypes: GUM007 contained no PBDEs, while GUM202 contained an abundance of PBDEs. To examine differences in secondary metabolite biosynthetic potential in the two sponge symbionts, we undertook a hybrid sequencing and assembly approach. The enriched cyanobacterial fractions yielded sufficient quantity and quality of DNA for PacBio RSII® sequencing. Additionally, unenriched metagenomes of each whole sponge were sequenced in parallel (Figure 11). Hybrid assemblies using binned Illumina reads and phylogenetically classified PacBio reads resulted in high quality draft genomes (Table 3). GUM007_hs is contained on 64 scaffolds, with an average length of 97 kb, and 91.82% completeness, according to CheckM[47]. GUM202_hs is contained on 70 scaffolds, with an average length of 98 kb and 93.64% completeness. Additionally, the genomes were checked for thirty housekeeping marker genes commonly used in building bacterial phylogenetic trees[49,75]. Both genomes had all thirty genes in one copy, further suggesting high quality assemblies. For comparison, 210 genomes out of 442 cyanobacterial genomes (47.5%) in JGI had all thirty genes in one copy, suggesting that just under half of the cyanobacterial genomes in JGI pass this low bar for genome completeness. The two GUM_hs genomes have an ANI of 96.18%, indicating that they belong to the same species, but diverge significantly with regard to synteny. While there are regions of the genomes that are syntenic, the two GUM_hs genomes have different organization overall, with no two scaffolds being completely syntenic.

Table 3. Assembly and Quality Statistics for GUM_hs Genomes.

General assembly statistics for the metagenome assembled genomes for *H. spongeliae* from GUM007 and GUM202 are displayed. CheckM results are also shown, confirming that these assemblies are roughly 92-94% complete. A two-way average nucleotide identity (ANI) shows that these two genomes are about 96% similar.

	GUM007__hs	GUM202_hs
# of scaffolds	64	70
Avg. length of scaffolds (bp)	97,372	98,137
Longest scaffold (bp)	345,613	304,315
Est. genome size (Mb)	6.2	6.8
N50	169,688	161,164
%GC	47.8%	47.5%
CheckM total markers	650	650
CheckM markers detected	477	477
CheckM completeness (%)	91.82	93.64
CheckM Contamination	7.56	9.73
CheckM Strain Heterogeneity	13.70	4.55
Two-way ANI: 96.18%		

A well supported MLSA phylogenetic tree made using 25 diverse housekeeping genes in single copy, extracted from 197 cyanobacteria genomes, shows how these unusual symbionts are related to other cyanobacteria (Figure 12). Both *H. spongeliae* strains clade together and are most closely related to multiple *Roseofilum* strains that are coral pathogens found in corals afflicted with Black Band disease[76-78]. Although *Roseofilum* spp. have been cultured in the lab, they are found in microbial assemblage mats that infect corals and have not yet been found free-living in the environment. Comparative genomics of five *Roseofilum* spp. has shown that they are reliant on other members of the microbial consortium, as many filamentous cyanobacteria are, and as *H. spongeliae* appear to be. They are also rich in secondary metabolism biosynthetic gene clusters[76]. Other close relatives include *Desertifilum* spp., which have been isolated from biological desert crusts and other microbial mat assemblages[79,80].

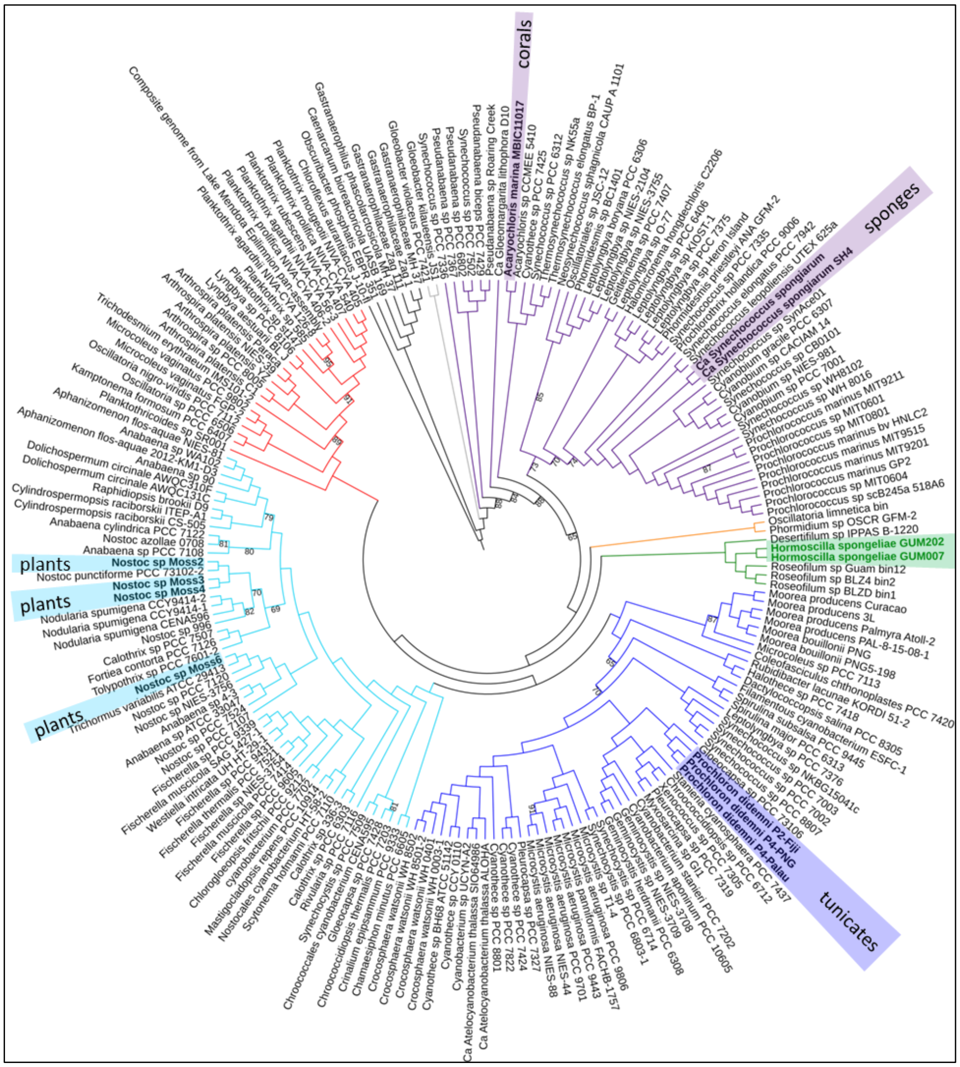

corals
Acaryochloris marina MBIC11017
sponges
Ca Synechococcus spongiarum
Ca Synechococcus spongiarum SH4
Hormoscilla spongeliae GUM202
Hormoscilla spongeliae GUM007
plants
Nostoc sp Moss2
plants
Nostoc sp Moss3
Nostoc sp Moss4
plants
Nostoc sp Moss6
tunicates
Prochloron didemni P2-Fiji
Prochloron didemni P4-PNG
Prochloron didemni P3-Palau

Figure 12. Multi-Locus Sequence Analysis (MLSA) of Diverse Cyanobacteria.

197 cyanobacteria genomes containing twenty-five housekeeping genes were used to construct this MLSA. *H. spongeliae* are highlighted in green. Their nearest sister clade are the *Roseofilum* spp., known coral pathogens. Other symbiotic cyanobacteria are highlighted throughout the tree. Only bootstrap values between 65-95 are shown, all other bootstrap values are above 95.

4.3.2 Genomic Hallmarks of a Symbiont Lifestyle

In an effort to understand why *Hormoscilla spongeliae* has defied laboratory cultivation, we compared metabolic pathways and gene essentiality with well characterized free-living cyanobacteria. A well curated genome-scale model (GEM) has recently been developed for the model cyanobacterium, *Synechococcus elongatus* PCC 7942[57]. This GEM is informed not just by genomic metabolic modeling data, but also by genome wide gene essentiality analysis, determined from ~250,000 transposon mutants tested for survival and growth and sequenced using random barcode transposon site sequencing (RB-TnSeq) to locate mutation sites and assign genes as essential, beneficial, or nonessential[58]. Although the free-living, single celled *S. elongatus* and the symbiotic, filamentous *H. spongeliae* are different in many ways, and differences in gene content are expected, the comparison allows us to leverage the extensive, experimentally validated data about gene essentiality in a free-living photosynthetic organism and potentially expose missing essential genes in a seemingly obligate symbiont. An initial BLAST comparison of the GUM_hs genes against the *S. elongatus* GEM revealed a combined list of 155 missing genes in both GUM_hs genomes. Missing genes were then converted to corresponding KO and pfam models for a more robust search. Any models that remained unfound were cross-referenced with gene essentiality data for *S. elongatus* PCC 7942, resulting in seven KOs/pfams that are both missing and essential for *S. elongatus* (Table 2). In order to determine if alternative enzymes are used by *H. spongeliae* in the incomplete metabolic pathways, we examined the comparative metabolic maps generated by ec2kegg[59] using *S. elongatus* PCC 7924 as a reference genome and each GUM_hs as a query genome. Additionally, we used the *Moorea producens* reference genome provided by ec2kegg as a secondary reference, as *M. producens* and *H. spongeliae* are more closely related and share distinct features including a filamentous morphology and large secondary metabolism repertoires[81].

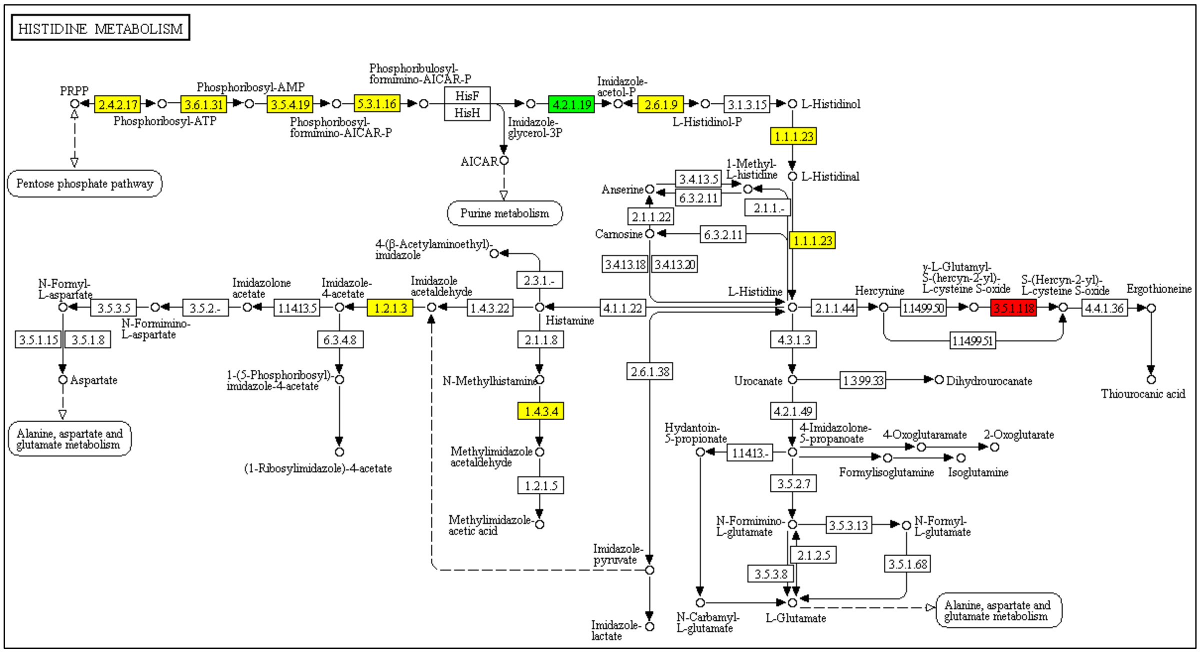


Figure 13. Histidine Metabolism in *H. spongeliae* vs *S. elongatus* PCC7942

Comparison of the biosynthetic pathway for the amino acid, histidine, in S. *elongatu*s, GUM007_hs, and GUM202_hs shows an apparent lack of an essential enzyme, imidazoleglycerol-phosphate dehydratase (EC 4.2.1.19), which performs the sixth step in histidine biosynthesis.

Both strains of *H. spongeliae* are prototrophic for all amino acids except histidine. The missing essential gene analysis exposed the lack of imidazoleglycerol-phosphate dehydratase (*hisB*), the enzyme responsible for the sixth step in histidine biosynthesis, in both GUM_hs genomes (Figure 13). There does not appear to be an alternative pathway for histidine biosynthesis, and both *S. elongatus* and *M. producens* harbor *hisB*. Additionally, a BLAST search using the *M. producens hisB* did not turn up any results in the GUM_hs genomes. Three additional assembled genomes for *H. spongeliae* symbionts of related *L. herbacea* specimens collected in Guam (GUM098, GUM102, and SP12—personal communication, Jessica Blanton) also lacked *hisB*. While it is possible that this gene was not assembled from the sequence data generated, it may also indicate that *H. spongeliae* is a histidine auxotroph, requiring histidine from its host or other members of the microbiome. Alternatively, there could be an unrelated gene encoding a novel dehydratase to perform this reaction.

A small subunit of the cytochrome b6-f complex, PetN, also appears to be missing in both GUM_hs genomes. While PetN was determined to be essential in *S. elongatus* PCC7942, *Synechocystis* sp. PCC 6803, and *Nostoc* sp. PCC 7120, it does not appear to be essential for *M. producens*, as this reference genome also lacks PetN[82,83]. The four large subunits of the cytochrome b6-f complex, PetA, PetB, PetC, and PetD, are present in both *H. spongeliae* genomes and *M. producens*. Although their exact function in cyanobacteria remains largely unknown, three small subunits of the cytochrome b6-f complex, PetG, PetL, and PetN, have been implicated in the assembly and stability of the whole complex in plants[84]. As the cytochrome b6-f complex plays an essential role in electron transfer between photosystem II and photosystem I, it is not likely that this enzyme is dysfunctional in *M. producens*, but rather that *M. producens* does not require PetN for its cytochrome b6-f complex to function[85]. We therefore conclude that this missing gene does not indicate that the cytochrome b6-f complex is missing or dysfunctional in *H. spongeliae*, but that *H. spongeliae* and the related *M. producens* may not need PetN for a functioning cytochrome b60-f complex, or it may be replaced by a yet unknown protein.

We found alternative pathways in *H. spongeliae* for four other missing essential genes. The gene encoding thymidylate synthase, *thyx* (EC 2.1.1.148), is missing in both GUM_hs genomes, but upon further interrogation of the pyrimidine metabolism map (Figure 14), an alternative gene, *thyA* (EC 2.1.1.45), can perform the same transformation and is present in both genomes. Upon investigation of the missing essential lycopene beta-cyclase gene, *lcyB* (EC 5.5.1.19), we found that *M. producens* is also missing this gene. *M. producens* and both *H.*

spongeliae genomes possess two other genes that encode lycopene cyclases, CruA and CruP, that can perform the same function of LcyB[86].

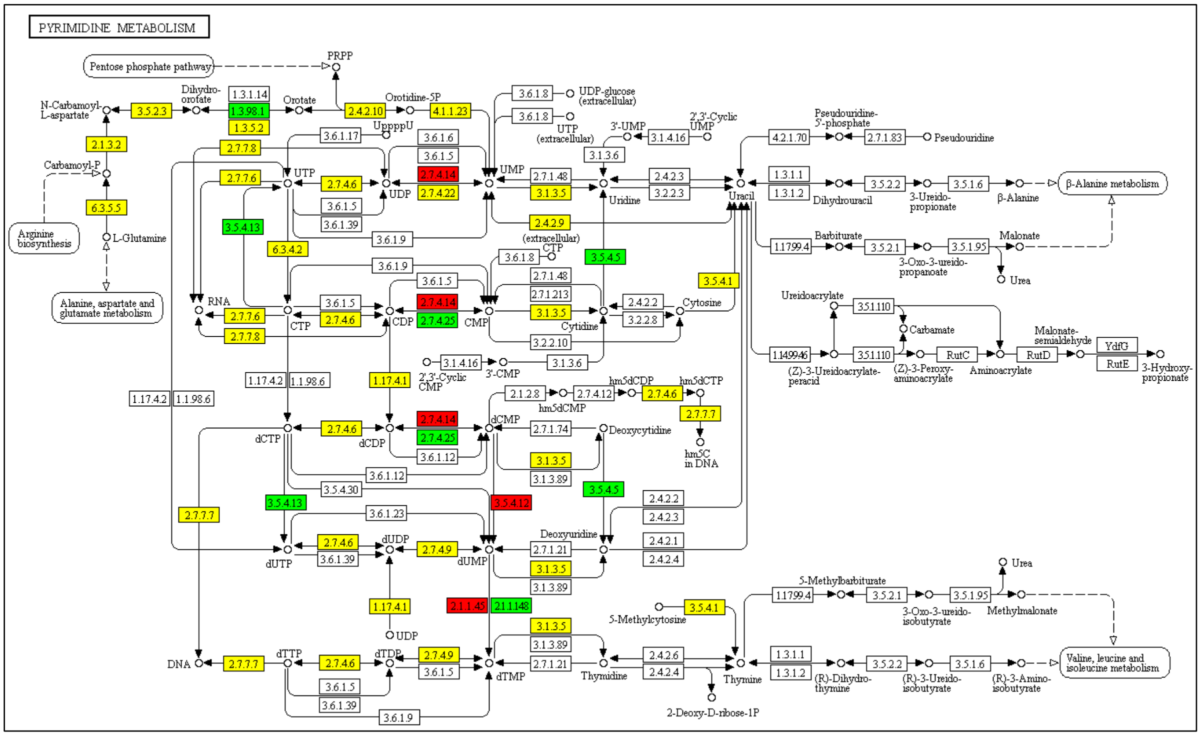


Figure 14. Pyrimidine Metabolism in *H. spongeliae* vs *S. elongatus*

Comparison of the biosynthetic pathways for making pyrimidines in the reference genome, *S. elongatus* and query genomes, GUM007_hs and GUM202_hs, shows that GUM_hs genomes have alternative enzymes to complete pyrimidine biosynthesis.

The missing putative membrane protein (K08973), as annotated in *S. elongatus*, is annotated as protoporphyrinogen oxidase HemJ in *M. producens*, and has been shown to be an enzyme that replaces HemG or HemY in heme biosynthesis[87]. The *H. spongeliae* genomes possess HemY (EC 1.3.3.4) to catalyze the formation of protoporphyrin IX, and thus do not require HemJ as *S. elongatus* and *M. producens* do (Figure 15). Finally, the missing gene, *cynT*, encodes the enzyme carbonic anhydrase, which acts as a key player in the carbon dioxide concentrating mechanism of cyanobacteria. This type of β-carbonic anhydrase resides in the carboxysome, where in converts bicarbonate to carbon dioxide for use in photosynthesis[88]. Once genomes of β-cyanobacteria began to be sequenced, it was noticed that they lacked a cytosolic β-carbonic anhydrase, but contained a CcmM protein with domains homologous to γ-carbonic anhydrases[89]. The elucidation of the structure of a cyanobacterial CcmM and subsequent experiments confirmed that it acts as a carbonic anhydrase in cyanobacteria lacking a conventional β-carbonic anhydrase[90]. Both *H. spongeliae* genomes contain CcmM, suggesting that it plays the role of the essential β-carbonic anhydrase in *S. elongatus*.

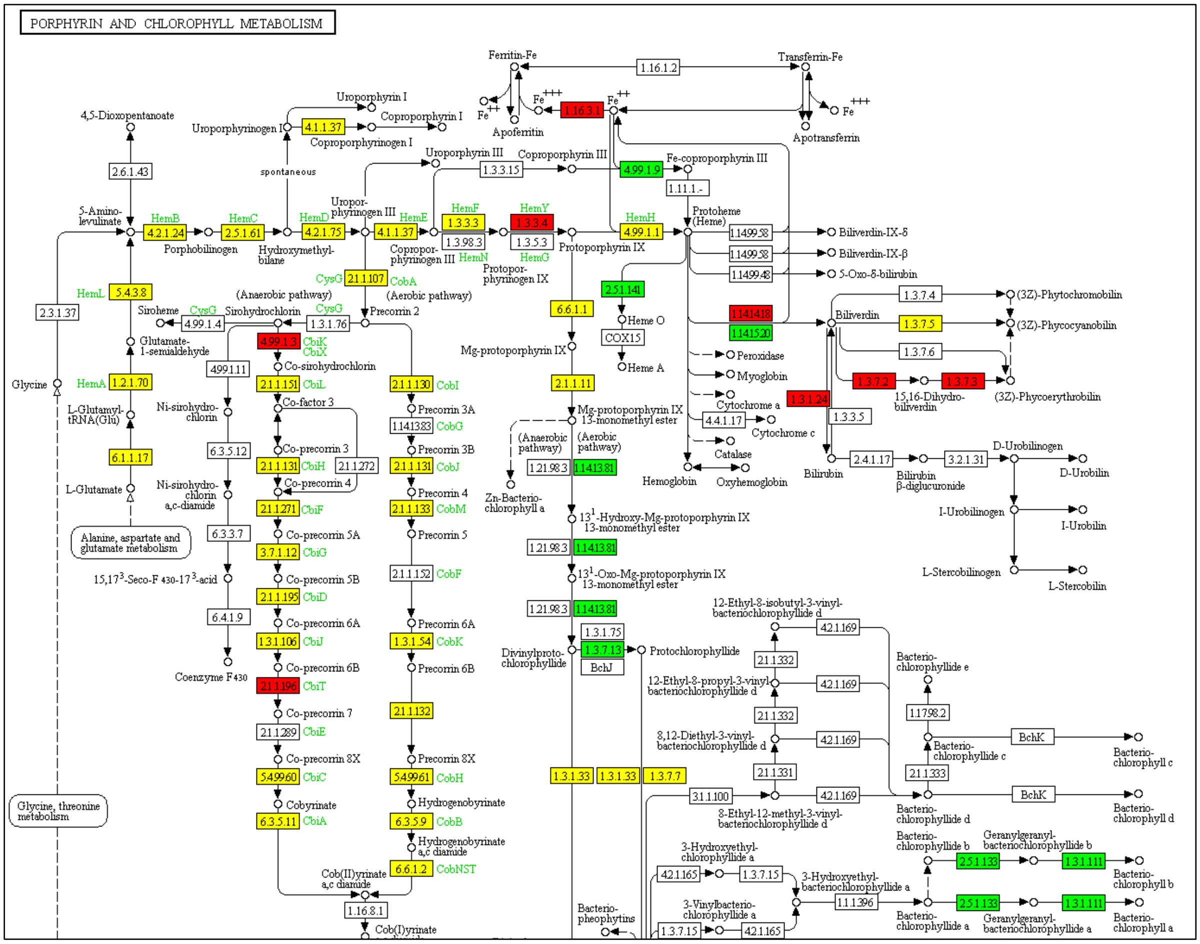


Figure 15. Porphyrin and Chlorophyll Metabolism in *H. spongeliae* vs *S. elongatus*

The missing putative membrane protein from *S. elongatus* (K08973) is annotated as HemJ, a protoporphyrinogen oxidase, in *M. producens*. HemJ can replace HemG or HemY, as it does in both reference genomes, but both *H. spongelie* have HemY to complete this step in protoporphyrinogen biosynthesis.

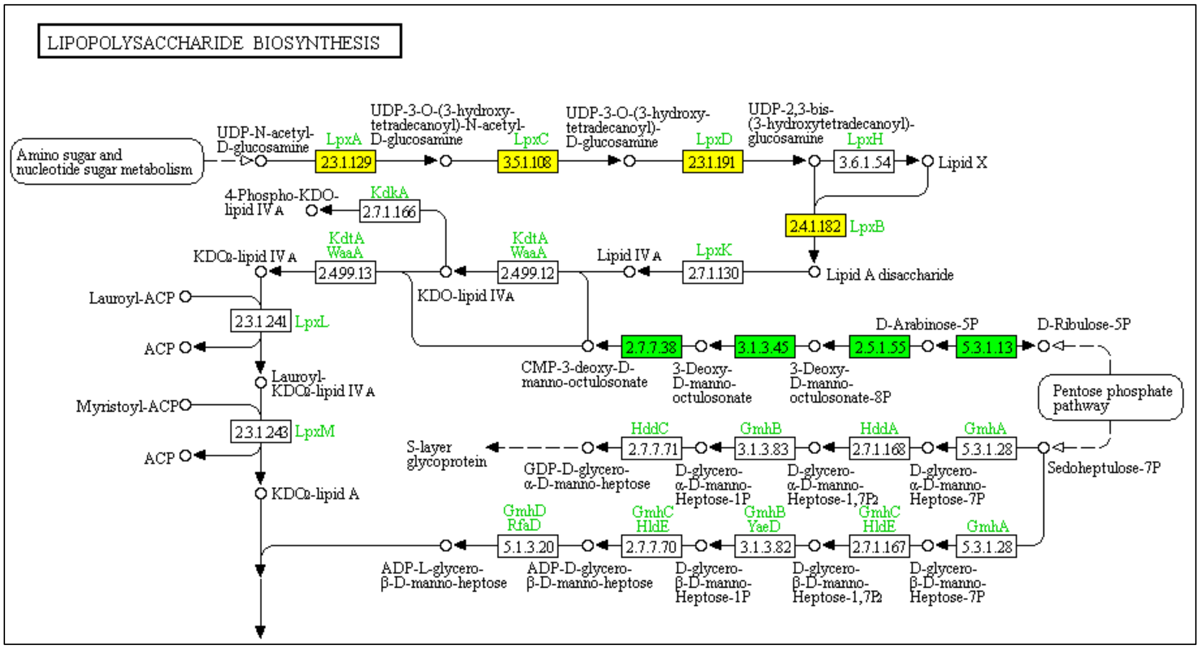

LIPOPOLYSACCHARIDE BIOSYNTHESIS
Amino sugar and nucleotide sugar metabolism
UDP-N-acetyl-D-glucosamine
LpxA
2.3.1.129
UDP-3-O-(3-hydroxy-tetradecanoyl)-N-acetyl-D-glucosamine
LpxC
3.5.1.108
UDP-3-O-(3-hydroxy-tetradecanoyl)-D-glucosamine
LpxD
2.3.1.191
UDP-2,3-bis-(3-hydroxytetradecanoyl)-glucosamine
LpxH
3.6.1.54
Lipid X
2.4.1.182
LpxB
Lipid A disaccharide
LpxK
2.7.1.130
Lipid IVA
KdtA
WaaA
2.4.99.12
KDO-lipid IVA
KdkA
2.7.1.166
4-Phospho-KDO-lipid IVA
KdtA
WaaA
2.4.99.13
KDO2-lipid IVA
Lauroyl-ACP
2.3.1.241
LpxL
ACP
Lauroyl-KDO2-lipid IVA
Myristoyl-ACP
2.3.1.243
LpxM
ACP
KDO2-lipid A
D-Arabinose-5P
D-Ribulose-5P
2.7.7.38
3.1.3.45
2.5.1.55
5.3.1.13
CMP-3-deoxy-D-manno-octulosonate
3-Deoxy-D-manno-octulosonate
3-Deoxy-D-manno-octulosonate-8P
Pentose phosphate pathway
S-layer glycoprotein
HddC
2.7.7.71
GmhB
3.1.3.83
HddA
2.7.1.168
GmhA
5.3.1.28
Sedoheptulose-7P
GDP-D-glycero-α-D-manno-heptose
D-glycero-α-D-manno-Heptose-1P
D-glycero-α-D-manno-Heptose-1,7P2
D-glycero-α-D-manno-Heptose-7P
GmhD
RfaD
5.1.3.20
GmhC
HldE
2.7.7.70
GmhB
YaeD
3.1.3.82
GmhC
HldE
2.7.1.167
GmhA
5.3.1.28
ADP-L-glycero-β-D-manno-heptose
ADP-D-glycero-β-D-manno-heptose
D-glycero-β-D-manno-Heptose-1P
D-glycero-β-D-manno-Heptose-1,7P2
D-glycero-β-D-manno-Heptose-7P

Figure 16. Lipopolysaccharide Metabolism in *H. spongeliae* vs. *S. elongatus*

The four enzymes in green boxes are responsible for the biosynthesis of ketodeoxyoctonate (Kdo), the sugar component of lipopolysaccharides. The pathway is present in *S. elongatus* and *M. producens*, but absent in *H. spongeliae*. This pathway is essential for *S. elongatus*, but it may not be essential for *H. spongeliae*, as it may not use Kdo in its lipopolysaccharide structure.

The essential missing gene analysis identified the *kdsB* gene (EC 2.7.7.38) involved in lipopolysaccharide biosynthesis, and indeed the entire Kds pathway, as missing in *H. spongeliae* (Figure 16). The Kds pathway, comprised of a suite of four enzymes, KdsA-D, is responsible for turning D-ribulose-5P into CMP-3-deoxy-D-manno-octulosonate (ketodeoxyoctonate or Kdo). *M. producens* also has the genes *kdsA-C*, and the pathway is largely conserved between plants and bacteria[91]. While Kdo is often present in the sugar component of lipopolysaccharides in Gram-negative bacteria, there have been multiple cyanobacteria found to not contain Kdo[92]. It is likely that *H. spongeliae* has different sugar variants in its lipopolysaccharide structure and therefore does not require Kdo. Lipopolysaccharides have been implicated in establishing symbiont-host symbioses in a variety of systems[93]. In a legume-endosymbiont association, it was found that symbionts with a mutated lipopolysaccharide had defective communication with their plant hosts, which were unable to form nodules[94]. A similar phenomenon was observed in the bean bug and its gut symbionts of the genus *Burkholderia*. In this system, the O-antigen component of the *Burkholderia* lipopolysaccharide is crucial for initial colonization of the bean bug, and mutations in the core oligosaccharide resulted in lower colonization rates and decreased host fitness[95]. This interesting interaction of symbiont and host through lipopolysaccharides has not been explored in *H. spongeliae* and would be an interesting avenue for further study.

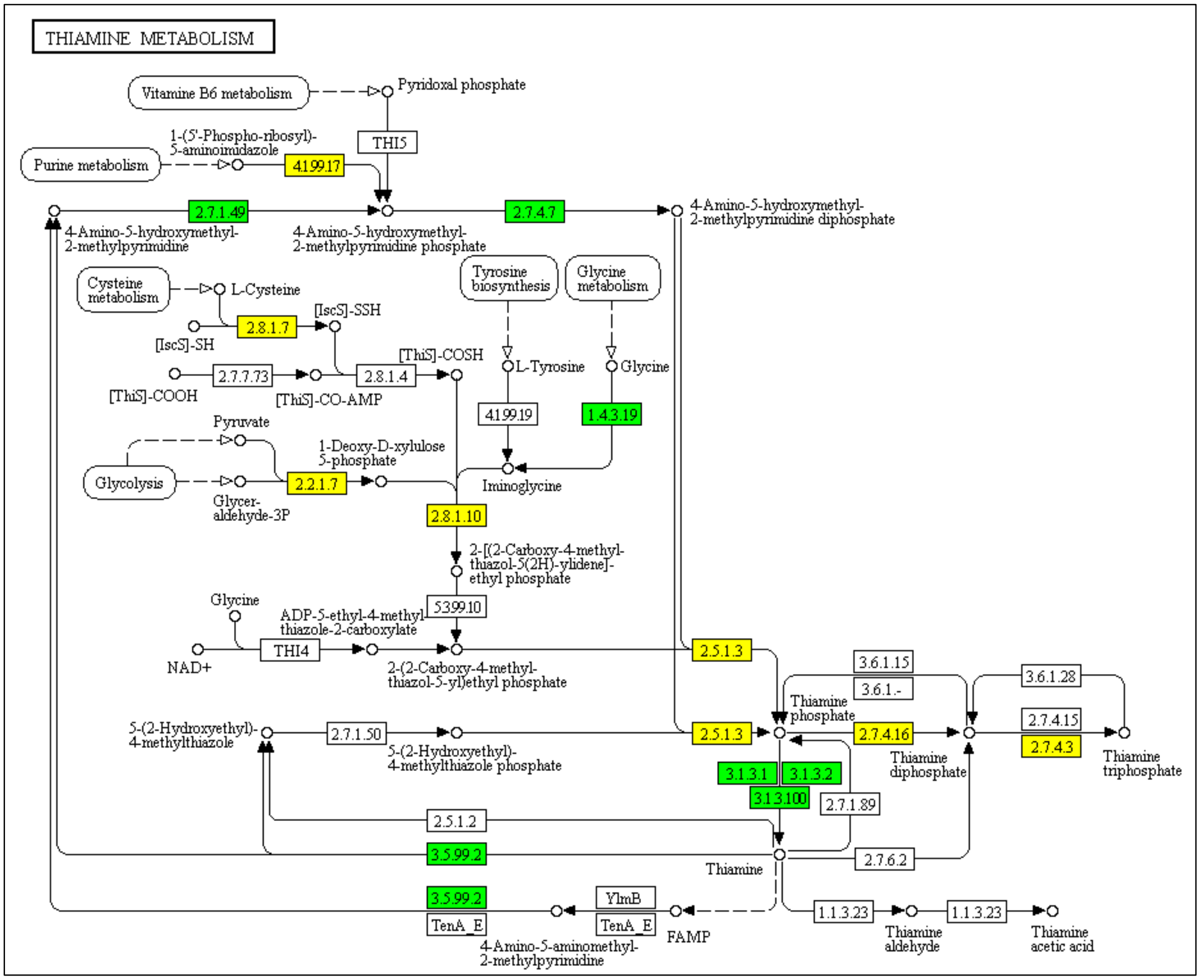


Figure 17. Thiamine Metabolism in *H. spongeliae* vs *S. elongatus*

Both GUM007_hs and GUM202_hs genomes appear to be lacking the three enzymes involved in the last step of thiamine biosynthesis (ECs 3.1.3.1, 3.1.3.2, and 3.1.3.100).

Of the apparent missing genes in both *H. spongeliae* genomes that were experimentally determined as essential for *S. elongatus* PCC 7942, the missing imidazoleglycerol-phosphate dehydratase appears to be the most problematic for a free-living lifestyle. Additionally, we examined the ability of *H. spongeliae* to produce important co-factors, such as thiamine and biotin. In the case of both GUM_hs genomes, the enzymes performing the last step in thiamine biosynthesis are missing (EC 3.1.3.1, EC 3.1.3.2, EC 3.1.3.100), but are present in both *S. elongatus* and *M. producens* (Figure 17). This suggests that *H. spongeliae* is unable to make thiamine, an important co-factor. Likewise, two *Entotheonella* spp., uncultivated symbionts of the marine sponge *Theonella swinhoei*, also appear to have an incomplete pathway for thiamine biosynthesis[96].

Also absent in the *Entotheonella* genomes was the pathway for biotin biosynthesis. While *H. spongeliae* do not completely lack the biotin pathway, they are missing key enzymes present *in S. elongatus* (EC 2.1.1.197, EC 3.1.1.85, EC 2.3.1.47, EC 2.6.1.62), two of which, BioC and BioH, are involved in making pimeloyl-CoA, and the remaining two, BioF and BioA, are involved in the core steps of biotin biosynthesis from pimeloyl-CoA. However, three of these enzymes (BioC, BioH, and BioA) also appear to be missing in *M. producens* (Figure 18). As biotin is a cofactor for acetyl-CoA carboxylase and essential for the production of polyketides and fatty acids, it is hard to reconcile *M. producens* being unable to make biotin. However, three enzyme orthologues, BioG, BioK, and BioJ, have all been shown to replace BioH in pimeloyl-CoA biosynthesis[97,98], and there appear to be homologues of BioK in *M. producens* and both *H. spongeliae*. These are annotated as "taurine-2-oxoglutarate transaminase," with about 32% similarity to a published *bioK* sequence from *Jeotgalibacillus marinus* DSM 1297. Furthermore, low homology is seen between two cyanobacterial *bioK* sequences (35%) in *Prochlorococcus* *marinus* MIT-9211 and *Synechococcus* sp. CC9902, so even with low similarity, these genes could reasonably be *biok* homologues[97]. While this may provide an alternative to the canonical BioH, this does not reconcile the missing core enzymes in the biotin pathway.

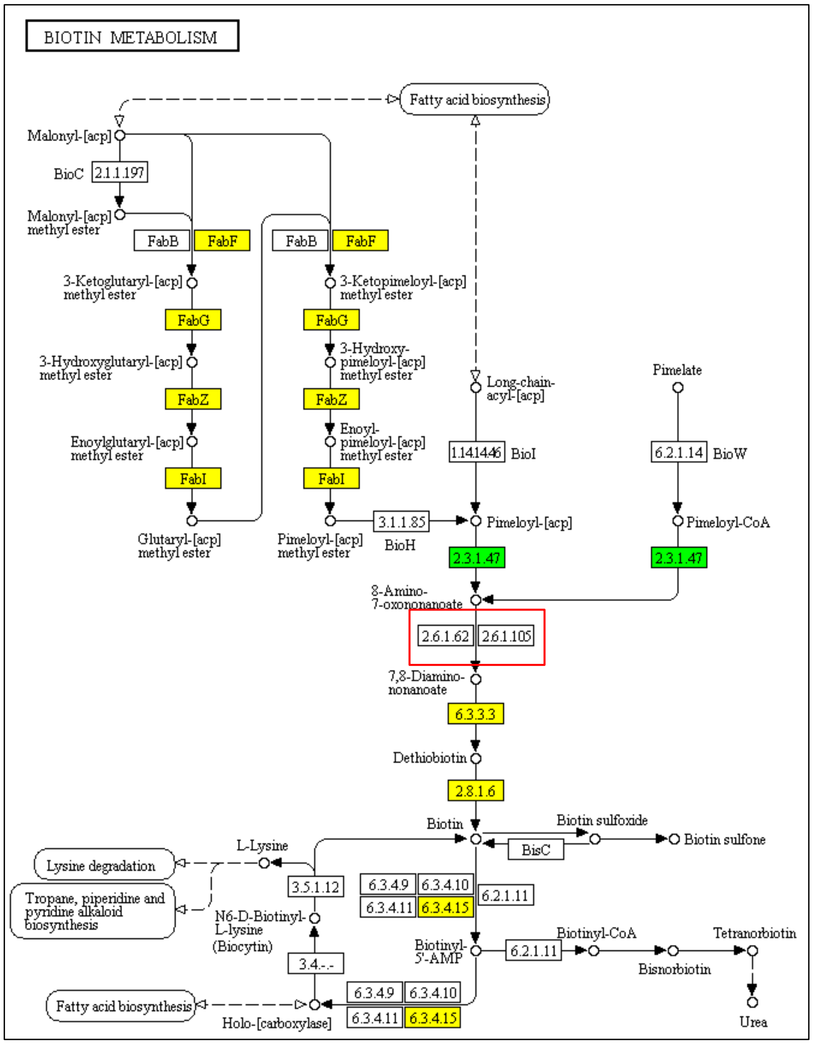


Figure 18. Biotin Metabolism in *H. spongeliae* vs *M. producens*.

The biotin pathway appears incomplete in both *M. producens* and *H. spongeliae*. Putative homologues of BioK, which can replace BioH in pimeloyl-CoA biosynthesis, were found in *M. producens* and *H. spongeliae*. However, no homologues of BioA (ECs 2.6.1.62 and 2.6.1.105, red box) have been found in either species.

BioA (Figure 18 red box) is missing in the *M. producens* genome used as a reference in ec2kegg, and all other *M. producens* genomes searched in JGI. However, *bioA*, and indeed the complete biotin pathway from pimloyl-CoA to biotin, is present in an assembly of a heterotrophic bacteria living with *M. producens* from Palmyra Atoll (JGI Genome ID: 2630968267)[99]. Interestingly, *M. producens,* and other filamentous cyanobacteria, are unable to be cultured axenically, and always grow with heterotrophic bacteria (personal communication, William Gerwick). Thus, it is possible that neither *M. producens* nor *H. spongeliae* can make biotin on their own and rely on heterotrophic bacteria for provision of this important co-factor.

While it is hard to pin-point missing essential genes from unfinished genomes that would lead to a specific reason why a symbiont may be uncultivable, our analysis provided a few possibilities. Future cultivation efforts should include addition of histidine, thiamine, and biotin to address possible deficits in these biosynthetic pathways. When comparing the *H. spongeliae* and *L. herbacea* symbiosis to other well studied bacterial symbioses, like those seen in plants and insects, many of the canonical symbiotic traits are absent. Often obligate symbionts undergo extreme genome reduction, such as that seen in Ca. *Synechococcus spongarium*, a cyanobacterial symbiont of diverse sponges[71]. However, *H. spongeliae* does not appear to have a minimized genome, with estimated genome sizes between 6.2 – 6.8 Mb, comparable to other free-living filamentous cyanobacteria. The same trend was seen in the uncultivatable *Prochlorococcus didemni* symbionts of tunicates[100]. Also observed in *P. didemni* was a normal GC content for cyanobacteria (41-42%), which we see in *H. spongeliae* as well (46-47%), while obligate symbionts usually have a higher AT content than their free-living counterparts[101]. In addition to reduced genome size and higher AT content, a lack of mobile elements is seen in long-standing symbioses[102]. However, in recently established symbiotic systems, an abundance of such mobile elements is common and *H. spongeliae* does indeed have an abundance of mobile genetic elements: GUM007_hs has 226 genes annotated as a transposase and GUM202_hs has 318. Genome minimization is postulated to take place in two separate stages, the first of which retains transposable elements. It is possible then, that the *H. spongeliae – L. herbacea* symbiosis is a recently established one and *H. spongeliae* has yet to undergo extreme genome reduction.

It has also been suggested that symbionts may have begun their symbiosis with a host as a pathogen, due to the shared molecular mechanisms between pathogens and symbionts[103]. These genomic hallmarks are especially seen in secondary symbionts, which are recently acquired and facultative from the host's perspective. Examples of shared molecular mechanisms between pathogens and symbionts include protein secretion systems, such as the type III secretion system (T3SS), dynamic processes, such as an abundance of transposons and bacteriophage sequences, and the presence of toxins[103]. While the *H. spongeliae* genomes do not contain a complete T3SS, they do contain a complete type I secretion system (T1SS), and retain genetic hallmarks from other secretion systems. Additionally, *H. spongeliae* genomes carry an abundance of mobile genetic elements and the capacity to make toxins. Furthermore, *H. spongeliae*'s closest sequenced relatives belong to the *Roseofilum* genus, known coral pathogens. It is therefore possible that *H. spongeliae* and *Roseofilum* spp. share a pathogenic lifestyle, with *H. spongeliae* recently establishing a symbiosis with *L. herbacea*.

4.3.3 Secondary Metabolite Biosynthetic Gene Clusters

The completeness of the two genomes afforded us a comprehensive look at the identifiable biosynthetic gene clusters within these uncultured *H. spongeliae* specimens. AntiSMASH 4.1.0 without ClusterBlast results for both GUM007_hs and GUM202_hs are shown in Figure 19a. A total of fifteen clusters (after splitting an unlikely NRPS/bacteriocin hybrid cluster) were found in GUM007_hs and eighteen in GUM202_hs. A gene cluster similarity network of all antiSMASH identified gene clusters reveals that the genomes share six clusters, roughly one third of their biosynthetic gene clusters, making two thirds of the gene clusters unique to each organism (Figure 19b). Of those shared, two are terpene clusters likely encoding carotenoid pigments, two are bacteriocins / ribosomally synthesized and post-translationally modified peptides (RiPPs), one is a T1PKS common in other filamentous cyanobacteria, and one NRPS cluster (Figure 19c). The shared NRPS cluster has all the hallmarks of being a mycosporine-like amino acid (MAA) with possibly a novel structure.

MAAs are known to be common metabolites of cyanobacteria, algae, fungi, and lichen, with UV protective properties, and may also contribute to environment amelioration as compatible osmolytes and antioxidants[104]. The biosynthetic pathway for shinorine, one such MAA, has been elucidated, and it was found that a core set of four genes are necessary to make shinorine: an NRPS-like gene, an ATP grasp gene, an O-methyltransferase gene, and a DHQS-like gene[105]. The adenylation (A) domain of the NRPS-like protein in the shinorine pathway is predicted to activate a serine, and this was observed to be incorporated into the shinorine structure by radio-labeled feeding experiments[105]. As MAAs are a common UV protecting metabolite in cyanobacteria, it is not surprising that we found almost identical putative MAA gene clusters in both GUM007_hs and GUM202_hs. In the case of both putative MAA gene clusters, all four essential genes determined from the characterized shinorine cluster[105] are present, suggesting the necessary machinery to make a MAA compound is existent. The A domain in the NRPS gene in both GUM007_hs and GUM202_hs clusters is predicted to have a specificity for proline, as determined by consensus of all A domain predictor software used by antiSMASH 4.1.0. Additionally, the putative MAA cluster in GUM202 may contain a second

NRPS gene, with an A domain selective for serine. To our knowledge, no cyanobacterial MAA compound incorporating proline has been described[106]. However, the structure of mycosporine-2, discovered in fungi, does contain a proline moiety[107]. We did not attempt to isolate an MAA molecule from our extracts, but the characteristic UV trace was seen in HPLC chromatograms for both samples (data not shown).

The remaining NRPS gene clusters in GUM007_hs and GUM202_hs are unique to their respective genomes. Additionally, all but one NRPS cluster (excluding those encoding putative MAAs) have no significant homology to known clusters in MIBiG, or to other sequenced organisms, as determined by MultiGeneBlast run in antiSMASH. The one NRPS cluster with homology to a characterized pathway is highly similar to the characterized aeruginoside cluster[108]. Dysinosins A-D have been isolated from *L. herbacea* sponges collected in Australia and are structurally related to the aeruginosides isolated from numerous cyanobacteria. We provide evidence for this putative dysinosin cluster and further discussion in 4.3.6.

The shared T1PKS cluster consists of one (GUM007_hs) or two (GUM202_hs) PKS genes with four total domains: KS, AT, KR, and DH. The KS domain is determined to be an enediyne KS, and the gene cluster seems to be partially conserved in multiple filamentous cyanobacteria (Figure 20a). Enedyine natural products are rare, with few characterized examples, most of which are isolated from actinomycetes, with two isolated from marine ascidians[109].

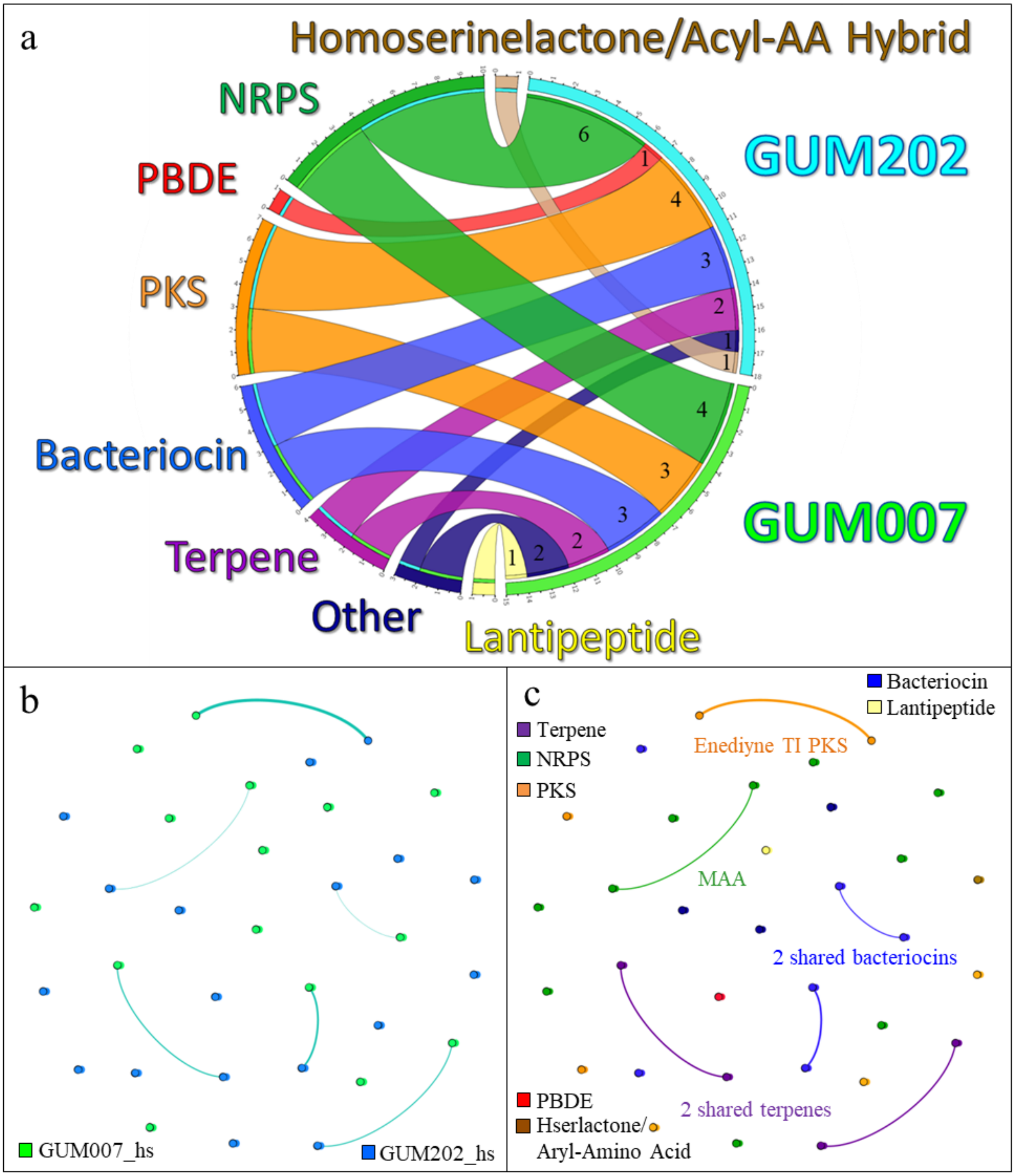


Figure 19. Secondary Metabolite Biosynthetic Gene Clusters of two *Hormoscilla spongeliae*

Figure 19a is a Circos diagram showing the number and classes of biosynthetic gene clusters in the GUM007_hs genomes. Figure 19b is a gene cluster similarity network made with BiG-SCAPE. Each node represents a gene cluster and those with similarity over a 0.4 threshold are connected by a line. The weight of the line indicates higher similarity for bolder lines. Figure 19c is the same network with each node colored according to antiSMASH classification.

The only enedyine natural product isolated from cyanobacteria is fischerellin A, from *Fischerella muscicola*[110]. Fischerellin A contains an uncyclized enedyine moiety, while most other characterized enediyne containing molecules belong to nine or ten membered ring systems[109]. The KS domains from the GUM_hs genomes were analyzed using NaPDoS[65] and form a sister clade with enediyne KS domains, but do not clade directly with them (Figure 20b). The enediyne KS domains and the GUM_hs KS domains are all sister clades to KS domains involved in poly-unsaturated fatty acid (PUFA) biosynthesis. As the GUM_hs KS domains appear phylogenetically distinct from characterized enediyne KS domains, it is possible that they are involved in novel enediyne PKS biosynthesis, and perhaps the formation of non-cyclized enedyine molecules like fischerellin A.

The GUM_hs genomes both contain two terpene clusters, which share similarity and likely are involved in making carotenoid-like pigments. Both clusters contain a phytoene synthase and accompanying tailoring enzymes, including terpene cyclases and dehydrogenases. These terpene clusters do not appear to be responsible for the diverse sesquiterpene molecules isolated from *L. herbacea* sponges, which have long been postulated to be produced by the sponge host. Finally, the two shared RiPP clusters have no predicted products based on bioinformatic analysis. One of the shared RiPP clusters has similar elements to the gene clusters responsible for making the many and varied cyclic cyanobactins[111], including both a YcaO protein and a SagB-like dehydrogenase. The remaining gene organization, however, does not highly resemble canonical cyanobactin pathways, and both pathways are closely flanked by a PyrG superfamily CTP synthase homologue, likely involved in primary metabolism. Cyanobactin-like molecules have never been isolated from *L. herbacea* sponges and these RiPP clusters may represent active cyanobactin-like gene clusters, or perhaps cyanobactin-like gene clusters in the process of deactivation.

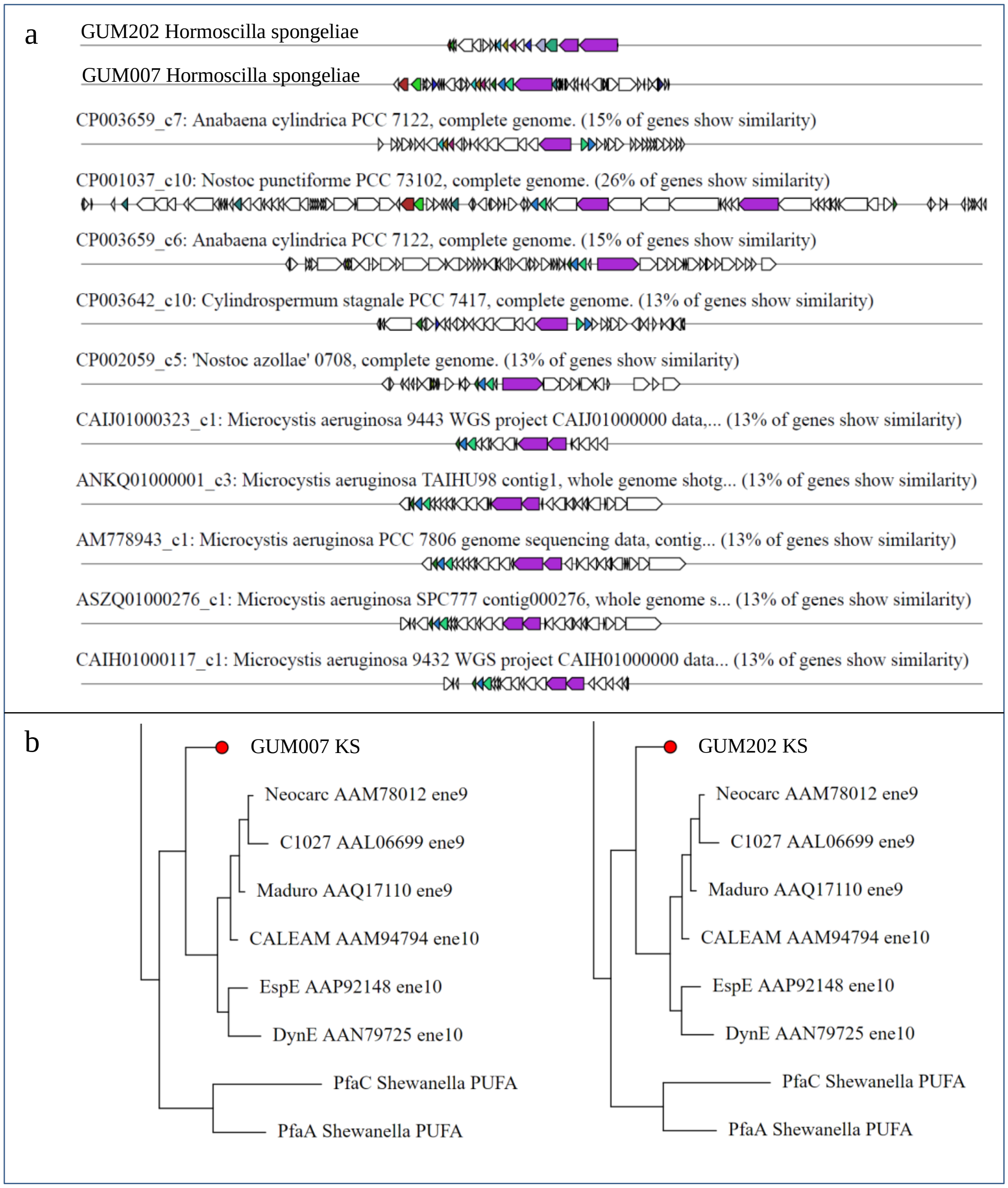

a
GUM202 Hormoscilla spongeliae
GUM007 Hormoscilla spongeliae
CP003659_c7: Anabaena cylindrica PCC 7122, complete genome. (15% of genes show similarity)
CP001037_c10: Nostoc punctiforme PCC 73102, complete genome. (26% of genes show similarity)
CP003659_c6: Anabaena cylindrica PCC 7122, complete genome. (15% of genes show similarity)
CP003642_c10: Cylindrospermum stagnale PCC 7417, complete genome. (13% of genes show similarity)
CP002059_c5: 'Nostoc azollae' 0708, complete genome. (13% of genes show similarity)
CAIJ01000323_c1: Microcystis aeruginosa 9443 WGS project CAIJ01000000 data,... (13% of genes show similarity)
ANKQ01000001_c3: Microcystis aeruginosa TAIHU98 contig1, whole genome shotg... (13% of genes show similarity)
AM778943_c1: Microcystis aeruginosa PCC 7806 genome sequencing data, contig... (13% of genes show similarity)
ASZQ01000276_c1: Microcystis aeruginosa SPC777 contig000276, whole genome s... (13% of genes show similarity)
CAIH01000117_c1: Microcystis aeruginosa 9432 WGS project CAIH01000000 data... (13% of genes show similarity)
b
GUM007 KS
Neocarc AAM78012 ene9
C1027 AAL06699 ene9
Maduro AAQ17110 ene9
CALEAM AAM94794 ene10
EspE AAP92148 ene10
DynE AAN79725 ene10
PfaC Shewanella PUFA
PfaA Shewanella PUFA
GUM202 KS
Neocarc AAM78012 ene9
C1027 AAL06699 ene9
Maduro AAQ17110 ene9
CALEAM AAM94794 ene10
EspE AAP92148 ene10
DynE AAN79725 ene10
PfaC Shewanella PUFA
PfaA Shewanella PUFA

Figure 20. Shared Enediyne Type 1 PKS

In Figure 20a, the MultiGeneBlast output (as run in antiSMASH 4.1.0) shows a shared T1PKS biosynthetic gene cluster among multiple strains of cyanobacteria. The NaPDoS analysis in Figure 20b shows that the GUM_hs KS domains form a sister clade with characterized enediyne KS domains. Both these clades form a sister clade with PUFA KS domains.

Overall, the biosynthetic gene clusters in the GUM_hs genomes are of varying classes, with a few bioinformatically predicted structures: structurally varied PBDEs, aeruginosin-like dysinosins, MAAs, and carotenoid-like pigments. These predicted structures correspond with molecules isolated from *L. herbacea* sponges that have been historically attributed to *H. spongeliae*. This new genetic evidence provides sound reasoning for attributing these compounds to the symbiont, *H. spongeliae*.

4.3.4 Genetic Expansion in *hs_bmp* Cluster Leads to Structural Variety of PBDEs

Previous investigations of three clades of *H. spongeliae* (Clades Ia, Ib, and IV), which are prolific producers of PBDEs, have revealed that PBDE biosynthesis is encoded on the semi-variable *hs_bmp* gene cluster[34]. At the very minimum, three core enzymes are needed to assemble PBDEs: Bmp5, a flavin-dependent brominase, Bmp6, a chorismate lyase, and Bmp7, a cytochrome P450 that couples the two brominated phenolic rings[112]. Using these fundamental biosynthetic features, we queried the genomes of GUM202_hs and GUM007_hs. While the *hs_bmp* gene cluster is absent from GUM007_hs, in accordance with its secondary metabolite profile, GUM202_hs possess an expanded *bmp* gene cluster. Previously, we observed a variable genomic region between *hs_bmp6* and *hs_bmp7* in the three sequenced *hs_bmp* clusters, one of which contains hs_Bmp12, a cytochrome P450 hydroxylase that was shown to be responsible for doubly hydroxylated PBDEs exclusive to Clade Ia (Figure 21). The variable region in the GUM202 *hs_bmp* pathway also contains a homologous gene for the extra P450 hydroxylase, hs_Bmp12 from the Clade Ia *H. spongelia* (Figure 21) as well as a gene encoding a second putative flavin-dependent halogenase, hs_Bmp18, with 78% pairwise nucleotide identity to the GUM202 hs_Bmp5.

Based on the genomic analysis of the *hs_bmp* cluster in GUM202_hs, we predicted that PBDEs isolated from this sponge would have higher degrees of halogenation, due to the extra putative halogenase, and could be doubly hydroxylated due to the extra cytochrome p450 hydroxylase. Consistent with this genotype, the major PBDEs isolated from GUM202 show a higher degree of bromination from previously isolated PBDEs in Guamanian *L. herbacea* specimens, and are doubly hydroxylated, matching the biosynthetic capacity seen in the GUM202 *hs_bmp* pathway. Notably, penta- and hexabromination had not been seen in our Guam sponge collections previously.

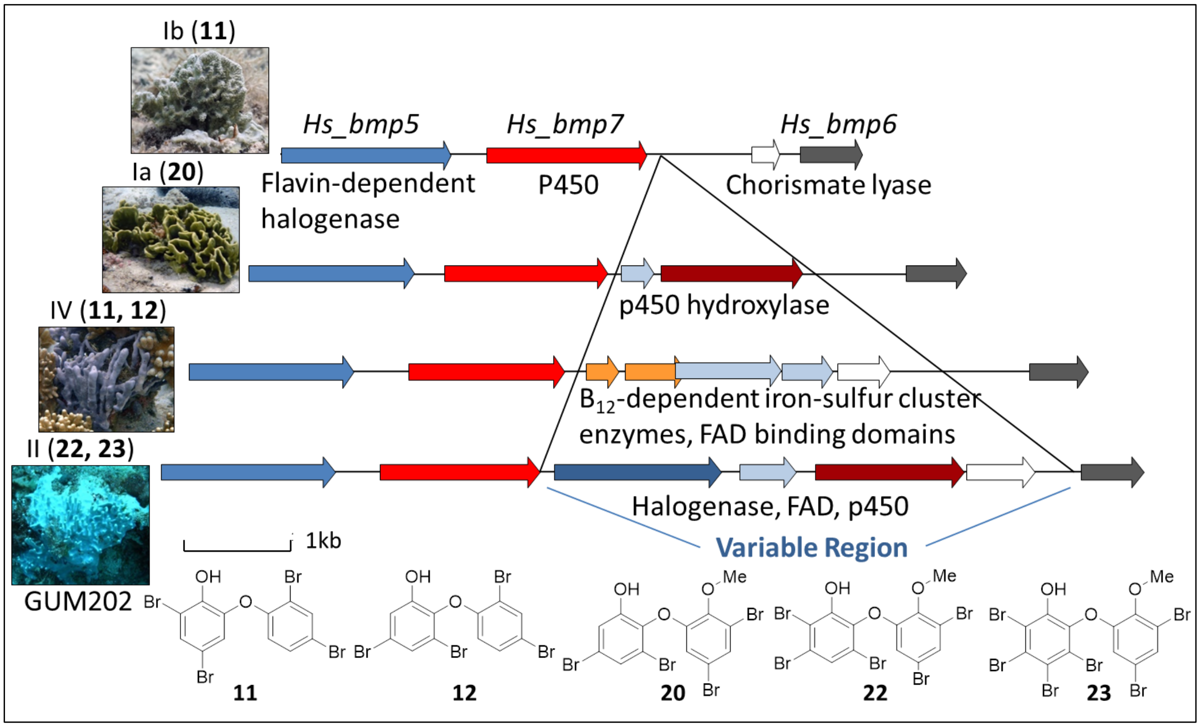


Figure 21. Hs_bmp clusters in four *Hormoscilla spongeliae* genomes

Four *hs_bmp* clusters from *H. spongeliae* genomes from four morphologically distinct Dysideidae sponges. The variable region of the gene cluster contains extra genes that correspond to the chemistry produced by each strain. The additional p450 hydroxylase is responsible for the extra hydroxylation in the molecules isolated from the clade Ia sponge (**20**), and the GUM202_hs p450 hydroxylase is 99% identical, corresponding with the double hydroxylation observed (**22, 23**). The GUM202_hs cluster contains an extra putative halogenase that is possibly responsible for the extra bromination seen in the PBDEs isolated from this sample.

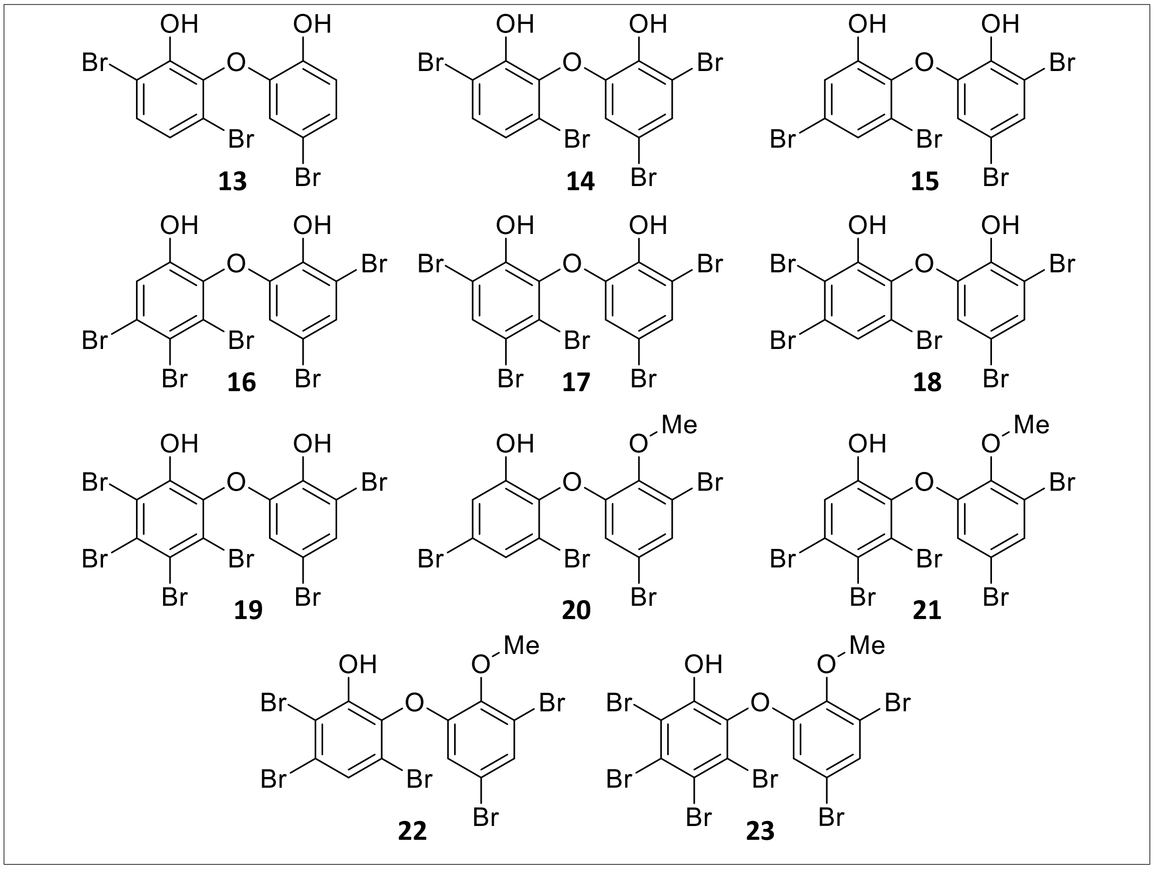


Figure 22. Wide variety of poly-brominated diphenyl ethers (PBDEs) isolated from GUM202

The eleven PBDEs isolated from GUM202 are all doubly hydroxylated, consistent with the presence of a p450 hydroxylase in the GUM202 *hs_bmp* cluster. A second putative halogenase is also encoded in the cluster and is likely responsible for the high degree of bromination observed. Additionally, the tetra-, penta-, and hexabrominated scaffolds all have O-methylated structures (**20-23**).

Eleven PBDEs were isolated from GUM202 (Figure 22) and structurally characterized by tandem mass spectrometry and 1D and 2D NMR spectroscopy in accordance to previously reported NMR chemical shift trends (Chapter 4 Appendix)[68]. Typically, extracts from GUM202 were dominated by a hexabrominated, doubly hydroxylated PBDE (**19**), which is consistent with the additional halogenation potential within the variable region of the *hs_bmp* cluster. Among the remaining PBDEs isolated, all are twice hydroxylated, (one hydroxyl group on each ring) with varying amounts of bromine substitution ranging from three (**13**) to six (**19, 23**). For the tetra-, penta-, and hexa-brominated molecules, the corresponding O-methylated products were also identified (**20-23**).

4.3.5 Potential for Halogenated Pathways in GUM007_hs

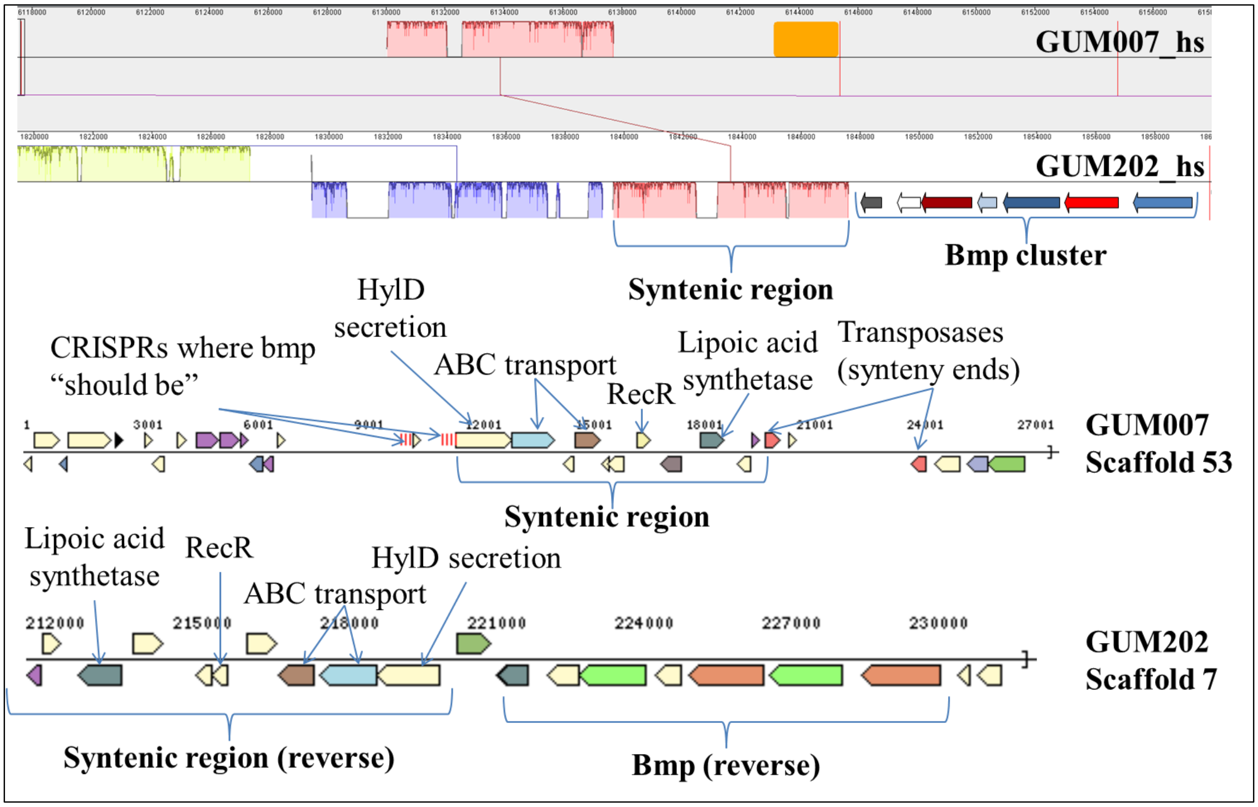


Figure 23. Genomic context of *hs_bmp* cluster in GUM202_hs and syntenic region in GUM007_hs

The genomic region upstream of the *hs_bmp* cluster in GUM202 corresponds to a syntenic region in GUM007_hs. This region contains ABC transport and secretion genes. In GUM007_hs, the syntenic region is flanked by two CRISPR arrays where the *hs_bmp* cluster "should be" and two transposases where synteny ends.

With GUM007 lacking the *hs_bmp* pathway and therefore PBDEs, we wanted to examine what the corresponding region of the genome looks like in GUM007_hs (Figure 23). The region directly preceding the *hs_bmp* cluster in GUM202 has a roughly 8 kb region of synteny with GUM007_hs. In the area where the *hs_bmp* pathway "should be" in GUM007, there are two series of CRISPR arrays, suggesting a mobile region of the genome where the *hs_bmp* cluster can move in or out of the genome. On the other side of the syntenic region, two transposases mark the end of synteny between GUM007_hs and GUM202_hs.

PBDEs are not the only halogenated natural products from *L. herbacea* sponges; a variety of polychlorinated amino acid-derived compounds have also been reported. Over the last four decades, multiple groups have reported that when PBDEs are found, polychlorinated amino acid derivatives are not found, and vice versa[37]. However, in the isolation of dysidenin (**3**), a hexachlorinated amino acid-derived metabolite from an Australian *Dysidea herbacea*, the authors describe two major fractions, the first one being a PBDE, and the second being dysidenin[11]. To our knowledge, this is the only reported case where both PBDEs and polychlorinated amino acid-derived compounds are present together. It is possible that the specimen was a mixed population of sponges, but without further evidence, this cannot be known. Regardless, co-occurrence of the two types of metabolites may be a rare case, as this hasn't been seen since the 1977 isolation of dysinenin.

We did not observe any polychlorinated molecules in the extracts of GUM007, which lacks PBDEs, but we queried the genome for halogenases nevertheless. We searched the GUM007_hs genome using the *dysB1* putative halogenase gene, amplified previously from a *H. spongeliae* containing *L. herbacea* sponge that makes a leucine-derived hexachlorinated molecule, (-)-neodysidinin, as a query[41]. Two matches were found using a tblastn search against the GUM007_hs genome, however, they have low homology to *dysB1*, with 28.3% and 23.2% pairwise identity. The similarity between *dysB1* and *barB1*, from a free-living cyanobacteria that makes the chlorinated leucine-derived barbamide, was 93%, suggesting that the halogenases found by BLAST in GUM007_hs are not of the same class that halogenates leucine in barbamide and similar molecules. They do, however, appear to be in the genomic context of possible secondary metabolite gene clusters, as both are identified within putative gene clusters when antiSMASH4 is run with clusterblast enabled (Table 4). A third putative halogenase is annotated by JGI as a tryptophan halogenase, and also appears to be in the context of a putative biosynthetic gene cluster, although ClusterFinder did not designate the genomic region around it as a putative gene cluster (Table 4). However, BLAST analysis of the surrounding genes, reveals ORFs with high homology to a phenylacetate-CoA ligase, a methyltransferase, a thioredoxin, and a multi-drug resistance efflux transporter, all suggestive of a secondary metabolite gene cluster. GUM007 is a unique *L. herbacea* sponge, in that no halogenated products are seen from this collection, and we confirmed that the hs_bmp cluster is not present in the GUM007_hs genome. There are, however, three putative halogenases that do not have high homology to the *dysB1* gene, but may be involved in the production of other halogenated secondary metabolites.

Table 4. Halogenase genes from GUM007

Three putative halogenases were found in the GUM007 genome. They all reside in operons that have hallmarks of secondary metabolite biosynthetic gene clusters. ClusterFinder (CF) identified two of these putative secondary metabolite clusters.

JGI Gene ID	JGI Annotation	BLAST hit [organism] (cov%/id%)	CF	Genomic Context
Ga0115830_10222	Ectoine hydroxylase-related dioxygenase, (PhyH) family	phytanoyl-CoA dioxygenase family protein [*Planktothrix tepida*] (85%/52%)	Yes	Methyltransferase (MT), dioxygenase (DO), epimerase (E), transaminase (TA)
Ga0115830_102250	non-haem Fe2+, alpha-ketoglutarate-dependent halogenase	halogenase [*Nostoc* sp. 'Peltigera membranacea cyanobiont'] (98%/52%)	Yes	Phenylproionate DO, acetyltransferase (AT), E, MT, decarboxylase
Ga0115830_11198	Tryptophan halogenase	tryptophan 7-halogenase [*Nostoc* sp. PCC 7524] (95%/53%)	No	Phenylacetate CoA ligase, deoxyribosyltransferase, MT, multidrug resistance protein

4.3.6 Genome Mining Leads to Novel Dysinosins

The advent of bacterial whole genome sequencing uncovered a wealth of cryptic biosynthetic gene clusters which spurred the development of the field of genome mining. Additionally, advances in synthetic biology have allowed researchers to interrogate and characterize numerous biosynthetic gene clusters encoding natural products. Curation of known gene clusters aids in the identification of related compounds through gene cluster similarity comparisons. Genome mining identified a putative dysinosin pathway in the *H. spongeliae* strain from GUM007. The dysinosins have been reported from *L. herbacea* sponges previously and share the majority of their structure with aeruginosides[28,29]. Dysinosins are potent inhibitors of the blood coagulation cascade factor VIIa and the serine protease thrombin[28]. Activity assays showed that the desulfated dysinosin D has ten times less activity against both targets, while the glycosylation seen in dysinosin B increased activity against factor VIIa and had a slight loss in selectivity for thrombin[29]. Structural studies showed that the hydroxyl group on the 2-carboxy-6-hydroxyoctahydroindole (Choi) moiety lies within the P2 pocket of thrombin and modification of this region could have the potential for improved activity and selectivity. Understanding the biosynthetic basis of these bioactive molecules would provide a platform for rational molecule design for more effective and selective inhibitors in the blood coagulation cascade.

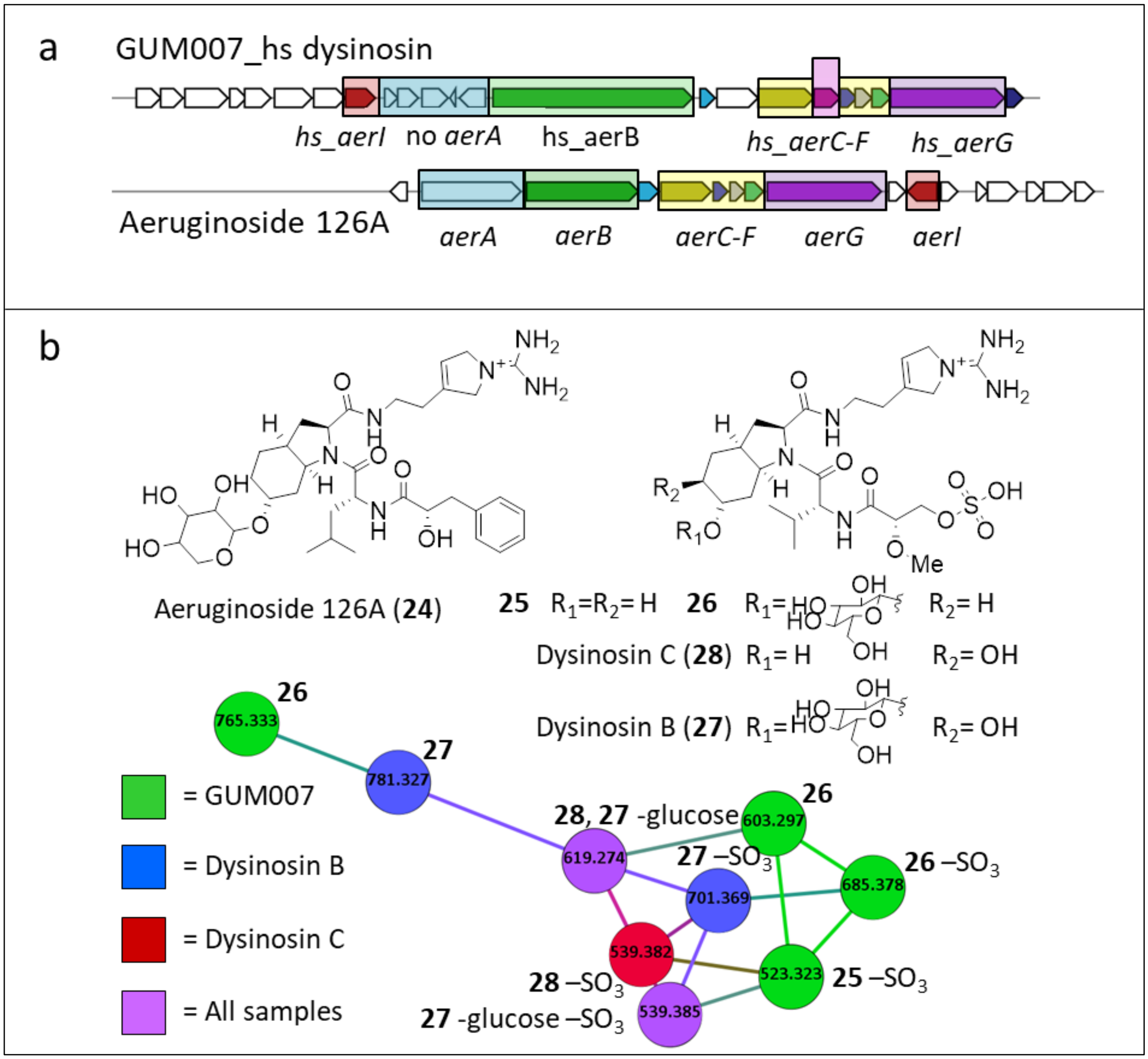


Figure 24. Gene cluster comparison and molecular network of dysinosins

Figure 24a shows the gene cluster found in GUM007_hs compared with the gene cluster for aeruginoside 126A (**26**). In GUM007_hs *aerA* is missing and *aerB* is expanded and contains a sulfotransferase domain. GUM007_hs also contains an extra gene shown in a pink box that is annotated as an isomerase. The molecular network includes dysinosins B (**27**) and C (**28**), and the two new dysinosins (**25, 26**) in GUM007. Parent masses and the neutral mass loss of sulfate for each structure are displayed inside each node. Nodes are colored according to their source sample.

An NRPS cluster in GUM007_hs has very high homology to the aeruginoside cluster deposited in MIBiG (BGC0000297) (Figure 24a). The GUM007_hs NRPS cluster contains homologous genes to *aerB*, an NRPS gene with an A domain selective for leucine, *aerC-F*, genes responsible for assembling the unique Choi moiety, *aerG*, another NRPS gene, predicted to activate and incorporate the Choi moiety, and *aerI*, a putative glycosyltransferase[108]. The only notable gene missing is *aerA*, the NRPS loading domain, which activates phenylpyruvate. When comparing the structures of aeruginoside and the related dysinosins, isolated from Australian *L. herbacea* specimens, the major difference is in the starting units: in aeruginoside, the unusual phenyllactic acid and in the dysinosins, a sulfated glyceric acid. The *aerB* homologue in GUM007_hs, is likely responsible for loading the sulfated glyceric acid, as a sulfotransferase domain is present. Additionally, adenomethyltransferase and Fkbh domains are present in *aerB*, which can also be implicated in constructing the unusual sulfated glyceric acid moiety. AerB contains two peptidyl carrier protein (PCP) domains, one condensation (C) domain, an A domain with no consensus specificity predicted by bioinformatic tools, and an epimerization (E) domain. Taken all together, AerB, a large multimodular NRPS-type complex, appears to be responsible for the activation of a sulfated glyceric acid and addition of an unspecified D-amino acid. The remainder of the biosynthetic gene clusters, and their corresponding molecules, are nearly identical. We therefore believe that this represents a putative biosynthetic gene cluster for a dysinosin-like molecule. Isolation attempts are underway, and comparison of GUM007 extracts with dysinosin standards led us to two masses for two putative novel dysinosins **25** and **26** (Figure 6b).

The logical bioinformatic basis for a putative dysinosin cluster led us to examine GUM007 extracts for dysinosin-like molecules. To aid in our LC-MS/MS-based molecular search, we obtained two dysinosin standards from the Quinn Lab at Griffith University for comparison. We discovered masses representing two new dysinosins (**25, 26**). These molecules are identical to dysinosins B (**27**) and C (**28**) except they have one hydroxylation on the Choi moiety instead of two. Further interrogation of MS/MS data revealed that in each of the standards, a neutral mass loss can be observed, in source and via collision induced dissociation, which is consistent with the loss of sulfate (79.95) from the terminal glyceric acid residue. We also observed this neutral loss of 79.95 in the new GUM007 dysinosins. GNPS[69] molecular networking produced a network of parent masses and corresponding neutral mass loss for both standards and the two new dysinosins (Figure 6b). The molecular network shows that all three samples contain m/z 619.274 and m/z 539.385, corresponding to dysinosin C with and without sulfate and dysinosin B without glucose, with and without sulfate, as is expected. The presence of these masses is not consistent across multiple GUM007 extracts. It is possible that there was some contamination of the extract, carryover in the MS column, or minute amounts of these known dysinosins may have only been picked up in this sample. The new dysinosins were found to have masses of 603.2806 (0.17 ppm error from theoretical m/z 603.2807) and 765.3327 (1.04 ppm error from theoretical m/z 765.3335).

Further examination of the MS^2 spectrum of these masses gave fragmentation patterns consistent with (**25**) (Figure 25) and (**26**) (Figure 26). As noted previously, the in source and collision induced loss of sulfate can be seen in MS^1 and MS^2 for both dysinosins, and is the most abundant species in both cases. Most fragmentation on this molecule occurs in the guanidyl group. The next major fragment in **25** represents the de-sulfated core structure with a loss of the two terminal amino groups on the guanadyl moiety (m/z 481.2934). The next major fragments observed are the Choi and guanadyl core with successive loss of terminal amino groups (m/z 322.2230, 305.1908, 280.2031). This successive loss is seen again in just the guanadyl fragment (m/z 155.1001, 140.1049, 113.1090). The ‘a ion’ produced when the valine is fragmented from the Choi can also be seen (m/z 174.1119), as can the ‘b ion’ (m/z 202.1006). Finally, the valine immonium ion is seen (m/z 72.0831). Similar fragmentation is seen in the MS^2 for **26**. Again, the main ion is the loss of sulfate (m/z 685.3779), followed by a loss of glucose (m/z 523.3259). The glycosylated fragment with a loss of sulfate and loss of both termainl amino groups from the guanadyl moiety can be seen (m/z 643.3323). The glycosylated Choi and guanadyl core is seen (m/z 484.2889) as well as the loss of both terminal amino groups (m/z 442.2736). The remaining fragments labelled with an asterisk match those seen in **25** (Figure 25). Although we have not yet isolated enough of the new dysinosins for full structural elucidation via NMR, we are confident that the MS/MS data, molecular networking, and genomic information support the discovery of two new dysinosins. This discovery represents the first example of genome mining in a metagenome assembled genome of an uncultivated symbiont resulting in new structures.

Figure 25. MS/MS spectra of new dysinosin (25)

The upper panel shows the MS_1 of **25**. The lower panel shows the MS_2 for the 603.2806 parent mass, with structures of observed fragment ions. The desulfated mass was always observed as both a neutral mass loss in source and via collision induced dissociation.

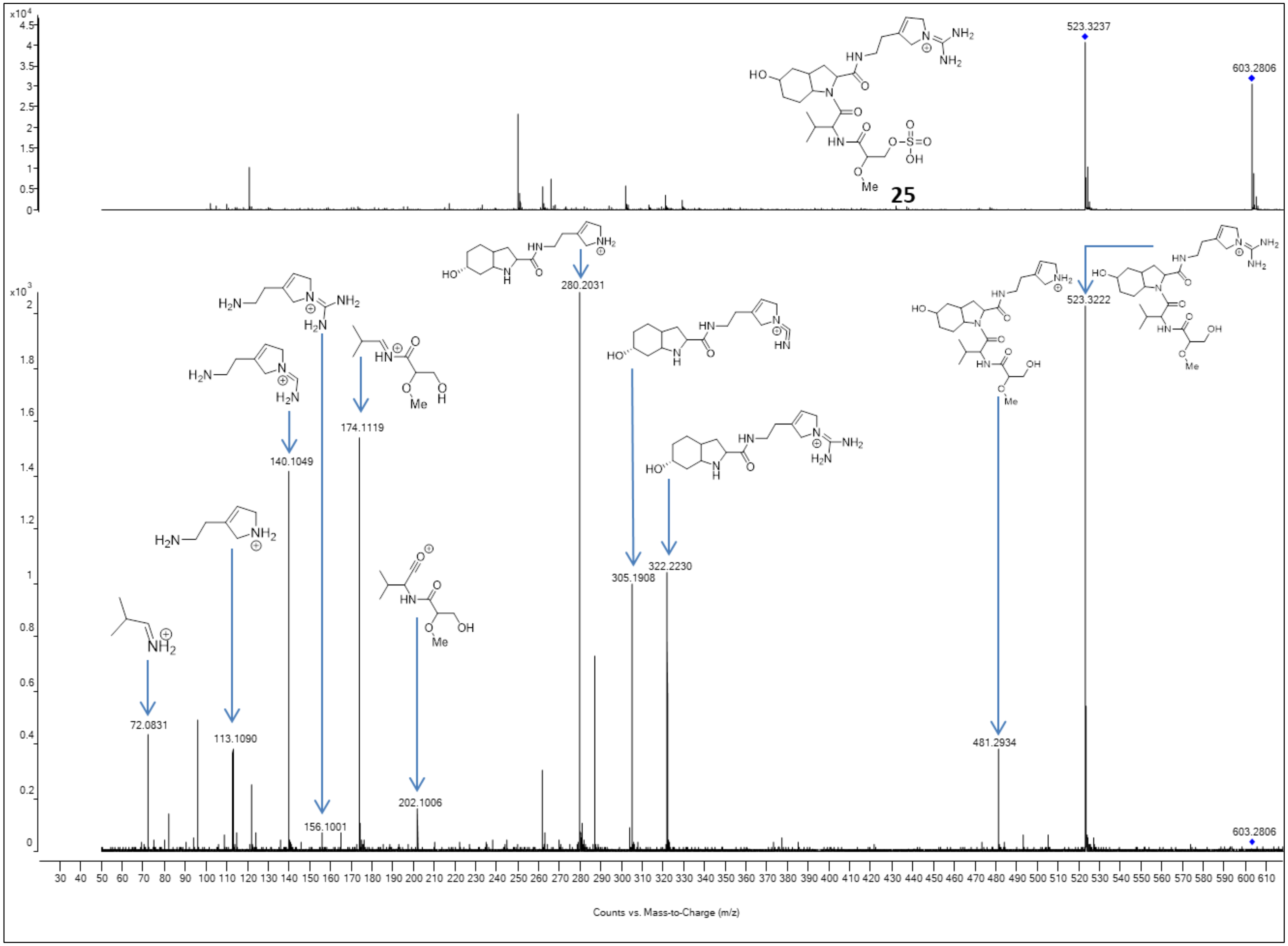
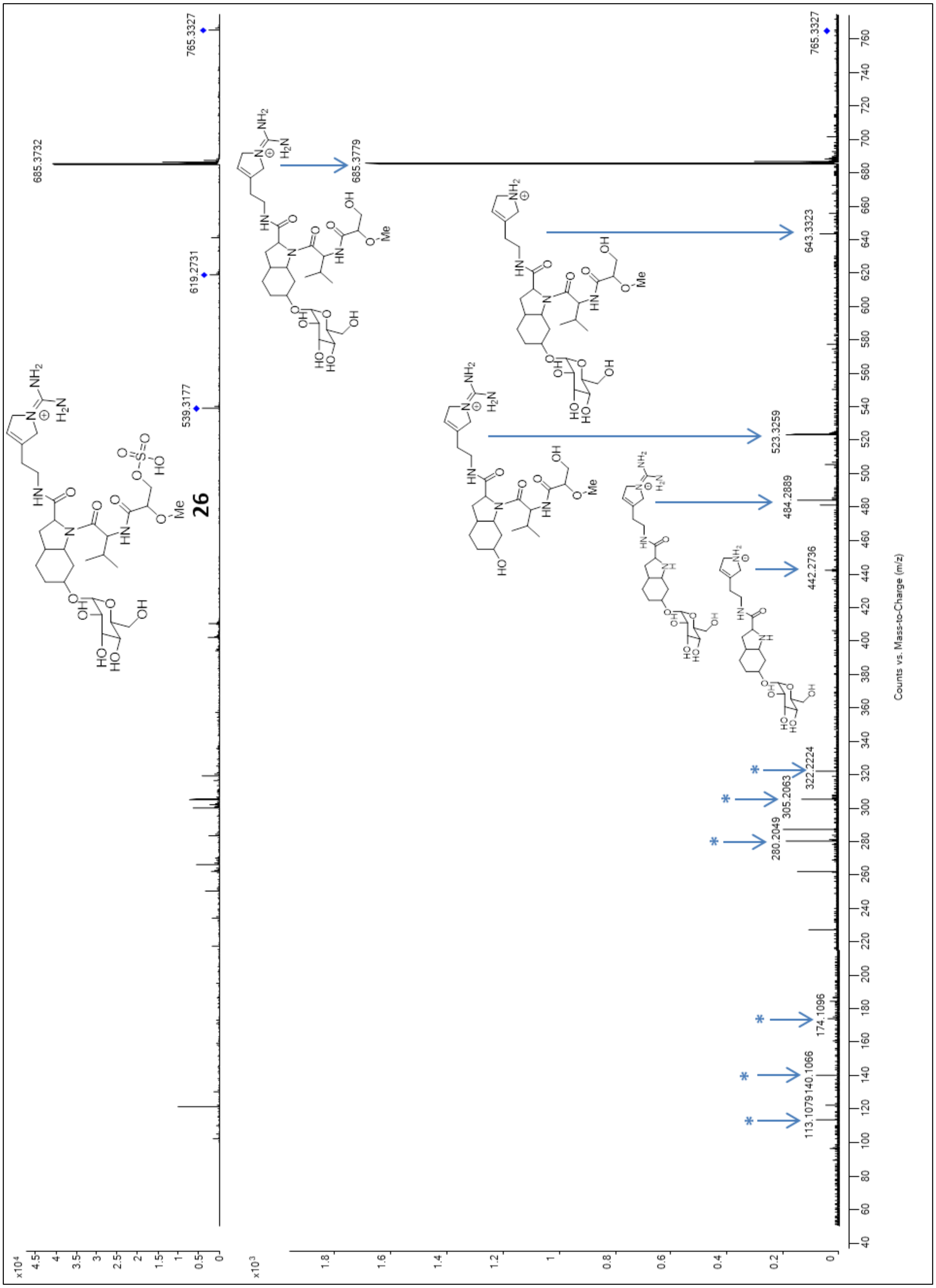

x10^4
4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
523.3237
603.2806
25
x10^3
2
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
72.0831
113.1090
140.1049
156.1001
174.1119
202.1006
280.2031
305.1908
322.2230
481.2934
523.3222
603.2806
30 40 50 60 70 80 90 100 110 120 130 140 150 160 170 180 190 200 210 220 230 240 250 260 270 280 290 300 310 320 330 340 350 360 370 380 390 400 410 420 430 440 450 460 470 480 490 500 510 520 530 540 550 560 570 580 590 600 610
Counts vs. Mass-to-Charge (m/z)

Figure 26. MS/MS spectra of new glycosylated dysinosin (26)

The upper panel shows the MS^1 of **26**. The lower panel shows the MS^2 for the 765.3327 parent mass, which is the glycosylated verson of **25**. Observed ions are drawn with arrows to their corresponding m/z. Multiple fragment ions also seen in Figure 25 are marked with an asterisk


x10^4
4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
765.3327
685.3732
619.2731
539.3177
26
x10^3
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
765.3327
685.3779
643.3323
523.3259
484.2889
442.2736
322.2224
305.2063
280.2049
174.1096
113.1079
140.1066
*
*
*
*
*
*
40
60
80
100
120
140
160
180
200
220
240
260
280
300
320
340
360
380
400
420
440
460
480
500
520
540
560
580
600
620
640
660
680
700
720
740
760
Counts vs. Mass-to-Charge (m/z)

4.4 Conclusions

Symbiotic associations between microbes and their eukaryotic hosts provide fascinating systems to study interdependence and the use of chemical entities in these holobiont systems. *L. herbacea* sponges have a long history of natural product discovery and an even longer association with their symbiotic cyanobacteria, *H. spongeliae*. *H. spongeliae* have long been implicated as a source for many of the varied natural products isolated from *L. herbacea*. The high quality draft genomes generated for two *H. spongeliae* populations afforded a complete look at the biosynthetic capacity of these uncultivatable symbionts. The genetic evidence presented here is the strongest yet for cyanobacterial production of multiple classes of compounds isolated from these sponges. Furthermore, we found the largest expansion yet of the *hs_bmp* gene cluster in the GUM202_hs genome, which corresponds with the most varied set of PBDEs isolated from the Guamanian *L. herbacea* GUM202. Genome mining and identification of a putative dysinosin pathway informed the search for dysinosins in the GUM007 extracts. Indeed, this genomic information spurred chemical search lead to two new dysinosins. Comparison of *H. spongeliae* genomes with well characterized free-living cyanobacteria exposed multiple avenues for exploring these symbionts' recalcitrance to cultivation in the lab. We determined that an incomplete histidine pathway may be the likely culprit, along with incomplete thiamine and biotin pathways, both necessary cofactors. Additionally, *H. spongeliae's* relation and shared genomic features with known coral pathogens suggests a shared pathogenic lifestyle, which for *H. spongeliae* morphed into a symbiosis with sponges. Overall, the metagenome assembled genomes for these captivating symbionts revealed a wealth of information into their lifestyle and capacity to produce bioactive natural products. Further efforts to cultivate these cyanobacteria can now be informed by genomic information and numerous orphan gene clusters are awaiting exploration.

4.5 Acknowledgements

The work presented in Chapter 4 is the result of collaborations with Eric Allen's lab at the Scripps Institution of Oceanography and Jason Biggs at the University of Guam Marine Lab. Eric Allen, Sheila Podell, and Jessica Blanton all aided in the sequencing and assembly of metagenomes and genomes. Specifically, Sheila Podell assembled the GUM007_hs genome. Jessica Blanton assembled an Illumina metagenome of GUM202 and binned GUM202_hs reads. Michelle Schorn binned PacBio reads and used the binned GUM202_hs Illumina reads for a hybrid assembly. The UCSD IGM Genomics Center prepared sequencing libraries sequenced all samples. Jason Biggs provided lab facilities and collection of sponge samples. Jason Biggs, Jessica Blanton, Vinayak Agarwal, Kate Bauman, and Michelle Schorn collected sponge samples and isolated enriched cell fractions in Guam. Peter Jordan directed the chemical isolation and LCMS, HPLC, and NMR characterization of PBDEs and dysinosins. Mohammad Alanjary helped with phylogenetic analysis and gene cluster networking. Jared Broddrick and David Welkie assisted in metabolic model comparisons and provided helpful discussions. Ron Quinn provided isolated dysinosin standards.

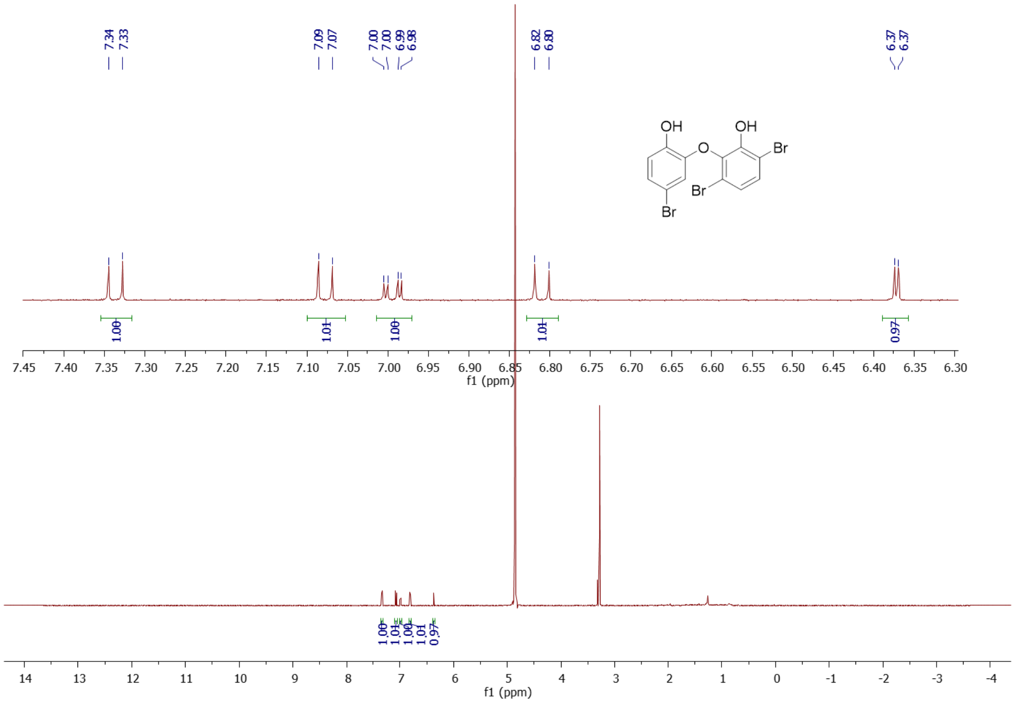
Chapter 4, in part, is currently being prepared for submission of the material. Schorn, M. A., Jordan, P., Podell, S., Blanton, J. M., Agarwal, V., Biggs, J. S., Allen, E. E., and Moore, B. S. The dissertation author is the primary investigator and author of this paper.

4.6 Chapter 4 Appendix

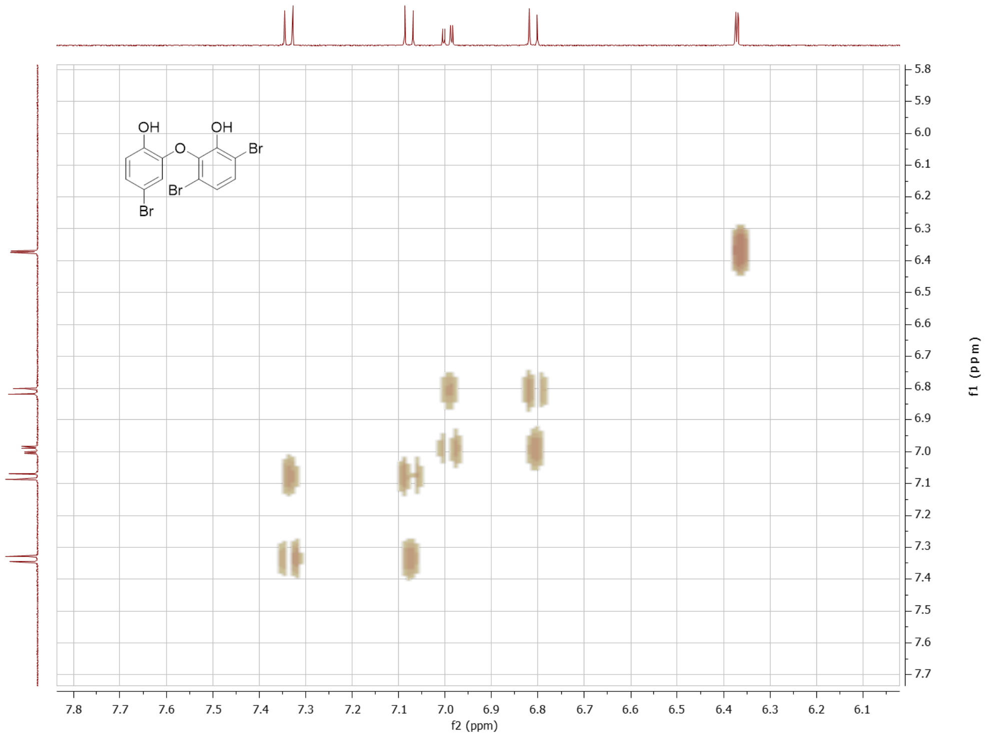
The following spectra were used to characterize the eleven PBDEs isolated from GUM202. Not all distinct peaks yielded enough pure compound to perform NMR, so efforts were focused on isolating PBDEs exclusively seen in GUM202 and not in previously collected Dysideidae sponges from Guam. Compounds 14 - 15 were characterized based on LCMS m/z, fragmentation, and characteristic bromination patterns. Compounds 13 and 16 - 23 were purified in enough quantity to characterize via NMR. Four experiments were run in deuterated methanol for each compound: ^{1}H-NMR, COSY, HSQC, and HMBC.

NMR spectra of compound 13; 1H-NMR, COSY, HSQC, HMBC

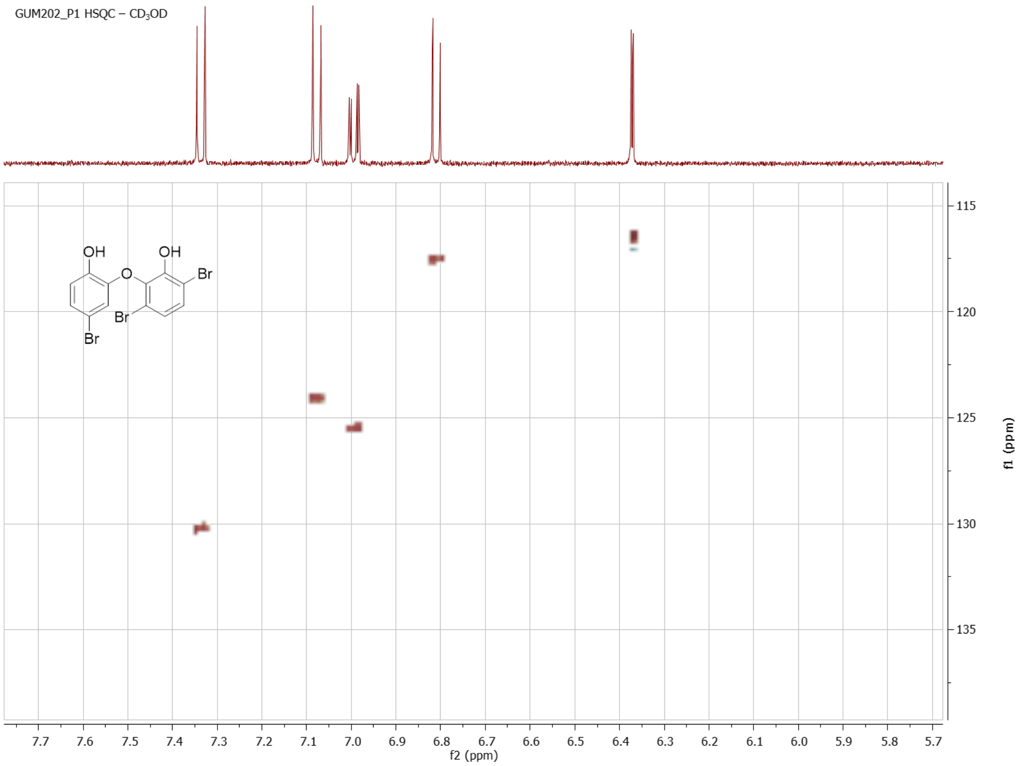
GUM202_P1 ^{1}H-NMR – CD_3OD



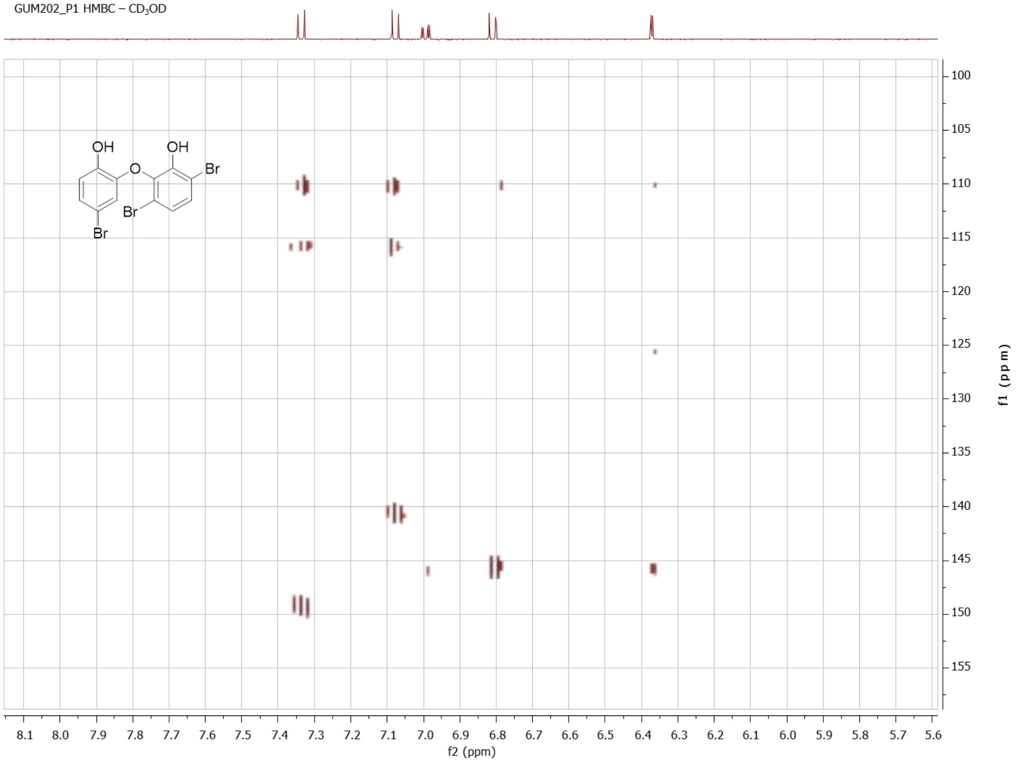
GUM202_P1 COSY – CD_3OD



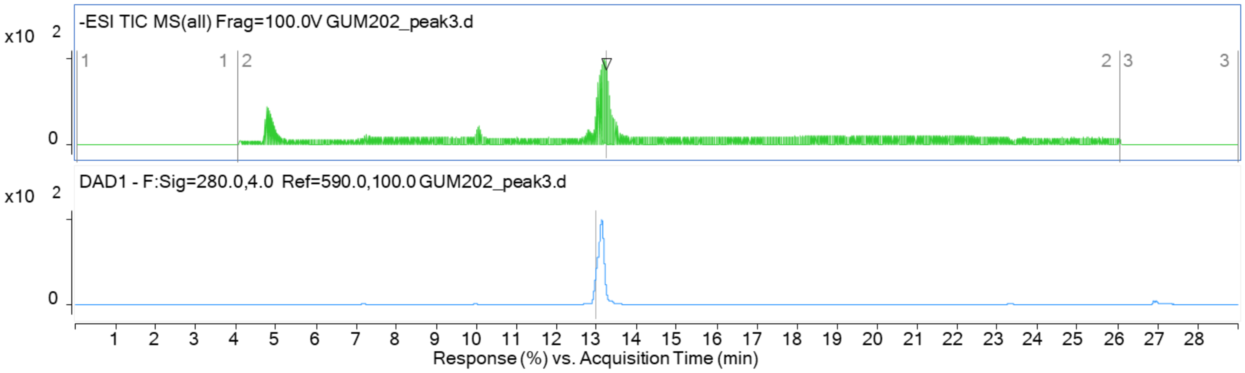
GUM202_P1 HSQC – CD_3OD

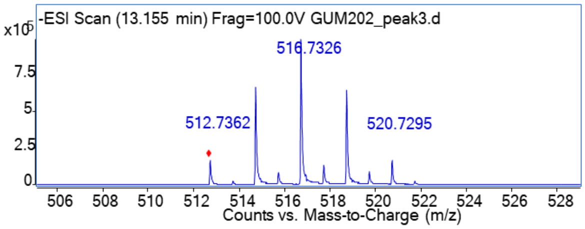


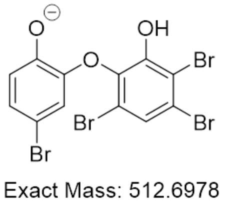
GUM202_P1 HMBC – CD_3OD

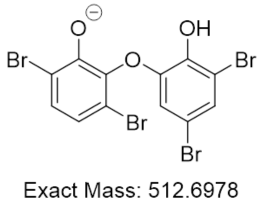


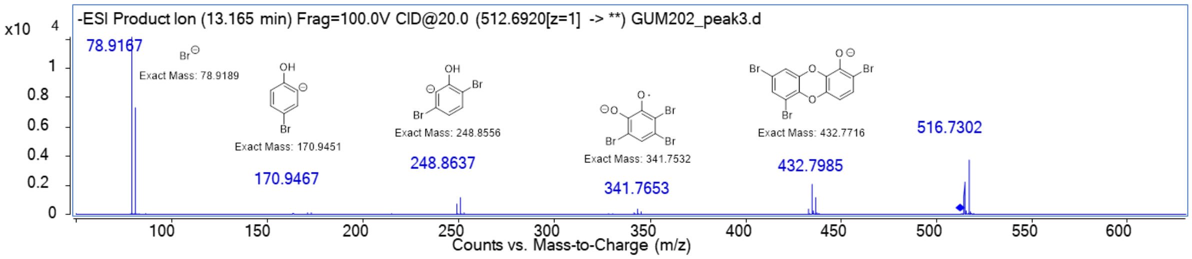
Hi-Res LC/MS of compound 14



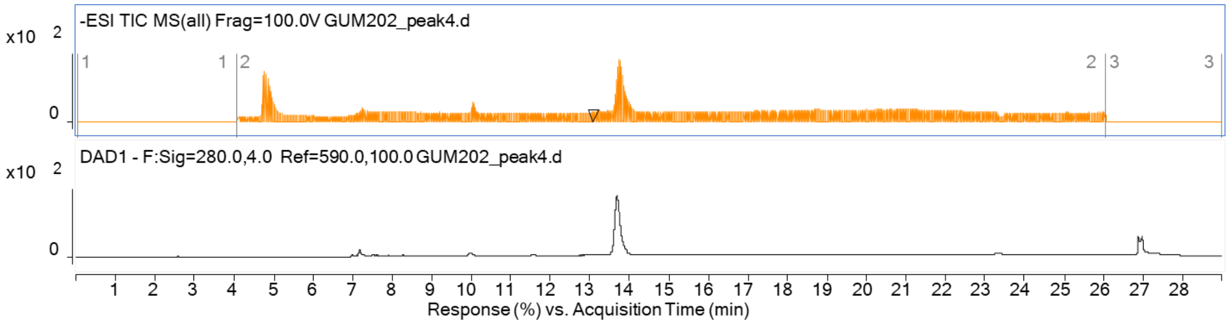


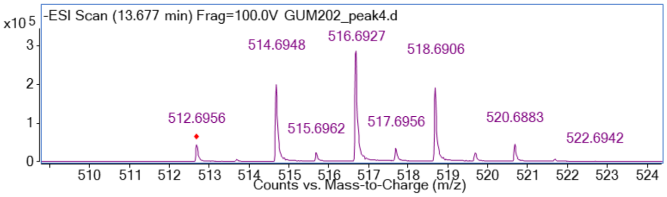


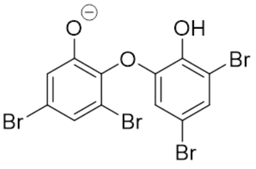




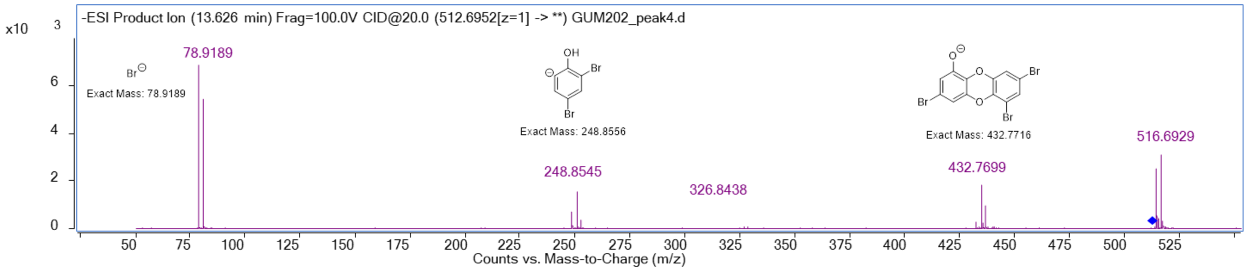
Hi-Res LC/MS of compound 15



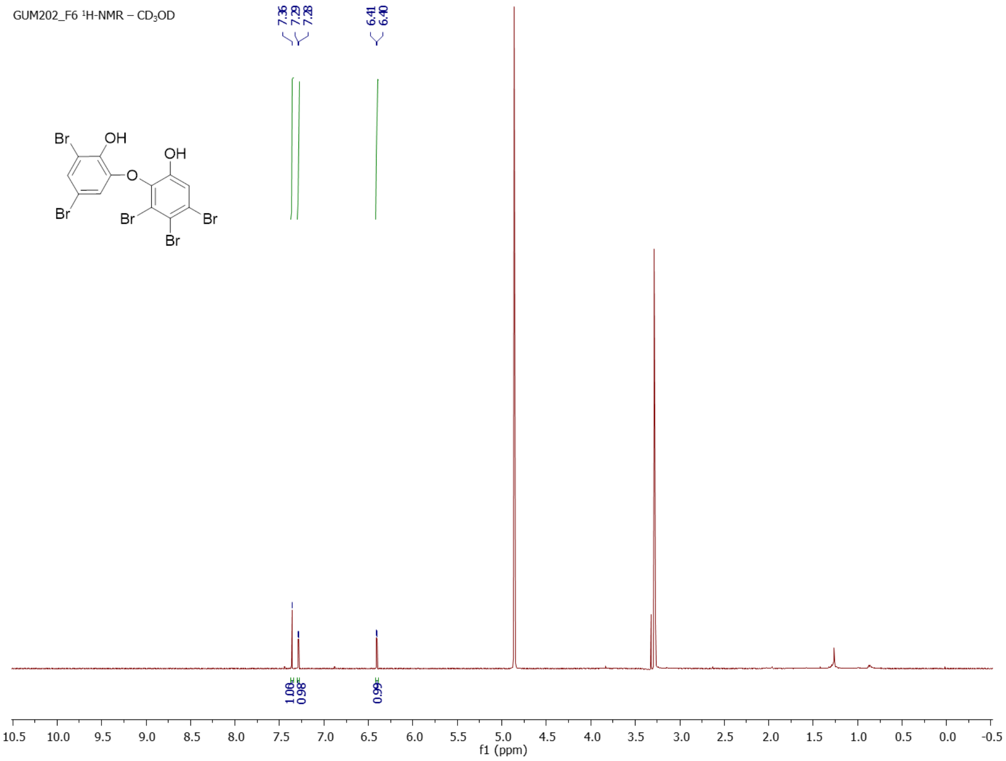


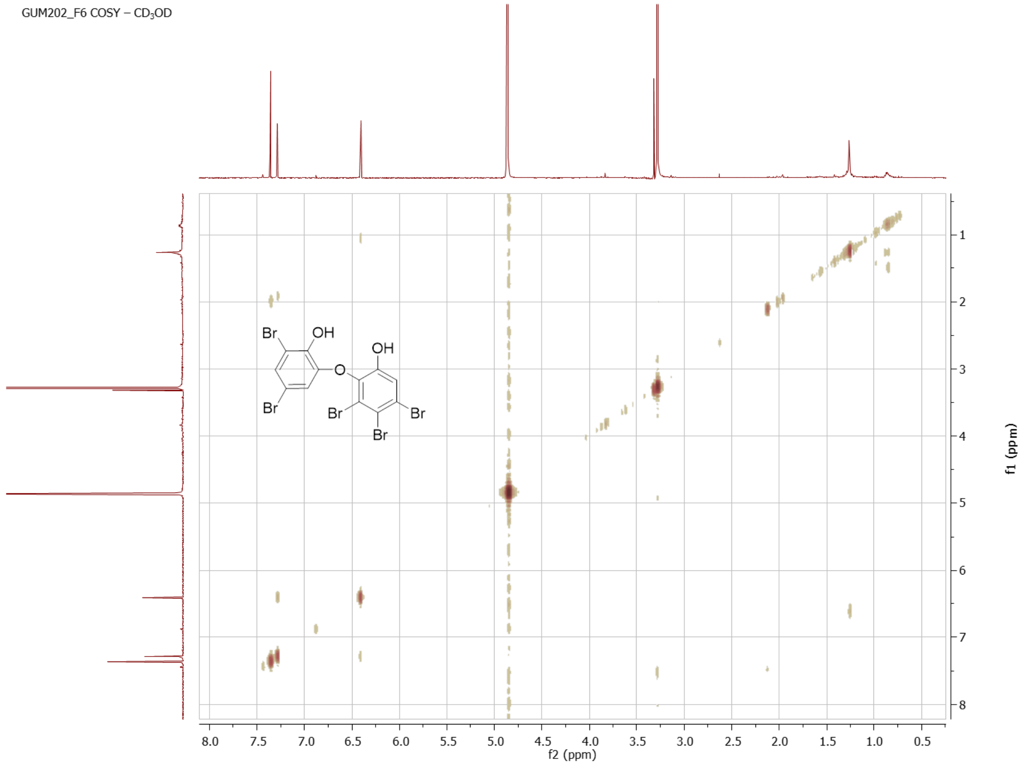


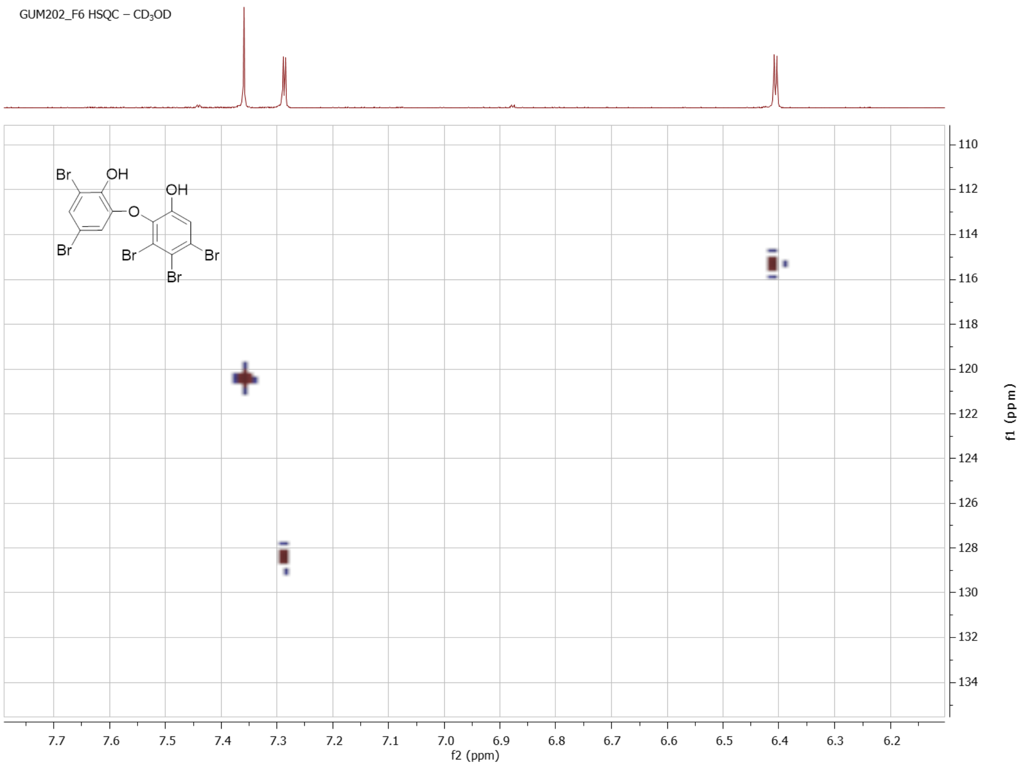
Exact Mass: 512.6978



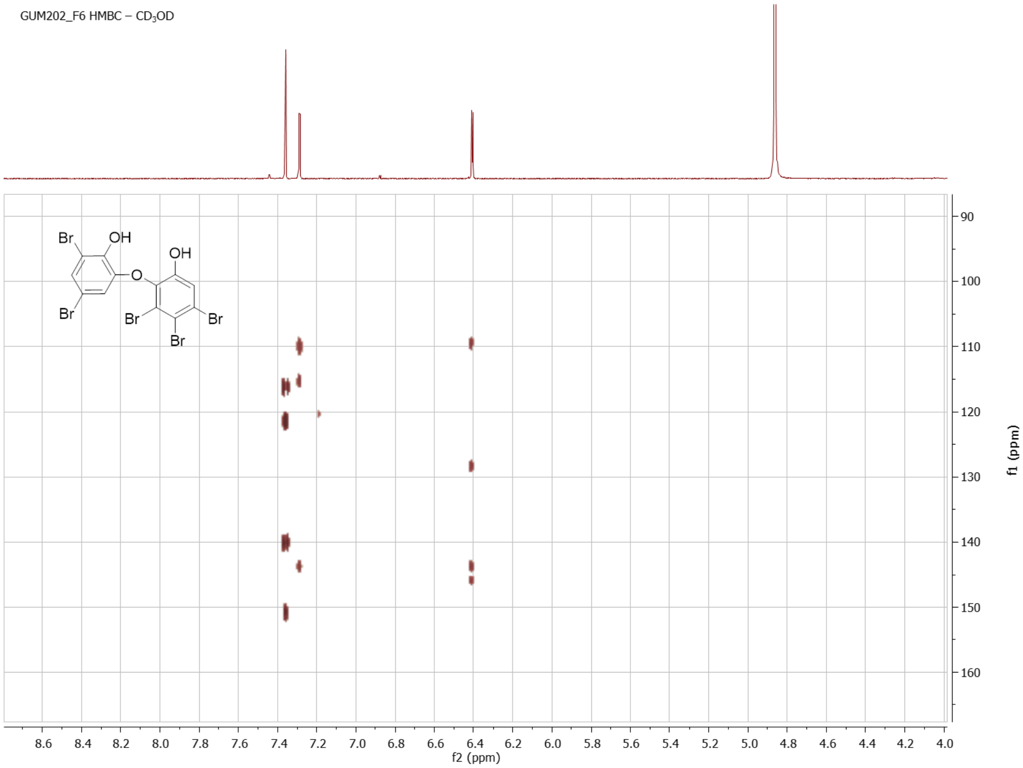
NMR spectra of compound 16; 1H-NMR, COSY, HSQC, HMBC





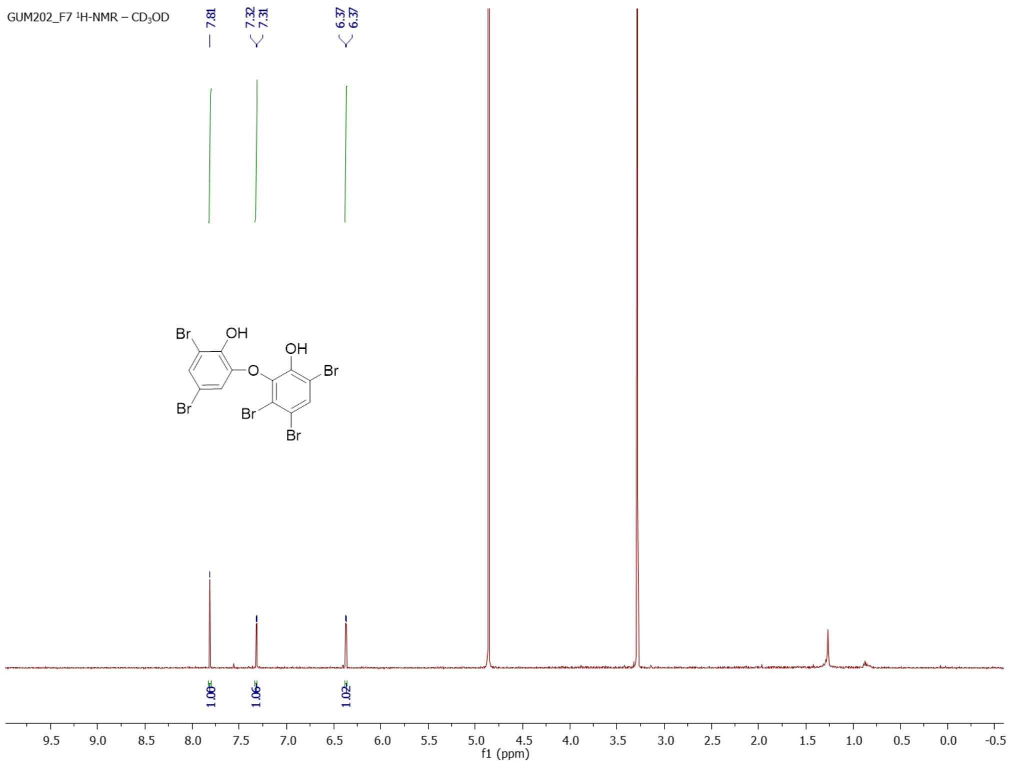


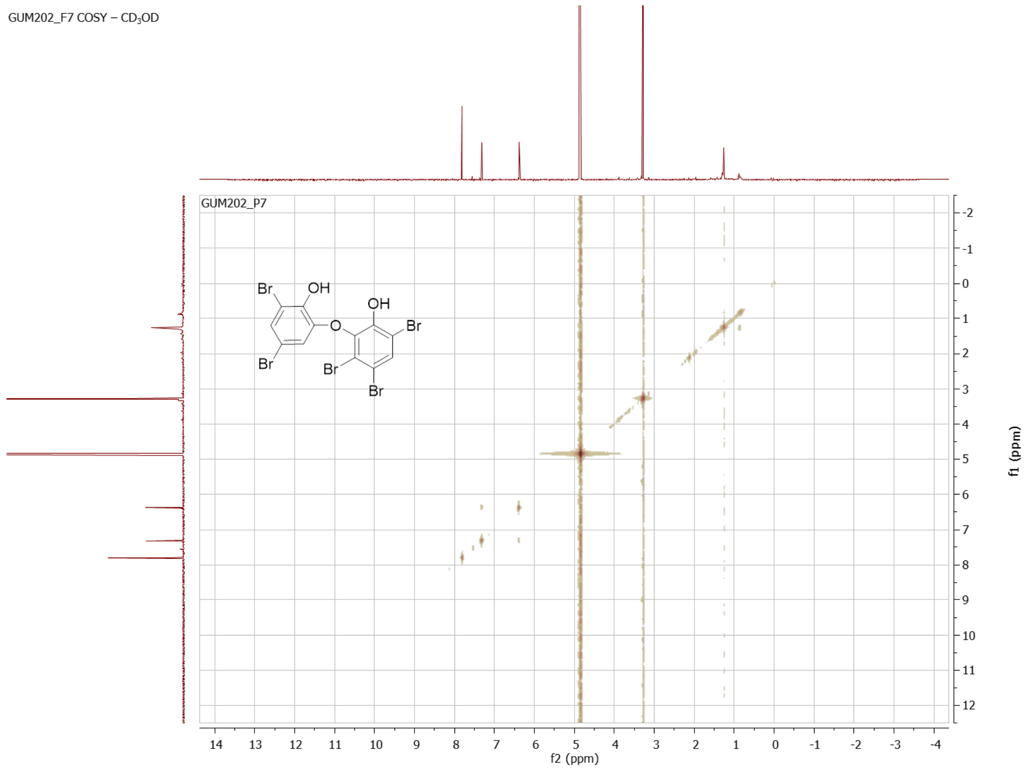
GUM202_F6 HSQC – CD_3OD

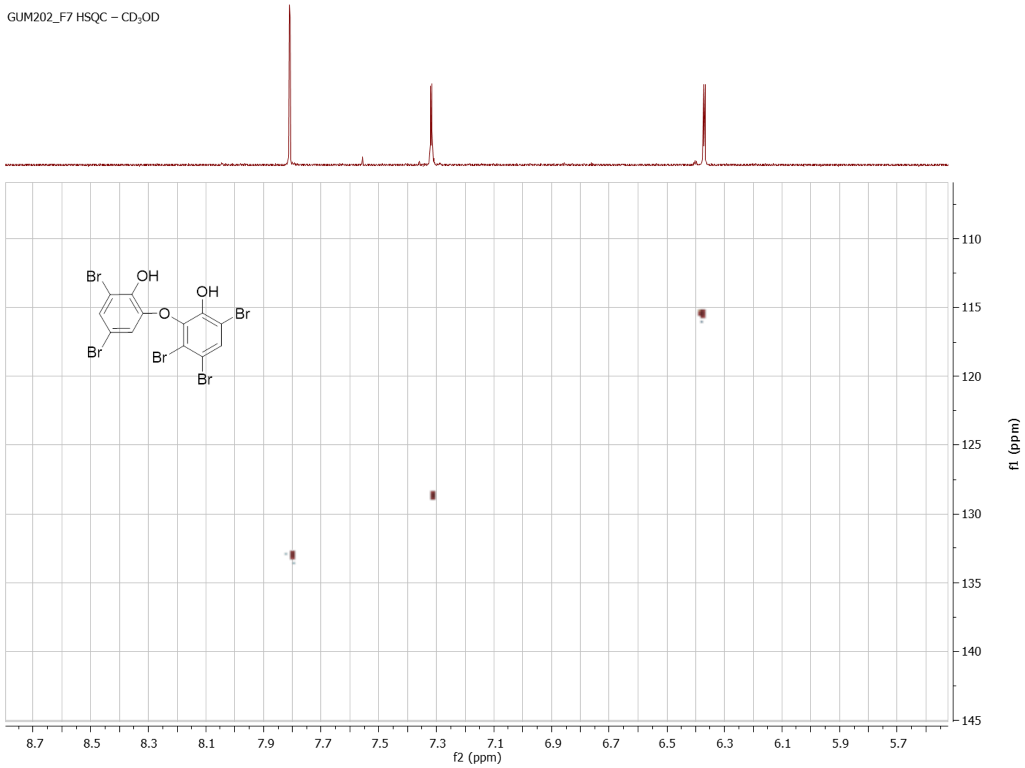


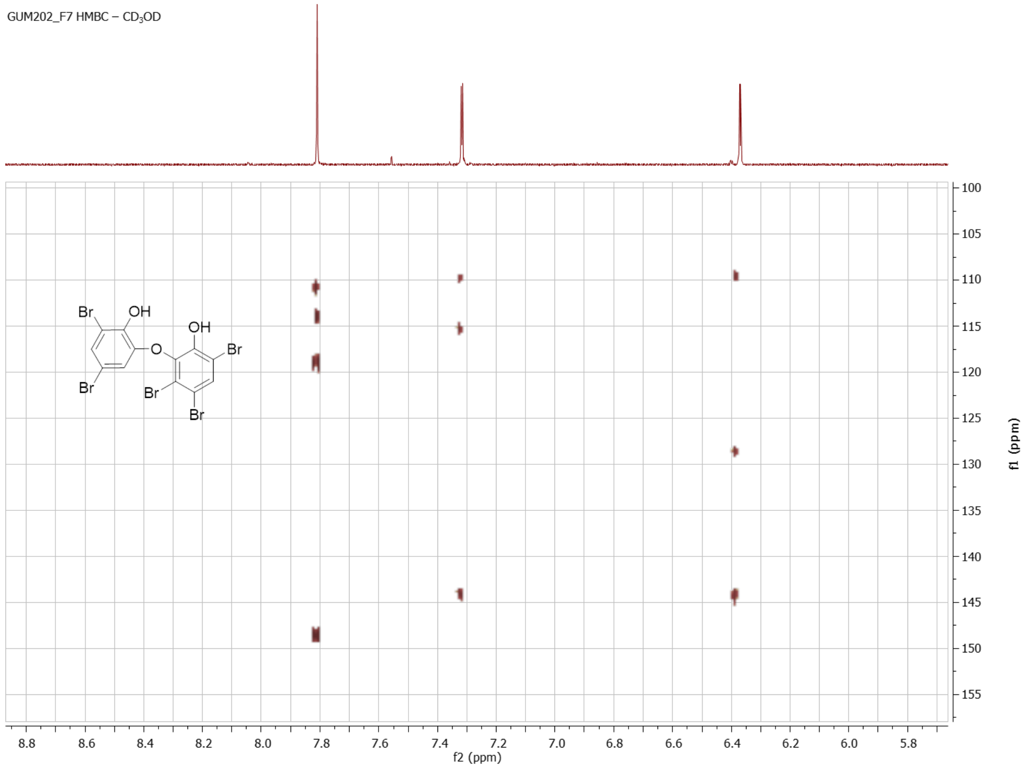
GUM202_F6 HMBC – CD_3OD

NMR spectra of compound 17; 1H-NMR, COSY, HSQC, HMBC

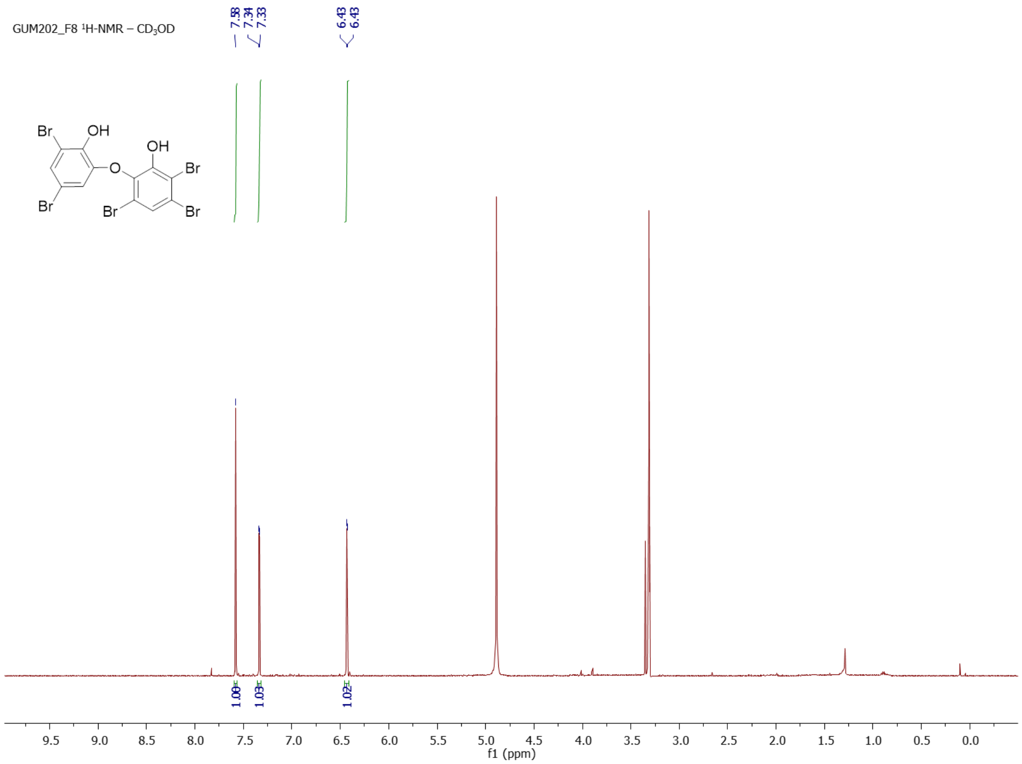


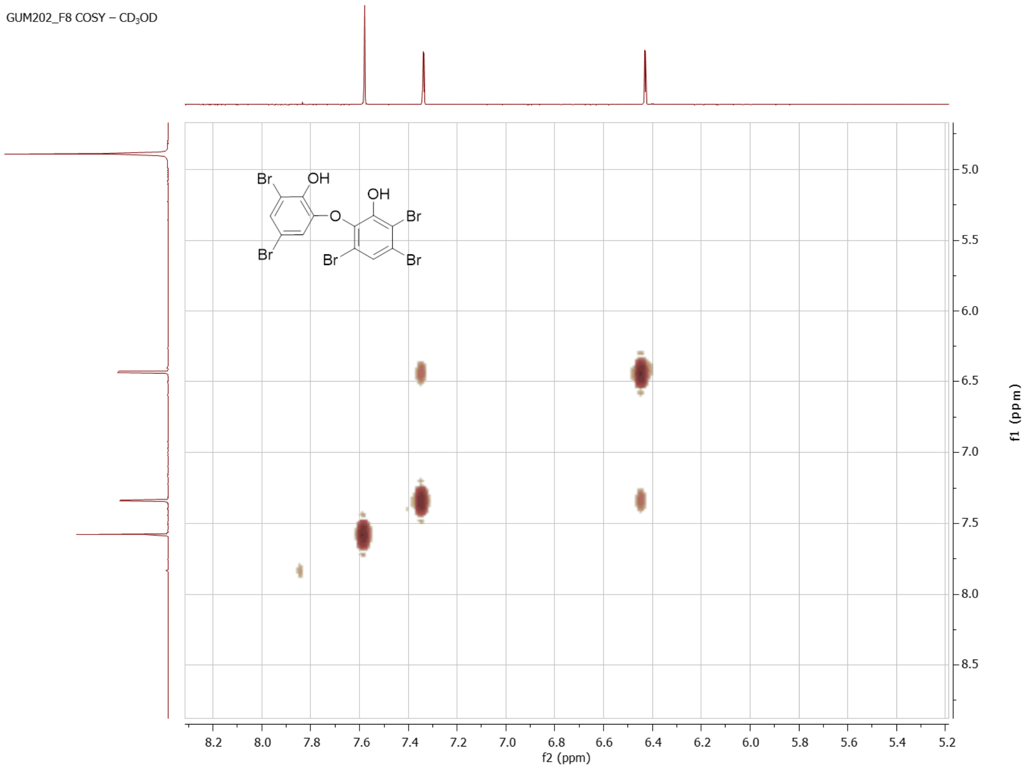


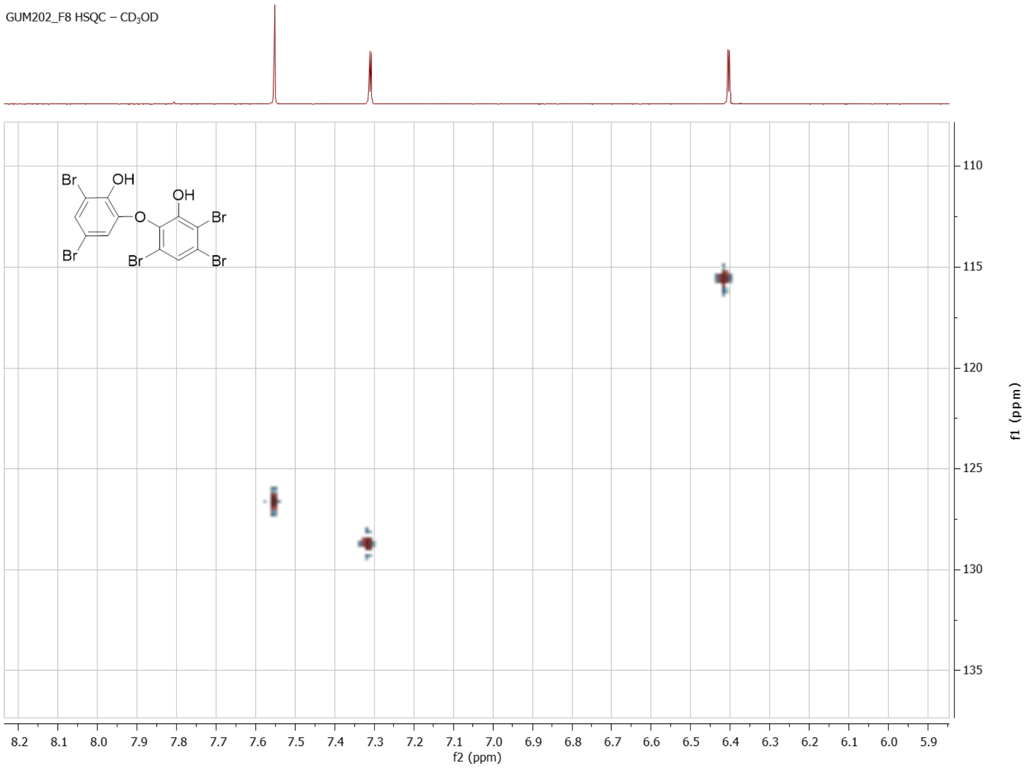

GUM202_F7 HSQC – CD$_3$OD
Br
OH
OH
O
Br
Br
Br
Br
110
115
120
125
130
135
140
145
f1 (ppm)
8.7
8.5
8.3
8.1
7.9
7.7
7.5
7.3
7.1
6.9
6.7
6.5
6.3
6.1
5.9
5.7
f2 (ppm)

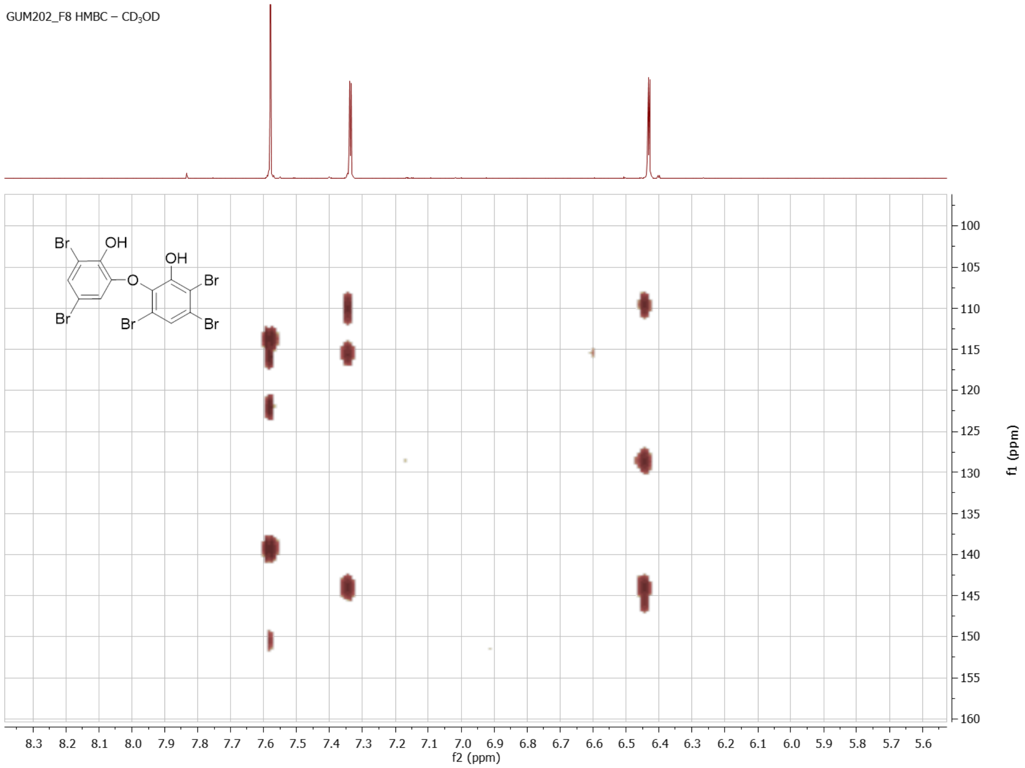

GUM202_F7 HMBC – CD$_3$OD
Br
OH
OH
O
Br
Br
Br
Br
100
105
110
115
120
125
130
135
140
145
150
155
f1 (ppm)
8.8
8.6
8.4
8.2
8.0
7.8
7.6
7.4
7.2
7.0
6.8
6.6
6.4
6.2
6.0
5.8
f2 (ppm)

NMR spectra of compound 18; 1H-NMR, COSY, HSQC, HMBC

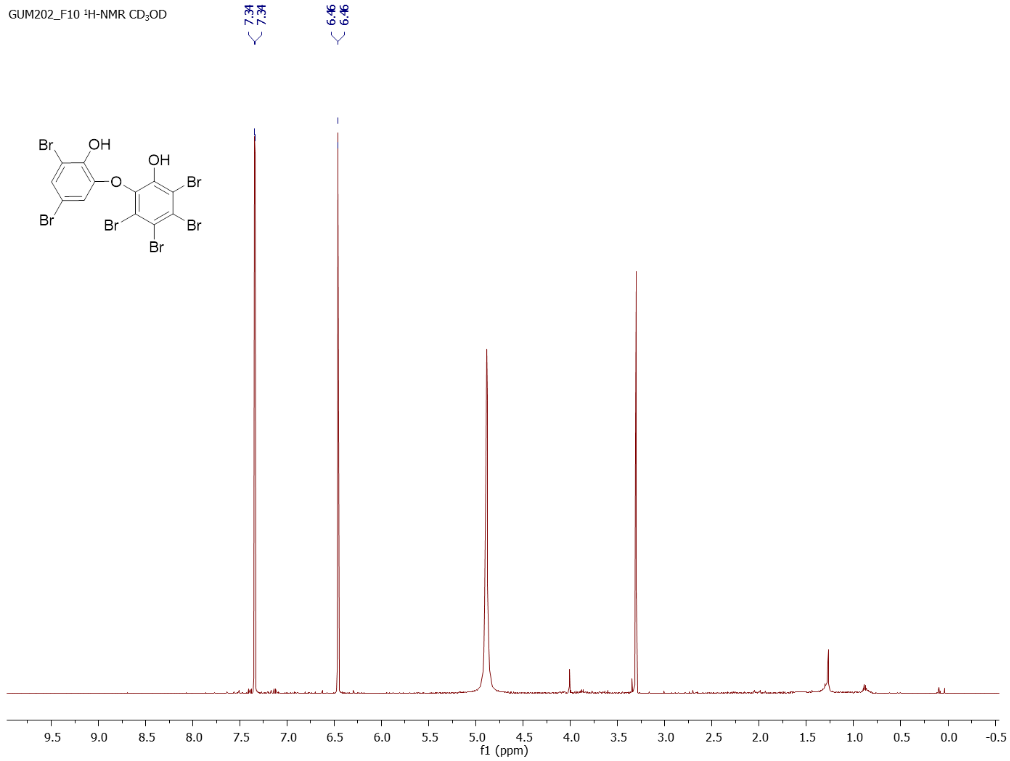


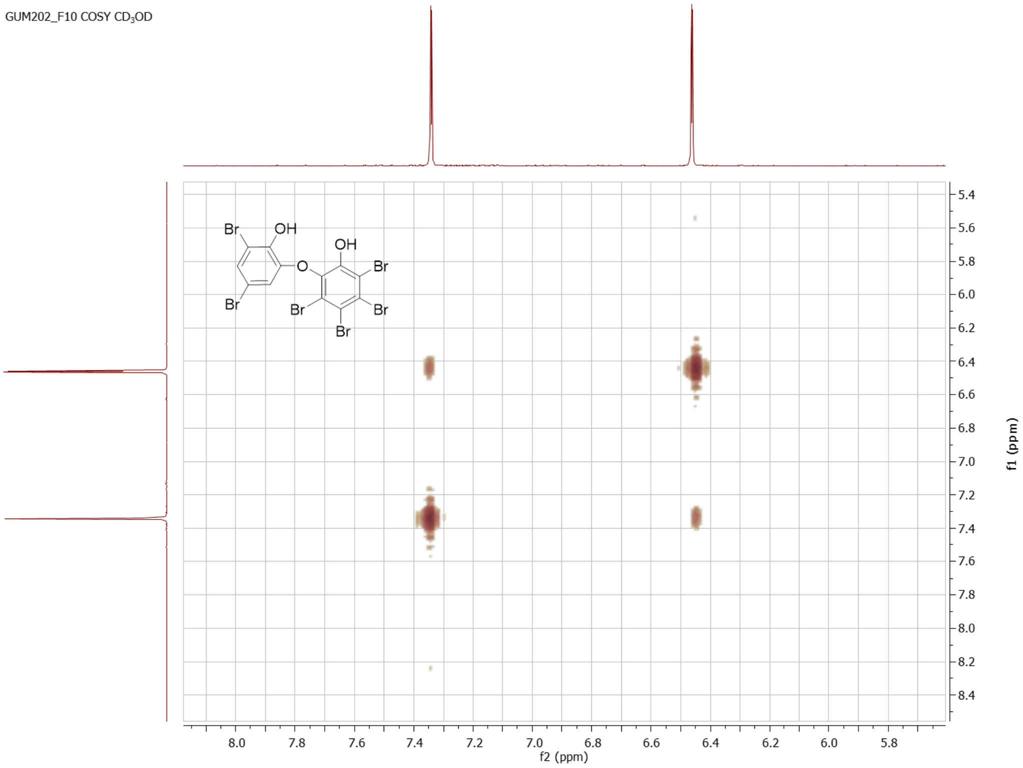


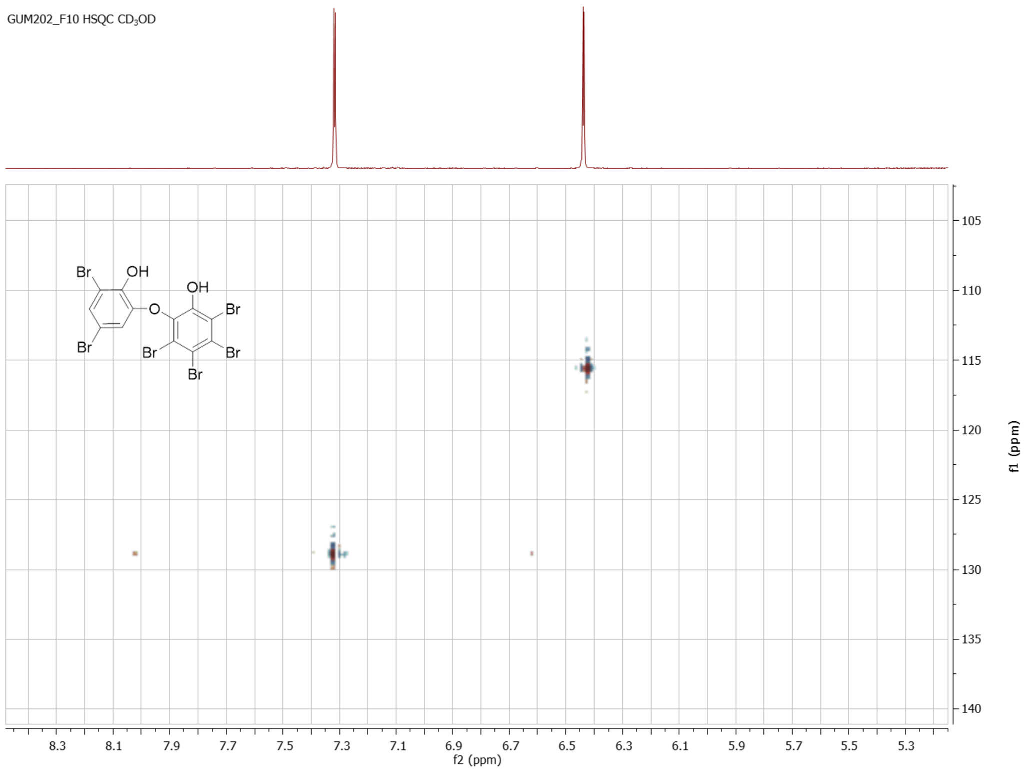

GUM202_F8 HSQC – CD$_3$OD
Br
OH
O
OH
Br
Br
Br
Br
110
115
120
125
130
135
f1 (ppm)
8.2 8.1 8.0 7.9 7.8 7.7 7.6 7.5 7.4 7.3 7.2 7.1 7.0 6.9 6.8 6.7 6.6 6.5 6.4 6.3 6.2 6.1 6.0 5.9
f2 (ppm)

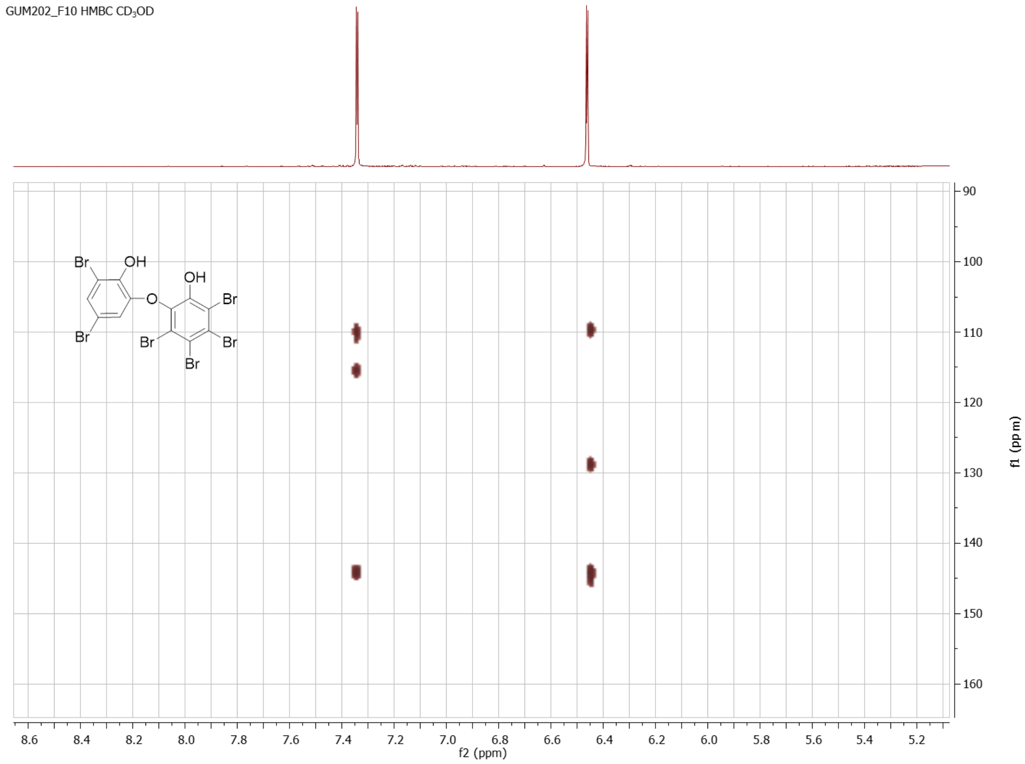

GUM202_F8 HMBC – CD$_3$OD
Br
OH
O
OH
Br
Br
Br
Br
100
105
110
115
120
125
130
135
140
145
150
155
160
f1 (ppm)
8.3 8.2 8.1 8.0 7.9 7.8 7.7 7.6 7.5 7.4 7.3 7.2 7.1 7.0 6.9 6.8 6.7 6.6 6.5 6.4 6.3 6.2 6.1 6.0 5.9 5.8 5.7 5.6
f2 (ppm)

NMR spectra of compound 19; 1H-NMR, COSY, HSQC, HMBC

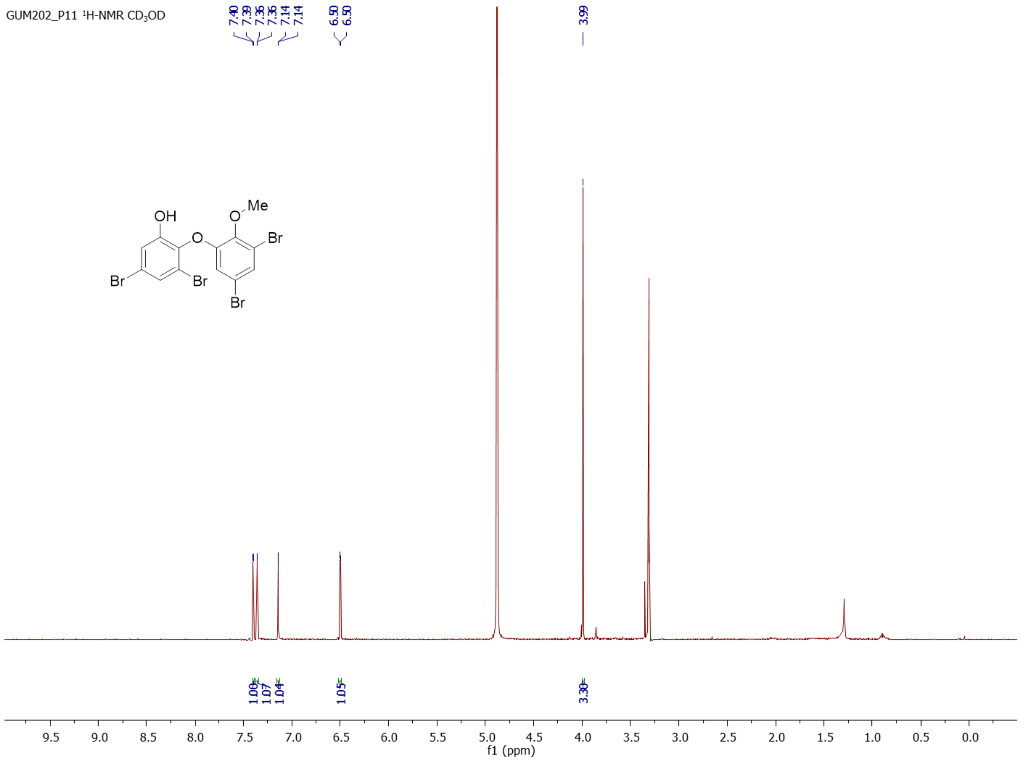


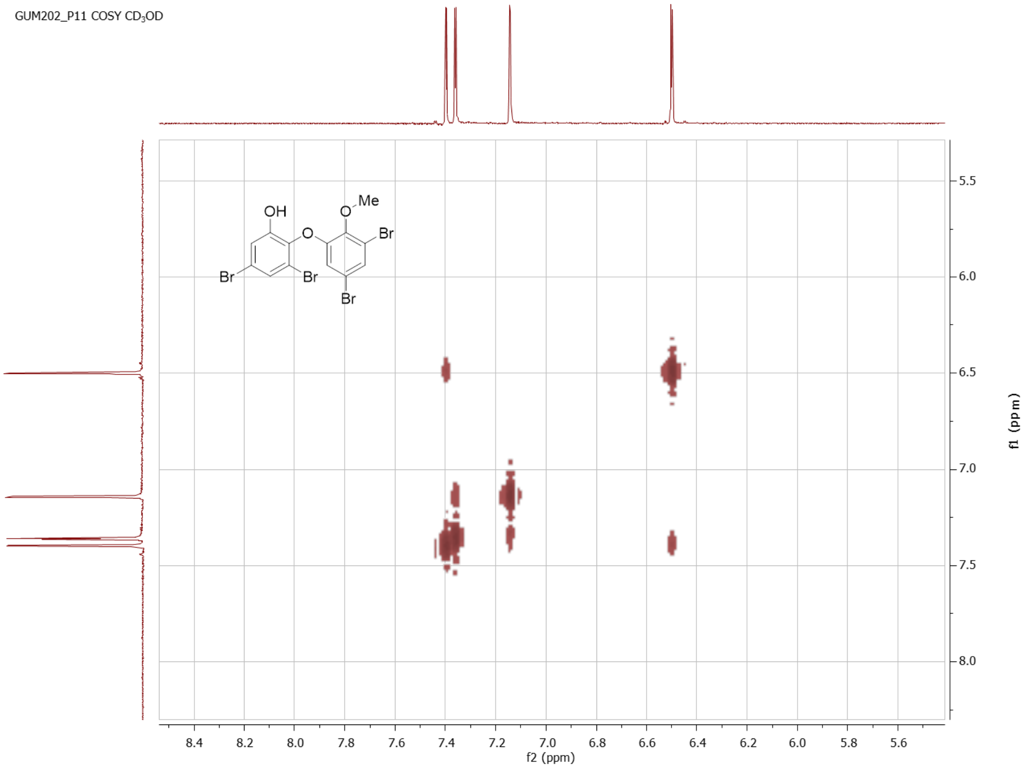


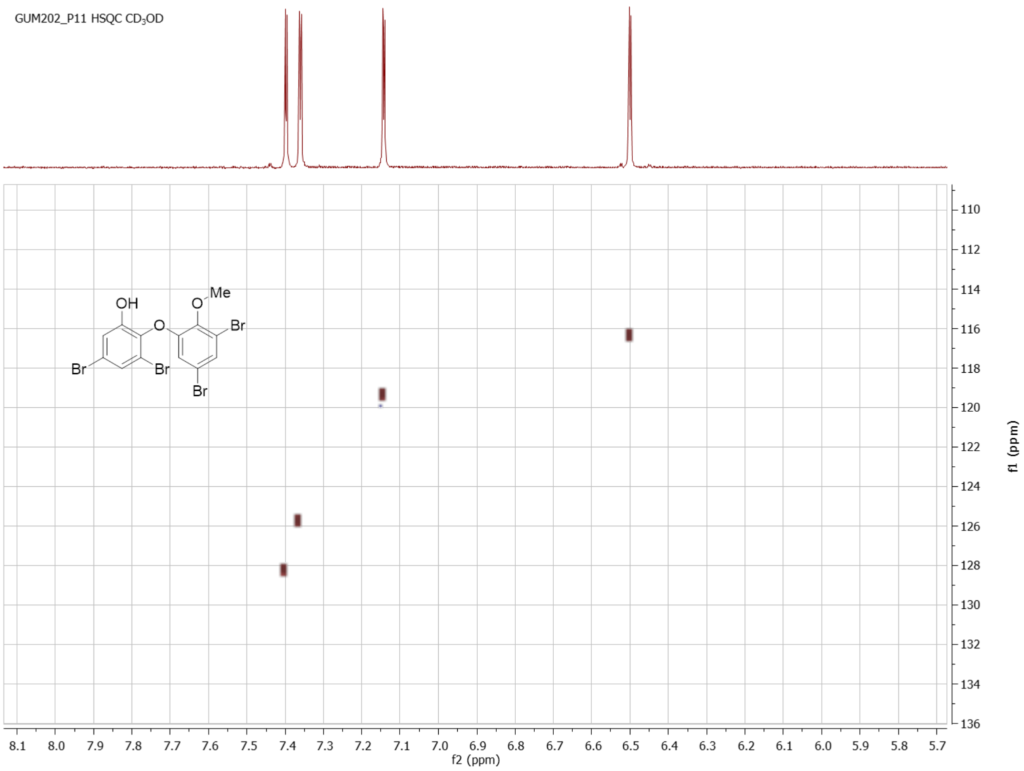

GUM202_F10 HSQC CD$_3$OD
Br
OH
OH
O
Br
Br
Br
Br
Br
105
110
115
120
125
130
135
140
f1 (ppm)
8.3
8.1
7.9
7.7
7.5
7.3
7.1
6.9
6.7
6.5
6.3
6.1
5.9
5.7
5.5
5.3
f2 (ppm)

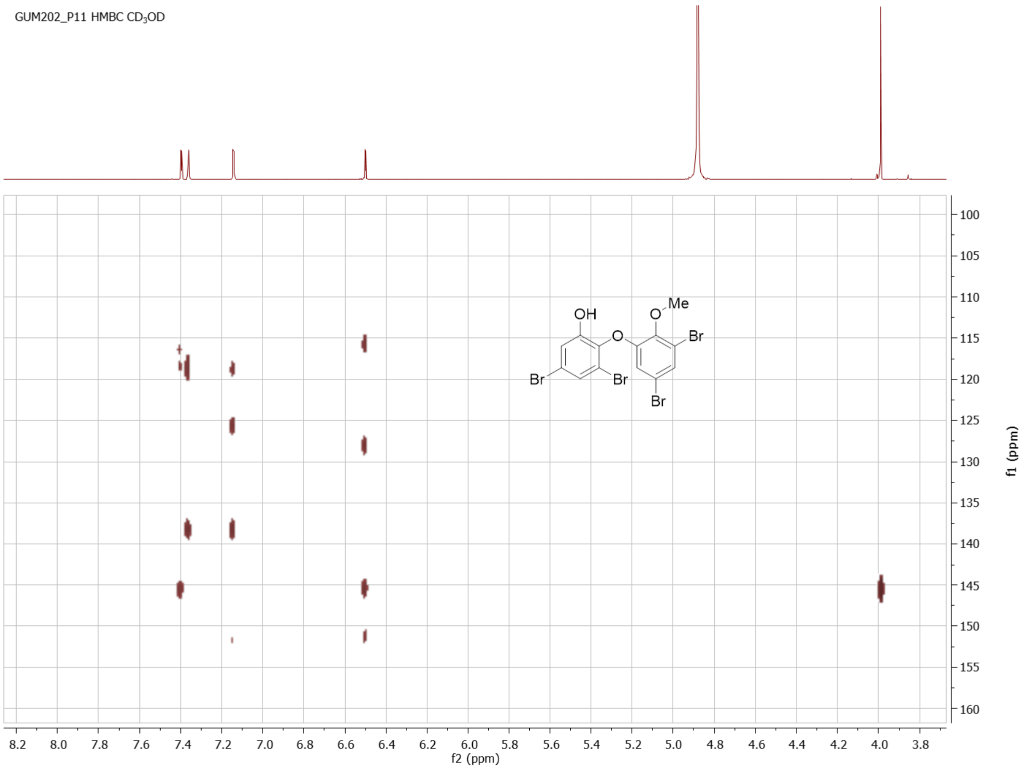

GUM202_F10 HMBC CD$_3$OD
Br
OH
OH
O
Br
Br
Br
Br
Br
90
100
110
120
130
140
150
160
f1 (ppm)
8.6
8.4
8.2
8.0
7.8
7.6
7.4
7.2
7.0
6.8
6.6
6.4
6.2
6.0
5.8
5.6
5.4
5.2
f2 (ppm)

NMR spectra of compound 20; 1H-NMR, COSY, HSQC, HMBC

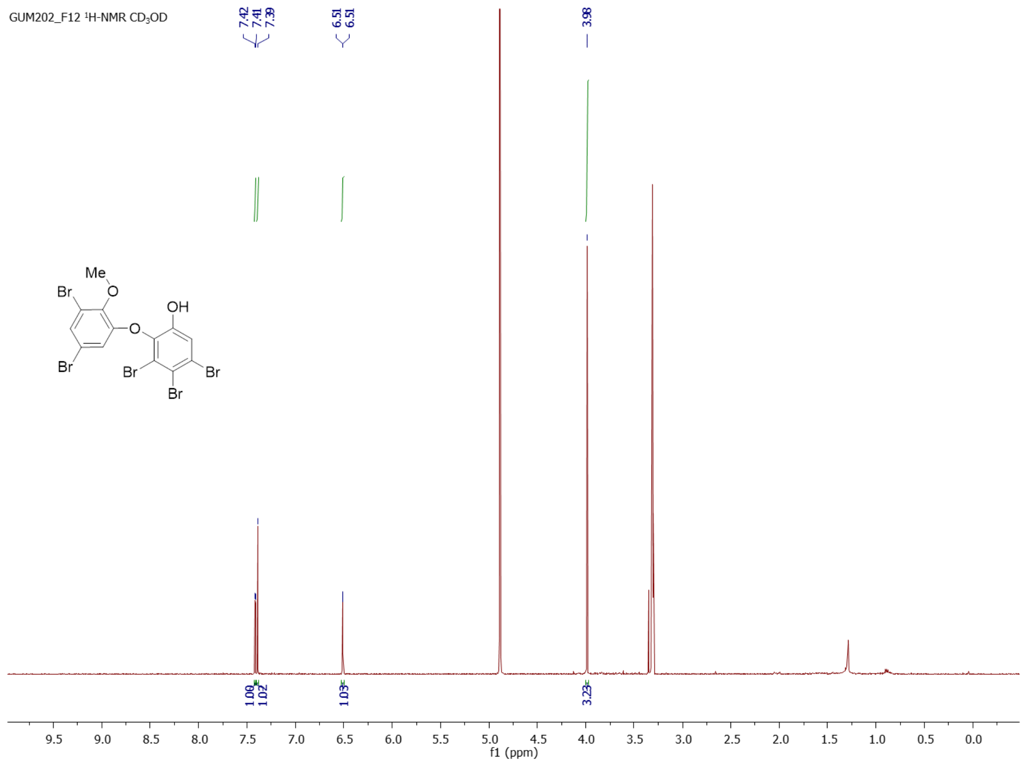


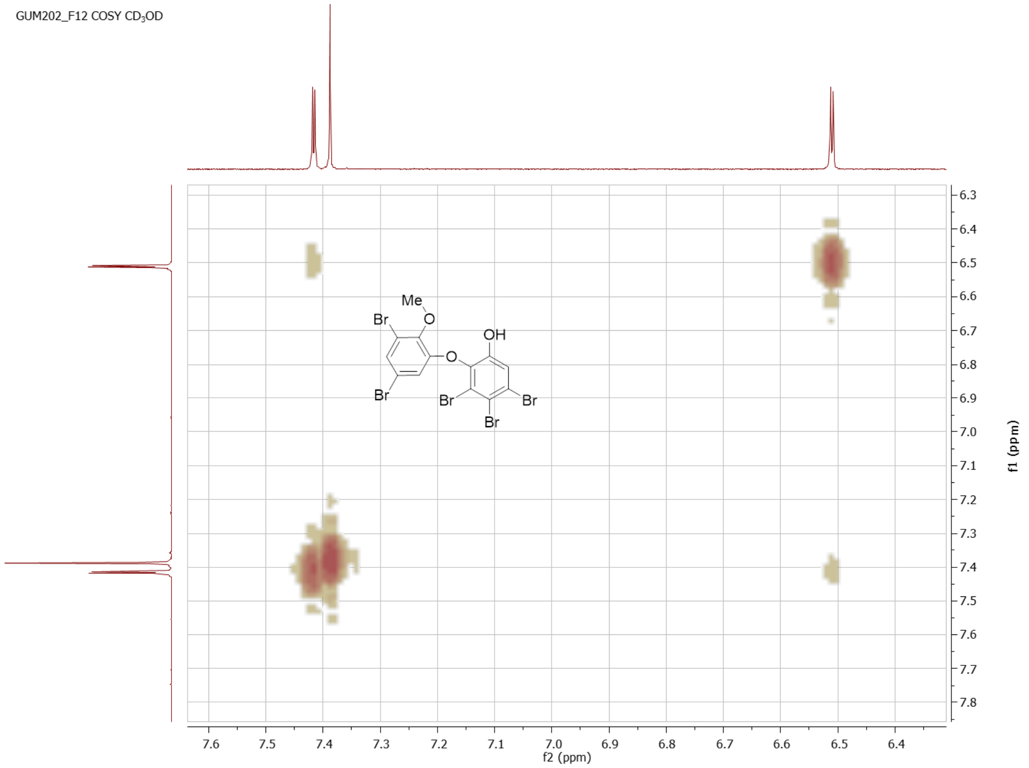


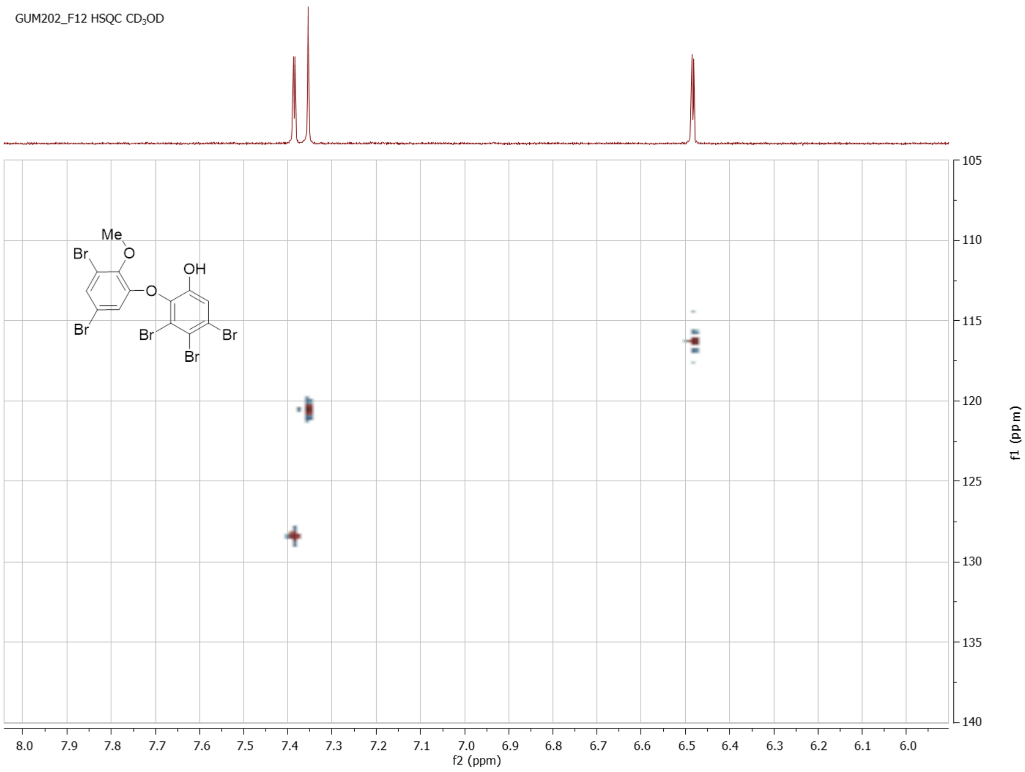

GUM202_P11 HSQC CD$_3$OD
OH
O
O
Me
Br
Br
Br
Br
110
112
114
116
118
120
122
124
126
128
130
132
134
136
f1 (ppm)
8.1 8.0 7.9 7.8 7.7 7.6 7.5 7.4 7.3 7.2 7.1 7.0 6.9 6.8 6.7 6.6 6.5 6.4 6.3 6.2 6.1 6.0 5.9 5.8 5.7
f2 (ppm)


GUM202_P11 HMBC CD$_3$OD
OH
O
O
Me
Br
Br
Br
Br
100
105
110
115
120
125
130
135
140
145
150
155
160
f1 (ppm)
8.2 8.0 7.8 7.6 7.4 7.2 7.0 6.8 6.6 6.4 6.2 6.0 5.8 5.6 5.4 5.2 5.0 4.8 4.6 4.4 4.2 4.0 3.8
f2 (ppm)

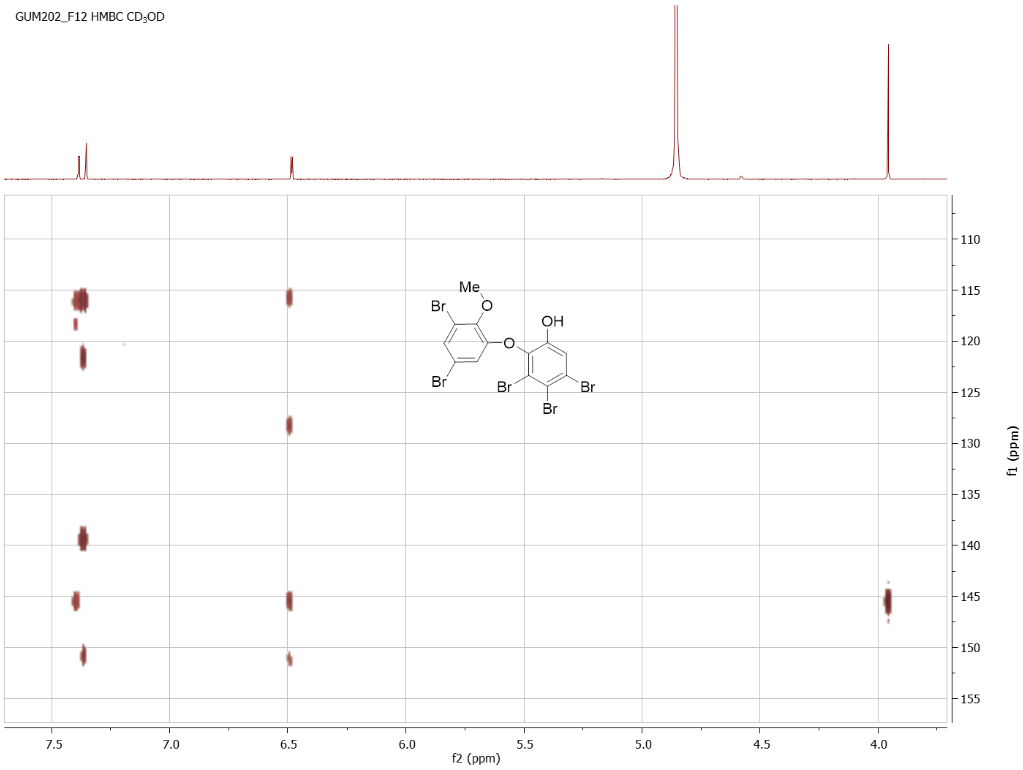
NMR spectra of compound 21; 1H-NMR, COSY, HSQC, HMBC





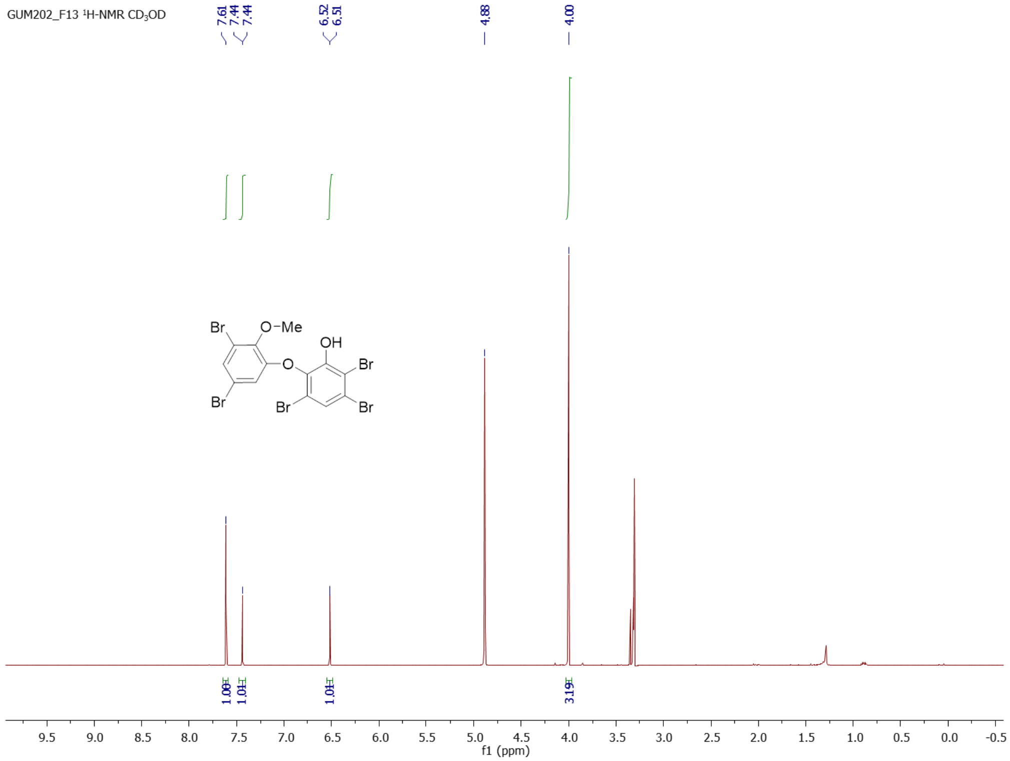


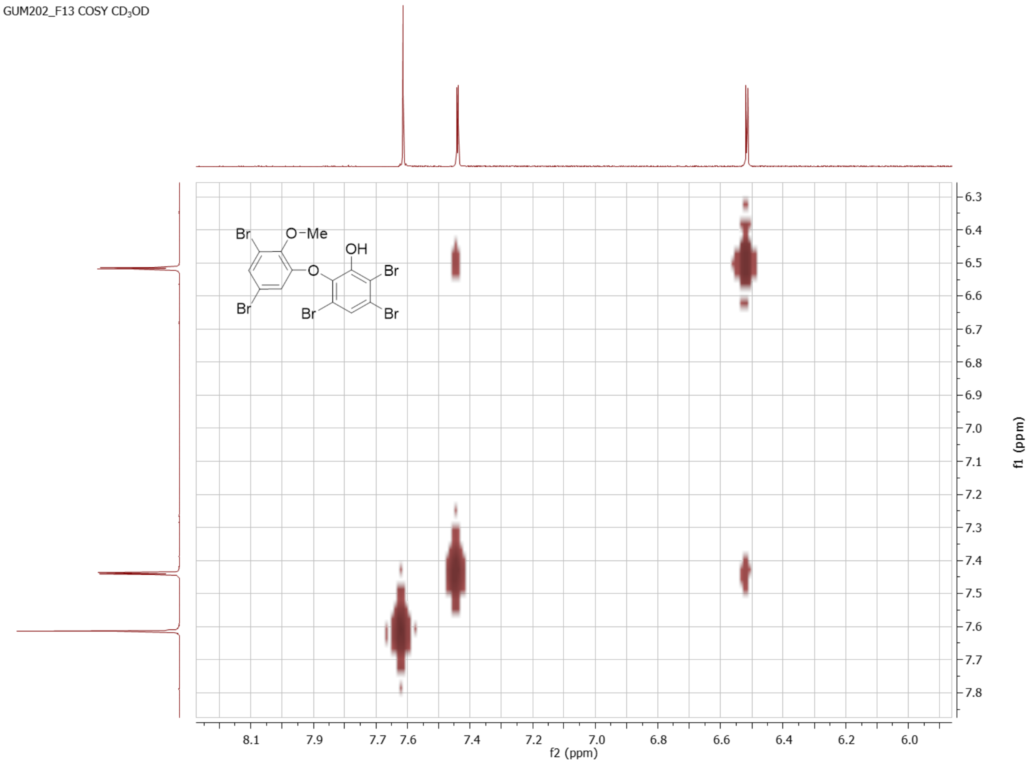
GUM202_F12 HSQC CD_3OD
Me
O
Br
OH
O
Br
Br
Br
Br
105
110
115
120
125
130
135
140
f1 (ppm)
8.0
7.9
7.8
7.7
7.6
7.5
7.4
7.3
7.2
7.1
7.0
6.9
6.8
6.7
6.6
6.5
6.4
6.3
6.2
6.1
6.0
f2 (ppm)

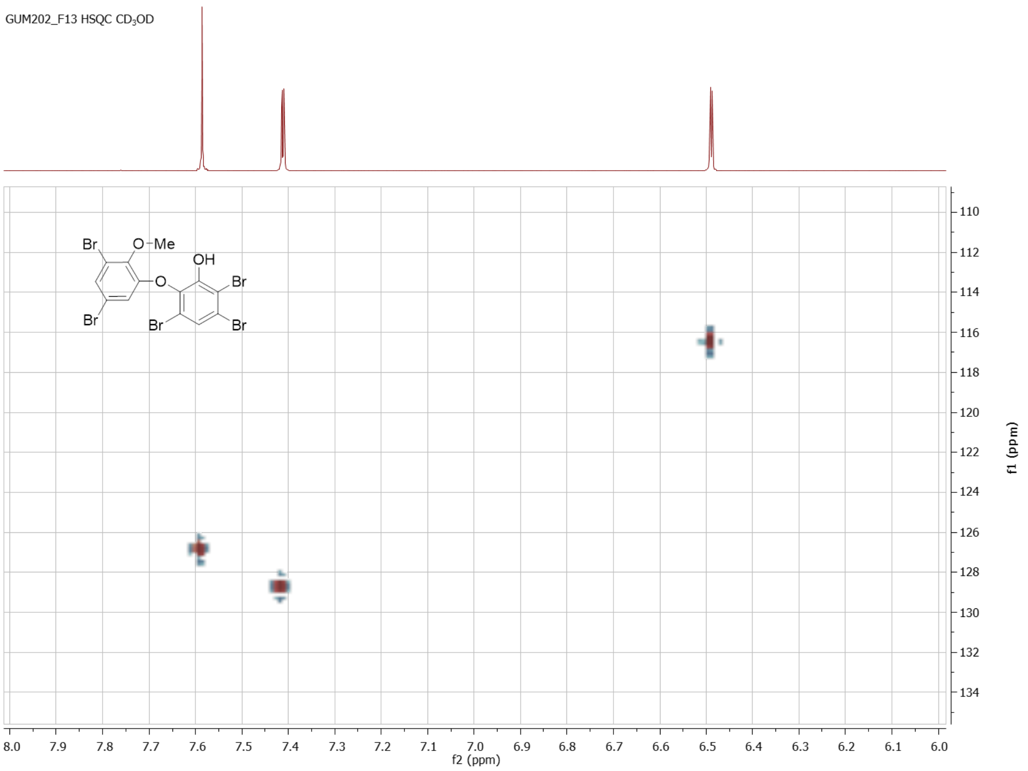


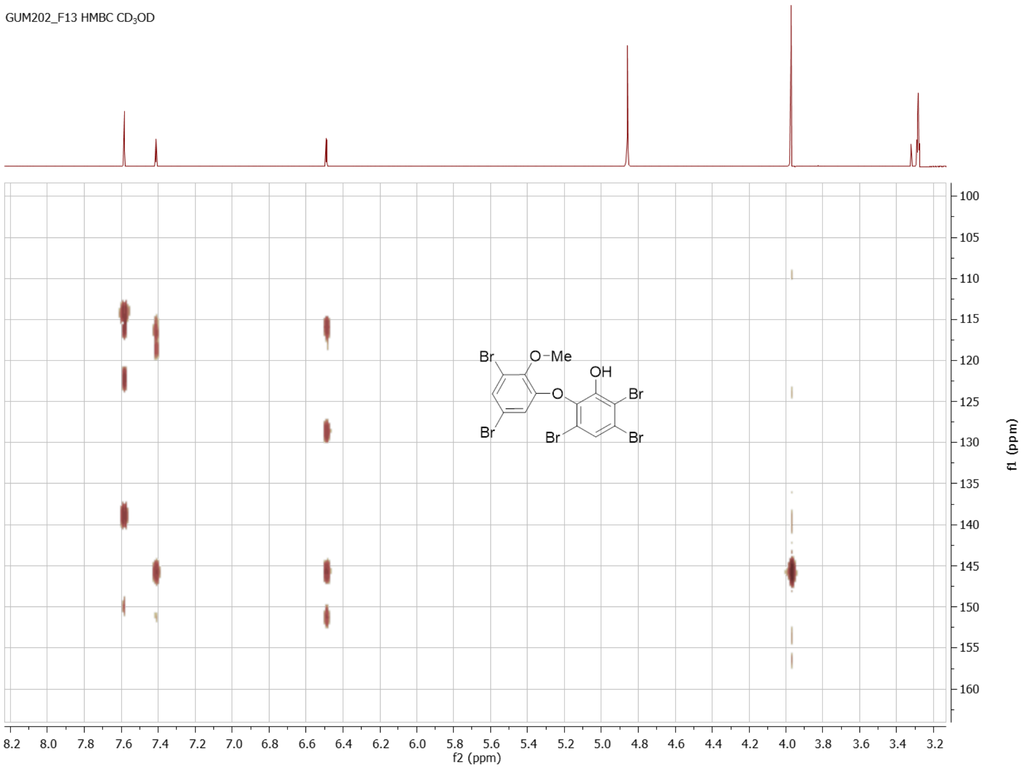
GUM202_F12 HMBC CD_3OD
Me
O
Br
OH
O
Br
Br
Br
Br
110
115
120
125
130
135
140
145
150
155
f1 (ppm)
7.5
7.0
6.5
6.0
5.5
5.0
4.5
4.0
f2 (ppm)

NMR spectra of compound 22; 1H-NMR, COSY, HSQC, HMBC

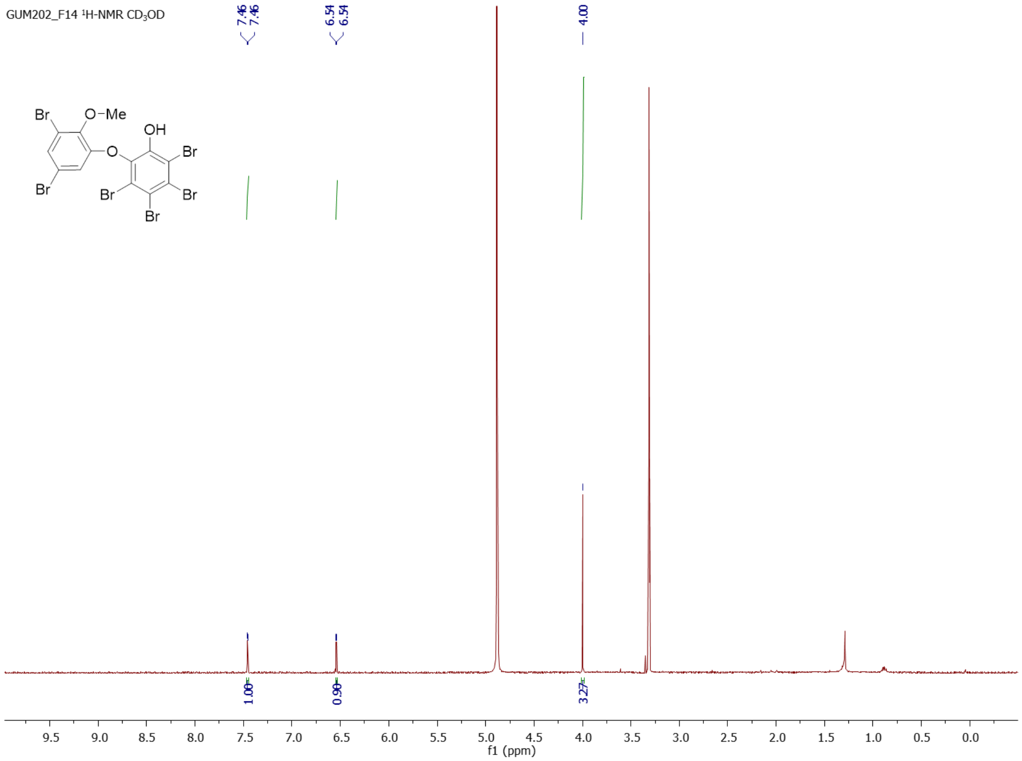


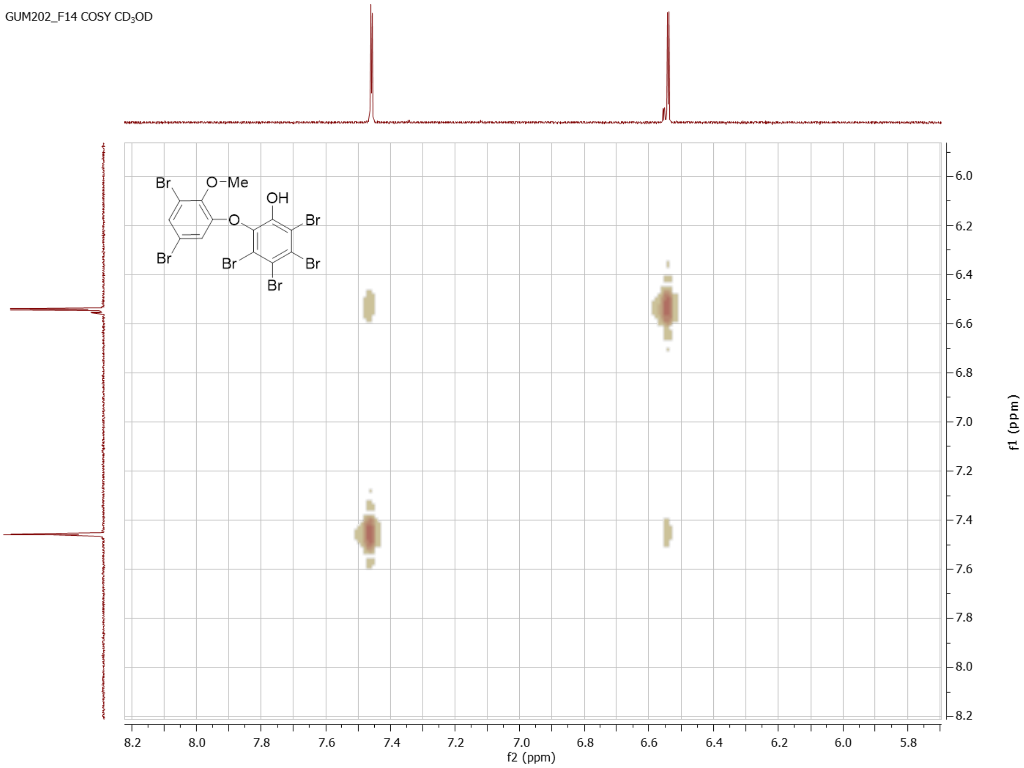


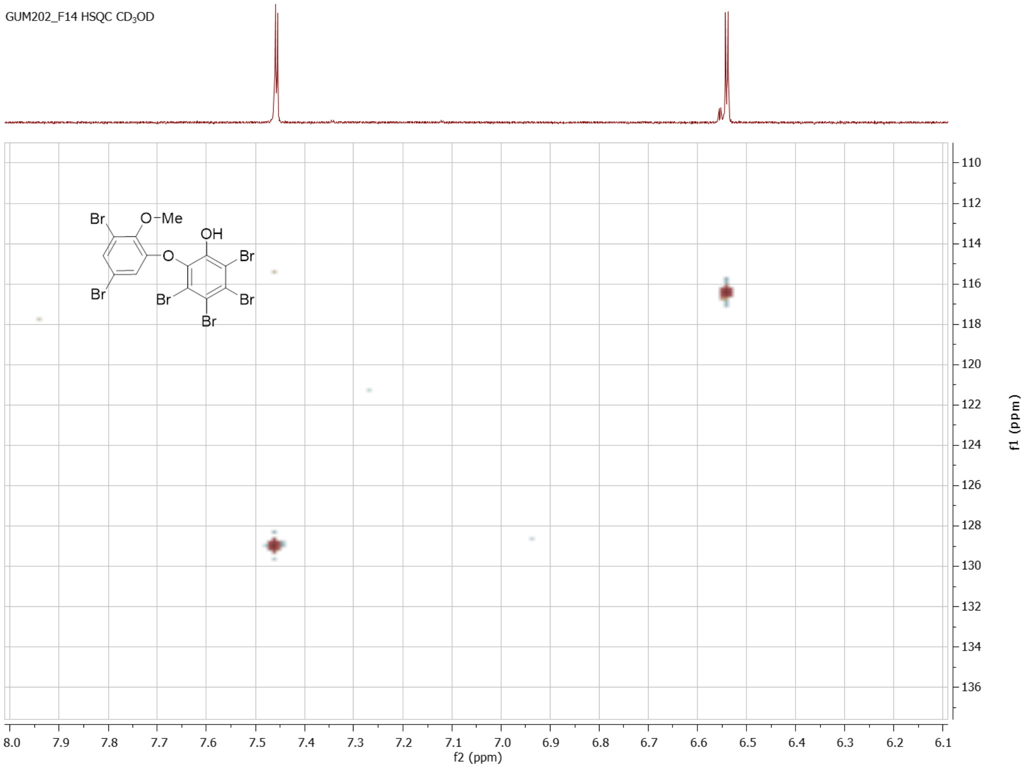

GUM202_F13 HSQC CD_3OD

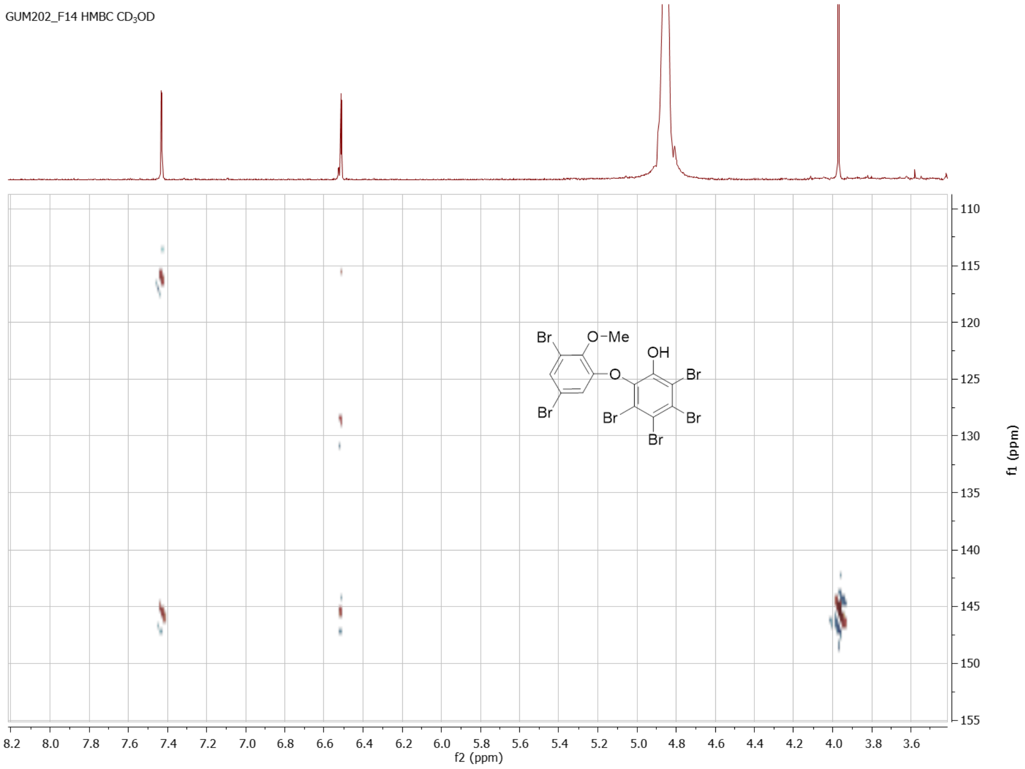

GUM202_F13 HMBC CD_3OD

NMR spectra of compound 23; 1H-NMR, COSY, HSQC, HMBC






GUM202_F14 HSQC CD_3OD


GUM202_F14 HMBC CD_3OD

4.7 References

1 Mehbub, M. F., Perkins, M. V., Zhang, W. & Franco, C. M. M. New marine natural products from sponges (Porifera) of the order Dictyoceratida (2001 to 2012); a promising source for drug discovery, exploration and future prospects. *Biotechnol Adv* **34**, 473-491, 2016. PMID: 26802363.

2 Sharma, G. M. & Vig, B. Studies on the antimicrobial substances of sponges. VI. Structures of two antibacterial substances isolated from the marine sponge Dysidea herbacea. *Tetrahedron Letters* **13**, 1715-1718, 1972.

3 Carté, B. & Faulkner, D. J. Polybrominated diphenyl ethers from Dysidea herbacea, Dysidea chlorea and Phyllospongia foliascens. *Tetrahedron* **37**, 2335-2339, 1981.

4 Fu, X., Schmitz, F. J., Govindan, M., Abbas, S. A., Hanson, K. M., Horton, P. A., Crews, P., Laney, M. & Schatzman, R. C. Enzyme inhibitors: new and known polybrominated phenols and diphenyl ethers from four Indo-Pacific Dysidea sponges. *J Nat Prod* **58**, 1384-1391, 1995. PMID: 7494145.

5 Handayani, D., Edrada, R. A., Proksch, P., Wray, V., Witte, L., Van Soest, R. W., Kunzmann, A. & Soedarsono. Four new bioactive polybrominated diphenyl ethers of the sponge Dysidea herbacea from West Sumatra, Indonesia. *J Nat Prod* **60**, 1313-1316, 1997. PMID: 9463111.

6 Zhang, H., Skildum, A., Stromquist, E., Rose-Hellekant, T. & Chang, L. C. Bioactive polybrominated diphenyl ethers from the marine sponge Dysidea sp. *J Nat Prod* **71**, 262-264, 2008. PMID: 18198840.

7 Novriyandi Hanif, Junichi Tanaka, †, Andi Setiawan, Agus Trianto, Nicole J. de Voogd, Anggia Murni, Chiaki Tanaka, a. & Higa†, T. Polybrominated Diphenyl Ethers from the Indonesian Sponge Lamellodysidea herbacea⊥. *J Nat Prod* **70**, 432-435, 2007.

8 Agrawal, M. S. & Bowden, B. F. Marine Sponge Dysidea herbacea revisited: Another Brominated Diphenyl Ether. *Mar Drugs* **3**, 9-14, 2005.

9 Unson, M. D., Rose, C. B., Faulkner, D. J., Brinen, L. S., Steiner, J. R. & Clardy, J. New polychlorinated amino acid derivatives from the marine sponge Dysidea herbacea. *J Org Chem* **58**, 6336-6343, 1993.

10 Hofheinz, W., Pharmazeutische Forschungsabteilung und Zentrale Forschungseinheiten der , C. B., Oberhänsli, W. E. & Pharmazeutische Forschungsabteilung und Zentrale Forschungseinheiten der , C. B. Dysidin, ein neuartiger, chlorhaltiger Naturstoff aus dem Schwamm Dysidea herbacea. *Helvetica Chimica Acta* **60**, 660-669, 1977.

11 Kazlauskas, R., Lidgard, R. O., Wells, R. J. & Vetter, W. A novel hexachloro-metabolite from the sponge dysidea herbacea. *Tetrahedron Letters* **18**, 3183-3186, 1977.

12 Lee, G. M. & Molinski, T. F. Herbaceamide, a chlorinated N-acyl amino ester from the marine sponge, Dysidea herbacea. *Tetrahedron Letters* **33**, 7671-7674, 1992.

13 George G. Harrigan, †, Gilles H. Goetz, Hendrik Luesch, Shengtian Yang, a. & Likos†, J. Dysideaprolines A−F and Barbaleucamides A−B, Novel Polychlorinated Compounds from a Dysidea Species. *J Nat Prod* **64**, 1133-1138, 2001.

14 Clark, W. D. & Crews, P. A novel chlorinated ketide amino acid, herbamide A, from the marine sponge Dysidea herbacea. *Tetrahedron Letters* **36**, 1185-1188, 1995.

15 Kazlauskas, R., Murphy, P. T. & Wells, R. J. A diketopiperazine derived from trichloroleucine from the sponge Dysidea herbacea. *Tetrahedron Letters* **19**, 4945-4948, 1978.

16 Flowers, A. E., Garson, M. J., Webb, R. I., Dumdei, E. J. & Charan, R. D. Cellular origin of chlorinated diketopiperazines in the dictyoceratid sponge Dysidea herbacea (Keller). *Cell Tissue Res* **292**, 597-607, 1998. PMID: 9582417.

17 Dumdei, E. J., Simpson, J. S., Garson, M. J., Byriel, K. A. & Kennard, C. H. L. New Chlorinated Metabolites from the Tropical Marine Sponge <emph type="2">Dysidea herbacea</emph>. *Australian Journal of Chemistry* **50**, 139-144, 1997.

18 Kazlauskas, R., Murphy, P. T. & Wells, R. J. A new sesquiterpene from the sponge Dysidea herbacea. *Tetrahedron Letters* **19**, 4949-4950, 1978.

19 Charles, C., Braekman, J. C., Daloze, D., Tursch, B., Declercq, J. P., Germain, G. & van Meerssche, M. Chemical studies of marine invertebrates. XXXIV(1). Herbadysidolide and herbasolide, two unusual sesquiterpenoids from the sponge dysidea herbacea(2). *Bulletin des Sociétés Chimiques Belges* **87**, 481-486, 1978.

20 Torii, M., Kato, H., Hitora, Y., Angkouw, E. D., Mangindaan, R. E. P., de Voogd, N. J. & Tsukamoto, S. Lamellodysidines A and B, Sesquiterpenes Isolated from the Marine Sponge Lamellodysidea herbacea. *J Nat Prod* **80**, 2536-2541, 2017. PMID: 28841316.

21 Venkateswarlu, Y., Biabani, M. A. F., Reddy, M. V. R., Chavakula, R. & Rao, J. V. A New Sesquiterpene from the Andaman Sponge Dysidea herbacea. *J Nat Prod* **57**, 827-828, 1994.

22 Dunlop, R., Kazlauskas, R., March, G., Murphy, P. & Wells, R. New furano-sesquiterpenes from the sponge Dysidea herbacea. *Australian Journal of Chemistry* **35**, 95-103, 1982.

23 Sera, Y., Kyoko Adachi, Nishida, F. & Shizuri, Y. A New Sesquiterpene as an Antifouling Substance from a Palauan Marine Sponge, Dysidea herbacea. *J Nat Prod* **62**, 395-396, 1999.

24 Sakai, R., Suzuki, K., Shimamoto, K. & Kamiya, H. Novel betaines from a micronesian sponge Dysidea herbacea. *J Org Chem* **69**, 1180-1185, 2004. PMID: 14961668.

25 Isaacs, S., Berman, R., Kashman, Y., Gebreyesus, T. & Yosief, T. New Polyhydroxy Sterols, Dysidamides, and a Dideoxyhexose from the Sponge Dysidea herbacea. *J Nat Prod* **54**, 83-91, 1991.

26 Capon, R. J. & Faulkner, D. J. Herbasterol, an ichthyotoxic 9,11-secosterol from the sponge Dysidea herbacea. *J Org Chem* **50**, 4771-4773, 1985.

27 Bandaranayake, W. M., Bemis, J. E. & Bourne, D. J. Ultraviolet absorbing pigments from the marine sponge Dysidea herbacea: Isolation and structure of a new mycosporine. *Comparative Biochemistry and Physiology Part C: Pharmacology, Toxicology and Endocrinology* **115**, 281-286, 1996.

28 Carroll, A. R., Pierens, G. K., Fechner, G., De Almeida Leone, P., Ngo, A., Simpson, M., Hyde, E., Hooper, J. N., Bostrom, S. L., Musil, D. & Quinn, R. J. Dysinosin A: a novel inhibitor of Factor VIIa and thrombin from a new genus and species of Australian sponge of the family Dysideidae. *J Am Chem Soc* **124**, 13340-13341, 2002. PMID: 12418859.

29 Carroll, A. R., Buchanan, M. S., Edser, A., Hyde, E., Simpson, M. & Quinn, R. J. Dysinosins B-D, inhibitors of factor VIIa and thrombin from the Australian sponge Lamellodysidea chlorea. *J Nat Prod* **67**, 1291-1294, 2004. PMID: 15332844.

30 Ishida, K., Christiansen, G., Yoshida, W. Y., Kurmayer, R., Welker, M., Valls, N., Bonjoch, J., Hertweck, C., Börner, T., Hemscheidt, T. & Dittmann, E. Biosynthetic pathway and structure of aeruginosides 126A and 126B, cyanobacterial peptides bearing a 2-carboxy-6-hydroxyoctahydroindole moiety. *Chem Biol* **14**, 565-576, 2007. PMID: 17524987.

31 Crawford, J. M. & Clardy, J. Bacterial symbionts and natural products. *Chem Commun (Camb)* **47**, 7559-7566, 2011. PMID: 21594283.

32 Schmidt, E. W., Nelson, J. T., Rasko, D. A., Sudek, S., Eisen, J. A., Haygood, M. G. & Ravel, J. Patellamide A and C biosynthesis by a microcin-like pathway in Prochloron didemni, the cyanobacterial symbiont of Lissoclinum patella. *Proc Natl Acad Sci U S A* **102**, 7315-7320, 2005. PMID: 15883371.

33 Wilson, M. C., Mori, T., Ruckert, C., Uria, A. R., Helf, M. J., Takada, K., Gernert, C., Steffens, U. A., Heycke, N., Schmitt, S., Rinke, C., Helfrich, E. J., Brachmann, A. O., Gurgui, C., Wakimoto, T., Kracht, M., Crusemann, M., Hentschel, U., Abe, I., Matsunaga, S., Kalinowski, J., Takeyama, H. & Piel, J. An environmental bacterial taxon with a large and distinct metabolic repertoire. *Nature* **506**, 58-62, 2014. PMID: 24476823.

34 Agarwal, V., Blanton, J. M., Podell, S., Taton, A., Schorn, M. A., Busch, J., Lin, Z., Schmidt, E. W., Jensen, P. R., Paul, V. J., Biggs, J. S., Golden, J. W., Allen, E. E. & Moore, B. S. Metagenomic discovery of polybrominated diphenyl ether biosynthesis by marine sponges. *Nat Chem Biol* **13**, 537-543, 2017. PMID: 28319100.

35 Berthold, R. J., Borowitzka, M. A. & Mackay, M. A. The ultrastructure of Oscillatoria spongeliae, the blue-green algal endosymbiont of the sponge Dysidea herbacea. *Phycologia* **21**, 327-335, 1982.

36 Thacker, R. W. & Starnes, S. Host specificity of the symbiotic cyanobacterium Oscillatoria spongeliae in marine sponges, Dysidea spp. *Marine biology* **142**, 2003.

37 Ridley, C. P., Bergquist, P. R., Harper, M. K., Faulkner, D. J., Hooper, J. N. A. & Haygood, M. G. Speciation and Biosynthetic Variation in Four Dictyoceratid Sponges and Their Cyanobacterial Symbiont, Oscillatoria spongeliae. *Chemistry & Biology* **12**, 397-406, 2005.

38 Hinde, R., Pironet, F. & Borowitzka, M. A. Isolation of *Oscillatoria spongeliae*, the filamentous cyanobacterial symbiont of the marine sponge *Dysidea herbacea*. *Marine Biology* **119**, 99-104, 1994.

39 Unson, M. D. & Faulkner, D. J. Cyanobacterial symbiont biosynthesis of chlorinated metabolites from Dysidea herbacea (Porifera) | SpringerLink. *Experientia* **49**, 349-353, 1993.

40 Unson, M. D., Holland, N. D. & Faulkner, D. J. A brominated secondary metabolite synthesized by the cyanobacterial symbiont of a marine sponge and accumulation of the crystalline metabolite in the sponge tissue. *Marine Biology* **119**, 1-11, 1994.

41 Flatt, P., T. Gautschi, J., Thacker, R., Musafija-Girt, M., Crews, P. & Gerwick, W. *Identification of the cellular site of polychlorinated peptide biosynthesis in the marine sponge Dysidea (Lamellodysidea) herbacea and symbiotic cyanobacterium Oscillatoria spongeliae by CARD-FISH analysis*. (2005).

42 Schmidt, E. W. & Donia, M. S. Chapter 23. Cyanobactin ribosomally synthesized peptides--a case of deep metagenome mining. *Methods Enzymol* **458**, 575-596, 2009. PMID: 19374999.

43 Podell, S. & Gaasterland, T. DarkHorse: a method for genome-wide prediction of horizontal gene transfer. *Genome Biol* **8**, R16, 2007. PMID: 17274820.

44 Boetzer, M. & Pirovano, W. SSPACE-LongRead: scaffolding bacterial draft genomes using long read sequence information. *BMC Bioinformatics* **15**, 211, 2014. PMID: 24950923.

45 Antipov, D., Korobeynikov, A., McLean, J. S. & Pevzner, P. A. hybridSPAdes: an algorithm for hybrid assembly of short and long reads. *Bioinformatics* **32**, 1009-1015, 2016. PMID: 26589280.

46 Nadalin, F., Vezzi, F. & Policriti, A. GapFiller: a de novo assembly approach to fill the gap within paired reads. *BMC Bioinformatics* **13 Suppl 14**, S8, 2012. PMID: 23095524.

47 Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res* **25**, 1043-1055, 2015.

48 Darling, A. E., Mau, B. & Perna, N. T. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *PLoS One* **5**, e11147, 2010. PMID: 20593022.

49 Wu, M. & Eisen, J. A. A simple, fast, and accurate method of phylogenomic inference. *Genome Biol* **9**, R151, 2008. PMID: 18851752.

50 Mistry, J., Finn, R. D., Eddy, S. R., Bateman, A. & Punta, M. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Res* **41**, e121, 2013. PMID: 23598997.

51 Katoh, K. & Standley, D. M. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol Biol Evol* **30**, 772-780, 2013. PMID: 23329690.

52 Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972-1973, 2009.

53 Nguyen, L. T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol* **32**, 268-274, 2015. PMID: 25371430.

54 Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q. & Vinh, L. S. UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Mol Biol Evol* **35**, 518-522, 2018. PMID: 29077904.

55 Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A. & Jermiin, L. S. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat Methods* **14**, 587-589, 2017. PMID: 28481363.

56 Letunic, I. & Bork, P. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Research*, 2016.

57 Broddrick, J. T., Rubin, B. E., Welkie, D. G., Du, N., Mih, N., Diamond, S., Lee, J. J., Golden, S. S. & Palsson, B. O. Unique attributes of cyanobacterial metabolism revealed by improved genome-scale metabolic modeling and essential gene analysis. *Proc Natl Acad Sci U S A* **113**, E8344-e8353, 2016. PMID: 27911809.

58 Rubin, B. E., Wetmore, K. M., Price, M. N., Diamond, S., Shultzaberger, R. K., Lowe, L. C., Curtin, G., Arkin, A. P., Deutschbauer, A. & Golden, S. S. The essential gene set of a photosynthetic organism. *Proc Natl Acad Sci U S A* **112**, E6634-6643, 2015. PMID: 26508635.

59 Porollo, A. EC2KEGG: a command line tool for comparison of metabolic pathways. *Source Code for Biology and Medicine* **9**, 19, 2014.

60 Abby, S. S., Cury, J., Guglielmini, J., Néron, B., Touchon, M. & Rocha, E. P. C. Identification of protein secretion systems in bacterial genomes. *Sci Rep* **6**, 2016. PMID: 26979785.

61 Afgan, E., Baker, D., van den Beek, M., Blankenberg, D., Bouvier, D., Cech, M., Chilton, J., Clements, D., Coraor, N., Eberhard, C., Gruning, B., Guerler, A., Hillman-Jackson, J., Von Kuster, G., Rasche, E., Soranzo, N., Turaga, N., Taylor, J., Nekrutenko, A. & Goecks, J. The Galaxy platform for accessible, reproducible and collaborative

biomedical analyses: 2016 update. *Nucleic Acids Res* **44**, W3-w10, 2016. PMID: 27137889.

62 Weber, T., Blin, K., Duddela, S., Krug, D., Kim, H. U., Bruccoleri, R., Lee, S. Y., Fischbach, M. A., Muller, R., Wohlleben, W., Breitling, R., Takano, E. & Medema, M. H. antiSMASH 3.0-a comprehensive resource for the genome mining of biosynthetic gene clusters. *Nucleic Acids Res* **43**, W237-243, 2015. PMID: 25948579.

63 Medema, M. H., Kottmann, R., Yilmaz, P., Cummings, M., Biggins, J. B., Blin, K., de Bruijn, I., Chooi, Y. H., Claesen, J., Coates, R. C., Cruz-Morales, P., Duddela, S., Dusterhus, S., Edwards, D. J., Fewer, D. P., Garg, N., Geiger, C., Gomez-Escribano, J. P., Greule, A., Hadjithomas, M., Haines, A. S., Helfrich, E. J., Hillwig, M. L., Ishida, K., Jones, A. C., Jones, C. S., Jungmann, K., Kegler, C., Kim, H. U., Kotter, P., Krug, D., Masschelein, J., Melnik, A. V., Mantovani, S. M., Monroe, E. A., Moore, M., Moss, N., Nutzmann, H. W., Pan, G., Pati, A., Petras, D., Reen, F. J., Rosconi, F., Rui, Z., Tian, Z., Tobias, N. J., Tsunematsu, Y., Wiemann, P., Wyckoff, E., Yan, X., Yim, G., Yu, F., Xie, Y., Aigle, B., Apel, A. K., Balibar, C. J., Balskus, E. P., Barona-Gomez, F., Bechthold, A., Bode, H. B., Borriss, R., Brady, S. F., Brakhage, A. A., Caffrey, P., Cheng, Y. Q., Clardy, J., Cox, R. J., De Mot, R., Donadio, S., Donia, M. S., van der Donk, W. A., Dorrestein, P. C., Doyle, S., Driessen, A. J., Ehling-Schulz, M., Entian, K. D., Fischbach, M. A., Gerwick, L., Gerwick, W. H., Gross, H., Gust, B., Hertweck, C., Hofte, M., Jensen, S. E., Ju, J., Katz, L., Kaysser, L., Klassen, J. L., Keller, N. P., Kormanec, J., Kuipers, O. P., Kuzuyama, T., Kyrpides, N. C., Kwon, H. J., Lautru, S., Lavigne, R., Lee, C. Y., Linquan, B., Liu, X., Liu, W., Luzhetskyy, A., Mahmud, T., Mast, Y., Mendez, C., Metsa-Ketela, M., Micklefield, J., Mitchell, D. A., Moore, B. S., Moreira, L. M., Muller, R., Neilan, B. A., Nett, M., Nielsen, J., O'Gara, F., Oikawa, H., Osbourn, A., Osburne, M. S., Ostash, B., Payne, S. M., Pernodet, J. L., Petricek, M., Piel, J., Ploux, O., Raaijmakers, J. M., Salas, J. A., Schmitt, E. K., Scott, B., Seipke, R. F., Shen, B., Sherman, D. H., Sivonen, K., Smanski, M. J., Sosio, M., Stegmann, E., Sussmuth, R. D., Tahlan, K., Thomas, C. M., Tang, Y., Truman, A. W., Viaud, M., Walton, J. D., Walsh, C. T., Weber, T., van Wezel, G. P., Wilkinson, B., Willey, J. M., Wohlleben, W., Wright, G. D., Ziemert, N., Zhang, C., Zotchev, S. B., Breitling, R., Takano, E. & Glockner, F. O. Minimum Information about a Biosynthetic Gene cluster. *Nat Chem Biol* **11**, 625-631, 2015. PMID: 26284661.

64 Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J Mol Biol* **215**, 403-410, 1990. PMID: 2231712.

65 Ziemert, N., Podell, S., Penn, K., Badger, J. H., Allen, E. & Jensen, P. R. The natural product domain seeker NaPDoS: a phylogeny based bioinformatic tool to classify secondary metabolite gene diversity. *PLOS ONE* **7**, 2012.

66 Yeong, M. *BiG-SCAPE: exploring biosynthetic diversity through gene cluster similarity networks* Msc thesis, Wageningen University and Research, (2016).

67 Bastian, M., Heymann, S. & Jacomy, M. in *Third International AAAI Conference on Weblogs and Social Media.*

68 Calcul, L., Chow, R., Oliver, A. G., Tenney, K., White, K. N., Wood, A. W., Fiorilla, C. & Crews, P. NMR strategy for unraveling structures of bioactive sponge-derived oxy-polyhalogenated diphenyl ethers. *J Nat Prod* **72**, 443-449, 2009. PMID: 19323567.

69 Wang, M., Carver, J. J., Phelan, V. V., Sanchez, L. M., Garg, N., Peng, Y., Nguyen, D. D., Watrous, J., Kapono, C. A., Luzzatto-Knaan, T., Porto, C., Bouslimani, A., Melnik, A. V., Meehan, M. J., Liu, W. T., Crusemann, M., Boudreau, P. D., Esquenazi, E., Sandoval-Calderon, M., Kersten, R. D., Pace, L. A., Quinn, R. A., Duncan, K. R., Hsu, C. C., Floros, D. J., Gavilan, R. G., Kleigrewe, K., Northen, T., Dutton, R. J., Parrot, D., Carlson, E. E., Aigle, B., Michelsen, C. F., Jelsbak, L., Sohlenkamp, C., Pevzner, P., Edlund, A., McLean, J., Piel, J., Murphy, B. T., Gerwick, L., Liaw, C. C., Yang, Y. L., Humpf, H. U., Maansson, M., Keyzers, R. A., Sims, A. C., Johnson, A. R., Sidebottom, A. M., Sedio, B. E., Klitgaard, A., Larson, C. B., P, C. A. B., Torres-Mendoza, D., Gonzalez, D. J., Silva, D. B., Marques, L. M., Demarque, D. P., Pociute, E., O'Neill, E. C., Briand, E., Helfrich, E. J. N., Granatosky, E. A., Glukhov, E., Ryffel, F., Houson, H., Mohimani, H., Kharbush, J. J., Zeng, Y., Vorholt, J. A., Kurita, K. L., Charusanti, P., McPhail, K. L., Nielsen, K. F., Vuong, L., Elfeki, M., Traxler, M. F., Engene, N., Koyama, N., Vining, O. B., Baric, R., Silva, R. R., Mascuch, S. J., Tomasi, S., Jenkins, S., Macherla, V., Hoffman, T., Agarwal, V., Williams, P. G., Dai, J., Neupane, R., Gurr, J., Rodriguez, A. M. C., Lamsa, A., Zhang, C., Dorrestein, K., Duggan, B. M., Almaliti, J., Allard, P. M., Phapale, P., Nothias, L. F., Alexandrov, T., Litaudon, M., Wolfender, J. L., Kyle, J. E., Metz, T. O., Peryea, T., Nguyen, D. T., VanLeer, D., Shinn, P., Jadhav, A., Muller, R., Waters, K. M., Shi, W., Liu, X., Zhang, L., Knight, R., Jensen, P. R., Palsson, B. O., Pogliano, K., Linington, R. G., Gutierrez, M., Lopes, N. P., Gerwick, W. H., Moore, B. S., Dorrestein, P. C. & Bandeira, N. Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nat Biotechnol* **34**, 828-837, 2016. PMID: 27504778.

70 Gomez-Escribano, J. P., Alt, S. & Bibb, M. J. Next generation sequencing of Actinobacteria for the discovery of novel natural products. *Mar Drugs* **14**, 2016. PMID: 27089350.

71 Gao, Z. M., Wang, Y., Tian, R. M., Wong, Y. H., Batang, Z. B., Al-Suwailem, A. M., Bajic, V. B. & Qian, P. Y. Symbiotic adaptation drives genome streamlining of the cyanobacterial sponge symbiont "Candidatus Synechococcus spongiarum". *MBio* **5**, e00079-00014, 2014. PMID: 24692632.

72 Tian, R. M., Zhang, W., Cai, L., Wong, Y. H., Ding, W. & Qian, P. Y. Genome Reduction and Microbe-Host Interactions Drive Adaptation of a Sulfur-Oxidizing Bacterium Associated with a Cold Seep Sponge. *mSystems* **2**, 2017. PMID: 28345060.

73 Goodwin, S., McPherson, J. D. & McCombie, W. R. Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics* **17**, 333, 2016.

74 Slaby, B. M., Hackl, T., Horn, H., Bayer, K. & Hentschel, U. Metagenomic binning of a marine sponge microbiome reveals unity in defense but metabolic specialization. *Isme j* **11**, 2465-2478, 2017. PMID: 28696422.

75 Calteau, A., Fewer, D. P., Latifi, A., Coursin, T., Laurent, T., Jokela, J., Kerfeld, C. A., Sivonen, K., Piel, J. & Gugger, M. Phylum-wide comparative genomics unravel the diversity of secondary metabolism in Cyanobacteria. *BMC Genomics* **15**, 977, 2014. PMID: 25404466.

76 Meyer, J. L., Paul, V. J., Raymundo, L. J. & Teplitski, M. Comparative Metagenomics of the Polymicrobial Black Band Disease of Corals. *Front Microbiol* **8**, 618, 2017. PMID: 28458657.

77 Casamatta, D., Stanić, D., Gantar, M. & Richardson, L. L. Characterization of Roseofilum reptotaenium (Oscillatoriales, Cyanobacteria) gen. et sp. nov. isolated from Caribbean black band disease. *Phycologia* **51**, 489-499, 2012.

78 Richardson, L. L., Stanić, D., May, A., Brownell, A., Gantar, M. & Campagna, S. R. Ecology and Physiology of the Pathogenic Cyanobacterium Roseofilum reptotaenium. *Life (Basel)* **4**, 968-987, 2014. PMID: 25517133.

79 Dadheech, P. K., Abed, R. M. M., Mahmoud, H., Mohan, M. K. & Krienitz, L. Polyphasic characterization of cyanobacteria isolated from desert crusts, and the description of Desertifilum tharense gen. et sp. nov. (Oscillatoriales). *Phycologia* **51**, 260-270, 2012.

80 Sinetova, M. A., Bolatkhan, K., Sidorov, R. A., Mironov, K. S., Skrypnik, A. N., Kupriyanova, E. V., Zayadan, B. K., Shumskaya, M. & Los, D. A. Polyphasic characterization of the thermotolerant cyanobacterium Desertifilum sp. strain IPPAS B-1220. *FEMS Microbiology Letters* **364**, 2017.

81 Engene, N., Rottacker, E. C., Kaštovský, J., Byrum, T., Choi, H., Ellisman, M. H., Komárek, J. & Gerwick, W. H. Moorea producens gen. nov., sp. nov. and Moorea bouillonii comb. nov., tropical marine cyanobacteria rich in bioactive secondary metabolites. *Int J Syst Evol Microbiol* **62**, 1171-1178, 2012. PMID: 21724952.

82 Schneider, D., Volkmer, T. & Rogner, M. PetG and PetN, but not PetL, are essential subunits of the cytochrome b6f complex from Synechocystis PCC 6803. *Res Microbiol* **158**, 45-50, 2007. PMID: 17224258.

83 Baniulis, D., Yamashita, E., Whitelegge, J. P., Zatsman, A. I., Hendrich, M. P., Hasan, S. S., Ryan, C. M. & Cramer, W. A. Structure-Function, Stability, and Chemical Modification of the Cyanobacterial Cytochrome b6f Complex from Nostoc sp. PCC 7120*. *J Biol Chem* **284**, 9861-9869, 2009. PMID: 19189962.

84 Schwenkert, S., Legen, J., Takami, T., Shikanai, T., Herrmann, R. G. & Meurer, J. Role of the low-molecular-weight subunits PetL, PetG, and PetN in assembly, stability, and dimerization of the cytochrome b6f complex in tobacco. *Plant Physiol* **144**, 1924-1935, 2007. PMID: 17556510.

85 Baniulis, D., Zhang, H., Zakharova, T., Hasan, S. S. & Cramer, W. A. Purification and crystallization of the cyanobacterial cytochrome b6f complex. *Methods Mol Biol* **684**, 65-77, 2011. PMID: 20960122.

86 Maresca, J. A., Graham, J. E., Wu, M., Eisen, J. A. & Bryant, D. A. Identification of a fourth family of lycopene cyclases in photosynthetic bacteria. *Proc Natl Acad Sci U S A* **104**, 11784-11789, 2007. PMID: 17606904.

87 Kato, K., Tanaka, R., Sano, S., Tanaka, A. & Hosaka, H. Identification of a gene essential for protoporphyrinogen IX oxidase activity in the cyanobacterium Synechocystis sp. PCC6803. *Proc Natl Acad Sci U S A* **107**, 16649-16654, 2010. PMID: 20823222.

88 Badger, M. The roles of carbonic anhydrases in photosynthetic CO2 concentrating mechanisms | SpringerLink. *Photosynthesis Research* **77**, 2003.

89 Badger, M. R. & Price, G. D. CO2 concentrating mechanisms in cyanobacteria: molecular components, their diversity and evolution. *J Exp Bot* **54**, 609-622, 2003. PMID: 12554704.

90 Pena, K. L., Castel, S. E., de Araujo, C., Espie, G. S. & Kimber, M. S. Structural basis of the oxidative activation of the carboxysomal gamma-carbonic anhydrase, CcmM. *Proc Natl Acad Sci U S A* **107**, 2455-2460, 2010. PMID: 20133749.

91 Smyth, K. M. & Marchant, A. Conservation of the 2-keto-3-deoxymanno-octulosonic acid (Kdo) biosynthesis pathway between plants and bacteria. *Carbohydr Res* **380**, 70-75, 2013. PMID: 23974348.

92 Durai, P., Batool, M. & Choi, S. Structure and Effects of Cyanobacterial Lipopolysaccharides. *Mar Drugs* **13**, 4217-4230, 2015. PMID: 26198237.

93 Perret, X., Staehelin, C. & Broughton, W. J. Molecular basis of symbiotic promiscuity. *Microbiol Mol Biol Rev* **64**, 180-201, 2000. PMID: 10704479.

94 Mathis, R., Van Gijsegem, F., De Rycke, R., D'Haeze, W., Van Maelsaeke, E., Anthonio, E., Van Montagu, M., Holsters, M. & Vereecke, D. Lipopolysaccharides as a communication signal for progression of legume endosymbiosis. *Proc Natl Acad Sci U S A* **102**, 2655-2660, 2005. PMID: 15699329.

95 Kim, J. K., Jang, H. A., Kim, M. S., Cho, J. H., Lee, J., Di Lorenzo, F., Sturiale, L., Silipo, A., Molinaro, A. & Lee, B. L. The lipopolysaccharide core oligosaccharide of Burkholderia plays a critical role in maintaining a proper gut symbiosis with the bean bug Riptortus pedestris. *J Biol Chem* **292**, 19226-19237, 2017. PMID: 28972189.

96 Lackner, G., Peters, E. E., Helfrich, E. J. & Piel, J. Insights into the lifestyle of uncultured bacterial natural product factories associated with marine sponges. *Proc Natl Acad Sci U S A* **114**, E347-e356, 2017. PMID: 28049838.

97 Shapiro, M. M., Chakravartty, V. & Cronan, J. E. Remarkable diversity in the enzymes catalyzing the last step in synthesis of the pimelate moiety of biotin. *PLoS One* **7**, e49440, 2012. PMID: 23152908.

98 Feng, Y., Napier, B. A., Manandhar, M., Henke, S. K., Weiss, D. S. & Cronan, J. E. A Francisella virulence factor catalyses an essential reaction of biotin synthesis. *Mol Microbiol* **91**, 300-314, 2014. PMID: 24313380.

99 Cummings, S. L., Barbe, D., Leao, T. F., Korobeynikov, A., Engene, N., Glukhov, E., Gerwick, W. H. & Gerwick, L. A novel uncultured heterotrophic bacterial associate of the cyanobacterium Moorea producens JHB. *BMC Microbiol* **16**, 198, 2016. PMID: 27577966.

100 Donia, M. S., Fricke, W. F., Partensky, F., Cox, J., Elshahawi, S. I., White, J. R., Phillippy, A. M., Schatz, M. C., Piel, J., Haygood, M. G., Ravel, J. & Schmidt, E. W. Complex microbiome underlying secondary and primary metabolism in the tunicate-Prochloron symbiosis. *Proc Natl Acad Sci U S A* **108**, E1423-1432, 2011. PMID: 22123943.

101 Wernegreen, J. J. Genome evolution in bacterial endosymbionts of insects. *Nat Rev Genet* **3**, 850-861, 2002. PMID: 12415315.

102 Moya, A., Pereto, J., Gil, R. & Latorre, A. Learning how to live together: genomic insights into prokaryote-animal symbioses. *Nat Rev Genet* **9**, 218-229, 2008. PMID: 18268509.

103 Dale, C. & Moran, N. A. Molecular interactions between bacterial symbionts and their hosts. *Cell* **126**, 453-465, 2006. PMID: 16901780.

104 Klisch, M. & Häder, D. P. Mycosporine-Like Amino Acids and Marine Toxins - The Common and the Different. *Mar Drugs* **6**, 147-163, 2008. PMID: 18728764.

105 Balskus, E. P. & Walsh, C. T. The genetic and molecular basis for sunscreen biosynthesis in cyanobacteria. *Science* **329**, 1653-1656, 2010. PMID: 20813918.

106 Sinha, R. P., Singh, S. P. & Hader, D. P. Database on mycosporines and mycosporine-like amino acids (MAAs) in fungi, cyanobacteria, macroalgae, phytoplankton and animals. *J Photochem Photobiol B* **89**, 29-35, 2007. PMID: 17826148.

107 Arpin, N., Favre-Bonvin, J. & Thivend, S. Structure de la mycosporine 2, nouvelle molecule, isolee de Botrytis cinerea. *Tetrahedron Letters* **18**, 819-820, 1977.

108 Ishida, K., Christiansen, G., Yoshida, W. Y., Kurmayer, R., Welker, M., Valls, N., Bonjoch, J., Hertweck, C., Borner, T., Hemscheidt, T. & Dittmann, E. Biosynthesis and structure of aeruginoside 126A and 126B, cyanobacterial peptide glycosides bearing a 2-carboxy-6-hydroxyoctahydroindole moiety. *Chem Biol* **14**, 565-576, 2007. PMID: 17524987.

109 Rudolf, J. D., Yan, X. & Shen, B. Genome Neighborhood Network Reveals Insights into Enediyne Biosynthesis and Facilitates Prediction and Prioritization for Discovery. *J Ind Microbiol Biotechnol* **43**, 261-276, 2016. PMID: 26318027.

110 Hagmann, L. & Jüttner, F. Fischerellin A, a novel photosystem-II-inhibiting allelochemical of the cyanobacterium Fischerella muscicola with antifungal and herbicidal activity. *Tetrahedron Letters* **37**, 6539-6542, 1996.

111 Donia, M. S., Ravel, J. & Schmidt, E. W. A global assembly line for cyanobactins. *Nat Chem Biol* **4**, 341-343, 2008. PMID: 18425112.

112 Agarwal, V., El Gamal, A. A., Yamanaka, K., Poth, D., Kersten, R. D., Schorn, M., Allen, E. E. & Moore, B. S. Biosynthesis of polybrominated aromatic organic compounds by marine bacteria. *Nat Chem Biol* **10**, 640-647, 2014. PMID: 24974229.