UNIVERSITY OF CALIFORNIA SAN DIEGO

Experimental and phylogenetic approaches to understand natural product biosynthetic gene cluster evolution in the marine actinomycete genus *Salinispora*

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

in

Marine Biology

by

Kaitlin Emma Creamer

Committee in charge:

Professor Paul R. Jensen, Chair
Professor Eric E. Allen
Professor Rachel Dutton
Professor Susan Golden
Professor Bradley S. Moore

2022

The Dissertation of Kaitlin Emma Creamer is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2022

DEDICATION

To my parents and brother.
Thank you for your unwavering love & support,
and for always believing in me.

IN MEMORY OF

Sean P. Bush (1994-2019)

EPIGRAPH

"A human

Microbiome is all the writhing forms on & inside this body.

Drafted under our life.

We are not me—

We are we.

Call us

What we carry."

-Amanda Gorman, *Call Us What We Carry*

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

"I sketch out possible projects. Alchemy, for one. If I could achieve that today, I could graduate tomorrow. // If alchemy doesn't work, I will move on to desalinating all of our oceans and providing freshwater to the people." -Weike Wang, *Chemistry*. First, a big thank you to Paul, for the opportunity to be a graduate student your lab, and for the adventures during my Ph.D. from travels across the country for coursework and conferences to the currents of the California Borderlands on the *E/V Nautilus*, I'm very grateful for your support. I've grown as a scientist in both research and writing; and that came from the flexibility and freedom you provided in supporting my research interests and letting me take full leadership of projects. I hope you will continue to lean into tough conversations to make the research experience in CMBB a better, more equitable environment that challenges and improves beyond the status quo. And, I'm thankful I didn't have to figure out alchemy to graduate, though it sometimes felt like that with NaPDoS2.

I would like to thank my committee members for their guidance, feedback, help, and support—Prof. Paul Jensen, Prof. Eric Allen, Prof. Rachel Dutton, Prof. Susan Golden, and Prof. Bradley Moore. I look up to each of you as scientists and leaders, and I appreciate all the advice you have shared along this journey.

Thank you to my coauthors on my research projects; I learned so much from each collaboration: the NaPDoS1 and 2 teams, especially Leesa Klau, Sheila Podell, Hans Singh, Alyssa Demko; the salinipostin team, especially Yuta Kudo; the new *Salinispora* team, especially Victoria Vasilat, David Vereau-Gorbitz, Alyssa Demko; and the pacificamide team, especially Gabriel Castro-Falcón. I want to thank the students that I've had the privilege to mentor—Maggie Lara, Victoria Vasilat, David Vereau-Gorbitz— thank you for your hard work and patience, it has been a joy to see you each accomplish great things.

I want to thank the SIO Grad Office for all of their work behind the scenes, your work does not go unappreciated: Denise Darling, Maureen McGreevy, Shelley Weisel, Gilbert Bretado, Tim DeBold, Olivia Padilla. Thank you also to the CMBB and MBRD business and fiscal office, especially for all the help battling Oracle and Concur. Thank you to SIO FM, especially Jackie for always checking in, and to Dejan Ristic & his team for all of their help over the years.

This dissertation would not have been possible without the help, support, friendship, and "that's fantastic" vibes, both in and out of the lab, from past and present members of the Jensen lab; I feel extremely lucky to be able to call such talented colleagues friends: Dulce Guillén Matus, Alyssa Demko, Leesa Klau, Julia Busch, Hans Singh, Henrique Ramalho Machado, Victoria Vasilat, Krystle Chavarria, Natalie Millán-Aguiñaga, Johanna Gutleben, and Gabriel Castro-Falcón. As it says on our inspirational supply closet poster, perhaps doubling as our lab motto: "All you need is the right equipment and a slight chemical imbalance in the brain". And thank you, I suppose, to those who perhaps unbeknownst to them were powerful motivators. I continued forward in this Ph.D. journey despite your doubts, your words, your actions. Being told "the only reason you got that fellowship/award/position was because you are a woman; it's so easy for you" and so many other countless iterations—thank you. Your words and actions that tried to break us even when & after we spoke up were the tinder to the fire. Nevertheless, she persisted.

I want to also thank the broader SIO, CMBB, and MBRD community at SIO, especially the S^3 2017-2018 co-organizers, the RPTP crew, the past & present Moore and Fenical lab Scholander Hall members, and everyone in the SIO-262/291 audiences over the years. Thank you to Prof. Brian Palnik, I learned so much from being a TA for your class even while trying to adapt to remote instruction. Thanks to Carlos Bauer, fellow Mercer St. researcher extraordinaire, for providing a house that truly became home over the past six years, especially through the trials and

tribulations of the COVID-19 global pandemic lockdown. And thanks to the entire CSHL ABG ’19 crew for the most memorable summer of research, friendship, and fun.

I would not have discovered microbiology and the joy of scientific research if it wasn’t for my Indian Creek & Severn science and math teachers and instrumental Kenyon professors and advisors, especially Chip Voros, Lara Tukarski, Joan Slonczewski, Keith Martinez, Michelle Clark, all *E. coli* lab members, Mark Q. Martindale, Dave Simmons, and Sasha Greenspan, among many others. Thank you for helping me develop my confidence and voice as a strong, independent scientist.

I am forever grateful to the friends that have been there since day one of this journey—Dulce Guillén Matus, Kate Bauman, Kate Nesbit, Mikayla Ortega, among many others in my cohort and SIO community. And to the friends who have been there long before day one—Amanda He, Hillary Chavez, Jenny Sledge, Emmett Carstens— thank you for all your love and support. A special thank you to Patty Crandall for sending lovely cards over the past six years, making tough days less lonely. This dissertation is dedicated in memory of the friend who taken too soon—Sean, you’d be proud of how many CPU hours I burned through building countless trees and genomes; thank you for the adventure that started in Higley 322.

And finally, thank you to my mom, dad, and brother for all their love and support over the years. I owe so much to the values I learned from your example—persistence, strength, confidence, hard work, empathy, kindness, and a love for learning. Thank you for the countless letters, comics, phone calls, and for taking care of Hermit. Thanks Henry for the hiking adventures, I’ve been so proud to watch you accomplish your dreams. Even with a global pandemic making the distance feel further, I am eternally grateful that y’all were there through the highs and the lows, ready to listen and provide support, I love you lots.

Chapter 2, in full, is a reprint of the materials as it was submitted to the *Journal of Biological Chemistry*. Klau, L.J.; Podell, S.; Creamer, K.E.; Demko, A.M.; Singh, H.W.; Allen, E.E.; Moore, B.S.; Ziemert, N.; Letzel, A.C.; Jensen, P.R., 2022. The dissertation author was one of three equally contributing primary authors of this manuscript.

Chapter 3 is coauthored with Hans W. Singh, Dr. Sheila Podell, Dr. Leesa J, Klau, and Dr. Paul R. Jensen. The dissertation author was the primary co-investigator and co-author of this chapter with Hans W. Singh.

Chapter 4 (Section 4.3), in full, is a reprint of the material as it appears in *Microbial Genomics* 7(5), Creamer, K.E.; Kudo, Y; Moore, B.S.; Jensen, P.R., 2021. The dissertation author was the primary investigator and author of this paper.

Chapter 5 is coauthored with Victoria Vasilat, David Vereau-Gorbitz, Alyssa M. Demko, and Paul R. Jensen. The dissertation author was the primary investigator and author of this chapter.

Chapter 5 Appendix is coauthored with David Vereau-Gorbitz, and Paul R. Jensen. The dissertation author was the primary investigator and author of this chapter.

Chapter 6 is coauthored with Paul R. Jensen. The dissertation author was the primary investigator and author of this chapter.

Appendix A (Section A.3), in full, is a reprint of the material as it appears in the *Journal of Natural Products Genomics* 85(4), 980-986. Castro-Falcón, G.; Creamer, K.E.; Chase, A.B.; Kim, M.C.; Sweeney, D.; Glukhov, E.; Fenical, W.; and Jensen, P.R., 2022. "Structure and Candidate Biosynthetic Gene Cluster of a Manumycin-Type Metabolite from *Salinispora pacifica*". The dissertation author was the second author of this paper.

VITA

2012 – 2016	Bachelor of Arts in Biology Kenyon College. Gambier, OH
2018	Master of Science in Marine Biology Scripps Institution of Oceanography University of California San Diego. La Jolla, CA
2016 – 2022	Doctor of Philosophy in Marine Biology Scripps Institution of Oceanography University of California San Diego. La Jolla, CA

PUBLICATIONS

Harden, MM*, He, A*, **Creamer, K***, Clark, MW, Hamdallah, I, Martinez, KA, Kresslein, RL, Bush, SP and Slonczewski, JL. (2015). Acid-adapted strains of *Escherichia coli* K-12 obtained by experimental evolution. *Applied and Environmental Microbiology*, 81(6):1932-1941.

Creamer, KE*, Ditmars, FS*, Basting, PJ*, Kunka, KS, Hamdallah, IN, Bush, SP, Scott, Z, He, A, Penix, SR, Gonzales, AS, Eder, EK, Camperchioli, DW, Berndt, A, Clark, MW, Rouhier KA and Slonczewski, JL. (2017). Benzoate-and salicylate-tolerant strains of *Escherichia coli* K-12 lose antibiotic resistance during laboratory evolution. *Applied and Environmental Microbiology*, 83(2):e02736-16.

He, A, Penix, SR, Basting, PJ, Griffith, JM, **Creamer, KE**, Camperchioli, D, Clark, MW, Gonzales, AS, Chavez Erazo, JS, George, NS, Bhagwat, AA and Slonczewski, JL. (2017). Acid evolution of *Escherichia coli* K-12 eliminates amino acid decarboxylases and reregulates catabolism. *Applied and Environmental Microbiology*, 83(12):e00442-17.

Hamdallah, I, Torok, N, Bischof, KM, Majdalani, N, Chadalavada, S, Mdluli, N, **Creamer, KE**, Clark, M, Holdener, C, Basting, PJ, Gottesman, S and Slonczewski, JL. (2018). Experimental evolution of *Escherichia coli* K-12 at high pH and with RpoS induction. *Applied and Environmental Microbiology*, 84(15):e00520-18.

Schlawis, C, Harig, T, Ehlers, S, Guillén-Matus, DG, **Creamer, KE**, Jensen, PR and Schulz, S. (2020). Extending the salinilactone family. *ChemBioChem*, 21(11):1629-1632.

Kudo, Y, Awakawa, T, Du, YL, Jordan, PA, **Creamer, KE**, Jensen, PR, Linington, RG, Ryan, KS and Moore, BS. (2020). Expansion of gamma-butyrolactone signaling molecule biosynthesis to phosphotriester natural products. *ACS Chemical Biology*, 15(12):3253-3261.

Raineault, NA, J. Flanders, J and Niiler E, eds. New Frontiers in Ocean Exploration: The E/V *Nautilus*, NOAA Ship *Okeanos Explorer*, and R/V *Falkor* 2020 Field Season. (2021). *Oceanography.* 34(1), supplement. (Levin, L, Jensen, P, Rouse, G, Mizell, K, Castro-Falcón, G, Pearson, K, **Creamer, K**, Vlach, D, Auscavitch, S and Coleman, DF. "Biodiversity Baselines and Biopharmaceutical Potential for the Borderland").

Creamer, KE, Kudo, Y, Moore, BS and Jensen, PR. (2021). Phylogenetic analysis of the salinipostin γ-butyrolactone gene cluster uncovers new potential for bacterial signalling-molecule diversity. *Microbial Genomics*, 7(5).

Aceves, CM, Aksenov, AA, Aleti, G, Allevato, MM, Ames, JG, Arang, N, Aron, AT, Bauermeister, A, Bittremieux, W, Bogdanov, A, Cancelada, L, Caraballo-Rodríguez, AM, Chu, T, **Creamer**, **KE,** *et al...*, Bandeira, N, and Dorrestein, PC. (2021). Public Metabolomics Data - An Underutilized Resource For Data Mining and Hypothesis Formulation. (*ChemRxiv*).

Castro-Falcón, G, **Creamer, KE**, Chase, AB, Kim, MC, Sweeney, D, Glukhov, E, Fenical, W, and Jensen, PR. (2022). Structure and Candidate Biosynthetic Gene Cluster of a Manumycin-Type Metabolite from *Salinispora pacifica*. *Journal of Natural Products*. 85(4).

Capone, D, Chamberlain, EJ, **Creamer, KE**, Matus, DG, Plummer, S, Romero, SL, Ufer, SL, Yin, HZ. (March 2022). "Let's talk about science: why and how we need to do better." *Explorations Now*.

Klau, LJ*, Podell, S*, **Creamer, KE***, Demko, AM, Singh, HW, Allen, EE, Moore, BS, Ziemert, N, Letzel, AC and Jensen, PR. (2022). The Natural Product Domain Seeker version 2 (NaPDoS2): Relating ketosynthase phylogeny to biosynthetic function. (*In review at the Journal of Biological Chemistry*).

Singh, HW, **Creamer, KE**, Chase, AB, Klau, LJ, Podell, S, and Jensen, PR. (2022). Metagenomic data reveals type I polyketide synthase distributions across biomes. (*Submitted to mBio*).

ABSTRACT OF THE DISSERTATION

Experimental and phylogenetic approaches to understand natural product biosynthetic gene cluster evolution in the marine actinomycete genus *Salinispora*

by

Kaitlin Emma Creamer

Doctor of Philosophy in Marine Biology

University of California San Diego, 2022

Professor Paul R. Jensen, Chair

Evolution is the fundamental process by which natural selection, genetic variation, and fitness adaptations can shape diversity in nature. At the foundation of natural ecosystems, small molecule chemical compounds, called specialized metabolites, play important ecological roles and mediate complex community interactions. Specialized metabolites are genetically encoded in biosynthetic gene clusters (BGCs) in the genomes of the producing organisms; thus, BGCs can undergo evolutionary diversification resulting in nature's incredible chemical diversity. However, the process by which this diversification occurs is unknown. We can use phylogenetic methods to assess the diversity and distribution of specialized metabolite BGCs to better understand the

dynamics of evolutionary chemical innovation. The goal of this dissertation was to develop new tools to identify novel targets of BGC biosynthetic potential and investigate the contrasting evolutionary history of different BGCs in the genus *Salinispora* to gain insight into the drivers of chemical diversity. First, I helped develop the updated NaPDoS2 webtool which identifies and characterizes polyketide and non-ribosomal peptide biosynthetic potential based on phylogenetically conserved domains in genomic, metagenomic, and targeted-amplicon sequencing data. Next, I used NaPDoS2 to assess the polyketide biosynthetic potential in 620,000 bacterial, archaeal, viral, plasmid, fungal, plant, agal, protist, and animal genomes across the tree of life. The second goal of this dissertation was to investigate the evolutionary patterns BGC diversity in the marine obligate actinomycete *Salinispora*. First, I uncovered an unexpected distribution and diversity of the *Salinispora* salinipostin (*spt*) BGC across all bacteria, including evidence that the entire *spt* BGC was exchanged from *Salinispora arenicola* to *S. tropica* in a location-dependent manner. Next, I applied a similar phylogenetic approach to the pacificamide (*pac*) BGC and found that the *Salinispora pac* had unique gene organization and limited distribution in three divergent Actinomycetia families. To expand my comparative analyses, I isolated and genome-sequenced 99 new "microscale" *Salinispora* strains—including three species of *S. arenicola, S. oceanensis*, and *S. pacifica*— from a $1m^2$ quadrant of marine sediment in Fiji. I found that there was significant genomic diversity within the microscale genomes, including *S. arenicola* sub-species diversification and unique BGCs. To investigate this further, I compared the lanthipeptide ribosomally synthesized and post-translationally modified (RiPP) BGC precursor peptide products, uncovering evidence of evolutionary radiation of diverse potential lanthipeptides. Finally, I explored possible mechanisms of BGC exchange in *Salinispora* by purifying and visualizing *Salinispora* plasmids. In conclusion, the results of this dissertation provide significant

advancements in our ability to detect and classify biosynthetic potential across the tree of life. Additionally, the description of 99 new *Salinispora* genomes and the application of evolutionary phylogenetic methods to specialized metabolite BGC diversification on two different spatial scales contribute to our understanding of *Salinispora* genomic diversity and patterns of BGC-mediated chemical innovation.

CHAPTER 1. Introduction

1.1 Microbial abundance and diversity on Earth

Life on Earth originated with microbes and the single chemical building blocks comprising the original cell's genetic material. Over 3.8 billion years, microbes have continued to adapt to their environment, shaping the world as we know it: Earth's atmosphere, oceans, geology, and the biogeochemical processes powering all ecosystems (Cavicchioli *et al.*, 2019). Adaptation to various changing environments over time has resulted in an incredible amount of microbial diversity, much of which we are still discovering today (Hug *et al.*, 2016). The term "microbe" can encompass many life forms, including bacteria, archaea, fungi, protists, and viruses. We now understand that all macroscopic organisms interact with microbes—no organism solely exists as an individual, but instead as a host to millions of associated microbial cells called the microbiome (Rosenberg and Zilber-Rosenberg, 2016). The microbes in the human gut microbiome, for example, are essential not only for healthy nutritional functioning and robust immune systems, but also thought to influence brain chemistry, affecting how we act and think (Qin *et al.*, 2010; Kau *et al.*, 2011; Launer, 2013). In the environment, microbes truly are the world's life support system, playing key roles in nutrient and carbon cycles, biological productivity, and the global food web; supporting animal and plant life; and crucial for humanity's industry, agriculture, biotechnology, and health purposes (Cavicchioli *et al.*, 2019; Flemming and Wuertz, 2019). Microbiology, the study of microbes, is truly an evolving science where advancing methods of genomic sequencing technology let us explore not only the bacteria we can culture and grow in laboratory settings, but also the un-culturable microbes whose genomes, and thus potential functions, can be captured

(Nayfach *et al.*, 2020; Almeida *et al.*, 2021). This has revolutionized the way we study microbes, especially bacteria, and understanding the role they play in the environment.

Current estimates predict that there are ~10^{30} bacteria cells on Earth, which is a billion times (9 orders of magnitude) larger than the number of stars in our universe's galaxies (Flemming and Wuertz, 2019). Using sequencing-based approaches to estimate the number of bacteria cells in different ecosystems, marine sediment was found to be the most microbially-enriched ecosystem (2,900 x 10^{26} cells), followed by soil (2,560 x 10^{26} cells), terrestrial subsurface (2,500 x 10^{26} cells), seawater (1,010 x 10^{26} cells), freshwater (1.3 x 10^{26} cells), plant hosts (1 x 10^{26} cells), and animal hosts (0.2 x 10^{26} cells) (Lloyd *et al.*, 2018). These estimates support the previous discoveries that revolutionized the study of marine microbiology: the ocean comprising >70% of our truly blue planet, is a microbial soup. In just 1mL of seawater, there are ~10 million viruses, ~1 million bacteria, and ~1,000 protists (Patel *et al.*, 2007). Marine microbes are the ecosystem engineers of the ocean, serving as keystones in the microbial loop, biogeochemical cycling, climate regulation, global oxygen production, and symbioses (Azam *et al.*, 1983). When investigating the role that microbial communities in marine sediments might play, scientists found that a single sand grain was covered with a diverse bacterial community comprised of over ~100,000 bacterial cells (Probandt *et al.*, 2018). We have not cultured this type of microbial diversity and abundance in laboratory conditions, as shown in a recent study where only 3% of a marine sediment community was captured using culture-dependent approaches compared to culture-independent techniques (Demko *et al.*, 2021). This leads us to a defining question of this dissertation—if the marine sediment microbial community is 10,000 times denser than the ocean water column, what is mediating the evolutionary adaptations, interactions, and population dynamics in such complex and diverse microbial communities?

1.2 Chemicals mediate complex interactions

Microbes have been described as the "chief molecular innovators of the biosphere" (Ulvestad, 2009), and one process by which they adapt, persist, and innovate is the production of various chemical compounds. When thinking about the types of molecules that bacteria produce, we can group them into two categories: primary and secondary (specialized) metabolites. Primary metabolites are chemical compounds that are essential for growth, development, and short-term survival in bacteria. Many of these are similar across broad phylogenetic groups and generally involved in basic metabolic processes. Primary metabolites include amino acids (peptides and proteins); carbohydrates (mono- and polysaccharides); fatty acids (phospholipid membranes); and nucleotides (DNA and RNA molecules), along with cofactors and biological polymers, to name a few (Keller, 2019). While bacteria can differ from one another based on these primary metabolites, there is a second group of compounds generally referred to as secondary metabolites. Secondary metabolites were named "secondary" to distinguish them from growth-related primary metabolism based on laboratory observations that the metabolites were produced during late phases of bacterial growth (Osbourn, 2010). However, this definition has since been expanded, especially with our increasing understanding of the regulation, expression dynamics, and genetic and biochemical investment in the production of these metabolites and the need to account for complex environmental communities with slow growth rates (Chevrette *et al.*, 2020). In this dissertation, I use the term specialized metabolites which is synonymous with the terms "secondary metabolites" and "natural products". The term "specialized metabolites" is considered an appropriate classification of these molecules to emphasize their importance in microbial ecology, even if the specific roles of the molecules are not known. Specialized metabolites are organic chemical

compounds that are essential for long-term survival of species that provide improved adaptation or overall fitness. Oftentimes, the necessary precursor building blocks for specialized metabolites are borrowed from primary metabolic pathways. The functions of specialized metabolites are usually ecological in nature, including signaling, defense against predators, interspecies competition, nutrient acquisition, and respiration. Specialized metabolites are not only produced by microbes, but all life forms including plants, animal, fungi, and protists (Medema *et al.*, 2021).

Specialized metabolites have diverse biosynthetic origins and chemical structures which can be used to classify them into different groups (Hug *et al.*, 2020). Classes and types of natural products are always expanding as new biosynthetic scaffolds and mechanisms are discovered. Some well-known groups include: terpenes, polyketides (PKs), ribosomally synthesized and post-translationally modified peptides (RiPPs), non-ribosomal peptides (NRPs), alkaloids, glycosides, nucleosides, steroids, and shikimic-acid/other primary metabolite derived products (Hug *et al.*, 2020). Terpenes are one of the largest classes of natural products across all kingdoms of life, found in organisms like plants, corals, fungi, and bacteria (Schulz *et al.*, 2020). They are derived from 5-carbon isoprene units which are ultimately combined and converted into the final monoterpene, sesquiterpene, and diterpene products. Polyketides comprise a very diverse group of natural products and are divided into 3 main classes. Their encoding polyketide synthase (PKS) enzymes resemble fatty-acid synthases, and they employ distinct stepwise assembly-like processes of iterative decarboxylative Claisen thioester condensations of 2 or 3 carbon building blocks that are derived from acetyl-CoA or propionyl-CoA (Hertweck, 2009; Nivina *et al.*, 2019). Type I, II, and III PKS pathways can either be assembly-line like (type II), or iterative (type I, II, III) where specific ketoreduction, cyclization, and other modifications are catalyzed by distinct tailoring enzymes instead of modular domains within a single gene (Hertweck, 2009; Nivina *et al.*, 2019).

This class of natural products is described in detail in Chapter 2 and 3 of this dissertation. Peptide natural products, including RiPPs and NRPs are formed by the condensation of amino acids, and in many cases, diversity arises from the incorporation of non-proteinogenic amino acids (Sikandar and Koehnke, 2019). RiPPs are synthesized by the ribosome; this class of natural products is described in detail in Chapter 6 of this dissertation. NRPs are assembled from large multifunctional and modular enzymes called non-ribosomal peptide synthetases (NRPSs) comprised of linear catalytic domains that are responsible for different reaction steps during the biosynthetic assembly. Briefly, RiPP biosynthesis begins with the ribosomal generation of a precursor peptide containing a leader and core component. Next, various enzymes recognize elements in the leader peptide and install post-translational modifications. This is followed by subsequent proteolysis and export of the modified core peptide to complete the biosynthesis (Arnison *et al.*, 2013; Le and van der Donk, 2021; Montalbán-López *et al.*, 2021). Some natural product biosynthetic mechanisms are combined together to form unique hybrid products, as in the case of a recently discovered amino-acid derived natural product that starts from a RiPP precursor peptide but is subsequently modified by NRPS machinery (Ting *et al.*, 2019). Thus, specialized metabolite chemical diversity is a product of diverse and combinatorial biosynthetic origins produced by organisms across the tree of life.

The ocean has been a prolific source of natural products, both from macroscopic and microscopic organisms. Organic chemists in the early 1970s started studying marine organisms such as seaweeds and sponges when they quickly realized that every compound isolated from the marine environment was new. A unique feature of marine natural products is that due to the prevalence of halogens (existing as halides) in the marine environment—such as Fluorine (F-), Chlorine (Cl-), Bromine (Br-), and Iodine (I-)— many marine natural products are halogenated

with exciting biological activities. In fact, one of the first reported marine microbial antibiotics was pentabromopseudilin (characterized from the Gram-negative bacterium *Pseudomonas bromoutilis* isolated from the surface of a marine seagrass in 1966), which was found to be 70% bromine by weight (Burkholder *et al.*, 1966).

Marine natural products are not only present in ocean ecosystems, but they also play important roles in mediating complex interactions. In 1989, an anti-fungal molecule named isatin was isolated from a commensal marine *Alteromonas* bacterial strain that colonized *Palaemon macrodactylus* shrimp egg embryos (Gil-Turnes *et al.*, 1989). Isatin was found to be the causative agent protecting the eggs from fungal pathogens (Gil-Turnes *et al.*, 1989). This was an early example of a microbial specialized metabolite playing an important role in marine chemical ecology. Chemical ecology is the study of interactions of plants and animals based upon chemical signals. In the ocean, one can imagine that small organic molecules such as specialized metabolites are the primary mechanism of communication. Marine microbial specialized metabolites can be used for: signaling molecules, settlement cues, developmental cues, mate recognition, allelopathy, prey detection, defense, and predator avoidance. For example, the molecule tetrabromopyrrole produced by the marine bacterium *Pseudoalteromonas sp.* induces complete settlement, attachment, and metamorphosis in Caribbean coral *Porites asteoides* larvae (Sneed *et al.*, 2014). In another example, the filamentous cyanobacteria *Moorena bouillonii* was found to coexist in a symbiotic relationship with the snapping shrimp *Alpheus frontalis* that weaves protective tubes out of the specialized metabolite-enriched cyanobacteria (Leber *et al.*, 2021). Furthermore, some marine natural products with unknown ecological functions have other important implications for human health, including the potent neurotoxin domoic acid produced by the diatom *Pseudo-nitzschia australis* in large algal-blooms (Brunson *et al.*, 2018).

In addition to serving important ecological functions, to date there are seventeen marine natural products (or their derivatives) that are registered as drugs; and another twenty-three are currently in clinical trials (Voser *et al.*, 2021) (https://www.marinepharmacology.org/approved). It should be noted, however, that Indigenous peoples were the first to discover and apply the knowledge of natural products for thousands of years before the 19th century epoch of "modern" drug discovery and development (Veeresham, 2012). Natural products form the foundation of traditional medicinal healing practices and are culturally and historically significant to Indigenous peoples across the globe (Heinrich, 2000). While natural products have been instrumental in modern drug development and advancements in critical therapeutics, it must be recognized that oftentimes the natural sources and inspirations for these molecules have been exploitatively extracted from the natural environment in the name of "bioprospecting". Moreover, the subsequent drug patents and research findings from these expeditions are not typically shared with the people whose cultural knowledge informed the discovery and whose land (and marine resources) they were found on (Vierros *et al.*, 2016). This has been called parachute science, which is the practice where international scientists, typically from wealthy nations, conduct studies in other countries that are often poorer without meaningful communication or collaborations with local people (Stefanoudis *et al.*, 2021). Efforts to address these longstanding inequitable practices and murky ownership of marine resources has included implementation of programs like the Nagoya Protocol. The Nagoya Protocol set standards for legal acquisition of genetic resources through due diligence, traceability, risk assessment and mitigation, and national authority inspections in an attempt to ensure fair and equitable benefits from shared resources between ratifying countries (Vierros *et al.*, 2016; De Mol *et al.*, 2018; Loureiro *et al.*, 2018). There is a critical need for new drug sources with potent pharmacological properties, and while the ocean holds incredible potential, future

discovery efforts can only succeed if our marine natural resources are sustainably managed and the original traditional knowledge is respected, credited, and preserved (Vierros *et al.*, 2016).

Important medicines with marine natural product origins have been isolated from a variety of sources. The first marine drug to be approved by the FDA was Cytosar-U (cytarabine) in 1969, which was originally isolated from the Caribbean sponge *Cryptotheca crypta* (Schwartsmann *et al.*, 2001). Cytarabine is still used today for treatment of leukemia and lymphoma (Schwartsmann *et al.*, 2001). Another early drug from the sea was Prialt, (ziconotide) which is a ω-conotoxin peptide originally isolated from the tropical marine cone snail *Conus magus* (Olivera *et al.*, 1985, 1990). Prialt was approved in in the United States for the treatment of chronic pain in spinal cord injury in 2004 (McGivern, 2007). Halaven mesylate (eribulin) is an analog of the halichondrin B molecule that was isolated from the marine sponge *Halichondria okadai* (Hirata and Uemura, 1986). Halaven was approved as a microtubule-depolymerizing drug for the treatment of breast cancer in 2010 (Dybdal-Hargreaves *et al.*, 2015). The most recently approved marine drug in 2021 was TIVDAK, or tisotumab vedotin-tftv for the treatment of metastatic cervical cancer. Tisotumab vedotin-tftv is an antibody-drug conjugate with a payload drug of monomethyl auristatin E, which is an analog of the peptide dolastatin 10 that was originally isolated from the Indian Ocean mollusk *Dolabella auricularia* (de la Torre and Albericio, 2022). While the point is not to exhaustively list all the marine natural products that have been approved as drugs, the range of marine organisms that have been important drug sources to date include sponges, tunicates, mollusks, and fish (https://www.marinepharmacology.org/approved). However, the drug Marizomib (salinosporamide A), which is currently undergoing phase III clinical trials for the treatment of cancers like glioblastoma, was originally discovered in the marine actinomycete *Salinispora* (Feling *et al.*, 2003). Salinosporamide A irreversibly binds to and inhibits the 20S catalytic core

subunit of the proteasome, making it more potent and selective than other proteasome inhibitor drugs. Recently, the entire biosynthesis of salinosporamide A in *Salinispora* was solved (Eustáquio *et al.*, 2009; Bauman *et al.*, 2022). Approval of salinosporamide A (Marizomib) for treatment would be the very first drug from a marine bacterial source, illustrating the incredible potential of marine natural products from bacteria for future drug discovery efforts. Taking a step back from solely marine natural products, specialized metabolites as a whole have been a primary source of clinically approved antibiotics, anticancer drugs, immunosuppressants, and other therapeutically relevant molecules (Chevrette and Currie, 2019), and thus understanding how they are biosynthesized and undergo incredible diversification in various bacteria can help us discover the drugs of the future.

1.3 Biosynthetic gene clusters encode specialized metabolites

Many specialized metabolites confer a selective advantage against competing organisms for their producers. However, specialized metabolites are encoded by biosynthetic gene clusters (BGCs), and thus the adaption and evolution of BGCs can also result in new chemical diversity that can be selected upon in iterative processes (Chevrette *et al.*, 2020). In bacteria, it is commonly observed that genes involved in successive steps of a biosynthetic pathway are clustered together on the chromosome (Roth and Lawrence, 1996). Biosynthetic gene clusters (BGCs) include all or most of the genes that are responsible for biosynthesis of a specialized metabolite. What makes specialized metabolite BGCs interesting to study is that they include not only the core biosynthetic (or "signature") specialized metabolite genes and relevant tailoring enzymes, but also genes associated with the regulation and resistance to the small-molecule product(s) they encode.

Signature biosynthetic genes encode for enzymes that build the core skeleton structures of different specialized metabolite classes, including non-ribosomal peptide synthetases (NRPSs), polyketide synthases (PKSs), and terpene cyclases (Osbourn, 2010; Medema *et al.*, 2015). These core biosynthetic genes can be used as "hooks" when genome mining, or searching for distinct BGCs that have the potential to produce a certain type of compound (Bauman *et al.*, 2021).

Biosynthetic gene clusters have been reported in bacteria, fungi, plants, and even animals, highlighting that these BGCs and their encoded chemical compounds confer important selective advantages for the producing organisms in the natural environment. (Osbourn, 2010; Nützmann *et al.*, 2018). Linking BGCs to their cognate molecules can be challenging because gene clusters are not always colinearly expressed (Machado *et al.*, 2017). While groups of adjacent genes are oftentimes part of the same operon, that is, they are transcribed as a single molecule of mRNA, environmental conditions and developmental processes can both affect BGC expression. BGCs that have been identified in a genome but no molecule has been linked to them are called silent, or cryptic. To activate these cryptic BGCs, researchers genetically manipulate strains by adding constitutively expressed promoters (Bauman *et al.*, 2019) or other genetic modifications (Kuhl *et al.*, 2021), co-culture strains with other organisms (Traxler *et al.*, 2013; Netzker *et al.*, 2015; Nai and Meyer, 2017; Sung *et al.*, 2017), treating them with antibiotics (Onaka, 2017) and other transcriptional or translational inhibitors (Almabruk *et al.*, 2018), growing strains in a wide variety of culturing conditions ("one strain many compounds" OSMAC) (Romano *et al.*, 2018; Hernandez *et al.*, 2021; Soldatou *et al.*, 2021), or other high-throughput methods to link genomic and metabolomic potential (Okada and Seyedsayamdost, 2017; Moon *et al.*, 2019; van der Hooft *et al.*, 2020; McCaughey *et al.*, 2021; Schorn *et al.*, 2021).

To find new natural product drug scaffolds and molecules with interesting biological activities, phylogenetic methods that investigate the evolutionary history and homologous comparisons of BGCs can be used (Adamek *et al.*, 2019). There are many tools that have been developed for genome mining (Chevrette and Handelsman, 2021), and their applications and future potential for improvement are discussed in Chapter 2 and 3 of this dissertation.

As more examples of the diversity and distribution of BGCs arise from *in silico* comparative genomic analyses and *in vitro* biosynthetic characterization of bacterial BGCs, the explanations of evolutionary processes contributing to these patterns can be revisited (Fischbach *et al.*, 2008; Osbourn, 2010; Jensen, 2016; Ruzzini and Clardy, 2016). The process of horizontal gene transfer (HGT) is commonly reported as being essential to BGC diversification and spread, as adaptive logic predicts the acquired ability to produce a specialized metabolite can drastically impact strain fitness. Horizontal gene transfer is defined as any process in which an organism incorporates genetic material from another organism without being offspring of that organism (Osbourn, 2010; Soucy *et al.*, 2015). This is in contrast to vertical gene transfer which is where one organism inherits its genetic material from a parent, or dividing sister cell (Osbourn, 2010). Evolutionary analyses have been conducted looking at the patterns of diversity and distribution of fungal specialized metabolite BGCs which sometimes show clustering similar to bacterial BGCs; HGT is now recognized as an important source of genomic innovation of BGCs in filamentous fungi (Wisecaver and Rokas, 2015; Lind *et al.*, 2017; Hoogendoorn *et al.*, 2018).

While HGT is less commonly reported in eukaryotes compared to prokaryotes, a recent landmark study showed genomic and direct experimental evidence of a yeast that acquired and functionalized a bacterial siderophore BGC, illustrating that modern techniques allow the investigation of HGT and horizontal operon transfer “HOT” between branches on the tree of life

(Kominek *et al.*, 2019). The evolutionary patterns and trajectories of transferred BGCs show patterns from the maintenance of ancestral BGCs to examples of pseudogenization, gene loss, rearrangement, expression changes, and new gene recruitment when BGCs undergo diversifying selection (Wisecaver and Rokas, 2015). It is clear that the evolutionary focus of specialized metabolite BGCs is informed by the growing body of literature exploring the evolution of primary metabolic and general gene clustering in bacteria (Roth and Lawrence, 1996; Pál and Hurst, 2004; Hosseini and Wagner, 2018). There are many hypotheses of how primary bacterial metabolic and essential gene operons become and stay clustered, and similar questions can be applied to the patterns seen in specialized metabolite BGCs.

There are very few experiments that have empirically explored the barriers and facilitators of HGT of BGCs in prokaryotes and in fungi (de Reus *et al.*, 2019). While the facilitated transfer of BGCs has been developed as an essential tool in heterologous expression of BGCs (Zhang *et al.*, 2019), this process is likely different from exchange events occurring in the natural environment. This led to the questions that formed a framework for how I approached the work presented in this dissertation: What are the evolutionary patterns of BGCs observed in bacteria? What selection pressures or processes facilitate the transfer, maintenance, expression, and adaptation of BGCs in bacteria? Are there different mechanisms of BGC origin and persistence compared to other clustered bacterial operons and to fungal BGCs? How do we expand beyond previously recognized BGCs and look for novel BGCs assisted by evolutionary theory?

1.4 The marine actinomycete *Salinispora* as a model system

The bacterial genus *Salinispora* was the first obligate marine actinomycete genus to be described (Mincer *et al.*, 2002; Maldonado *et al.*, 2005). At present time, hundreds of *Salinispora* strains have been isolated from marine sediments (Mincer *et al.*, 2002, 2005; Jensen *et al.*, 2005), marine seaweed (Jensen *et al.*, 2005), and marine sponges (Kim *et al.*, 2005; Vidgen *et al.*, 2012). *Salinispora* are slow-growing, aerobic, Gram-positive high G+C content (~69%) actinomycetes that have orange branching, filamentous substrate mycelium and form black non-motile spores after extended growth in agar and liquid culture media (Maldonado *et al.*, 2005). Next-generation sequencing technology was used to sequence 118 *Salinispora* genomes (Millán-Aguiñaga *et al.*, 2017) and subsequent whole-genome average nucleotide identity (ANI) analyses confirmed that the *Salinispora* genus includes nine species: *Salinispora tropica, S. arenicola, S. pacifica, S. mooreana, S. cortesiana, S. fenicalii, S. vitiensis, S. goodfellowii,* and *S. oceanensis* (Román-Ponce *et al.*, 2020). There have been additional *Salinispora* genomes sequenced, including those isolated from deep-sea samples (Ulanova *et al.*, 2020), marine sediments in Hawaii (Schulze *et al.*, 2015), Brazil (Bauermeister *et al.*, 2018), the Federated States of Micronesia (Matsuda *et al.*, 2009), however many of these have not yet been included in comprehensive full-genome analyses for classification based on similarity to a *Salinispora* type species.

Salinispora is a prolific producer of specialized metabolites with important biological activities, including cancer cell cytotoxicity, antibiotic, and antifungal properties (Mincer *et al.*, 2002). Over 40 compounds have been characterized from *Salinispora* to date and the number of new molecules, derivatives, and analogs continue to grow (Jensen *et al.*, 2015). Complementing continued chemical work to isolate new compounds, *Salinispora* genome sequences have

facilitated an *in-silico* genome-mining approach to link BGCs to metabolites of interest (Bauermeister *et al.*, 2018). Co-culture methods (Patin *et al.*, 2015, 2018), direct genetic manipulation (Schlawis *et al.*, 2018), heterologous expression (Zhang *et al.*, 2018), comparative transcriptomics (Amos *et al.*, 2017), and even re-isolation of additional analogs and side-products from characterized biosynthetic pathways (Williams *et al.*, 2022) continue to add to the known *Salinispora* chemical repertoire. Additionally, investigations into the ecological functions contribute to our understanding of how *Salinispora* might use specialized metabolites in the natural environment. For example, for defense against invertebrate predation (Tuttle *et al.*, 2022), or protection against UV and oxidative stress (Jezkova *et al.*, 2021). In this dissertation, exploration of a *Salinispora* BGC encoding a putative signaling molecule is discussed (Chapter 4), and the discovery of a new *Salinispora* specialized metabolite, pacificamide (and the report of a known molecule not seen before in *Salinispora*, triacsin D), with weak antibiotic activity and cytotoxicity against lung cancer cell lines, respectively, is reported in Appendix A. Additional background on the *Salinispora* genus, its BGC biosynthetic potential, and proposed processes driving BGC diversification are explored in Chapter 5 and 6.

1.5 Overview of the Dissertation

The overarching goal of this dissertation was to investigate how BGC diversification over time contributes to nature's chemical diversity. This was divided into three connected research challenges:

1) We need tools to help find new natural products and patterns of specialized metabolite biosynthetic potential.

2) We do not know the evolutionary patterns and drivers of BGC diversification in bacteria, especially the marine obligate actinomycete *Salinispora*. Do *Salinispora* grow clonally or exist as multiple sub-populations within a microscale-sampled marine sediment quadrant? Do closely related *Salinispora* share the same BGC biosynthetic potential? Does HGT contribute to BGC diversification in *Salinispora*? Do different types of BGCs undergo different rates of evolution?
3) What could be the mechanism for *Salinispora* BGC diversification, including the exchange and acquisition of BGCs? Could this be mediated by plasmids or a BGC "mobilome"?

To answer these questions, I developed new tools to access biosynthetic potential in all genomes across the tree of life, whole-genome sequenced a "microscale" collection of 99 *Salinispora* marine actinomycete strains, preliminarily characterized the presence of plasmids in *Salinispora*, and analyzed a class of BGC and two specific *Salinispora* BGCs with known products to discover their unique evolutionary history. This dissertation is divided into five main research chapters and one appendix research chapter, as follows:

In Chapter 2, I contributed to the development of the NaPDoS2 webtool that detects and classifies polyketide and non-ribosomal peptide biosynthetic potential. Webtools like NaPDoS2 are essential tools in the arsenal of genome mining techniques to identify biosynthetic gene clusters of interest. To benchmark the accuracy of NaPDoS2, I created a first-of-its-kind negative and positive control dataset of structurally and functionally relevant sequences. Finally, I led the testing of the webtool where I analyzed multiple biological sequencing datasets, including genes, bacterial genomes, fungal genomes, animal genomes, metagenomes, and KS amplicon sequences. By

working closely with the two main developers of the webtool database and classification scheme (Dr. Leesa Klau) and the computational backend of the tool (Dr. Sheila Podell), we significantly improved the analysis capability and user accessibility of the NaPDoS2 webtool, thus providing an excellent tool to the community for identifying promising biosynthetic potential. Chapter 2 is currently in review at the *Journal of Biological Chemistry* and thus remains largely unchanged from its manuscript form in this dissertation.

Chapter 3 seeks to apply the NaPDoS2 webtool developed in Chapter 2 to explore the distribution and diversity of polyketide potential across the genome-sequenced tree of life. For this chapter, my collaborator (Hans W. Singh) and I analyzed 620,000 genomes across the tree of life to map taxa-specific distributions of polyketide ketosynthase (KS) biosynthetic potential. We selected representative genomes comprehensively across all major phyla, including bacterial, fungal, animal, plant, protist, algae, archaeal, CPR, viral, and plasmid taxa. By using NaPDoS2 to detect and classify all KS domains from these genomes in a phylogeny-driven classification scheme, we were able to map specific polyketide class/subclass biosynthetic potential in taxa at a level of detail that has not been possible to date. We observed unexpected taxa with KS biosynthetic potential and divergent taxa sharing similar KSs. Phylogenetic analyses demonstrate where KSs we identified were unlike known BGC KSs, and thus future efforts can target natural product discovery from these taxa. To our knowledge, this is the first KS dataset across the entire tree of life of its type, and thus this analysis presents a unique opportunity for future dereplication methods to find additional polyketide potential. This chapter is currently being written into a manuscript for publication in collaboration with co-first author Hans W. Singh.

Chapter 4 seeks to explore the specific evolutionary diversity and distribution of the *Salinispora* salinipostin BGC. After the biosynthesis of the salinipostin gene cluster was fully

characterized, we sought to explore if other bacteria had the potential to produce salinipostin-like molecules. I used the essential *spt9* gene and the entire salinipostin *spt* BGC as search hooks to identify 500 top *spt9* homologs in diverse gene neighborhoods and 91 salinipostin-like BGCs. By further investigating the *spt* BGC in the *Salinispora* genus, I uncovered evidence of a location-dependent recombination event of the entire *spt* BGC being shared from a *Salinispora arenicola* to *S. tropica*, which is incongruent with the *Salinispora* species phylogeny. The diversity and distribution of *spt* BGCs across the bacterial tree of life has implications for bacterial communication as salinipostin shares key structural similarities to the known A-factor signaling molecule and a key homologous biosynthetic gene. Thus, this work proposed a new distribution and diversity of salinipostin-like potential signaling molecules across bacteria. Chapter 4 has been published in *Microbial Genomics* and is a reprint of the publication.

In Chapter 5, I sought to explore BGC evolutionary diversification by creating a new set of *Salinispora* genomes that were isolated on a different scale from our current collection of *Salinispora* genomes. New *Salinispora* strains were selectively cultured from sediment samples collected in a 1-meter by 1-meter quadrant plot in Fiji. Whole-genome sequencing of the 99 "microscale" strains from each of the 16 sub-quadrants revealed that we had isolated three species – *Salinispora arenicola, S. oceanensis,* and *S. pacifica*. Comparisons with the current "macroscale" *Salinispora* genomes from worldwide isolation sites revealed these new microscale strains covered most of the known *S. arenicola* diversity and showed no pattern of sub-quadrant strains being closer related to each other compared to strains from other sub-quadrants. I observed new BGC diversity in the microscale *Salinispora* and uncovered previously rare *Salinispora* BGCs in this new dataset. This contributes to our knowledge of *Salinispora* population and biosynthetic diversity from one sampled location in comparison to worldwide isolation efforts. Additionally,

we documented for the first time the putative presence of plasmids in the new microscale *Salinispora*. Plasmids could be the mechanism by which BGCs (and other genes) are exchanged between *Salinispora* and related bacteria, contributing to both biosynthetic and overall genetic diversity. Appendix 1 of Chapter 5 includes plasmid purification and visualization protocols that were developed for *Salinispora* as part of this work. This chapter is currently being written into a manuscript for publication.

Chapter 6 builds on the 99 "microscale" whole-genome sequenced *Salinispora* where I sought to characterize the evolutionary dynamics of lanthipeptide ribosomally synthesized and post-translationally modified peptides (RiPPs) BGCs. RiPPs are unique BGCs because they require a precursor peptide gene that encodes two separate peptide parts: the leader which enzymes recognize to install post-translational modifications; and the core which becomes the peptide product backbone. With the microscale genomes, there was a unique opportunity to compare the evolutionary diversity of lanthipeptide products on two spatial scales. By comparing the sequence similarity of the precursor peptides, I discovered that there is a very large potential *Salinispora* lanthipeptide chemical diversity that is unlike characterized RiPP lanthipeptide products. For each group of similar precursor peptides, there was varying diversity observed either in the leader or core part of the peptide in species-specific patterns, suggesting that there is selection for different types of lanthipeptide RiPPs in different species of *Salinispora*. This comprehensive description of the distribution and diversity of lanthipeptide BGCs and precursor peptides in *Salinispora* will facilitate future efforts to characterize the first RiPP product from the *Salinispora* genus. This chapter is currently being written into a manuscript for publication.

Building on my work from Chapter 4, Appendix A describes the characterization of the new manumycin-like pacificamide compound from *Salinispora pacifica*. It also describes the

report of a previously described compound, triacsin D, in the *Salinispora* genus for the first time. In collaboration with lead author Dr. Gabriel Castro-Falcón, I led the evolutionary analysis of the *pac* BGC across all bacteria. We discovered that the *Salinispora pac* BGC was found in only two *Salinispora* strains and that it had a unique gene organization compared to the top 30 homologous BGCs, including other known manumycin-like BGCs. We observed that BGCs with known manumycin-like molecules generally shared similar organization in a genus-specific pattern, yet there were no BGCs with similar organization to *pac*. Furthermore, the *pac*-like BGCs were only observed in the divergent *Streptomycetaceae, Micromonosporaceae,* and *Pseudonocardiacea* bacterial families, indicating that this BGC could have been horizontally exchanged between these taxa. From the comparison of the top *pac*-like BGCs, the biosynthesis of pacificamide was proposed. Appendix A has been published in the *Journal of Natural Products* and thus is a reprint of the publication.

Finally, Chapter 7 discusses the significant findings and general conclusions brought forth through this dissertation. Future directions related to this research are discussed.

1.6 References

Adamek, M., Alanjary, M., and Ziemert, N. (2019) Applied evolution: phylogeny-based approaches in natural products research. *Nat Prod Rep*.

Almabruk, K.H., Dinh, L.K., and Philmus, B. (2018) Self-Resistance of Natural Product Producers: Past, Present, and Future Focusing on Self-Resistant Protein Variants. *ACS Chem Biol* **13**: 1426–1437.

Almeida, A., Nayfach, S., Boland, M., Strozzi, F., Beracochea, M., Shi, Z.J., Pollard, K.S., Sakharova, E., Parks, D.H., Hugenholtz, P., Segata, N., Kyrpides, N.C., and Finn, R.D. (2021) A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nat Biotechnol* **39**: 105–114.

Amos, G.C.A., Awakawa, T., Tuttle, R.N., Letzel, A.-C., Kim, M.C., Kudo, Y., Fenical, W., Moore, B.S., and Jensen, P.R. (2017) Comparative Transcriptomics as a Guide to Natural Product Discovery and Biosynthetic Gene Cluster Functionality. *PNAS* **114**: E11121–E11130.

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

Azam, F., Fenchel, T., Field, J.G., Gray, J.S., Meyer-Rei, L.A., and Thingstad, F. (1983) The Ecological Role of Water-Column Microbes in the Sea. *Mar Ecol Prog Ser* **10**: 257–263.

Bauermeister, A., Velasco-Alzate, K., Dias, T., Macedo, H., Ferreira, E.G., Jimenez, P.C., Lotufo, T.M.C., Lopes, N.P., Gaudêncio, S.P., and Costa-Lotufo, L. V. (2018) Metabolomic Fingerprinting of Salinispora From Atlantic Oceanic Islands. *Front Microbiol* **9**: 1–13.

Bauman, K.D., Butler, K.S., Moore, B.S., and Chekan, J.R. (2021) Genome mining methods to discover bioactive natural products. *Nat Prod Rep*.

Bauman, K.D., Li, J., Murata, K., Mantovani, S.M., Dahesh, S., Nizet, V., Luhavaya, H., and Moore, B.S. (2019) Refactoring the Cryptic Streptophenazine Biosynthetic Gene Cluster Unites Phenazine, Polyketide, and Nonribosomal Peptide Biochemistry. *Cell Chem Biol* 1–13.

Bauman, K.D., Shende, V. V., Chen, P.Y.-T., Trivella, D.B.B., Gulder, T.A.M., Vellalath, S., Romo, D., and Moore, B.S. (2022) Enzymatic assembly of the salinosporamide γ-lactam-β-lactone anticancer warhead. *Nat Chem Biol*.

Brunson, J.K., McKinnie, S.M.K., Chekan, J.R., McCrow, J.P., Miles, Z.D., Bertrand, E.M., Bielinski, V.A., Luhavaya, H., Oborník, M., Smith, G.J., Hutchins, D.A., Allen, A.E., and Moore, B.S. (2018) Biosynthesis of the neurotoxin domoic acid in a bloom-forming diatom. *Science (80-)* **361**: 1356–1358.

Burkholder, P.R., Pfister, R.M., and Leitz, F.H. (1966) Production of a pyrrole antibiotic by a marine bacterium. *Appl Microbiol* **14**: 649–653.

Cavicchioli, R., Ripple, W.J., Timmis, K.N., Azam, F., Bakken, L.R., Baylis, M., Behrenfeld, M.J., Boetius, A., Boyd, P.W., Classen, A.T., Crowther, T.W., Danovaro, R., Foreman, C.M., Huisman, J., Hutchins, D.A., Jansson, J.K., Karl, D.M., Koskella, B., Mark Welch, D.B.,

Martiny, J.B.H., Moran, M.A., Orphan, V.J., Reay, D.S., Remais, J. V., Rich, V.I., Singh, B.K., Stein, L.Y., Stewart, F.J., Sullivan, M.B., van Oppen, M.J.H., Weaver, S.C., Webb, E.A., and Webster, N.S. (2019) Scientists' warning to humanity: microorganisms and climate change. *Nat Rev Microbiol* **17**: 569–586.

Chevrette, M.G. and Currie, C.R. (2019) Emerging evolutionary paradigms in antibiotic discovery. *J Ind Microbiol Biotechnol* **46**: 257–271.

Chevrette, M.G., Gutiérrez-García, K., Selem-Mojica, N., Aguilar-Martínez, C., Yañez-Olvera, A., Ramos-Aboites, H.E., Hoskisson, P.A., and Barona-Gómez, F. (2020) Evolutionary dynamics of natural product biosynthesis in bacteria. *Nat Prod Rep* **37**: 566–599.

Chevrette, M.G. and Handelsman, J. (2021) Needles in haystacks: Reevaluating old paradigms for the discovery of bacterial secondary metabolites. *Nat Prod Rep* **38**: 2083–2099.

Demko, A.M., Patin, N. V., and Jensena, P.R. (2021) Microbial diversity in tropical marine sediments assessed using culture-dependent and culture-independent techniques. *Environ Microbiol* 1–50.

Dybdal-Hargreaves, N.F., Risinger, A.L., and Mooberry, S.L. (2015) Eribulin Mesylate: Mechanism of action of a unique microtubule-Targeting agent. *Clin Cancer Res* **21**: 2445–2452.

Eustáquio, A.S., McGlinchey, R.P., Liu, Y., Hazzard, C., Beer, L.L., Florova, G., Alhamadsheh, M.M., Lechner, A., Kale, A.J., Kobayashi, Y., Reynolds, K. a, and Moore, B.S. (2009) Biosynthesis of the salinosporamide A polyketide synthase substrate chloroethylmalonyl-coenzyme A from *S*-adenosyl-L-methionine. *Proc Natl Acad Sci U S A* **106**: 12295–300.

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

Fischbach, M.A., Walsh, C.T., and Clardy, J. (2008) The evolution of gene collectives: How natural selection drives chemical innovation. *PNAS* **105**: 4601–4608.

Flemming, H.-C. and Wuertz, S. (2019) Bacteria and archaea on Earth and their abundance in biofilms. *Nat Rev Microbiol*.

Gil-Turnes, M.S., Hay, M.E., and Fenical, W. (1989) Symbiotic marine bacteria chemically defend crustacean embryos from a pathogenic fungus. *Science (80-)* **246**: 116–118.

Heinrich, M. (2000) Ethnobotany and its role in drug development. *Phyther Res* **14**: 479–488.

Hernandez, A., Nguyen, L.T., Dhakal, R., and Murphy, B.T. (2021) The need to innovate sample collection and library generation in microbial drug discovery: A focus on academia. *Nat Prod Rep* **38**: 292–300.

Hertweck, C. (2009) The biosynthetic logic of polyketide diversity. *Angew Chemie - Int Ed* **48**:

4688–4716.

Hirata, Y. and Uemura, D. (1986) Halichondrins—antitumor polyether macrolides from a marine sponge. *Pure Appl Chem* **58**: 701–710.

van der Hooft, J.J.J., Mohimani, H., Bauermeister, A., Dorrestein, P.C., Duncan, K.R., and Medema, M.H. (2020) Linking genomics and metabolomics to chart specialized metabolic diversity. *Chem Soc Rev* **49**: 3297–3314.

Hoogendoorn, K., Barra, L., Waalwijk, C., Dickschat, J.S., van der Lee, T.A.J., and Medema, M.H. (2018) Evolution and diversity of biosynthetic gene clusters in Fusarium. *Front Microbiol* **9**: 1–12.

Hosseini, S.R. and Wagner, A. (2018) Genomic organization underlying deletional robustness in bacterial metabolic systems. *Proc Natl Acad Sci U S A* **115**: 7075–7080.

Hug, J.J., Krug, D., and Müller, R. (2020) Bacteria as genetically programmable producers of bioactive natural products. *Nat Rev Chem* **4**:.

Hug, L.A., Baker, B.J., Anantharaman, K., Brown, C.T., Probst, A.J., Castelle, C.J., Butterfield, C.N., Hernsdorf, A.W., Amano, Y., Ise, K., Suzuki, Y., Dudek, N., Relman, D.A., Finstad, K.M., Amundson, R., Thomas, B.C., and Banfield, J.F. (2016) A new view of the tree of life. *Nat Microbiol* **1**: 1–6.

Jensen, P.R. (2016) Natural Products and the Gene Cluster Revolution. *Trends Microbiol* **24**: 968–977.

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

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

Jezkova, Z., Schulzova, V., Krizova, I., Karabin, M., and Branyik, T. (2021) Influence of Cultivation Conditions on the Sioxanthin Content and Antioxidative Protection Effect of a Crude Extract from the Vegetative Mycelium of Salinispora tropica. *Mar Drugs* **19**:.

Kau, A.L., Ahern, P.P., Griffin, N.W., Goodman, A.L., and Gordon, J.I. (2011) Human nutrition, the gut microbiome and the immune system. *Nature* **474**: 327–336.

Keller, N.P. (2019) Fungal secondary metabolism: regulation, function and drug discovery. *Nat Rev Microbiol* **17**: 167–180.

Kim, T.K., Garson, M.J., and Fuerst, J.A. (2005) Marine actinomycetes related to the '*Salinospora*' group from the Great Barrier Reef sponge *Pseudoceratina clavata*. *Environ Microbiol* **7**: 509–518.

Kominek, J., Doering, D.T., Opulente, D.A., Shen, X.-X., Zhou, X., DeVirgilio, J., Hulfachor, A.B., Kurtzman, C.P., Rokas, A., and Hittinger, C.T. (2019) Eukaryotic acquisition of a bacterial operon. *Cell* **176**: 1356–1366.

Kuhl, M., Rückert, C., Gläser, L., Beganovic, S., Luzhetskyy, A., Kalinowski, J., and Wittmann, C. (2021) Microparticles enhance the formation of seven major classes of natural products in native and metabolically engineered actinobacteria through accelerated morphological development.

de la Torre, B.G. and Albericio, F. (2022) The Pharmaceutical Industry in 2021. An Analysis of FDA Drug Approvals from the Perspective of Molecules. *Molecules* **27**:.

Launer, J. (2013) Meet your microbiome. *Postgrad Med J* **89**: 367–8.

Le, T. and van der Donk, W.A. (2021) Mechanisms and Evolution of Diversity-Generating RiPP Biosynthesis. *Trends Chem* **3**: 266–278.

Leber, C.A., Reyes, A.J., Biggs, J.S., and Gerwick, W.H. (2021) Cyanobacteria-shrimp colonies in the Mariana Islands. *Aquat Ecol* **55**: 453–465.

Lind, A.L., Wisecaver, J.H., Lameiras, C., Wiemann, P., Palmer, J.M., Keller, N.P., Rodrigues, F., Goldman, G.H., and Rokas, A. (2017) Drivers of genetic diversity in secondary metabolic gene clusters within a fungal species. *PLoS Biol* **15**: 1–26.

Lloyd, K.G., Steen, A.D., Ladau, J., Yin, J., and Crosby, L. (2018) Phylogenetically Novel Uncultured Microbial Cells Dominate Earth Microbiomes. *mSystems* **3**: 1–12.

Loureiro, C., Medema, M.H., van der Oost, J., and Sipkema, D. (2018) Exploration and exploitation of the environment for novel specialized metabolites. *Curr Opin Biotechnol* **50**: 206–213.

Machado, H., Tuttle, R.N., and Jensen, P.R. (2017) Omics-based natural product discovery and the lexicon of genome mining. *Curr Opin Microbiol* **39**: 136–142.

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

Matsuda, S., Adachi, K., Matsuo, Y., Nukina, M., and Shizuri, Y. (2009) Salinisporamycin, a novel metabolite from salinispora arenicora. *J Antibiot (Tokyo)* **62**: 519–526.

McCaughey, C.S., van Santen, J.A., van der Hooft, J.J.J., Medema, M.H., and Linington, R.G. (2021) An isotopic labeling approach linking natural products with biosynthetic gene clusters. *Nat Chem Biol*.

McGivern, J.G. (2007) Ziconotide: A review of its pharmacology and use in the treatment of pain. *Neuropsychiatr Dis Treat* **3**: 69–85.

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., Düsterhus, 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.N., Hillwig, M.L., Ishida, K., Jones, A.C., Jones, C.S., Jungmann, K., Kegler, C., Kim, H.U., Kötter, P., Krug, D., Masschelein, J., Melnik, A. V, Mantovani, S.M., Monroe, E.A., Moore, M., Moss, N., Nützmann, 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-Gómez, 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.M., Ehling-Schulz, M., Entian, K.-D., Fischbach, M.A., Gerwick, L., Gerwick, W.H., Gross, H., Gust, B., Hertweck, C., Höfte, 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., Méndez, C., Metsä-Ketelä, M., Micklefield, J., Mitchell, D.A., Moore, B.S., Moreira, L.M., Müller, 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., Süssmuth, 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., and Glöckner, F.O. (2015) The Minimum Information about a Biosynthetic Gene cluster (MIBiG) specification. *Nat Chem Biol* **11**: 625–631.

Medema, M.H., de Rond, T., and Moore, B.S. (2021) Mining genomes to illuminate the specialized chemistry of life. *Nat Rev Genet*.

Millán-Aguiñaga, N., Chavarria, K.L., Ugalde, J.A., Letzel, A.-C., Rouse, G.W., and Jensen, P.R. (2017) Phylogenomic Insight into *Salinispora* (Bacteria, Actinobacteria) Species Designations. *Sci Rep* **7**: 3564.

Mincer, T.J., Fenical, W., and Jensen, P.R. (2005) Culture-dependent and culture-independent diversity within the obligate marine actinomycete genus *Salinispora*. *Appl Environ Microbiol* **71**: 7019–7028.

Mincer, T.J., Jensen, P.R., Kauffman, C.A., and Fenical, W. (2002) Widespread and Persistent Populations of a Major New Marine Actinomycete Taxon in Ocean Sediments. *Appl Environ Microbiol* **68**: 5005–5011.

De Mol, M.L., Snoeck, N., De Maeseneire, S.L., and Soetaert, W.K. (2018) Hidden antibiotics: Where to uncover? *Biotechnol Adv* **36**: 2201–2218.

Montalbán-López, M., Scott, T.A., Ramesh, S., Rahman, I.R., Van Heel, A.J., Viel, J.H., Bandarian, V., Dittmann, E., Genilloud, O., Goto, Y., Grande Burgos, M.J., Hill, C., Kim, S., Koehnke, J., Latham, J.A., Link, A.J., Martínez, B., Nair, S.K., Nicolet, Y., Rebuffat, S., Sahl,

H.G., Sareen, D., Schmidt, E.W., Schmitt, L., Severinov, K., Süssmuth, R.D., Truman, A.W., Wang, H., Weng, J.K., Van Wezel, G.P., Zhang, Q., Zhong, J., Piel, J., Mitchell, D.A., Kuipers, O.P., and Van Der Donk, W.A. (2021) New developments in RiPP discovery, enzymology and engineering. *Nat Prod Rep* **38**: 130–239.

Moon, K., Xu, F., Zhang, C., and Seyedsayamdost, M.R. (2019) Bioactivity-HiTES Unveils Cryptic Antibiotics Encoded in Actinomycete Bacteria. *ACS Chem Biol* **14**: 767–774.

Nai, C. and Meyer, V. (2017) From Axenic to Mixed Cultures: Technological Advances Accelerating a Paradigm Shift in Microbiology. *Trends Microbiol* 1–17.

Nayfach, S., Roux, S., Seshadri, R., Udwary, D., Varghese, N., Schulz, F., Wu, D., Paez-Espino, D., Chen, I.M., Huntemann, M., Palaniappan, K., Ladau, J., Mukherjee, S., Reddy, T.B.K., Nielsen, T., Kirton, E., Faria, J.P., Edirisinghe, J.N., Henry, C.S., Jungbluth, S.P., Chivian, D., Dehal, P., Wood-Charlson, E.M., Arkin, A.P., Tringe, S.G., Visel, A., Abreu, H., Acinas, S.G., Allen, E., Allen, M.A., Andersen, G., Anesio, A.M., Attwood, G., Avila-Magaña, V., Badis, Y., Bailey, J., Baker, B., Baldrian, P., Barton, H.A., Beck, D.A.C., Becraft, E.D., Beller, H.R., Beman, J.M., Bernier-Latmani, R., Berry, T.D., Bertagnolli, A., Bertilsson, S., Bhatnagar, J.M., Bird, J.T., Blumer-Schuette, S.E., Bohannan, B., Borton, M.A., Brady, A., Brawley, S.H., Brodie, J., Brown, S., Brum, J.R., Brune, A., Bryant, D.A., Buchan, A., Buckley, D.H., Buongiorno, J., Cadillo-Quiroz, H., Caffrey, S.M., Campbell, A.N., Campbell, B., Carr, S., Carroll, J.L., Cary, S.C., Cates, A.M., Cattolico, R.A., Cavicchioli, R., Chistoserdova, L., Coleman, M.L., Constant, P., Conway, J.M., Mac Cormack, W.P., Crowe, S., Crump, B., Currie, C., Daly, R., Denef, V., Denman, S.E., Desta, A., Dionisi, H., Dodsworth, J., Dombrowski, N., Donohue, T., Dopson, M., Driscoll, T., Dunfield, P., Dupont, C.L., Dynarski, K.A., Edgcomb, V., Edwards, E.A., Elshahed, M.S., Figueroa, I., Flood, B., Fortney, N., Fortunato, C.S., Francis, C., Gachon, C.M.M., Garcia, S.L., Gazitua, M.C., Gentry, T., Gerwick, L., Gharechahi, J., Girguis, P., Gladden, J., Gradoville, M., Grasby, S.E., Gravuer, K., Grettenberger, C.L., Gruninger, R.J., Guo, J., Habteselassie, M.Y., Hallam, S.J., Hatzenpichler, R., Hausmann, B., Hazen, T.C., Hedlund, B., Henny, C., Herfort, L., Hernandez, M., Hershey, O.S., Hess, M., Hollister, E.B., Hug, L.A., Hunt, D., Jansson, J., Jarett, J., Kadnikov, V. V., Kelly, C., Kelly, R., Kelly, W., Kerfeld, C.A., Kimbrel, J., Klassen, J.L., Konstantinidis, K.T., Lee, L.L., Li, W.J., Loder, A.J., Loy, A., Lozada, M., MacGregor, B., Magnabosco, C., Maria da Silva, A., McKay, R.M., McMahon, K., McSweeney, C.S., Medina, M., Meredith, L., Mizzi, J., Mock, T., Momper, L., Moran, M.A., Morgan-Lang, C., Moser, D., Muyzer, G., Myrold, D., Nash, M., Nesbø, C.L., Neumann, A.P., Neumann, R.B., Noguera, D., Northen, T., Norton, J., Nowinski, B., Nüsslein, K., O'Malley, M.A., Oliveira, R.S., Maia de Oliveira, V., Onstott, T., Osvatic, J., Ouyang, Y., Pachiadaki, M., Parnell, J., Partida-Martinez, L.P., Peay, K.G., Pelletier, D., Peng, X., Pester, M., Pett-Ridge, J., Peura, S., Pjevac, P., Plominsky, A.M., Poehlein, A., Pope, P.B., Ravin, N., Redmond, M.C., Reiss, R., Rich, V., Rinke, C., Rodrigues, J.L.M., Rossmassler, K., Sackett, J., Salekdeh, G.H., Saleska, S., Scarborough, M., Schachtman, D., Schadt, C.W., Schrenk, M., Sczyrba, A., Sengupta, A., Setubal, J.C., Shade, A., Sharp, C., Sherman, D.H., Shubenkova, O. V., Sierra-Garcia, I.N., Simister, R., Simon, H., Sjöling, S., Slonczewski, J., Correa de Souza, R.S., Spear, J.R., Stegen, J.C., Stepanauskas, R., Stewart, F., Suen, G., Sullivan, M., Sumner, D., Swan, B.K., Swingley, W., Tarn, J., Taylor, G.T., Teeling, H., Tekere, M., Teske, A., Thomas, T., Thrash, C., Tiedje, J., Ting, C.S., Tully, B., Tyson, G.,

Ulloa, O., Valentine, D.L., Van Goethem, M.W., VanderGheynst, J., Verbeke, T.J., Vollmers, J., Vuillemin, A., Waldo, N.B., Walsh, D.A., Weimer, B.C., Whitman, T., van der Wielen, P., Wilkins, M., Williams, T.J., Woodcroft, B., Woolet, J., Wrighton, K., Ye, J., Young, E.B., Youssef, N.H., Yu, F.B., Zemskaya, T.I., Ziels, R., Woyke, T., Mouncey, N.J., Ivanova, N.N., Kyrpides, N.C., and Eloe-Fadrosh, E.A. (2020) A genomic catalog of Earth's microbiomes. *Nat Biotechnol*.

Netzker, T., Fischer, J., Weber, J., Mattern, D.J., Konig, C.C., Valiante, V., Schroeckh, V., and Brakhage, A.A. (2015) Microbial communication leading to the activation of silent fungal secondary metabolite gene clusters. *Front Microbiol* **6**: 1–13.

Nivina, A., Yuet, K.P., Hsu, J., and Khosla, C. (2019) Evolution and Diversity of Assembly-Line Polyketide Synthases. *Chem Rev* **119**: 12524–12547.

Nützmann, H.W., Scazzocchio, C., and Osbourn, A. (2018) Metabolic gene clusters in eukaryotes. *Annu Rev Genet* **52**: 159–183.

Okada, B.K. and Seyedsayamdost, M.R. (2017) Antibiotic dialogues: induction of silent biosynthetic gene clusters by exogenous small molecules. *FEMS Microbiol Rev* **41**: 19–33.

Olivera, B.M., Gray, W.R., Zeikus, R., Mcintosh, J.M., Rivier, J., Santos, V. De, and Cruz, L.J. (1985) Peptide Neurotoxins from Fish-Hunting Cone Snails. *Science (80-)* **230**: 1338–1343.

Olivera, B.M., Rivier, J., Clark, C., Ramilo, C.A., Corpuz, G.P., Abogadie, F.C., Mena, E.E., Woodward, S.R., Hillyard, D.R., and Cruz, L.J. (1990) Diversity of Conus Neuropeptides. *Science (80)* **249**: 257–263.

Onaka, H. (2017) Novel antibiotic screening methods to awaken silent or cryptic secondary metabolic pathways in actinomycetes. *Nat Publ Gr* **70**: 865–870.

Osbourn, A. (2010) Secondary metabolic gene clusters: Evolutionary toolkits for chemical innovation. *Trends Genet* **26**: 449–457.

Pál, C. and Hurst, L.D. (2004) Evidence against the selfish operon theory. *Trends Genet* **20**: 232–234.

Patel, A., Noble, R.T., Steele, J.A., Schwalbach, M.S., Hewson, I., and Fuhrman, J.A. (2007) Virus and prokaryote enumeration from planktonic aquatic environments by epifluorescence microscopy with SYBR Green I. *Nat Protoc* **2**: 269–276.

Patin, N., Floros, D., Hughes, C.C., Dorrestein, P., and Jensen, P. (2018) The role of inter-species interactions in Salinispora specialized metabolism. *Submitted* 1–10.

Patin, N. V, Duncan, K.R., Dorrestein, P.C., and Jensen, P.R. (2015) Competitive strategies differentiate closely related species of marine actinobacteria. *ISME J* 1–13.

Probandt, D., Eickhorst, T., Ellrott, A., Amann, R., and Knittel, K. (2018) Microbial life on a sand grain: from bulk sediment to single grains. *ISME J* **12**: 623–633.

Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K.S., Manichanh, C., Nielsen, T., Pons, N., Levenez, F., Yamada, T., Mende, D.R., Li, J., Xu, J., Li, Shaochuan, Li, D., Cao, J., Wang, B., Liang, H., Zheng, H., Xie, Y., Tap, J., Lepage, P., Bertalan, M., Batto, J.-M., Hansen, T., Le Paslier, D., Linneberg, A., Nielsen, H.B., Pelletier, E., Renault, P., Sicheritz-Ponten, T., Turner, K., Zhu, H., Yu, C., Li, Shengting, Jian, M., Zhou, Y., Li, Y., Zhang, X., Li, Songgang, Qin, N., Yang, H., Wang, Jian, Brunak, S., Doré, J., Guarner, F., Kristiansen, K., Pedersen, O., Parkhill, J., Weissenbach, J., Antolin, M., Artiguenave, F., Blottiere, H., Borruel, N., Bruls, T., Casellas, F., Chervaux, C., Cultrone, A., Delorme, C., Denariaz, G., Dervyn, R., Forte, M., Friss, C., van de Guchte, M., Guedon, E., Haimet, F., Jamet, A., Juste, C., Kaci, G., Kleerebezem, M., Knol, J., Kristensen, M., Layec, S., Le Roux, K., Leclerc, M., Maguin, E., Melo Minardi, R., Oozeer, R., Rescigno, M., Sanchez, N., Tims, S., Torrejon, T., Varela, E., de Vos, W., Winogradsky, Y., Zoetendal, E., Bork, P., Ehrlich, S.D., and Wang, Jun (2010) A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464**: 59–65.

de Reus, E., Nielsen, M.R., and Frandsen, R.J.N. (2019) Metabolic and regulatory insights from the experimental Horizontal Gene Transfer of the aurofusarin and bikaverin gene clusters to *Aspergillus nidulans*. *Mol Microbiol* mmi.14376.

Román-Ponce, B., Millán-Aguiñaga, N., Guillen-Matus, D., Chase, A.B., Ginigini, J.G.M., Soapi, K., Feussner, K.D., Jensen, P.R., and Trujillo, M.E. (2020) Six novel species of the obligate marine actinobacterium *Salinispora*, *Salinispora cortesiana* sp. nov., *Salinispora fenicalii* sp. nov., *Salinispora goodfellowii* sp. nov., *Salinispora mooreana* sp. nov., ... *Int J Syst Evol Microbiol* **70**: 4668–4682.

Romano, S., Jackson, S., Patry, S., and Dobson, A. (2018) Extending the "One Strain Many Compounds" (OSMAC) Principle to Marine Microorganisms. *Mar Drugs* **16**: 244.

Rosenberg, E. and Zilber-Rosenberg, I. (2016) Microbes Drive Evolution of Animals and Plants: the Hologenome Concept. *MBio* **7**: e01395-15-.

Roth, J.R. and Lawrence, J.G. (1996) Selfish operons: horizontal transfer may drive the evolution of gene clusters. *Genetics* **143**: 1843–60.

Ruzzini, A.C. and Clardy, J. (2016) Gene Flow and Molecular Innovation in Bacteria. *Curr Biol* **26**: R859–R864.

Schlawis, C., Kern, S., Kudo, Y., Grunenberg, J., Moore, B., and Schulz, S. (2018) Structural Elucidation of Trace Components Combining GC/MS, GC/IR, DFT-Calculation and Synthesis - Salinilactones, Unprecedented Bicyclic Lactones from *Salinispora* Bacteria. *Angew Chemie Int Ed* **57**: 14921–14925.

Schorn, M.A., Verhoeven, S., Ridder, L., Huber, F., Acharya, D.D., Aksenov, A.A., Aleti, G., Moghaddam, J.A., Aron, A.T., Aziz, S., Bauermeister, A., Bauman, K.D., Baunach, M., Beemelmanns, C., Beman, J.M., Berlanga-Clavero, M.V., Blacutt, A.A., Bode, H.B., Boullie, A., Brejnrod, A., Bugni, T.S., Calteau, A., Cao, L., Carrión, V.J., Castelo-Branco, R., Chanana, S., Chase, A.B., Chevrette, M.G., Costa-Lotufo, L. V., Crawford, J.M., Currie, C.R., Cuypers, B., Dang, T., de Rond, T., Demko, A.M., Dittmann, E., Du, C., Drozd, C.,

Dujardin, J.C., Dutton, R.J., Edlund, A., Fewer, D.P., Garg, N., Gauglitz, J.M., Gentry, E.C., Gerwick, L., Glukhov, E., Gross, H., Gugger, M., Guillén Matus, D.G., Helfrich, E.J.N., Hempel, B.F., Hur, J.S., Iorio, M., Jensen, P.R., Kang, K. Bin, Kaysser, L., Kelleher, N.L., Kim, C.S., Kim, K.H., Koester, I., König, G.M., Leao, T., Lee, S.R., Lee, Y.Y., Li, X., Little, J.C., Maloney, K.N., Männle, D., Martin H, C., McAvoy, A.C., Metcalf, W.W., Mohimani, H., Molina-Santiago, C., Moore, B.S., Mullowney, M.W., Muskat, M., Nothias, L.F., O'Neill, E.C., Parkinson, E.I., Petras, D., Piel, J., Pierce, E.C., Pires, K., Reher, R., Romero, D., Roper, M.C., Rust, M., Saad, H., Saenz, C., Sanchez, L.M., Sørensen, S.J., Sosio, M., Süssmuth, R.D., Sweeney, D., Tahlan, K., Thomson, R.J., Tobias, N.J., Trindade-Silva, A.E., van Wezel, G.P., Wang, M., Weldon, K.C., Zhang, F., Ziemert, N., Duncan, K.R., Crüsemann, M., Rogers, S., Dorrestein, P.C., Medema, M.H., and van der Hooft, J.J.J. (2021) A community resource for paired genomic and metabolomic data mining. *Nat Chem Biol* **17**: 363–368.

Schulz, S., Biwer, P., Harig, T., Koteska, D., and Schlawis, C. (2020) Chemical Ecology of Bacterial Volatiles, 3rd ed. Elsevier Ltd.

Schulze, C.J., Navarro, G., Ebert, D., DeRisi, J., and Linington, R.G. (2015) Salinipostins A-K, long-chain bicyclic phosphotriesters as a potent and selective antimalarial chemotype. *J Org Chem* **80**: 1312–1320.

Schwartsmann, G., Da Rocha, A.B., Berlinck, R.G.S., and Jimeno, J. (2001) Marine organisms as a source of new anticancer agents. *Lancet Oncol* **2**: 221–225.

Sikandar, A. and Koehnke, J. (2019) The role of protein-protein interactions in the biosynthesis of ribosomally synthesized and post-translationally modified peptides. *Nat Prod Rep* **36**: 1576–1588.

Sneed, J.M., Sharp, K.H., Ritchie, K.B., and Paul, V.J. (2014) The chemical cue tetrabromopyrrole from a biofilm bacterium induces settlement of multiple Caribbean corals. *Proc R Soc B Biol Sci* **281**: 1–9.

Soldatou, S., Eldjárn, G.H., Ramsay, A., van der Hooft, J.J.J., Hughes, A.H., Rogers, S., and Duncan, K.R. (2021) Comparative Metabologenomics Analysis of Polar Actinomycetes. *Mar Drugs* **19**: 1–21.

Soucy, S.M., Huang, J., and Gogarten, J.P. (2015) Horizontal gene transfer: building the web of life. *Nat Rev Genet* **16**: 472–482.

Stefanoudis, P. V., Licuanan, W.Y., Morrison, T.H., Talma, S., Veitayaki, J., and Woodall, L.C. (2021) Turning the tide of parachute science. *Curr Biol* **31**: R184–R185.

Sung, A., Gromek, S., and Balunas, M. (2017) Upregulation and Identification of Antibiotic Activity of a Marine-Derived Streptomyces sp. via Co-Cultures with Human Pathogens. *Mar Drugs* **15**: 250.

Ting, C.P., Funk, M.A., Halaby, S.L., Zhang, Z., Gonen, T., and van der Donk, W.A. (2019) Use of a scaffold peptide in the biosynthesis of amino acid–derived natural products. *Science (80)* **365**: 280–284.

Traxler, M.F., Watrous, J.D., Alexandrov, T., Dorrestein, P.C., and Kolter, R. (2013) Interspecies interactions stimulate diversification of the Streptomyces coelicolor secreted metabolome. *MBio* **4**: 1–12.

Tuttle, R.N., Rouse, G.W., Castro-Falcón, G., Hughes, C.C., and Jensen, P.R. (2022) Specialized Metabolite-Mediated Predation Defense in the Marine Actinobacterium Salinispora . *Appl Environ Microbiol* **88**:.

Ulanova, D., Uenaka, Y., Sakama, M., and Sakurai, T. (2020) Draft Genome Sequence of Salinispora sp. Strain H7-4, Isolated from Deep-Sea Sediments of the Shikoku Basin . *Microbiol Resour Announc* **9**: 7–9.

Ulvestad, E. (2009) Cooperation and conflict in host-microbe relations. *Apmis* **117**: 311–322.

Veeresham, C. (2012) Natural products derived from plants as a source of drugs. *J Adv Pharm Technol Res* **3**: 200–201.

Vidgen, M.E., Hooper, J.N.A., and Fuerst, J.A. (2012) Diversity and distribution of the bioactive actinobacterial genus *Salinispora* from sponges along the Great Barrier Reef. *Antonie Van Leeuwenhoek* **101**: 603–618.

Vierros, M., Suttle, C.A., Harden-Davies, H., and Burton, G. (2016) Who Owns the Ocean? Policy Issues Surrounding Marine Genetic Resources. *Limnol Oceanogr Bull* **25**: 29–35.

Voser, T.M., Campbell, M.D., and Carroll, A.R. (2021) How different are marine microbial natural products compared to their terrestrial counterparts? *Nat Prod Rep*.

Williams, D.E., Morgan, K.D., Dalisay, D.S., Matainaho, T., Perrachon, E., Viller, N., Delcroix, M., Gauchot, J., Niikura, H., Patrick, B.O., Ryan, K.S., and Andersen, R.J. (2022) Natural Products Produced in Culture by Biosynthetically Talented Salinispora arenicola Strains Isolated from Northeastern and South Pacific Marine Sediments. *Molecules* **27**:.

Wisecaver, J.H. and Rokas, A. (2015) Fungal metabolic gene clusters-caravans traveling across genomes and environments. *Front Microbiol* **6**: 1–11.

Zhang, J.J., Moore, B.S., Tang, X., and Moore, B.S. (2018) Engineering Salinispora tropica for heterologous expression of natural product biosynthetic gene clusters.

Zhang, J.J., Tang, X., and Moore, B.S. (2019) Genetic platforms for heterologous expression of microbial natural products. *Nat Prod Rep*.

CHAPTER 2. The Natural Product Domain Seeker version 2 (NaPDoS2): Relating Ketosynthase Phylogeny to Biosynthetic Function

2.1 Context of the Project: Introduction to Chapter 2

In today's exciting world where genome sequencing has become more accessible and affordable, new tools that can rise to the challenge of identifying specialized metabolite biosynthetic potential in genomic data are needed. Genome sequencing technology has evolved from limited throughput sequencing methods to high-throughput next-generation sequencing technologies including Illumina short-reads, PacBio and Oxford Nanopore single-molecule real-time long reads, and 10X genomics synthetic long reads (Goodwin *et al.*, 2016). Coupled with advanced genome-sequencing technologies, computational tools have also advanced to transform raw sequencing reads into complete genomes, among many "omics" based experimental applications. While genome sequencing has become routine for smaller bacterial genomes, advanced next-generation sequencing technology also allows fungal, animal, plant, protist and even the complete human genome to be sequenced and assembled (Nurk *et al.*, 2022). These advanced sequencing technologies have revolutionized the options of experimental approaches and applications. For discovery of new natural product biosynthetic potential, we need tools that allow us to discern promising specialized metabolite leads from microbial, fungal, animal, and environmental sequencing data.

There have been many fantastic bioinformatic tools that have helped researchers to assess specialized metabolite biosynthetic potential in genome sequences. Popular *in silico* genome mining tools such as antiSMASH 6.0 (Blin *et al.*, 2021), PRISM 4 (Skinnider *et al.*, 2020), and DeepBGC (Hannigan *et al.*, 2019) identify many classes of BGCs, whereas other tools such as PKMiner (Kim and Yi, 2012), RiPP-Miner (Agrawal *et al.*, 2021) and transATor (Helfrich *et al.*, 2019) specifically predict and identify targeted classes of natural products. All of these tools are

essential components of the modern-day natural product scientist's toolbox as genomic predictions have been used to inform biosynthetic proposals (Kim *et al.*, 2018), construct enzymatic phylogenetic predictions (Gerlt, 2017; Robinson *et al.*, 2021), uncover new diversity from known genetic scaffolds (Yan *et al.*, 2016), and even reverse-engineer the functions of gene clusters in "retrobiosynthetic" approaches, using chemical scaffolds to search for relevant biosynthetic genes in producing organisms (Bauman *et al.*, 2021). In the future, the natural products field and the ongoing search for new chemical diversity will continue to rely on sequencing data, so tools to help prioritize biosynthetic leads are needed.

In 2012, the webtool NaPDoS—which stands for "Natural Product Domain Seeker" was released (Ziemert *et al.*, 2012). The NaPDoS webtool (napdos.ucsd.edu) detects and classifies ketosynthase (KS) and condensation (C) domains from polyketide synthase (PKS) and non-ribosomal peptide synthetase (NRPS) BGCs, respectively. NaPDoS uses a phylogeny-based classification scheme that enables functional predictions about the encoding PKS and NRPS BGCs that domains are part of. However, 10 years has passed since the tool's initial release and thus the NaPDoS webtool is due for an update to keep pace with the new functional diversity that has been discovered. Of particular necessity, NaPDoS needed to be updated so that its application and use could be expanded—including: adding newly discovered classes of polyketides and non-ribosomal peptides, thus allowing wider potential for discovery; and secondly, to expand the tool so that genomic, metagenomic, and targeted amplicon sequencing data could be analyzed. The goals for the NaPDoS2 update included: 1) update the database with broader taxonomic breadth of KS domains; 2) develop a new KS classification scheme that is detailed and specific, relevant to specific polyketide structural classes; 3) increase the performance of the webtool thus allowing large datasets to be analyzed; 4) assess the capability of the tool by statistical benchmarking

analyses; 5) use the tool to analyze a wide variety of biological datasets to ascertain from a user's perspective the utility and application of the webtool; and 6) update the web text on the website, create clear usage directions, provide a detailed users guide, and provide data and interpretation guidelines so users at any experience level and with any type of sequencing data can utilize the webtool.

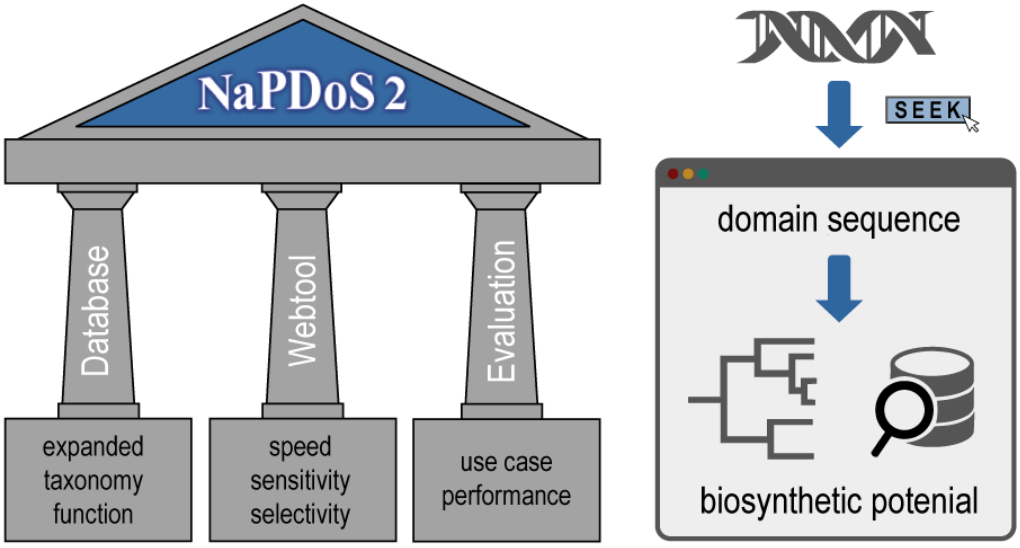


Figure 2.1. Graphical abstract of NaPDoS2.

Left: The three main pillars of the NaPDoS2 webtool update include database expansion, webtool upgrades, and application use case evaluations. Right: A stylized schematic of how NaPDoS2 works: sequencing data, depicted as DNA, is input into NaPDoS2, and domain sequences and predicted biosynthetic potential is the output.

We achieved these goals as a result of a collaboration with the three main co-first authors (Dr. Leesa J Klau, Dr. Sheila Podell, and myself) leading key pillars of the project (**Figure 2.1**). Dr. Klau led the creation of the expanded update database covering new taxonomy and polyketide synthase KS class/subclass function. Dr. Podell led the development of the webtool including the back-end updates and implemented updates to improve speed, sensitivity, and selectivity of the

tool. For this project, the role I played evolved over time to become the third pillar of work that was essential to the updated NaPDoS2 release. First, I led the development of a benchmarking method for the tool, including the development of a negative control dataset. To the best of our knowledge, a negative control dataset like the one I created has not been implemented for other *in silico* genome-mining tools. This was important because one of the biggest questions plaguing NaPDoS (and genome-mining tools like NaPDoS) is understanding how well they work. In the past, tools were assessed by seeing how well they could find known specialized metabolite BGCs, but ideally, we want to be able to know how well the tool works to finds new things that are true positive hits. We wanted to minimize the number of false positives (hits that aren't types of enzymes/classes that we are searching for), while also avoiding false negatives (domains that are true hits but aren't recognized because they are too distant from what we know). To do this, we needed to be sure that NaPDoS2 will not falsely identify gene clusters that we know are not involved in the biosynthesis of our specific targeted classes, while maintaining a good accuracy of finding true hits and filtering out true negatives (non-hits). Along these lines, I took the approach that ketosynthase enzymes belong to the larger "condensing enzymes" superfamily and thus share structural similarities yet catalyze distinct reactions. I realized that the ketosynthase domains that NaPDoS2 identifies from type I and type II polyketide and fatty acid synthases fall into distinct clades and share sister clades of unrelated domains like KSs from type III PKSs. This was the perfect opportunity to use these closely related yet functionally different domains as negative controls, and additionally populate our positive control dataset with additional type I and II PKS KS domains, along with additional KS domains from the NaPDoS2 database. For the first time, we were able to perform ROC analyses along with other statistics to calculate the sensitivity, selectivity, and specificity of the tool and showing how NaPDoS2 outperforms version 1 of the

webtool in speed and accuracy. These advances included a lot of benchmarking based on phylogenetic analyses of the condensing enzyme superfamily and other datasets. This combination of structure-function and phylogeny informed analyses was a very powerful method to ground truth the webtool.

Next, with the updated database and classification scheme by Dr. Klau and an updated webtool interface by Dr. Podell, I set out to test a wide variety of biological datasets using NaPDoS2. Through the process of three separate iterations of the webtool—where each iteration was improved upon from the results of my application use case tests—I was able to demonstrate how NaPDoS2 performed compared to the version 1 and the previously published tool eSNaPD2 (Reddy *et al.*, 2014). I illustrated the diversity of datasets that can now be analyzed with NaPDoS2, including single genes, bacterial genomes, fungal genomes, animal genomes, environmental metagenomic data, and targeted KS amplicon data. Various analyses were run for each dataset to understand how the NaPDoS2 settings affected the results. In this way, guidelines could be suggested for users to implement best practices for their analyses. As part of this, I authored nine different and detailed NaPDoS2 tutorials, including a "Quick Start" tutorial which is featured as a webpage tab on the webtool. These tutorials, included as part of the downloadable documentation PDF on the webtool site (and outlined on the project's OSF page: https://osf.io/uzhcp/wiki/Tutorial/) give detailed, step-by-step analysis instructions complete with tutorial objectives, details about each, interpretation of the expected results, and suggestions for further analyses.

As a whole, this project provided the perfect foundation for my continuing studies on using bioinformatic tools to detect, classify, and analyze biosynthetic genes in all types of sequencing data. Our NaPDoS2 webtool update provides a significant improvement to our ability to detect and

classify polyketide and non-ribosomal peptide biosynthetic diversity, which then we can use to study evolutionary patterns and target novel chemical clades in interesting taxa. The skills I learned in this project lead me to develop a larger scale NaPDoS2-mediated analysis of 620,000 genomes across the tree of life (Chapter 3) in collaboration with Hans W. Singh. Additionally, I collaborated with Hans W. Singh to use NaPDoS2 to analyze 240 Gbp of metagenomic data (submitted for publication, not included in this dissertation), where Hans uncovered new patterns of polyketide biosynthetic potential across the tree of life and environmental biomes on earth. Additionally, I was able to use NaPDoS2 to assist with the phylogenetic placement of the salinosporamide *salC* KS domain. By identifying other top *salC* homologs and analyzing them together with KSs in the NaPDoS2 database, it was clear that *salC* formed its own clade outside of other KSs within its class/subclass. Study lead Kate Bauman later showed from her structural and biochemical verification work, *salC* is indeed a novel class of KS that performs a unique reaction in salinosporamide biosynthesis (Bauman *et al.*, 2022).

Chapter 2, in full, is a reprint of the materials as it was submitted to the *Journal of Biological Chemistry*. Klau, L.J.; Podell, S.; Creamer, K.E.; Demko, A.M.; Singh, H.W.; Allen, E.E.; Moore, B.S.; Ziemert, N.; Letzel, A.C.; Jensen, P.R., 2022. The dissertation author was one of three equally contributing primary authors of this manuscript.

2.2 Abstract

The Natural Product Domain Seeker (NaPDoS) webtool detects and classifies ketosynthase (KS) and condensation (C) domains from genomic, metagenomic, and amplicon sequence data. Unlike other tools, a phylogeny-based classification scheme is used to make broader predictions about the polyketide synthase (PKS) and non-ribosomal peptide synthetase (NRPS) genes in which

these domains occur. NaPDoS is particularly useful for the analysis of incomplete biosynthetic genes or gene clusters, as are often observed in poorly assembled genomes and metagenomes, or when loci are not clustered, as in eukaryotic genomes. To help support the growing interest in sequence-based analyses of natural product biosynthetic diversity, we introduce version 2 of the webtool: NaPDoS2, available at http://napdos.ucsd.edu/napdos2. It includes the addition of 1,417 KS sequences, representing a major expansion of the taxonomic and functional diversity represented in the webtool database. The phylogeny-based KS classification scheme now recognizes 41 class and subclass assignments, including new type II PKS subclasses. Workflow modifications accelerate run times allowing larger datasets to be analyzed. Default parameters were established using statistical validation tests to maximize KS detection and classification accuracy while minimizing false positives. The applications of NaPDoS2 to assess PKS biosynthetic potential are demonstrated using genomic, metagenomic, and PCR amplicon datasets. These examples illustrate how NaPDoS2 can be used to predict biosynthetic potential and detect genes involved in the biosynthesis of specific structure classes or new biosynthetic mechanisms.

2.3 Introduction

Increased access to DNA sequence data coupled with a better understanding of the molecular genetics of natural product biosynthesis are driving major advances in natural products research. These advances have been facilitated by webtools such as antiSMASH 6.0 (Blin *et al.*, 2021) and PRISM 4 (Skinnider *et al.*, 2020) that identify natural product biosynthetic gene clusters (BGCs) from assembled sequence data and provide insight into the types of small molecules produced (Medema, 2021). These tools have proven instrumental for genome mining research (Medema *et al.*, 2021), while others have been developed to address more specific topics such as resistance-guided antibiotic discovery (Mungan *et al.*, 2020), biosynthetic gene biogeography

(Reddy *et al.*, 2014), and *trans*-acyl transferase (*trans*-AT) substrate specificity (Helfrich *et al.*, 2019). The Natural Product Domain Seeker (NaPDoS) is a specialized webtool used to assess biosynthetic diversity based on short sequence tags and thus does not require BGC assembly (Ziemert *et al.*, 2012). It targets ketosynthase (KS) and condensation (C) domains to make broader predictions about the polyketide synthase (PKS) and non-ribosomal peptide synthetase (NRPS) genes, respectively, in which they reside. NaPDoS detects these domains and classifies them using a phylogeny-based scheme that reflects well-supported biosynthetic knowledge and established PKS and NRPS function.

Here we introduce version 2 of the NaPDoS webtool (NaPDoS2), which includes an updated KS database and classification scheme that better reflects the expanded taxonomic distributions and functional diversity of PKSs. These enzymes produce structurally diverse polyketides ranging from lipids to macrolides, which represent an important source of compounds for pharmaceutical and other biotechnological applications (Hertweck, 2009). Polyketides also serve important ecological functions ranging from antioxidants (Schöner *et al.*, 2016) to allelochemicals (Wietz *et al.*, 2013) and thus can provide insight into how organisms interact with each other and the environment. PKSs share much in common with fatty acid synthases, generating compounds via the successive decarboxylative condensation and processing of acyl-CoA precursors (Fischbach and Walsh, 2006). Polyketide synthases have been broadly divided into three types based on their organization and function (Shen, 2003), of which NaPDoS2 detects and classifies KSs associated with types I and II. While canonical type I PKSs have a modular organization and function in an assembly line fashion, some function iteratively while others (e.g., *trans*-AT) lack a cognate acyl-transferase domain (Hertweck, 2009; Lohman *et al.*, 2015). Similarly, canonical type II PKSs function iteratively and were originally best known to produce

aromatic polyketides. Yet some type II PKSs function non-iteratively (Shen, 2003), while others produce linear specialized metabolites (Grammbitter *et al.*, 2019). A central feature of PKSs is the KS domain, which in most cases catalyzes a Claisen condensation between the extender unit and the growing, thioester-linked polyketide chain. In recent years, our knowledge of KS functional diversity has expanded significantly to reveal new enzymology and diverse product outcomes across biology. The broad distributions and functional specificities among type I and II PKSs can, in many cases, be resolved phylogenetically (Metsä-Ketelä *et al.*, 2002; Moffitt and Neilan, 2003; Jenke-Kodama *et al.*, 2005), with these evolutionary relationships forming the basis of the NaPDoS2 classification scheme.

The NaPDoS2 website, available at (http://napdos.ucsd.edu/napdos2/), includes many updated features that improve the usability of the tool for natural product discovery. Here we report database and pipeline modifications that provide broader taxonomic coverage, better resolution among functionally characterized PKSs, a new subclassification scheme for type II PKSs, an increased capacity for processing large datasets, and the enhanced detection of eukaryotic KSs. Statistical validation tests have been used to select parameters for optimizing sensitivity and specificity, including both detection and classification accuracy based on query sequence length. The upgraded webtool was used to demonstrate the utility of NaPDoS2 for predicting biosynthetic potential in genomic, metagenomic, and amplicon datasets.

2.4 Methods

2.4.1 Sequence database expansion.

Experimentally validated PKS biosynthetic gene clusters (BGCs) were selected from MIBiG (https://mibig.secondarymetabolites.org/) and published reports. Sequences were extracted from GenBank (https://www.ncbi.nlm.nih.gov/genbank/) and KS domains identified using annotations from both published literature and MIBiG. *In silico* predictions were made using the NRPS/PKS prediction tool (Bachmann and Ravel, 2009). KS domains were further annotated (e.g., type II non-iterative) according to experimentally verified functions and information derived from the associated gene or BGC. Metadata including BGC name, BGC type (e.g., PKS, NRPS), MIBiG accession number, PubMed reference, source organism, and example product name and structure were recorded for each BGC (see **Figure 2.S14** for an example structure from all KS classes/subclasses). The complete NaPDoS2 amino acid sequence database (KS and C domains in FASTA format), BGC table, and domain metadata can be downloaded from the NaPDoS2 website.

2.4.2 KS sequence alignment, phylogenetic analysis, and classification.

Phylogenies generated from the 1,877 KS database sequences (1,417 new and 460 existing) were used to establish class and subclass assignments based on correspondence between functional annotation and tree topology. Sequences were aligned using MAFFT (v 7.017) (Katoh and Standley, 2013) as implemented in Geneious (v 6.1.8) (Kearse *et al.*, 2012) using the FFT-NS-i x 1000 alignment algorithm, BLOSUM62 scoring matrix, and defaults for both gap open penalty and offset value (1.53 and 0.123, respectively). The alignments were trimmed, exported to

PHYLIP, and phylogenies generated using the PhyML online tool (http://www.atgc-montpellier.fr/phyml/) with smart model selection (SMS) enabled (Lefort *et al.*, 2017).

The type II aromatic PKS subclasses were established using a concatenated alignment of the KSα and KSβ subunits and phylogenies generated using the methods described above. Subclasses were delineated based on phylogenetic groupings and features of the core polyketide structure including the length of the poly-β-ketoacyl intermediate, the carbon position of the first ring cyclization, and the type of starter unit. The accuracy of select NaPDoS2 KS classifications from the use case analyses were assessed by analyzing the associated contig with antiSMASH 6.0 (Blin *et al.*, 2021), the NRPS/PKS prediction tool (Bachmann and Ravel, 2009), and transATor (Helfrich *et al.*, 2019) (**Figure 2.S6**).

2.4.3. KS reference trees.

A subset of 414 KS sequences representative of all class and subclass assignments were chosen to generate a KS reference tree (**Figure 2.2**, **Figure 2.S7**). Sequences were aligned using MAFFT (Katoh and Standley, 2013) (v 7.407, FFT-NS-i x 1000 alignment algorithm, BLOSUM62 scoring matrix, default gap open penalty=1.53, and offset value=0.123), and trimmed using trimAI (Capella-Gutiérrez *et al.*, 2009) (1.4.1 with automatic configuration). The trimmed alignment was used to estimate a maximum likelihood tree using FastTree (Lemoine *et al.*, 2019) (v 2.1.11, LG+G model of evolution) with 1000 bootstraps. Booster (Lemoine *et al.*, 2018) (v 0.3.1) was used to estimate support as Transfer Bootstrap Expectation (TBE) values. These programs were implemented in ngphylogeny.fr (Lemoine *et al.*, 2019) and run locally as a docker image. Two type II KS phylogenies were generated following the same steps. One comprising all 201 type II KS sequences in the database, eight FAS sequences, and three thiolase outgroup sequences (**Figure 2.3, Figure 2.S8**). The second used concatenated KSα and KSβ subunits from 59 type II aromatic

BGCs (**Figure 2.4, Figure 2.S9**). Trees were visualized using TreeViewer (https://treeviewer.org/) and annotations added using Adobe Illustrator.

2.4.4 The NaPDoS2 workflow.

Database sequences and associated metadata are stored in a back-end MySQL database linked to the NaPDoS2 web portal through CGI-scripting as previously described (Ziemert *et al.*, 2012). The transeq tool from the EMBOSS package (v 6.6) is used for 6-frame translations of nucleic acid queries (Rice *et al.*, 2000). BLAST queries of amino acid sequences are performed using DIAMOND v 0.9.29 (Buchfink *et al.*, 2015). Multiple sequence alignments are obtained with MUSCLE v 3.8 (Edgar, 2004) using the profile alignment feature to merge query sequences with previously aligned database sequences. Phylogenetic trees are constructed from trimmed amino acid sequences using FastTree v 2.2.1 (Price *et al.*, 2010). Graphical tree depictions are generated using Newick Utilities v 1.5.0.

2.4.5 Performance testing.

The 1,877 full-length amino acid sequences in the NaPDoS2 KS database were clustered at 50% identity using CD-HIT v 4.7 (Fu *et al.*, 2012), yielding 213 non-redundant positive controls. Using the CD-Search (Marchler-Bauer and Bryant, 2004) function of the curated NCBI Conserved Domain Database (Lu *et al.*, 2020), negative controls were selected from subfamilies within the condensing enzyme superfamily (cl09938) providing sister clades that are functionally related to type I and II KS domains but distinct from the positive control sequences. An additional 49 sequences (Jiang *et al.*, 2008)and 14 KSs from type III PKSs in the MIBiG 2.0 repository (Kautsar *et al.*, 2020) were added for a total of 697 sequences (**Table 2.S3**), which were also clustered at 50% identity using CD-HIT to obtain 308 non-redundant negative controls.

2.4.6 Cross-validation and receiver operating characteristic (ROC) curves.

Domain detection sensitivity and specificity were determined using BLASTP searches of full length positive and negative control sequences against the NaPDoS and NaPDoS2 databases, excluding self-matches, to generate leave-one-out cross validation values. ROC curves were constructed and area under the curve (AUC) values calculated from these results using easyROC v 1.3 (Goksuluk *et al.*, 2016). EasyROC data output tables were used to identify potential cutoff points based on the most restrictive cross-validation e-value at which 100% of true positives were detected, to maximize sensitivity with the minimum possible number of false positives.

2.4.7 KS detection and classification accuracy.

A custom perl script (sequence_subdivider.pl, available at https://github.com/spodell/NaPDoS2_website) was used to subdivide the 213 positive control KS sequences into test sets containing overlapping subsequences of 30, 50, 100, or 200 amino acids, each offset by a 10 amino acid sliding window start site. These size-selected test sets, which contained 8555, 8129, 7064, and 4934 subsequences, respectively, were analyzed using the NaPDoS2 workflow to assess the effects of query size on classification accuracy. Accuracy evaluations for each test set were based on the percentage of subsequences whose best non-self, BLASTP match had the same NaPDoS2 classification as the original, full-length sequence from which it was derived.

2.4.8 Application use cases.

Accession information for all sequences and datasets analyzed is provided in **Table S3**. All analyses used the following default settings unless noted otherwise: NaPDoS version 1: domain detection: HMM 1e-5, 200aa minimum alignment length, pathway assignment: e-value cutoff of 1e-5. NaPDoS2: e-value cutoff 1e-8 and 200aa minimum alignment length. Sequence files

containing >500,000 sequences or larger than 500MB, were split into smaller subunits using a custom perl script (serialize_seqs.pl, available at https://github.com/spodell/NaPDoS2_website).

Genomes and Metagenomes. *Salinispora* genome protein sequences 53 downloaded from NCBI and JGI IMG/MER (I.-M. A. Chen *et al.*, 2018) were concatenated into a single FASTA file for NaPDoS2 analysis. Fungal genome protein sequences were downloaded from NCBI; fungal PKS BGCs were extracted from the MIBiG 2.0 repository using the query "Kingdom: Fungi AND BGC type: pks". Coding sequences and predicted proteins for *Elysia chlorotica* were downloaded from NCBI (Cai *et al.*, 2019). Trimmed *E. chlorotica* KS sequences generated by NaPDoS2 were aligned with previously published EcPKS1, EcPKS2, and EcFAS sequences (Torres *et al.*, 2020) and used to construct a phylogenetic tree with the closest NaPDoS2 database hits. Marine sediment metagenomes were selected from the Paired Omics Data Platform (Schorn *et al.*, 2021) and downloaded from NCBI SRA.

Amplicons. KS amplicon sequences were analyzed using minimum alignment lengths of 50 aa in NaPDoS2 unless otherwise noted. Sequence accession and dataset references are listed in **Table S3**. When necessary, random subsets of query sequences were obtained using custom perl scripts (get_seq_info.pl, randomize_lines.pl, serialize_large_list.pl, and getseq_multiple.pl, available at https://github.com/spodell/NaPDoS2_website).

2.5 Results and Discussion

2.5.1 Pipeline efficiency and interface upgrades.

As in the original release (Ziemert *et al.*, 2012), NaPDoS2 detects and classifies ketosynthase (KS) and condensation (C) domains from nucleotide or amino acid sequence data.

While both versions follow the same general workflow, substitution of the more computationally efficient program DIAMOND (Buchfink *et al.*, 2015) for NCBI BLAST (Altschul *et al.*, 1990), eliminates the need for an extra Hidden Markov Model (HMM) pre-filtering step (**Figure 2.S1**). Speed improvements using DIAMOND are relatively modest for small jobs but can reach orders of magnitude for large datasets, especially those consisting of short query sequences (Buchfink *et al.*, 2015). The NaPDoS2 pipeline executes most rapidly on amino acid sequences, which do not need to be translated. Although processing times increase with total nucleotide sequence length and the number of database matches, results for microbial genomes, assembled metagenomes, and PCR KS amplicons containing thousands of hits can typically be obtained within seconds to minutes (**Table 2.S1**). These improvements enable users to perform large-scale analyses that were not feasible with the original NaPDoS release.

User interface upgrades include the addition of a "Domain Classification Summary" page that lists the total number of domains detected in the query data as grouped by their NaPDoS2 classification (**Figure 2.S2A**). This page provides the option to select classes of specific interest for more detailed investigation (**Figure 2.S2B**), which is particularly useful when large numbers of domains are detected. Each BGC represented in the database is linked to a summary page that includes a representative structure and the classification of each KS or C domain within that BGC (**Figure 2.S2C**). The independent classification of each domain is particularly useful given that a single gene or BGC can contain multiple domain types (Sigrist *et al.*, 2020). New webtool features also include quick start instructions outlining the NaPDoS2 workflow (**Figure 2.S3A**) and a downloadable documentation and tutorial file.

2.5.2 Database expansion.

The primary goals for NaPDoS2 were to expand the KS database and classification scheme to include biosynthetic functions that were not represented in the original release, to supplement poorly populated classes, and to provide greater taxonomic coverage. 1,417 new KS sequences were added, raising the database total to 1,877 (average length 418 ± 49 amino acids), an increase of 308% (**Figure 2.S4**). Most of the additional sequences were derived from the MIBiG repository of experimentally verified BGCs (Kautsar *et al.*, 2020), including new type II PKSs, type I fungal PKSs, and type I FASs from both bacterial and fungal sources. A few uncharacterized protist and metazoan sequences were included due to the scarcity of experimental verification among these groups. Although 84 C domain sequences were added, their classification scheme has not been updated and remains an important goal for future releases. The taxonomic breakdown of the current NaPDoS2 KS database sequences is 93.8% bacteria, 4.7% fungi, and 1.5% other eukaryotes (**Table 2.S2**), reflecting the taxonomic skew of available experimental data. To improve result interpretation, database (match) IDs now display the BGC name, domain number, class, and subclass identifiers for each domain.

2.5.3. Phylogeny-based KS classification.

The ability to predict PKS and NRPS biosynthetic potential based on KS or C domains distinguishes NaPDoS2 from other bioinformatic tools. The domain classification scheme is derived from sequence phylogenies and their relationships to established biochemical functions, gene architectures, and structural features of the compounds produced. While not all classes and subclasses are monophyletic, functionally coherent clades form the basis of the classification scheme. A phylogeny generated from the updated KS database allowed us to establish 41 class and subclass assignments associated with type I and II PKSs and FASs (**Table 2.S2**, **Figure 2.S5**).

This represents an increase of 410% over the original release. To verify the sufficiency of using KS domains to indicate broader PKS context, we assessed the genomic neighborhoods of KSs detected in a variety of genomic and metagenomic datasets. In all cases, the NaPDoS2 KS classification agreed with the antiSMASH 6.0 (Blin *et al.*, 2021) classification (**Figure 2.S6**). In some cases, NaPDoS2 provided more detailed information including the identification of type II subclasses, omega-3 poly-unsaturated fatty acid (PUFA) KSs, and highly, partially, and non-reducing fungal KSs. NaPDoS2 can also distinguish among different KS classes within a single gene or BGC (**Figure 2.S6**). With the classification scheme established, a simplified reference tree was generated using 414 sequences representing all class and subclass assignments (**Figure 2.2**).

Broadly, the KS reference phylogeny (**Figure 2.2**) delineates the well-established relationships between fatty acid synthases (FASs) and type I and II PKSs (Jenke-Kodama *et al.*, 2005). The enhanced classification of type I FAS KSs from bacteria/fungi, protists, and metazoa allows users to better distinguish between KSs associated with specialized metabolism and fatty acid biosynthesis across diverse taxa. The expanded eukaryotic type I coverage includes KSs from fungal iterative *cis*-AT PKSs and several protist (Phyla Amoebozoa and Apicomplexa) and metazoan (Phyla Chordata, Echinodermata, and Nematoda) PKSs, including those linked to characterized natural products from *Caenorhabditis elegans* (Feng *et al.*, 2021) and *Dictyostelium discoideum* (Austin *et al.*, 2006). The seven type I PKS classes described in the original NaPDoS release have been reorganized into three classes (modular *cis*-AT, iterative *cis*-AT, and *trans*-AT) and 14 subclasses in the NaPDoS2 release (**Table 2.S2**, **Figure 2.S5**, and user's guide/webtool documentation).

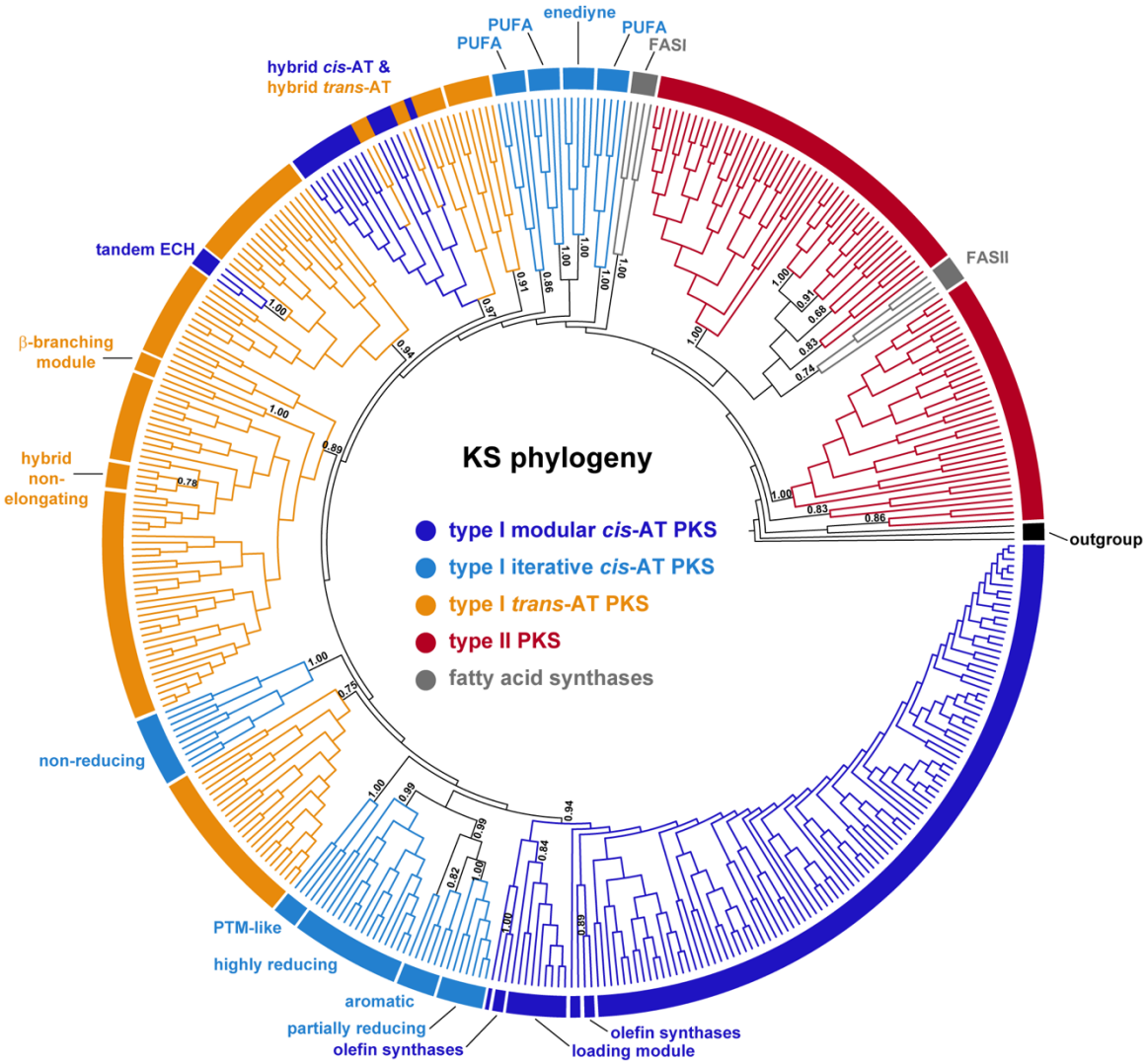


Figure 2.2. KS phylogeny-based classification.

Maximum likelihood phylogeny generated from 414 KS sequences. Clades are color-coded and labeled according to their NaPDoS2 classification. Transfer Bootstrap Expectation (TBE) bootstrap support estimated using Booster (Lemoine et al., 2018). The full name of each sequence can be viewed in Figure 2.S7, which can be used to link a query match to a specific location in the tree. Thiolases from *Escherichia coli*, *Zoogloea ramigera*, and *Streptomyces avermitilis* were used as outgroups. An expanded phylogeny of the type II KSs is presented in Figure 2.3.

Modular *cis*-AT KSs. The modular *cis*-AT class comprises KSs associated with the canonical assembly line PKSs (e.g., erythromycin biosynthesis) and now includes four subclasses of which two (olefin synthase and tandem ECH) are new to NaPDoS2 (**Figure 2.2**). The olefin synthase subclass is best known from cyanobacteria and is associated with the formation of a terminal olefin on a fatty acyl precursor (Coates *et al.*, 2014). These KS sequences form two clades in the reference tree, which reflects the sporadic distribution of the OLS pathway among cyanobacteria (Coates *et al.*, 2014). The tandem ECH subclass is associated with gene cassettes that introduce a branch at the β-keto position. In these PKSs, the KS domain is located immediately downstream of tandem enoyl-CoA hydratase (ECH) and enoyl reductase (ER) domains, which catalyze decarboxylation and reduction reactions, respectively, as seen in cylindrocyclophane biosynthesis (Nakamura *et al.*, 2012) and reviewed elsewhere in detail (Walker *et al.*, 2021). Most of the modular *cis*-AT KSs in the database are not associated with a specialized function and thus are not assigned to a subclass.

Iterative *cis*-AT KSs (bacteria). The iterative *cis*-AT class (iPKSs) maintains a modular organization, with multiple enzymatic domains on a single protein, yet instead of functioning as an assembly line these enzymes catalyze an iterative series of elongation steps (Herbst *et al.*, 2018). The expanded NaPDoS2 classification scheme now includes seven iPKS subclasses. The four observed in bacteria include the aromatic and polycyclic tetramate macrolactam (PTM) subclasses (Chen and Du, 2016), both of which are new to NaPDoS2, and the enediyne and PUFA subclasses, which were identified in the original release. PUFA KSs now form three clades in the reference tree, which correlate with the three KS domains typically present in PUFA PKSs (Metz *et al.*, 2001). Aromatic iPKSs generally produce simple mono- or bicyclic aromatic compounds that are distinct from enediynes and PUFAs (Chen and Du, 2016). The polycyclic tetramate macrolactam-

like (PTM) iPKSs produce complex compounds containing a macrocyclic lactam with an embedded tetramic acid moiety fused with a polycyclic system (i.e., 2-3 rings) derived from polyene chains (Chen and Du, 2016). The ability to quickly recognize PTM biosynthetic potential provides opportunities to expand on the number of compounds discovered in this unusual and often biologically active class (Cao *et al.*, 2010). New iterative type I PKSs continue to be discovered, including some that are unexpectedly abundant and widespread in the genus *Streptomyces* (Cao *et al.*, 2010), providing opportunities to expand this class in future updates beyond the seven subclasses currently recognized by NaPDoS2.

Iterative *cis*-AT KSs (fungi). Iterative *cis*-AT KSs are also observed in fungi and can be delineated into highly reducing (HR), partially reducing (PR), and non-reducing (NR) iPKS subclasses (Gallo *et al.*, 2013). These subclasses differ in the number and type of β-keto processing ketoreductase (KR), dehydratase (DH), and enoyl reductase (ER) domains present. Highly reducing (HR) iPKS BGCs, which contain all three β-keto processing domains, often contain a C-methylation domain responsible for α-carbon methylation (CMeT) and some are fused with NRPS modules (Chooi and Tang, 2012). Their products are usually linear or cyclic, non-aromatic compounds. Partially reducing (PR) iPKS BGCs lack ER, and sometimes also DH, domains. Due to their similarity to bacterial aromatic iPKSs, they have been hypothesized to originate from a horizontal gene transfer event between bacteria and fungi (Schmitt and Lumbsch, 2009). Non-reducing (NR) iPKS BGCs lack all three β-keto processing domains and have specialized domains that facilitate starter unit loading and product folding (Chooi and Tang, 2012).

***Trans*-AT KSs**. KS domains from *trans*-AT PKSs form multiple clades in the reference tree (**Figure 2.2**), which may reflect KS substrate-specificity (Nguyen *et al.*, 2008). Nonetheless, three functionally distinct subclasses can be recognized by NaPDoS2 of which those associated

with β-branching modules and hybrid non-elongating KSs are new to NaPDoS2. The β-branching module, comprising a KS domain, a cryptic "B" domain, and an acyl carrier protein (ACP) domain (KS-B-ACP), represents one of the mechanisms for the formation of β-branching in *trans*-AT PKSs (Bretschneider *et al.*, 2013; Hertweck, 2015). Hybrid, non-elongating KSs (KS^0) typically follow an NRPS module, lack the catalytic histidine required for elongation, and are proposed to facilitate the transfer of the growing polyketide to the next module (A. Chen *et al.*, 2018; Helfrich *et al.*, 2019). The hybrid *trans*-AT KS subclass clades with the hybrid *cis*-AT subclass and functions similarly in modules that lack AT domains.

A major aim for the NaPDoS2 upgrade was to expand the classification of type II KS domains based on established functions and evolutionary relationships (Hillenmeyer *et al.*, 2015). A phylogeny of the type II KS domains in the updated database (**Figure 2.3**) provided sufficient resolution to delineate five functionally defined classes and four sequences that are yet to be assigned a functional class (annotated as type II unclassified). This represents a major improvement over the original NaPDoS release, which simply identified a sequence as being associated with a type II PKS. The most highly populated of the five type II KS classes recognized by NaPDoS2 are associated with PKSs that function iteratively to produce reactive poly-β-ketoacyl intermediates that cyclize to polycyclic aromatic compounds. The type II aromatic class could be further delineated into nine functionally defined subclasses based on a concatenated alignment of the two KS subunits, which generated a better resolved phylogeny (**Figure 2.4**). These KSs function as a heterodimer, with the KSα subunit responsible for the iterative elongation of a nascent poly-β-keto chain, the KSβ subunit determining the chain length (also referred to as the chain-length factor or CLF), and both subunits mediating the first ring cyclization reaction (Tang *et al.*, 2003; A. Chen *et al.*, 2018). The KSα and KSβ subunits are implemented as individual sequences

in the NaPDoS2 workflow thus allowing them to be distinguished. To define these subclasses, the tree topology was compared with poly-β-keto chain length, carbon-carbon position of the first ring cyclization, type of starter unit, and the core and intermediate structures of the products. These subclasses are consistent with previous phylogenies (Metsä-Ketelä *et al.*, 2002; Komaki and Harayama, 2006) and chemotypes identified in PKMiner 42 and antiSMASH version 5.0 (Villebro *et al.*, 2019) based on full-length BGC analyses. KSs from several unusual type II aromatic PKSs that are poorly represented in the database (e.g., resistomycin biosynthesis) were not assigned subclasses and are instead annotated as "type II aromatic unclassified" in NaPDoS2.

The four remaining type II KS classes are more unusual and include one that is responsible for introducing β-branching in *trans*-AT PKSs. These KSs are associated with HMGS or HCS (3-hydroxy-3-methylglutaryl coenzyme A synthase) gene cassettes that encode several monodomain proteins including a decarboxylating KS that can be resolved in the type II KS phylogeny (**Figure 2.3**) (Walker *et al.*, 2021). Another unusual type II KS class recognized in the phylogeny functions non-iteratively. The KSs associated with these enzymes facilitate the incorporation of acetate, propionate, and succinate building blocks leading to the production of macrocyclic compounds such as the nonactin antibiotics, in which case they catalyze C-O bond formation (Kwon *et al.*, 2002). Non-iterative type II PKSs contain multiple type II and III KSs in which each domain is responsible for a single condensation step (Walczak *et al.*, 2000; Rebets *et al.*, 2015). The remaining type II classes represent KSs associated with the production of highly reduced polyenes or aryl polyenes (**Figure 2.3**), with the latter representing the most prominent biosynthetic family observed across a wide taxonomic distribution of bacterial genomes (Cimermancic *et al.*, 2014). Similar to type II aromatic PKSs, polyene and aryl polyene PKS also contain KSα and KSβ subunits with biosynthesis proceeding through a series of alternating elongation and reduction

steps (Grammbitter *et al.*, 2019; Du *et al.*, 2020; Lee *et al.*, 2021). Aryl polyene PKSs also possess homodimeric KSs that function to complement the heterodimer (Grammbitter *et al.*, 2019; Lee *et al.*, 2021). These are most similar to KSs observed in type II FASs (not shown in the tree) and are classified as such in the NaPDoS2 database.

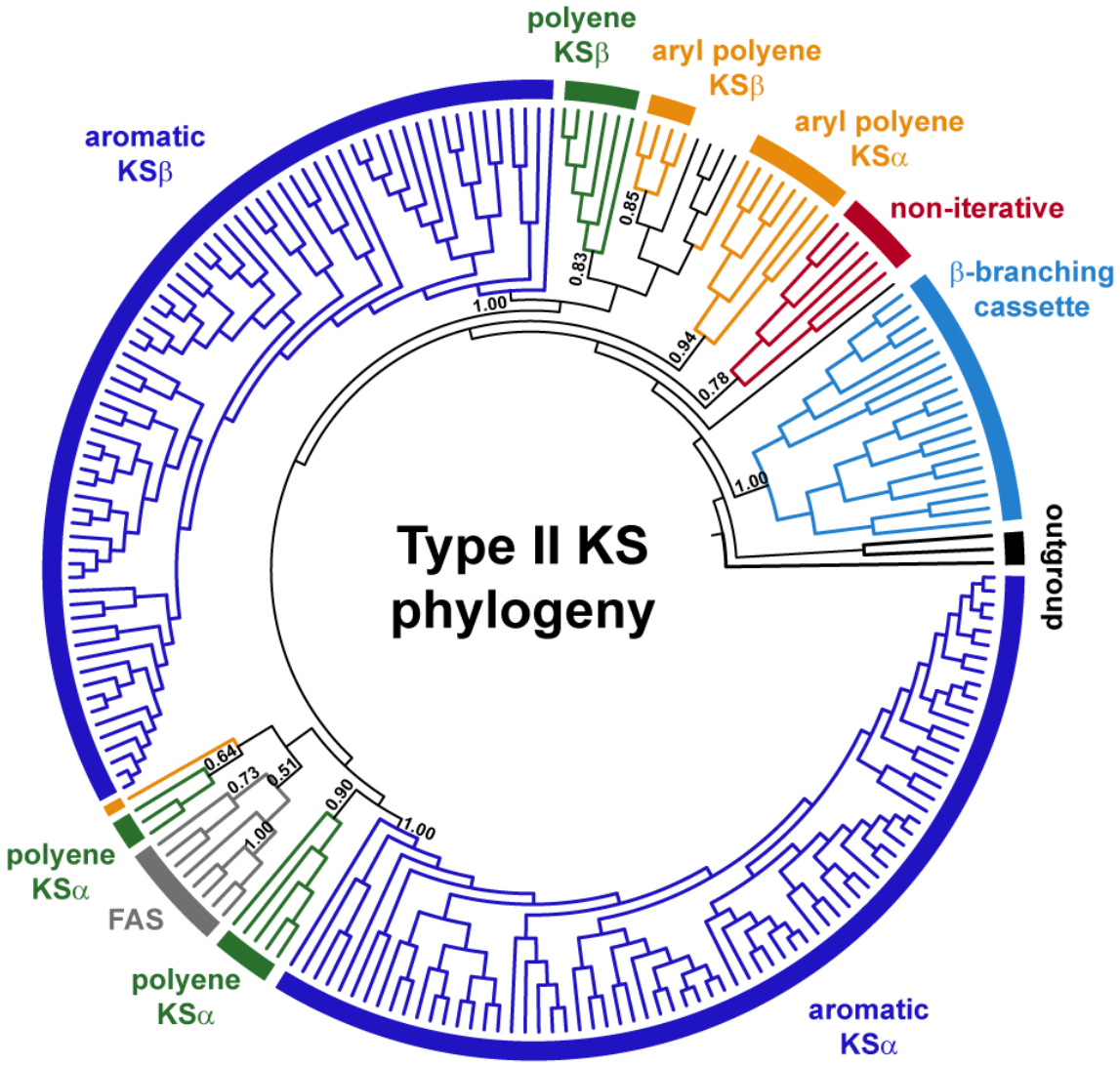


Figure 2.3. Type II KS phylogeny-based classification.

Maximum likelihood phylogeny of 212 KS sequences. Clades are color-coded and labeled according to their NaPDoS2 classification. TBE bootstrap support estimated using Booster (Lemoine *et al.*, 2018). See Figure 2.S8 for sequence annotations. Thiolases from *E. coli*, *Z. ramigera,* and *S. avermitilis* were used as outgroups. Sequences without annotations are type II KSs that have yet to be assigned a functional class.

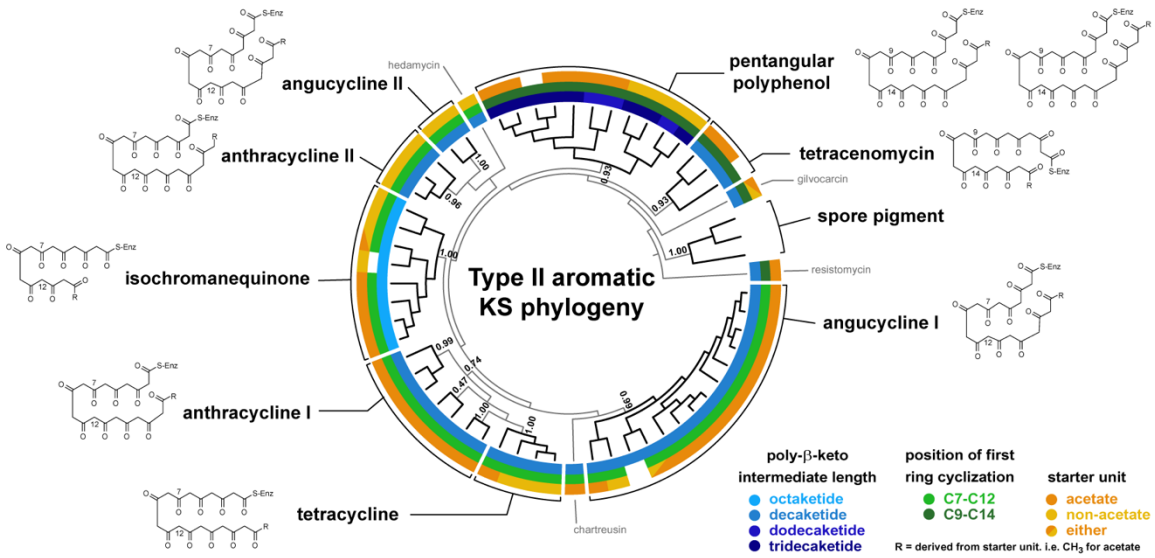


Figure 2.4. Type II aromatic KS phylogeny-based classification.

Maximum likelihood phylogeny of concatenated KSα and β subunits from 59 type II BGCs. Clades are annotated based on NaPDoS2 classification. Structural motifs associated with subclasses shown in colored rings (white, not determined). See Figure 2.S9 for sequence annotations and Table 2.S3 for biosynthetic references. TBE bootstrap support estimated using Booster (Lemoine *et al.*, 2018).

2.5.4. Performance evaluation.

Recognizing that computational prediction algorithms are seldom perfect, leave-one-out cross validation 50 and Receiver Operating Characteristic (ROC) curves (Fawcett, 2006) are powerful statistical tools for quantifying sensitivity and specificity, but require calibration using known positive and negative control datasets that maximize discriminatory power. A set of positive controls (213 sequences) for evaluating NaPDoS2 performance was obtained by clustering all full length, experimentally verified KS domains in the NaPDoS2 database at 50% amino acid identity. These sequences belong to five conserved domain families within the condensing enzyme superfamily of the NCBI Conserved Domain Database (**Figure 2.S10A, B**), which encompasses enzymes that catalyze decarboxylating or non-decarboxylating Claisen-like condensation

reactions for the synthesis and degradation of fatty acids and polyketides from all kingdoms of life (23% Eukaryota, 70% Bacteria, and 7% Archaea). Negative controls (308 sequences) were selected from condensing enzyme families falling immediately outside the NaPDoS2 clades, and similarly clustered at 50% amino acid identity (**Figure 2.S10A**). These negative controls, which include beta-ketoacyl-ACP synthases, ketoacyl-acyl carrier protein synthases III, type III chalcone and stilbene synthases, thiolases, and sterol carrier protein-associated thiolases, were augmented with additional sequences (Jiang *et al.*, 2008) (**Figure 2.S10C**) and KSs from type III PKSs retrieved from MIBiG 2.0 (**Figure 2.S10D).**

Leave-one-out cross-validation scores were generated for positive and negative control sequences based on the e-values of their closest non-self match using both the original NaPDoS and NaPDoS2 KS reference databases. Receiver Operating Characteristic (ROC) curves were calculated from these scores to generate area under the curve (AUC) values (**Figure 2.5A**), demonstrating the superior performance of NaPDoS2 (AUC=0.987) versus the original release (AUC=0.978). ROC curves were further used to establish e-value cutoff points that maximize sensitivity and minimize false positives, identifying optimum values as 1e-8 for NaPDoS2 and 4e-11 for the original NaPDoS release. Although these cutoff values provided equivalent sensitivity (detection of true positives) for their respective algorithms, the highest achievable specificity obtainable for NaPDoS was 93.0% (7% false positive rate) versus 97.2% for NaPDoS2 (2.8% false positives). This improvement is most likely explained by the expanded database underlying the NaPDoS2 pipeline. Consequences of these statistically identified differences in real-world use cases are presented in the "NaPDoS2 applications" section below.

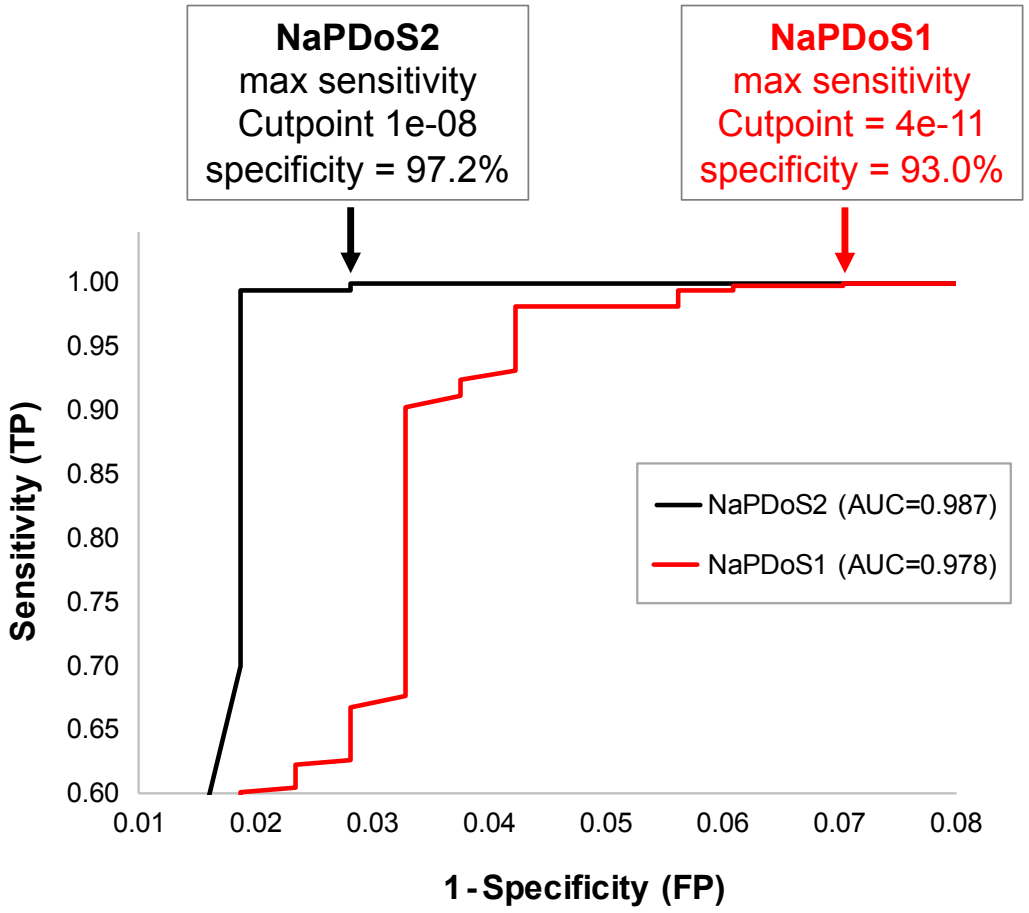


Figure 2.5. Receiver Operating Characteristic (ROC) curves for NaPDoS and NaPDoS2.

Non-redundant sets of 213 positive and 308 negative control KS domains were compared based on the BLASTP e-values of their closest non-self match in each NaPDoS database. Optimal cutoff values were selected to maximize sensitivity. FP, false positives; TP, true positives.

The effects of partial KS sequences on detection and classification accuracy were evaluated using both full-length database sequences (typically 425 amino acids) and shorter overlapping subsets of 30, 50, 100, and 200 amino acids (aa) covering the entire length of these sequences. These subsets were designed to mimic domain fragments encountered in draft genomes, metagenomic assemblies, or KS amplicon sequences. Using the previously established 1e-8 cutoff

for maximum sensitivity, NaPDoS2 detected 99% of the 200 aa length subsequences as KS domains, of which 85% were correctly classified (**Figure 2.6**). Detection and classification accuracy declined with shorter sequences, falling dramatically for sequences <50 amino acids. Length-dependent performance degradation was also observed using 1e-5 and 1e-10 e-values as cutoff scores (**Figure 2.S11**)**.** These results illustrate the difficulty of analyzing unassembled, next-generation sequencing reads and short contigs covering partial domain fragments.

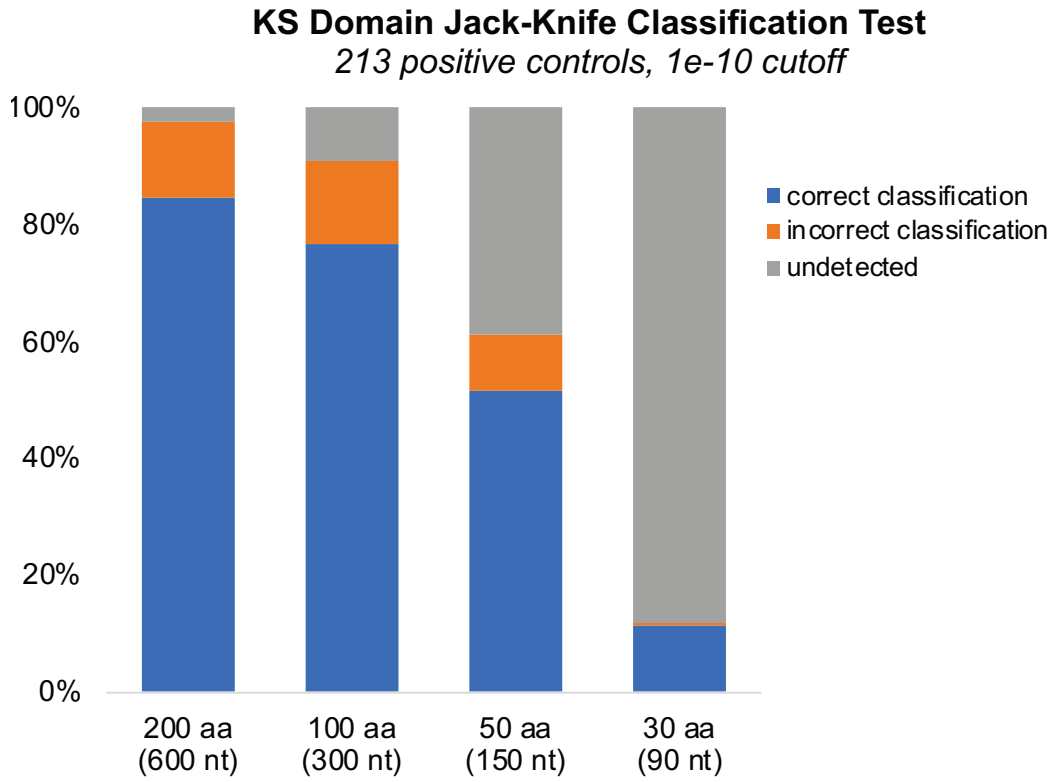


Figure 2.6. Effect of query size on KS detection and classification accuracy.

Classifications were based on a 1e-8 BLASTP e-value cutoff score for the closest non-self-database match. Test sequences of varying lengths were obtained as overlapping sliding window subsequences covering the full length of 213 non-redundant, positive control KS domains.

Based on performance evaluation results, default NaPDoS2 parameters were set at an e-value of 1e-8 and a minimum alignment length of 200 amino acids; however, users may choose to adjust these settings for their individual datasets. Sensitivity declines for partial KS domains can be partially offset by decreasing BLAST stringency, at the cost of increasing false positives and misclassifications. PCR-generated amplicons can provide better sensitivity than similarly sized random sequence fragments but may still be too short to obtain accurate classifications. Confident assignments to poorly populated classes or subclasses are particularly challenging given the sequence diversity observed within highly populated classes and subclasses (**Figure 2.S12**). Some ambiguities may be resolved by using NaPDoS2 sequence alignments to generate detailed phylogenetic trees, but others will remain until additional functional studies are reported.

2.5.5. NaPDoS2 applications.

Large-scale performance evaluations targeting genomic, metagenomic, and amplicon sequence data from a variety of bacterial, fungal, metazoa, and environmental sources were conducted to demonstrate the utility of NaPDoS2 in identifying polyketide biosynthetic potential (**Table 2.1**; accession numbers in **Table 2.S3**). These examples include the integration of NaPDoS2 output with other webtools (**Figure 2.S3B**).

Table 2.1. NaPDoS2 applications.

The utility of NaPDoS2 was demonstrated across a variety of data types and biological sources. Details for each analysis can be found in Supplementary Tables 2.S4-2.S12 as summarized below. Table 2.S3 lists accession numbers for all analyses.

Table #	Application	Data type	Biological source	Dataset	Ref.
2.S4 2.S5	Bacterial type II KS	genome	bacteria	118 *Salinispora* strains	(Millán-Aguiñaga *et al.*, 2017)
2.S6	Fungal KS & FAS	genome	fungi	27 fungal spp.	(Almeida *et al.*, 2019)
2.S7	Fungal KS & FAS	genome	fungi	159 fungal MIBiG 2.0 PKS BGCs	(Kautsar *et al.*, 2020)
2.S7	Environmental type II KS	amplicon	environmental DNA	147 KS clones from soil	(Wawrik *et al.*, 2005)
2.S8	Eukaryotic KS & FAS	genome	metazoa	*Elysia chlorotica*	(Cai *et al.*, 2019; Torres *et al.*, 2020)
2.S9	Environmental type I & II KS, FAS	metagenome	environmental DNA	20 marine sediment samples	(Schorn *et al.*, 2021)
2.S10	Environmental type I KS	amplicon	environmental DNA	eSNaPD v2.0 KS sequences from soil	(Owen *et al.*, 2013)
2.S11	Environmental type II KS	amplicon	environmental DNA	Type II KS sequences from 12 soil samples	(Borsetto *et al.*, 2019)
2.S12	Environmental and cultured type I & type II KS	amplicon	bacteria, environmental DNA	Type I and II KS sequences from lake sediment and enrichment cultures	(Elfeki, Alanjary, Stefan J. Green, *et al.*, 2018)

Genomes.

While the original NaPDoS release simply identified KSs as type II, the new release provides more sensitive detection and more detailed classifications. We analyzed 118 *Salinispora* genomes (Millán-Aguiñaga *et al.*, 2017) with NaPDoS2 and detected a total of 662 type II KS domains in contrast to 363 using the original release (**Table 2.S4**). The type II KSs detected by NaPDoS2 were delineated into seven functionally defined classes and subclasses; those that were unassigned may represent new functional diversity. A broader summary of all KSs detected in these genomes provided the first evidence that *S. arenicola* has the potential to produce polycyclic tetramate macrolactams (PTMs) (**Table 2.S5),** a class of structurally complex natural products that exhibits diverse biological activities (Zhang *et al.*, 2015). These results provide new insight into the biosynthetic potential of this marine actinomycete genus.

Another important NaPDoS2 improvement is the ability to detect and classify eukaryotic KS sequences. This is illustrated by the analysis of 27 taxonomically diverse fungal genomes (Almeida *et al.*, 2019), where KSs ranged from one in *Malassezia globosa* to 50 in *Aspergillus niger* (**Table 2.S6**) and the majority could be assigned to the highly reducing subclass. We next analyzed all 159 fungal PKS BGCs in the MIBiG 2.0 repository (Kautsar *et al.*, 2020) and identified 182 KS domains, all of which matched MIBiG 2.0 descriptions and literature reports (**Table 2.S7**). In contrast, the original NaPDoS release only identified 14 KS domains from these same BGCs. Although relatively few metazoan PKSs have been experimentally characterized, NaPDoS2 recovered the recently described FAS and FAS-like KSs from the *Elysia chlorotica* sacoglossan genome (**Table 2.S8**) (Cai *et al.*, 2019; Torres *et al.*, 2020) and correctly classified them as metazoan type I FASs. A phylogenetic tree generated using NaPDoS2 confirmed their divergence from previously characterized animal FASs (data not shown). In an exploratory search

for KSs in other eukaryotic genomes, NaPDoS2 detected 37 modular *cis*-AT domains, 17 type I FAS domains, 2 type II FAS domains, and one protist-type KS domain from the dinoflagellate *Symbodium minimus* (Beedessee *et al.*, 2015), and 6 type II FAS domains and one type II KSα aromatic anthracycline domain from the diatom *Nitzschia inconpicua* (Oliver *et al.*, 2021), thus further validating its utility for analyzing eukaryotic sequences. While these data show that NaPDoS2 can detect and classifying KS domains from complex metazoan datasets, it is best suited for predicted proteins, translated coding sequences, or transcriptomes, since it cannot excise introns associated with eukaryotic sequence data.

Metagenomes.

A single NaPDoS2 analysis can provide a simultaneous overview of both bacterial and eukaryotic polyketide and fatty acid biosynthetic potential in large metagenomic datasets. While it is not reliant on fully assembled BGCs, assembled contigs are recommended to avoid the reduced classification accuracies associated with short sequence reads (**Figure 2.6**, **Figure 2.S11**). To illustrate this application, we assessed 20 assembled marine sediment metagenomes deposited in the Paired Omics Data Platform (PoDP) (Schorn *et al.*, 2021). We observed a wide range in the numbers and types of KSs detected, which can provide important insight when selecting samples for further study (**Table 2.S9**). These results show how NaPDoS2 can be used to identify samples with the potential to produce compounds in rare but biologically active classes, such as enediynes and PTMs, to expand on poorly understood metazoan PKS diversity, and, in cases with low sequence similarity to database or BLAST matches, detect new functional diversity. Trimmed domains can be used as search queries to assess broader genomic context (**Figure 2.S6**), compare with previously reported BGCs (Kautsar *et al.*, 2020), and potentially identify the host organism when phylogenetic markers are encountered in the KS-containing contig.

Amplicons.

NaPDoS2 is particularly useful for the analysis of KS/C domain PCR amplicon datasets, where it can be used not only to classify sequences into specific functional categories and assign a top BGC product match, but also to assess primer specificity and remove non-target sequences prior to downstream analysis. We illustrate the applications of NaPDoS2 using four KS amplicon datasets, starting with 147 KS sequences cloned from soil eDNA using type II specific KS primers (Wawrik *et al.*, 2005). Both versions of NaPDoS identified all 147 KSs as type II, while NaPDoS2 further delineated them into three type II aromatic subclasses, which agrees with the original report (**Table 2.S7**). We next compared the NaPDoS2 output with that from eSNaPD v 2.0, which relates KS amplicon sequences to a database of characterized BGCs (Owen *et al.*, 2013; Reddy *et al.*, 2014). NaPDoS2 detected all 381 KS sequences in the eSNaPD v 2.0 New Mexico desert soil library "NM_KS_ARRY_LIB01" dataset (Owen *et al.*, 2013) and classified the vast majority as *cis*-AT modular (**Table 2.S10**)**.** Additionally, NaPDoS2 classified virtually all KS sequences within what eSNapD2 v 2.0 listed as novel clusters, providing new information about the biosynthetic diversity within this library (Owen *et al.*, 2013; Reddy *et al.*, 2014) (**Table 2.S10**).

Larger KS amplicon datasets generated from soil eDNA and lake sediment enrichment cultures using type I and II specific primers were also evaluated (Elfeki, Alanjary, Stefan J Green, *et al.*, 2018; Borsetto *et al.*, 2019). While the original analyses estimated biosynthetic potential based solely on the closest MIBiG repository match, NaPDoS2 further delineated the sequences into specific type I and type II classes and subclasses. Analysis of the type II soil amplicons (Borsetto *et al.*, 2019) revealed that many of the amplicons were not recognized as KSs and may represent off-target or too short amplifications. Of those identified as KSs, a number were classified by NaPDos2 as type II FASs and a few as type I PKSs (**Table 2.S11**). This illustrates

the application of NaPDoS2 to assess primer specificity and identify potential non-target KS sequences prior to downstream analyses.

Finally, we addressed the question of whether lowering the minimum alignment length for amplicons might increase the number of false positives, as previously observed for random domain fragments (**Figure 2.6**). To do this, we analyzed 40,000 randomly selected sequences from longer amplicon (602 bp) type I and II KS datasets from lake sediment enrichment cultures (Elfeki, Alanjary, Stefan J Green, *et al.*, 2018) using a range of minimum amino acid alignment lengths (5,000 of these sequences shown in **Table 2.S12**). We placed both hit and non-hit sequences in a phylogenetic context within the condensing enzyme superfamily tree (Jiang *et al.*, 2008) and mapped the conserved domains with TREND (Gumerov and Zhulin, 2020) (**Figure 2.S13**). Sequences identified by NaPDoS2 as KSs, regardless of alignment length (30-200aa), fall within the two clades associated with the NaPDoS2 database positive control sequences. Conversely, the sequences that NaPDoS2 did not identify as KSs fell outside of these clades and were associated with off-target, non-ketosynthase domains such as AMP-binding domains and multiple phage-related domains. These results confirm that shorter NaPDoS2 alignment length settings can be used to assess polyketide biosynthetic potential from amplicon datasets and highlight the value of verifying detection and classification accuracy using phylogenetic approaches and conserved domain architecture.

2.6 Conclusions

While several tools can detect and classify the gene clusters associated with natural product biosynthesis (Reddy *et al.*, 2014; Blin *et al.*, 2021), NaPDoS2 employs KS and C domains as sequence tags to predict biosynthetic potential. This approach makes it well-suited for non-

clustered or incomplete BGCs, amplicons, and eukaryotic genomes where other tools are less effective. The NaPDoS2 update features an expanded KS database and classification scheme that better reflects the broader taxonomic and functional diversity now recognized among type I and II PKSs. It provides a single workflow to detect and classify KS domains from diverse biological origins including bacteria, fungi, and other eukaryotes and to distinguish among those involved in fatty acid and specialized metabolite production. Updates to workflow efficiency can now accommodate the larger genomic, metagenomic, and amplicon datasets achievable with next-generation sequencing. NaPDoS2 provides a rapid method to identify microorganisms or environments with the potential to yield rare classes of compounds, such as those produced by non-iterative type II PKSs (Shen, 2003). It can be used to prioritize samples for cultivation and to identify potentially new biosynthetic mechanisms when sequences are phylogenetically distinct from those previously characterized, as recently demonstrated for the standalone KS (*salC*) that functions as an aldolase/β-lactone synthase in salinosporamide A biosynthesis (Bauman *et al.*, 2022). The NaPDoS2 upgrades expand on the PKS diversity that can be detected with this tool and provide a method to quickly assess biosynthetic potential in a manner that facilitates more targeted approaches to natural product discovery.

2.7 Data availability

Relevant alignment files, phylogenetic tree files, webtool documentation file/user's guide, example tutorials sequence files, and other supporting information can be found on the corresponding OSF project page: https://osf.io/uzhcp/. The code used to construct version 2 of the Natural Product Domain Seeker website can be found on the corresponding Github repository: https://github.com/spodell/NaPDoS2_website. All accession information and dataset references

can be found in **Table 2.S3** (separate Microsoft Excel file, not included in this dissertation but accessible on the OSF project page: https://osf.io/uzhcp/).

2.8 Funding sources

This research was supported by the National Institutes of Health (grant no. 5R01GM085770 to P.R.J. and B.S.M.), the National Science Foundation Graduate Research Fellowship Program (grant no. DGE-1650112 to K.E.C. and A.M.D.; grant no. DGE-2038238 to H.W.S.) and the National Science Foundation Division of Molecular and Cellular Biosciences (grant MCB-1149552 to E.E.A.). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation nor other funding providers.

2.9 Supplementary Figures and Tables

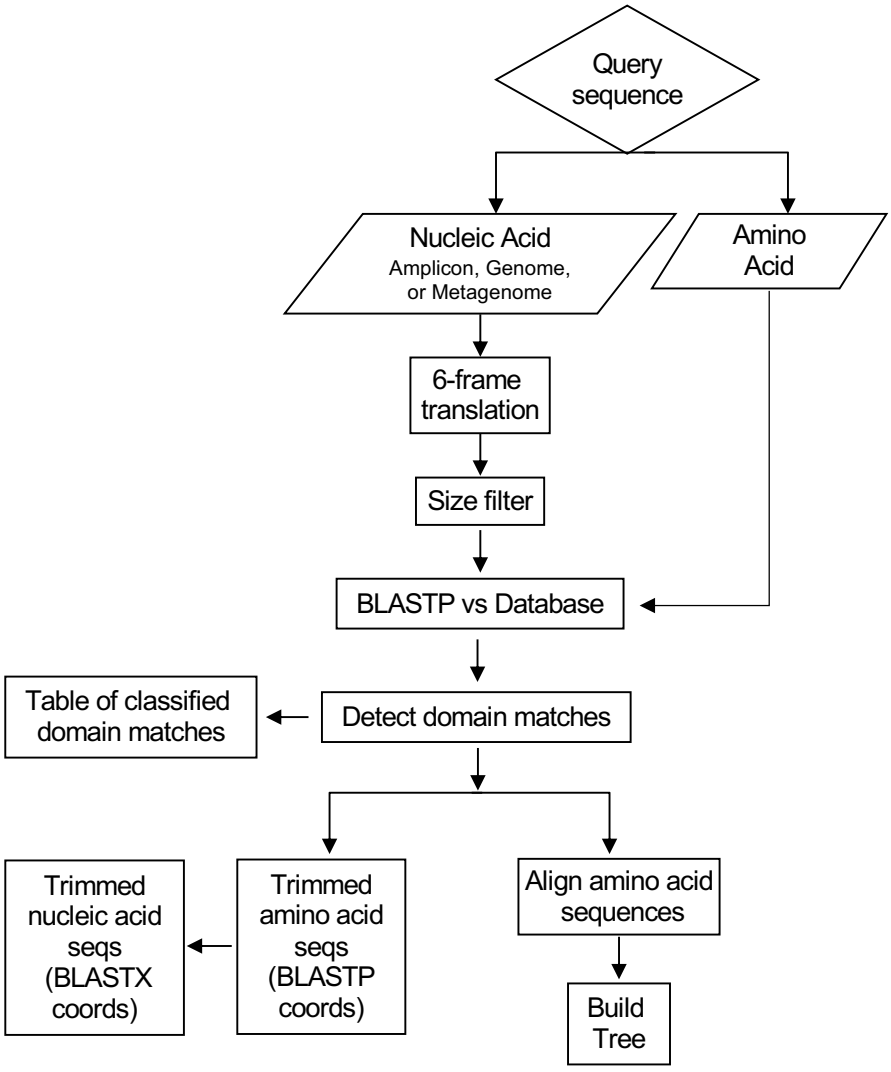


Figure S1. NaPDoS2 bioinformatic pipeline. Translated nucleotide and protein query sequences are compared to an internal database of KS and C domains using BLASTP. Detailed descriptions of individual steps can be found in the GitHub repository site: https://github.com/spodell/NaPDoS2_website

Figure 2.S1. NaPDoS2 bioinformatic pipeline.

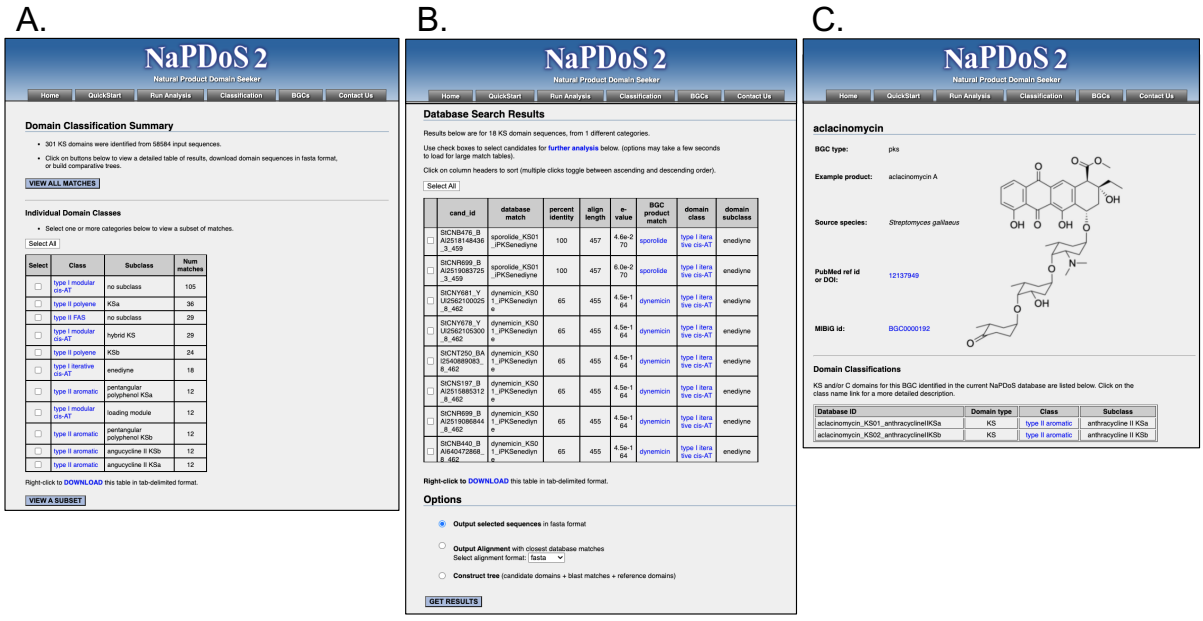
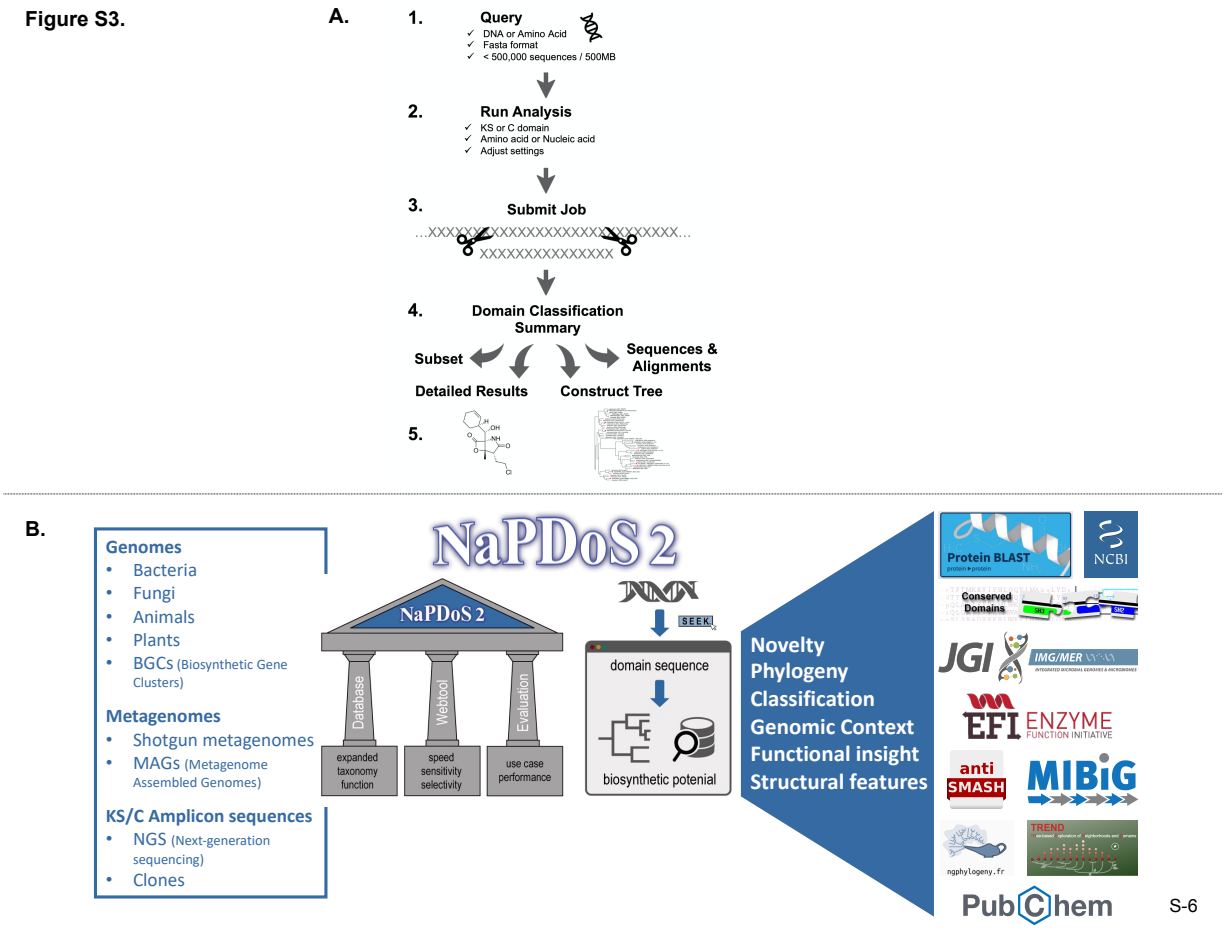


Figure S2. **Featured webtool updates**. A). Domain classification summary page. In this example, 301 KS domains were detected in twelve bacterial genomes (58,584 protein sequences). The number of domains detected in each class and subclass is indicated. B). Database search results page. Expanded view of the 18 type I iterative *cis*-AT enediyne domains from the search shown in (A). C). BGC page. Clicking on the BGC product match hyperlink from the Database Search Results in panel (B) provides the compound structure and details about the associated BGC and all corresponding KS classifications (C domains to be updated in a later release).

Figure 2.S2. Featured webtool updates.

Figure S3.



S-6

Figure S3. NaPDoS2 workflow and analysis roadmap. A) The NaPDoS2 user workflow consists of submitting a query sequence, selecting the type of analysis to run, submitting the job, and deciding what output to view or analyses to complete. B) A roadmap for the use of NaPDos2 starts with genomic, metagenomic, or KS/C domain amplicon sequences derived from a variety of sources. The expanded database and webtool improvements provide important analytical upgrades while extensive use testing demonstrates the applications of NaPDoS2 to assess biosynthetic potential by detecting and classifying KS and C domain sequences. NaPDoS2 output can be further analyzed using a variety of tools to assess novelty, phylogeny, classification, genomic context, function, and small molecule structural features. Detailed, step-by-step tutorial examples can be found in the downloadable “Documentation” PDF linked on the NaPDoS2 webpage.

Figure 2.S3. NaPDoS2 workflow and analysis roadmap.

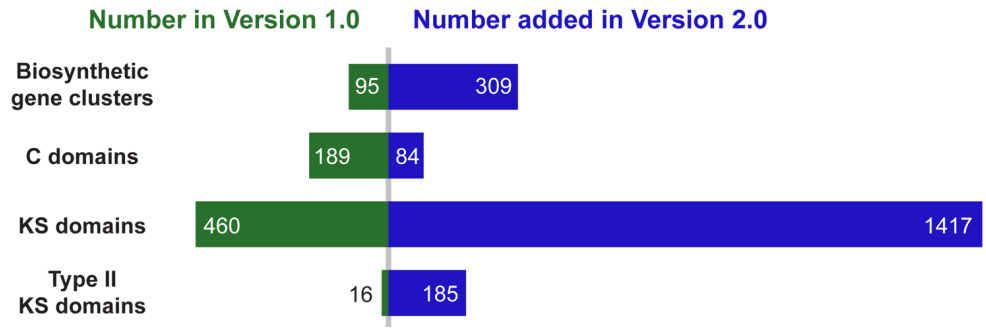
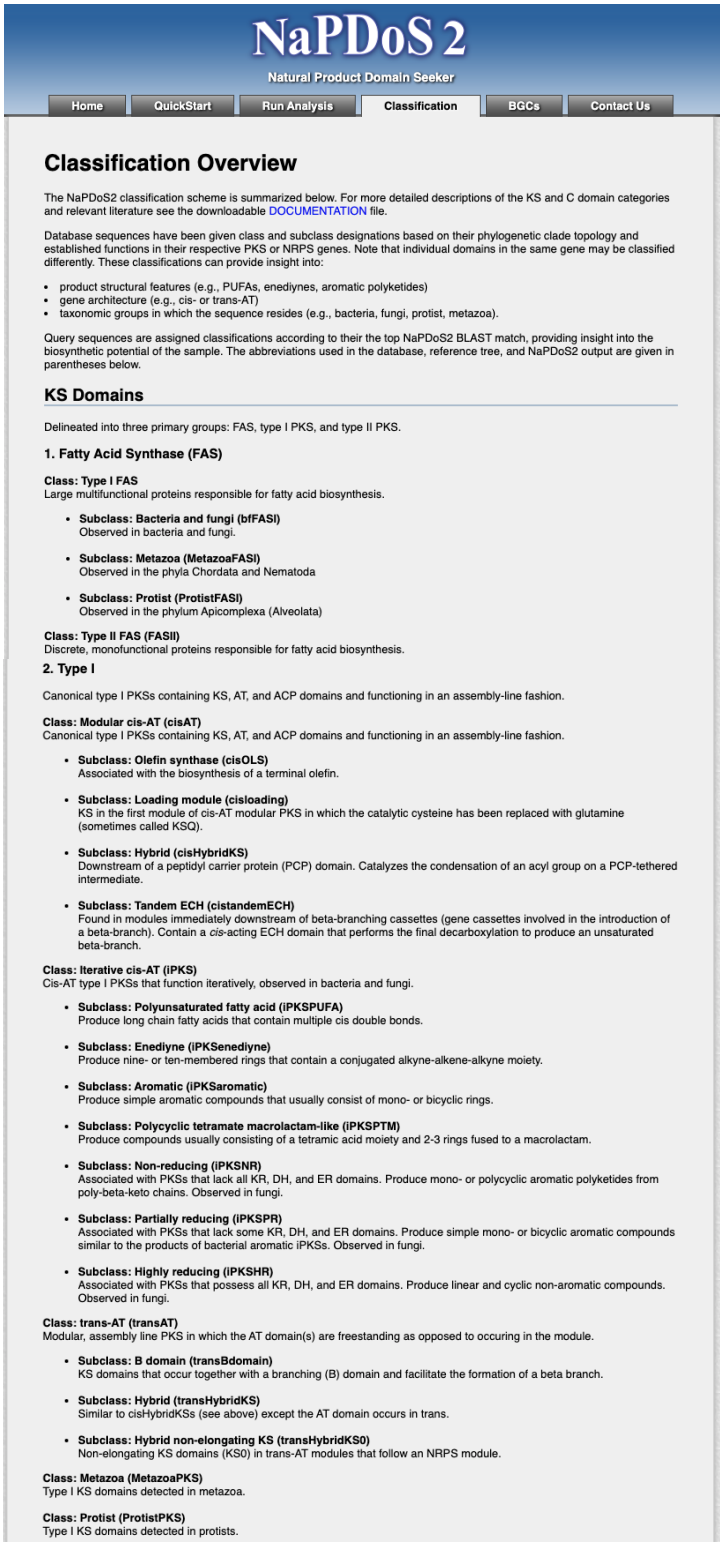


Figure S4. Comparison of the expanded NaPDoS2 database with the original NaPDoS version 1 release[1]. The NaPDoS2 reference database contains 273 C domains and 1,877 KS domains from a total of 404 BGCs.

Figure 2.S4. Comparison of the expanded NaPDoS2 database.

Figure S5.



NaPDoS 2

Natural Product Domain Seeker

Home | QuickStart | Run Analysis | Classification | BGCs | Contact Us

Classification Overview

The NaPDoS2 classification scheme is summarized below. For more detailed descriptions of the KS and C domain categories and relevant literature see the downloadable DOCUMENTATION file.

Database sequences have been given class and subclass designations based on their phylogenetic clade topology and established functions in their respective PKS or NRPS genes. Note that individual domains in the same gene may be classified differently. These classifications can provide insight into:

- product structural features (e.g., PUFAs, enediynes, aromatic polyketides)
- gene architecture (e.g., cis- or trans-AT)
- taxonomic groups in which the sequence resides (e.g., bacteria, fungi, protist, metazoa).

Query sequences are assigned classifications according to their the top NaPDoS2 BLAST match, providing insight into the biosynthetic potential of the sample. The abbreviations used in the database, reference tree, and NaPDoS2 output are given in parentheses below.

KS Domains

Delineated into three primary groups: FAS, type I PKS, and type II PKS.

1. Fatty Acid Synthase (FAS)

Class: Type I FAS
Large multifunctional proteins responsible for fatty acid biosynthesis.

- **Subclass: Bacteria and fungi (bfFASI)**
  Observed in bacteria and fungi.
- **Subclass: Metazoa (MetazoaFASI)**
  Observed in the phyla Chordata and Nematoda
- **Subclass: Protist (ProtistFASI)**
  Observed in the phylum Apicomplexa (Alveolata)

Class: Type II FAS (FASII)
Discrete, monofunctional proteins responsible for fatty acid biosynthesis.

2. Type I

Canonical type I PKSs containing KS, AT, and ACP domains and functioning in an assembly-line fashion.

Class: Modular cis-AT (cisAT)
Canonical type I PKSs containing KS, AT, and ACP domains and functioning in an assembly-line fashion.

- **Subclass: Olefin synthase (cisOLS)**
  Associated with the biosynthesis of a terminal olefin.
- **Subclass: Loading module (cisloading)**
  KS in the first module of cis-AT modular PKS in which the catalytic cysteine has been replaced with glutamine (sometimes called KSQ).
- **Subclass: Hybrid (cisHybridKS)**
  Downstream of a peptidyl carrier protein (PCP) domain. Catalyzes the condensation of an acyl group on a PCP-tethered intermediate.
- **Subclass: Tandem ECH (cistandemECH)**
  Found in modules immediately downstream of beta-branching cassettes (gene cassettes involved in the introduction of a beta-branch). Contain a *cis*-acting ECH domain that performs the final decarboxylation to produce an unsaturated beta-branch.

Class: Iterative cis-AT (iPKS)
Cis-AT type I PKSs that function iteratively, observed in bacteria and fungi.

- **Subclass: Polyunsaturated fatty acid (iPKSPUFA)**
  Produce long chain fatty acids that contain multiple cis double bonds.
- **Subclass: Enediyne (iPKSenediyne)**
  Produce nine- or ten-membered rings that contain a conjugated alkyne-alkene-alkyne moiety.
- **Subclass: Aromatic (iPKSaromatic)**
  Produce simple aromatic compounds that usually consist of mono- or bicyclic rings.
- **Subclass: Polycyclic tetramate macrolactam-like (iPKSPTM)**
  Produce compounds usually consisting of a tetramic acid moiety and 2-3 rings fused to a macrolactam.
- **Subclass: Non-reducing (iPKSNR)**
  Associated with PKSs that lack all KR, DH, and ER domains. Produce mono- or polycyclic aromatic polyketides from poly-beta-keto chains. Observed in fungi.
- **Subclass: Partially reducing (iPKSPR)**
  Associated with PKSs that lack some KR, DH, and ER domains. Produce simple mono- or bicyclic aromatic compounds similar to the products of bacterial aromatic iPKSs. Observed in fungi.
- **Subclass: Highly reducing (iPKSHR)**
  Associated with PKSs that possess all KR, DH, and ER domains. Produce linear and cyclic non-aromatic compounds. Observed in fungi.

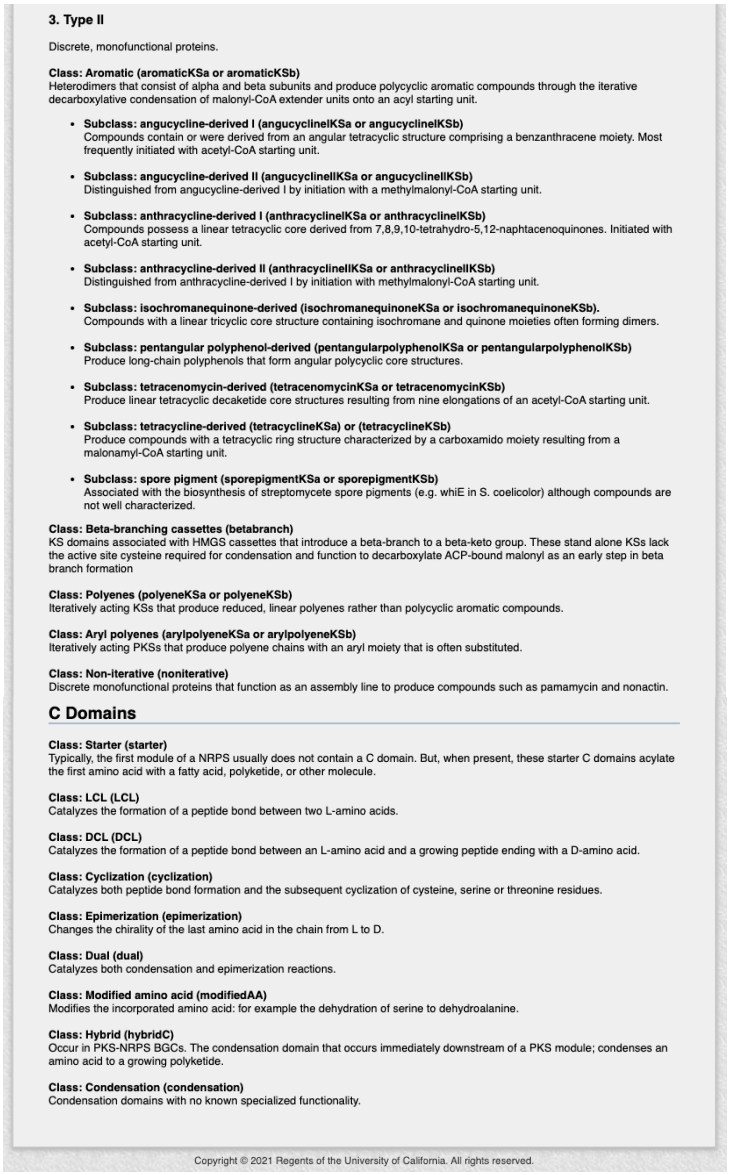
Class: trans-AT (transAT)
Modular, assembly line PKS in which the AT domain(s) are freestanding as opposed to occuring in the module.

- **Subclass: B domain (transBdomain)**
  KS domains that occur together with a branching (B) domain and facilitate the formation of a beta branch.
- **Subclass: Hybrid (transHybridKS)**
  Similar to cisHybridKSs (see above) except the AT domain occurs in trans.
- **Subclass: Hybrid non-elongating KS (transHybridKS0)**
  Non-elongating KS domains (KS0) in trans-AT modules that follow an NRPS module.

Class: Metazoa (MetazoaPKS)
Type I KS domains detected in metazoa.

Class: Protist (ProtistPKS)
Type I KS domains detected in protists.

Figure S5.



3. Type II

Discrete, monofunctional proteins.

Class: Aromatic (aromaticKSa or aromaticKSb)
Heterodimers that consist of alpha and beta subunits and produce polycyclic aromatic compounds through the iterative decarboxylative condensation of malonyl-CoA extender units onto an acyl starting unit.

- **Subclass: angucycline-derived I (angucyclinelKSa or angucyclinelKSb)**
  Compounds contain or were derived from an angular tetracyclic structure comprising a benzanthracene moiety. Most frequently initiated with acetyl-CoA starting unit.
- **Subclass: angucycline-derived II (angucyclinellKSa or angucyclinellKSb)**
  Distinguished from angucycline-derived I by initiation with a methylmalonyl-CoA starting unit.
- **Subclass: anthracycline-derived I (anthracyclinelKSa or anthracyclinelKSb)**
  Compounds possess a linear tetracyclic core derived from 7,8,9,10-tetrahydro-5,12-naphtacenoquinones. Initiated with acetyl-CoA starting unit.
- **Subclass: anthracycline-derived II (anthracyclinellKSa or anthracyclinellKSb)**
  Distinguished from anthracycline-derived I by initiation with methylmalonyl-CoA starting unit.
- **Subclass: isochromanequinone-derived (isochromanequinoneKSa or isochromanequinoneKSb).**
  Compounds with a linear tricyclic core structure containing isochromane and quinone moieties often forming dimers.
- **Subclass: pentangular polyphenol-derived (pentangularpolyphenolKSa or pentangularpolyphenolKSb)**
  Produce long-chain polyphenols that form angular polycyclic core structures.
- **Subclass: tetracenomycin-derived (tetracenomycinKSa or tetracenomycinKSb)**
  Produce linear tetracyclic decaketide core structures resulting from nine elongations of an acetyl-CoA starting unit.
- **Subclass: tetracycline-derived (tetracyclineKSa) or (tetracyclineKSb)**
  Produce compounds with a tetracyclic ring structure characterized by a carboxamido moiety resulting from a malonamyl-CoA starting unit.
- **Subclass: spore pigment (sporepigmentKSa or sporepigmentKSb)**
  Associated with the biosynthesis of streptomycete spore pigments (e.g. whiE in S. coelicolor) although compounds are not well characterized.

Class: Beta-branching cassettes (betabranch)
KS domains associated with HMGS cassettes that introduce a beta-branch to a beta-keto group. These stand alone KSs lack the active site cysteine required for condensation and function to decarboxylate ACP-bound malonyl as an early step in beta branch formation

Class: Polyenes (polyeneKSa or polyeneKSb)
Iteratively acting KSs that produce reduced, linear polyenes rather than polycyclic aromatic compounds.

Class: Aryl polyenes (arylpolyeneKSa or arylpolyeneKSb)
Iteratively acting PKSs that produce polyene chains with an aryl moiety that is often substituted.

Class: Non-iterative (noniterative)
Discrete monofunctional proteins that function as an assembly line to produce compounds such as pamamycin and nonactin.

C Domains

Class: Starter (starter)
Typically, the first module of a NRPS usually does not contain a C domain. But, when present, these starter C domains acylate the first amino acid with a fatty acid, polyketide, or other molecule.

Class: LCL (LCL)
Catalyzes the formation of a peptide bond between two L-amino acids.

Class: DCL (DCL)
Catalyzes the formation of a peptide bond between an L-amino acid and a growing peptide ending with a D-amino acid.

Class: Cyclization (cyclization)
Catalyzes both peptide bond formation and the subsequent cyclization of cysteine, serine or threonine residues.

Class: Epimerization (epimerization)
Changes the chirality of the last amino acid in the chain from L to D.

Class: Dual (dual)
Catalyzes both condensation and epimerization reactions.

Class: Modified amino acid (modifiedAA)
Modifies the incorporated amino acid: for example the dehydration of serine to dehydroalanine.

Class: Hybrid (hybridC)
Occur in PKS-NRPS BGCs. The condensation domain that occurs immediately downstream of a PKS module; condenses an amino acid to a growing polyketide.

Class: Condensation (condensation)
Condensation domains with no known specialized functionality.

Figure S5. NaPDoS2 classification overview. Class and subclass descriptions as described on the NaPDoS2 website. Additional details and relevant references can be found in the downloadable “Documentation” PDF linked at the top of the webpage.

Figure 2.S5. NaPDoS2 classification overview.

Figure S6.

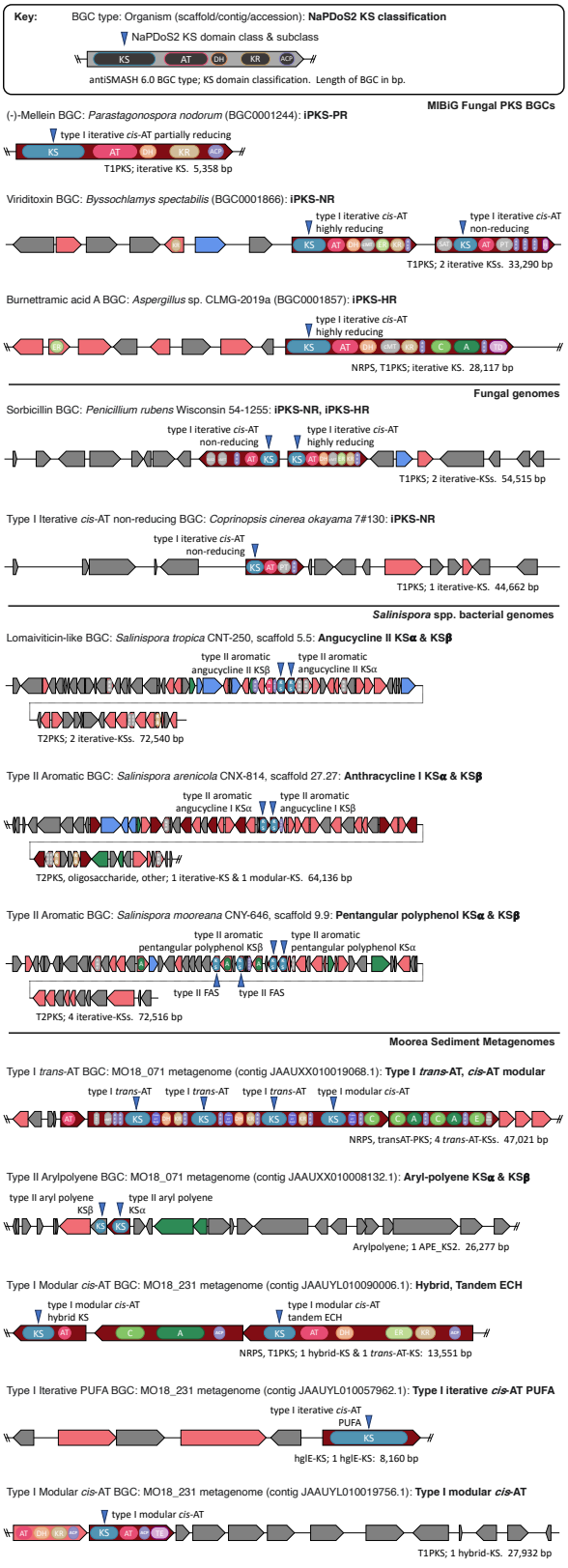


Figure 2.S6. Verification of NaPDoS2 KS domain classifications.

Figure S6. **Verification of NaPDoS2 KS domain classifications.** Biosynthetic gene cluster context of select KSs detected in the application use case analyses. For each KS, the associated scaffold or contig was extracted and run through antiSMASH 6.0[2] (https://antismash.secondarymetabolites.org/) transATor[3] (https://transator.ethz.ch/) , and the "PKS/NRPS Analysis Web-site"[4] (https://nrps.igs.umaryland.edu/) The BGCs were drawn and colored as determined by antiSMASH 6.0 (maroon, core biosynthetic gene; pink, additional biosynthetic gene; blue, transport-related genes; green, regulatory genes, grey, other genes). Domain position and function were drawn and colored according to antiSMASH 6.0[2], transATor[3], and "PKS/NRPS Analysis Web-site"[4] (blue, KS ketosynthase; pink, AT acyl transferase; sand, KR ketoreductase; pale purple ACP Phosphopantetheine acyl carrier protein, ACPS holo-ACP synthase; orange, DH dehydratase; light grey, cMT carbon methyltransferase, FkbH domain, NAD Male sterility protein, AmT aminotransferase, Cyc cyclase, GNAT GNAT domain, TIGR01720 NRPS domain of unknown function; light pink, TE thioesterase, TD Terminal reductase domain; pale green ER enoyl reductase; dark blue, *trans*-AT docking *trans*-acyltransferase docking domain; light green, C condensation domain of NRPS, E epimerization domain; dark green, A adenylation domain). Blue arrows point to KS hits that NaPDoS2 detected and classified in the BGC context; arrows are labeled with the NaPDoS2 KS domain classification. The antiSMASH 6.0[2] BGC type and KS domain classification, followed by the length of the entire BGC (in base pairs) is listed below each BGC, as indicated in the key.

We strategically chose diverse use case analyses for ground-truthing the contextual genomic evidence for KS domain-based classification. NaPDoS2 correctly classified KSs associated with partially reducing ((-)-Mellein), non-reducing (viriditoxin), and highly reducing (burnettramic acid A) fungal BGCs from MIBiG 2.0[5], as confirmed by literature reports and antiSMASH 6.0[2] output. Next, NaPDoS2 correctly classified the KSs in the sorbicillin BGC from *Penicillium rubens* fungal genome, which contains both non-reducing and highly reducing KS domains. NaPDoS2 also identified a type I iterative *cis*-AT non-reducing KS in an orphan BGC in the *Coprinopsis cinerea okayama* basidiomycete genome. Next, NaPDoS2 also correctly identified the KS domains in the lomaiviticin BGC as type II aromatic angucycline II, and detected the KSs associated with two *Salinispora* orphan BGCs as type II aromatic anthracycline I and pentangular polyphenol. Finally, NaPDoS2 correctly classified *trans*-AT, *cis*-AT, aryl-polyene, hybrid, trandem ECH, modular, and PUFA KS domains from metagenomic assemblies based on their respective BGC context. In many cases, the NaPDoS2 classification was more specific than antiSMASH 6.0[2] BGC domain predictions.

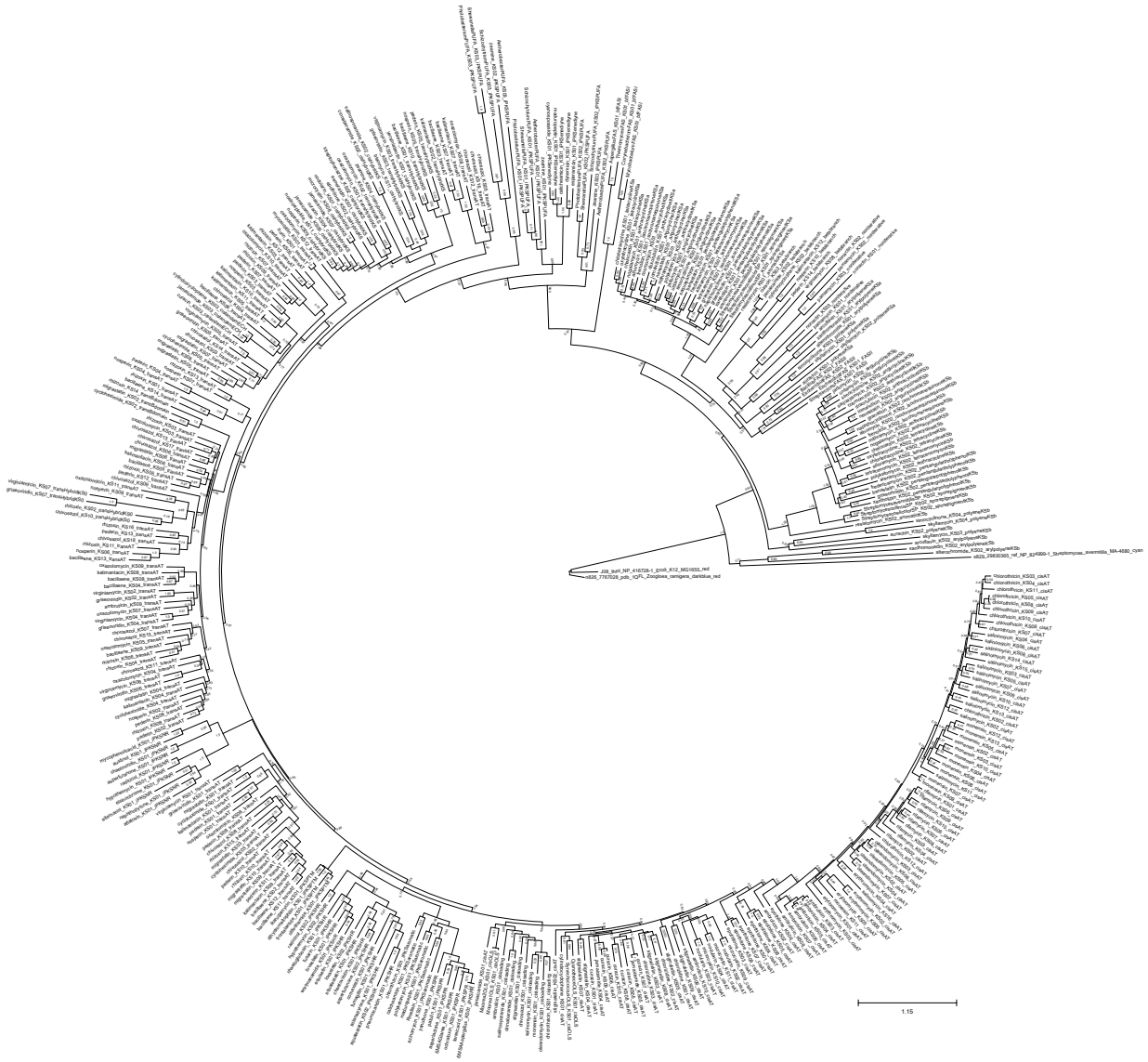


Figure S7. **Maximum likelihood KS phylogeny.** Tree includes 414 KS sequences and 3 thiolase sequences as outgroups. The full name of each sequence is listed on the branch tips, which can help link a query match to a specific location in the tree. Bootstrap support is listed for each node. Thiolases from *Escherichia coli* (NP_416728.1) and *Zoogloea ramigera* (1QFL_A) and a SCP-x thiolase from *Streptomyces avermitilis* (NP_824999.1) were used as outgroups.

Figure 2.S7. Maximum likelihood KS phylogeny.

Figure S8.

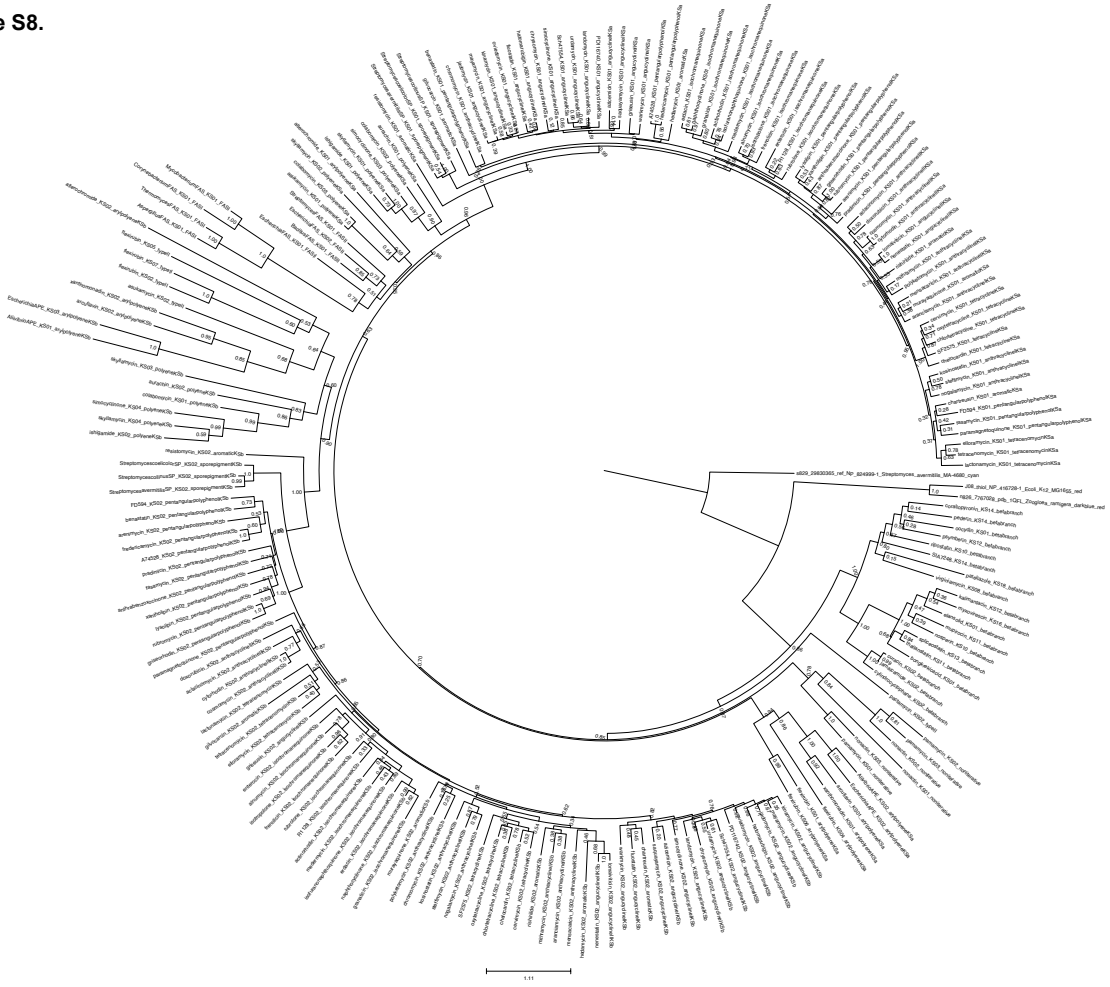


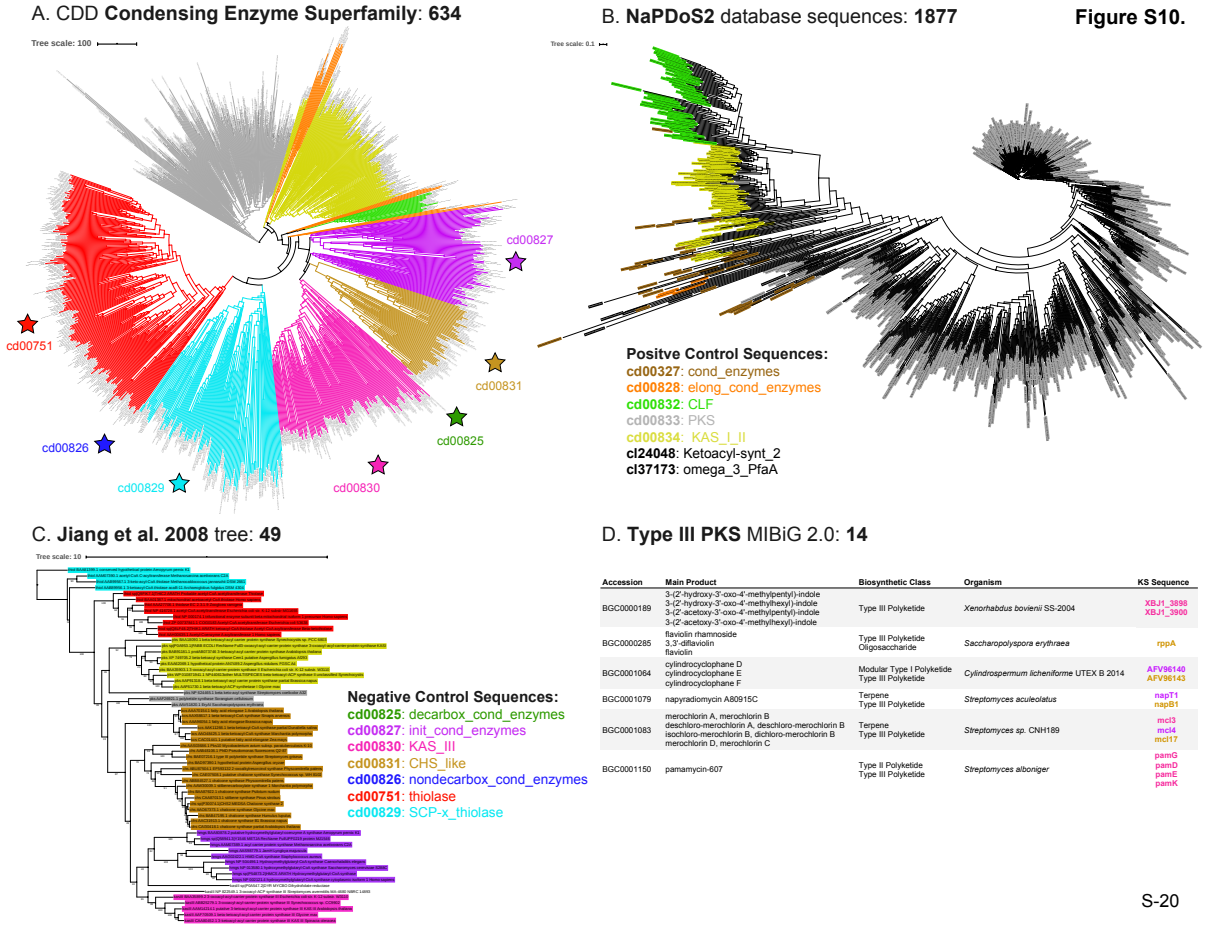
Figure S8. Maximum likelihood type II KS phylogeny. Tree includes 201 type II KS sequences, 8 FAS sequences and 3 thiolase sequences as outgroups. The full name of each sequence is listed on the branch tips, which can help link a query match to a specific location in the tree. Bootstrap support is listed for each node. Thiolases from *Escherichia coli* (NP_416728.1) and *Zoogloea ramigera* (1QFL_A) and a SCP-x thiolase from *Streptomyces avermitilis* (NP_824999.1) were used as outgroups.

Figure 2.S8. Maximum likelihood type II KS phylogeny.

Figure S9.

Figure S9. Maximum likelihood type II aromatic KS phylogeny of concatenated KSα and KSβ sequences from 59 biosynthetic gene clusters. The full name of each sequence is listed on the branch tips, which can help link a query match to a specific location in the tree. Bootstrap support is listed for each node.

Figure 2.S9. Maximum likelihood type II aromatic KSα and KSβ phylogeny.



Accession	Main Product	Biosynthetic Class	Organism	KS Sequence
BGC0000189	3-(2'-hydroxy-3'-oxo-4'-methylpentyl)-indole 3-(2'-hydroxy-3'-oxo-4'-methylhexyl)-indole 3-(2'-acetoxy-3'-oxo-4'-methylpentyl)-indole 3-(2'-acetoxy-3'-oxo-4'-methylhexyl)-indole	Type III Polyketide	*Xenorhabdus bovienii* SS-2004	XBJ1_3898 XBJ1_3900
BGC0000285	flaviolin rhamnoside 3,3'-diflaviolin flaviolin	Type III Polyketide Oligosaccharide	*Saccharopolyspora erythraea*	rppA
BGC0001064	cylindrocyclophane D cylindrocyclophane E cylindrocyclophane F	Modular Type I Polyketide Type III Polyketide	*Cylindrospermum licheniforme* UTEX B 2014	AFV96140 AFV96143
BGC0001079	napyradiomycin A80915C	Terpene Type III Polyketide	*Streptomyces aculeolatus*	napT1 napB1
BGC0001083	merochlorin A, merochlorin B deschloro-merochlorin A, deschloro-merochlorin B isochloro-merochlorin B, dichloro-merochlorin B merochlorin D, merochlorin C	Terpene Type III Polyketide	*Streptomyces sp.* CNH189	mcl3 mcl4 mcl17
BGC0001150	pamamycin-607	Type II Polyketide Type III Polyketide	*Streptomyces alboniger*	pamG pamD pamE pamK

Figure 2.S10. Negative control KS sequence selection.

Figure S10. **Negative control KS sequence selection.** All sequence accession information is listed in Table S13.

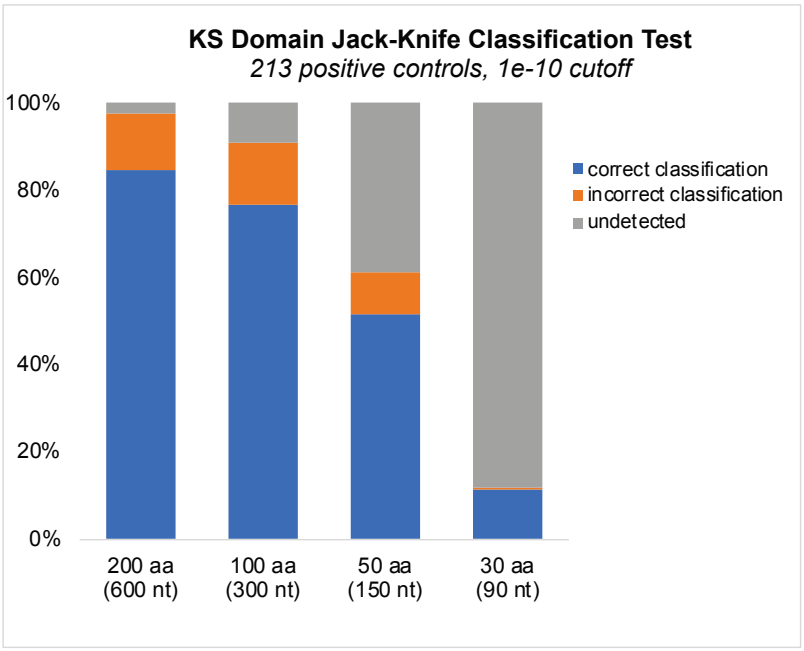
A) Distance matrix tree of 1,072 condensing enzyme superfamily (cl09938) sequences from the NCBI Conserved Domain Database[6] (CDD) tool CDtree[7], colored by conserved domain (CD) family in iTOL[8] (see panel B and C keys). Stars mark CD families not represented in the NaPDoS2 database, which were selected as negative controls (634 total). The negative control CD families comprise: initiating condensing enzymes (init_cond_enzymes, cd00827; n=84), chalcone and stilbene synthases (CHS_like, cd00831; n=67), decarboxylating condensing enzymes (decarbox_cond_enzymes, cd00825; n=3), ketoacyl-acyl carrier protein synthase III enzymes (KAS_III, cd00830; n=130), sterol carrier protein (SCP)-x isoform-associated thiolase domains (SCP-x_thiolase, cd00829; n=125), non-decarboxylating condensing enzymes (nondecarbox_cond_enzymes, cd00826; n=2), and thiolase enzymes (thiolase, cd00751; n=223).

B) Phylogenetic tree of the 1,877 NaPDoS2 KS sequences colored by CD family as determined by NCBI CD-Search[6] (see key). These sequences, which are associated with experimentally verified PKS and FAS biosynthetic gene clusters (BGCs), were selected as positive controls. The 1,877 NaPDoS2 KS sequences were aligned using MUSCLE[9]; the phylogenetic tree was calculated using FastTreeMP[10] on the CIPRES Science Gateway[11] and visualized in iTOL[8]. The NaPDoS2 database positive control CD families comprise: the condensing enzymes subfamily (cond_enzymes, cd00327; n=39), the elongating condensing enzyme subfamily (elong_cond_enzymes, cd00828; n=5), the chain-length factor subfamily (CLF, cd00832; n=70), the polyketide synthase PKS subfamily (PKS, cd00833; 1,649), the beta-ketoacyl-acyl carrier protein synthase (KAS) type I and II subfamily (KAS_I_II, cd00834; n=102), the N-terminal domain beta-ketoacyl synthase pfam13723 superfamily (cl24048; n=5), and the polyketide-type polyunsaturated fatty acid synthase omega-3 PfaA TIGR02813 superfamily (cl37173; n=7).

C) Phylogenetic tree of 61 condensing enzyme superfamily sequences from Jiang *et al* 2008[12], colored by conserved domain family as determined by NCBI CD-Search[6] (see panel B and C keys). Of these, 49 sequences (all but the 12 sequences colored yellow and grey) were added to the pool of negative control sequences. The 61 sequences were aligned using MUSCLE[9]; ProtTest 3.4.2[13] was used to define a model; the phylogeny was calculated with RAxML[14] WAG+G with 200 bootstraps; and the resulting tree visualized in iTOL[8].

D) Fourteen KS domains from the six experimentally characterized type III PKS BGCs in the MIBiG 2.0[5] repository were also added to the pool of negative control sequences. Sequence names are colored by their CD family as determined by NCBI CD-Search[6].

Figure S11.



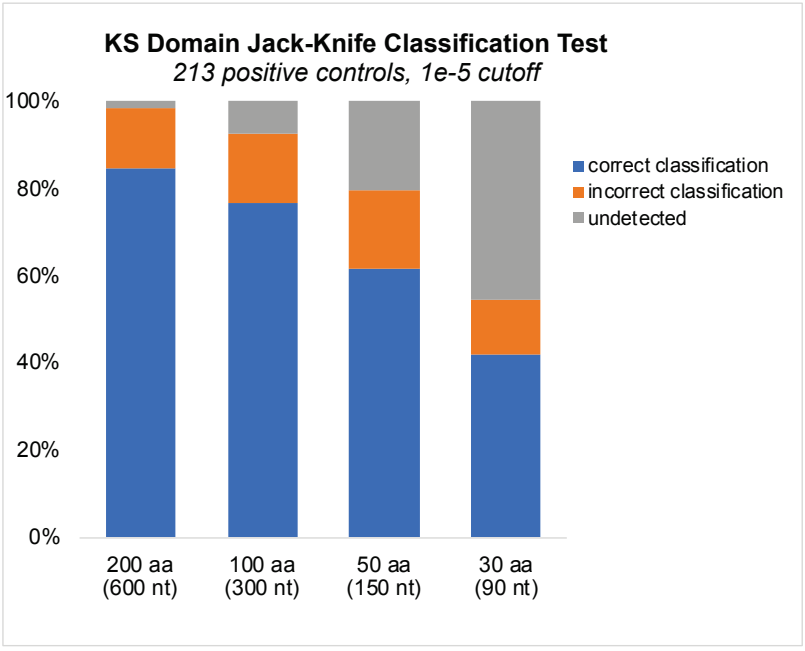


Figure S11. **Effect of query size on KS detection and classification accuracy at various E-values.** Classifications were based on varying BLASTP e-value cutoff scores as indicated for the closest non-self database match. Test sequences of varying lengths were obtained as overlapping sliding window subsequences covering the full length of 213 non-redundant, positive control KS domains.

Figure 2.S11. Effect of query size on KS detection and accuracy.

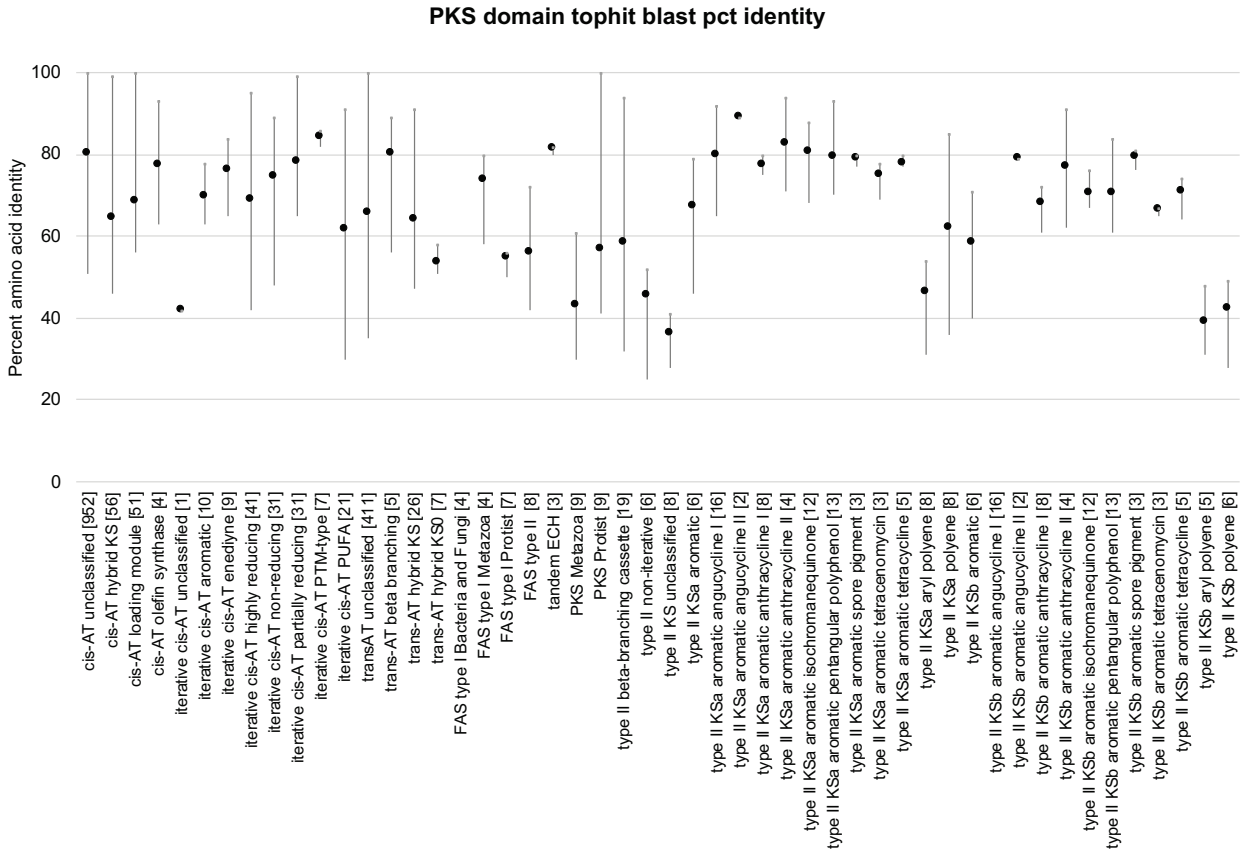


Figure S12. KS domain sequence diversity. Percent sequence identities were determined for each KS class in the NaPDoS2 database using an all-against-all BLASTP comparison. Mean values (points) and ranges (vertical lines) are shown for top non-self matches within each NaPDoS2 class (x-axis). Bracketed values indicate the total number of database sequences in that class.

Figure 2.S12. KS domain sequence diversity.

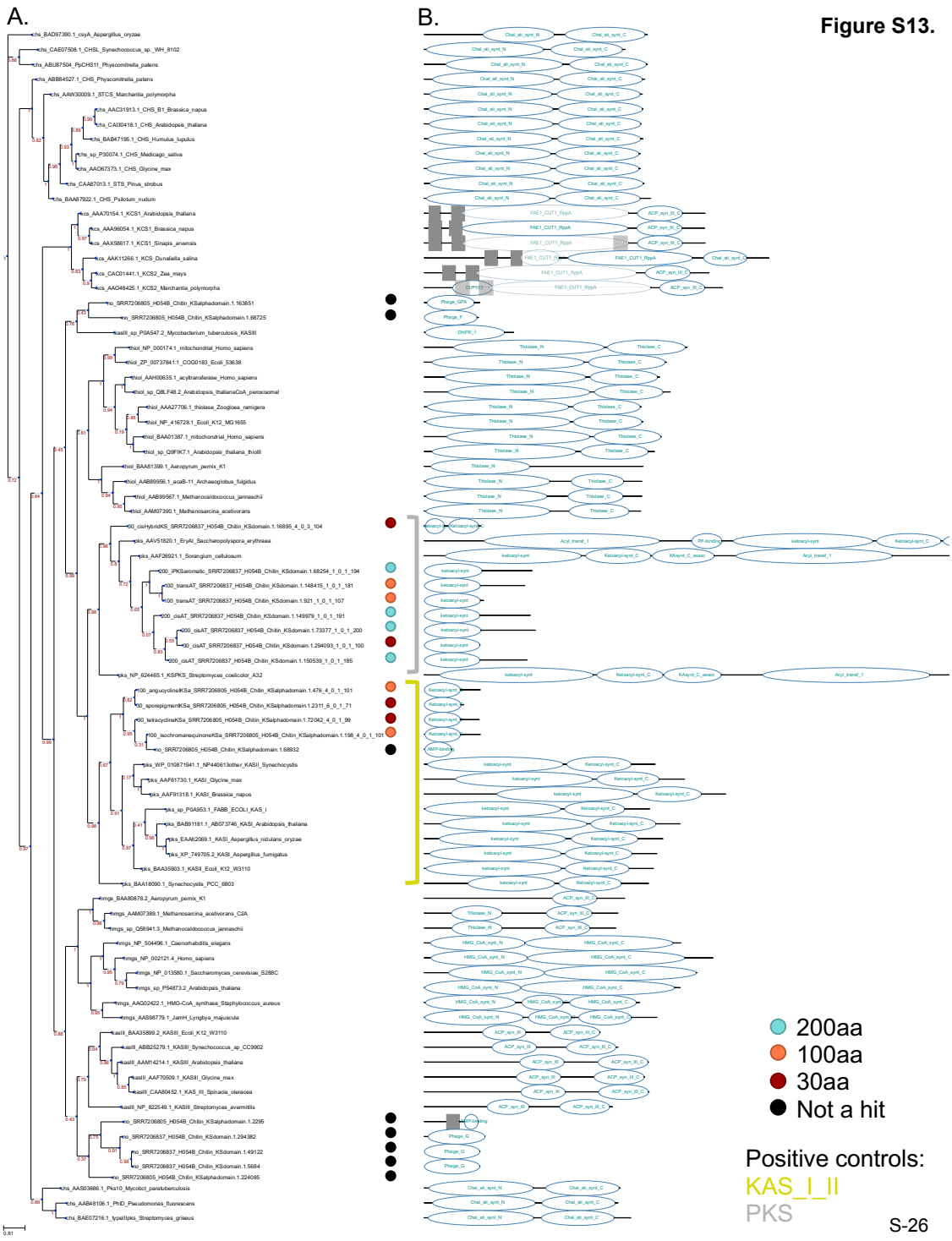
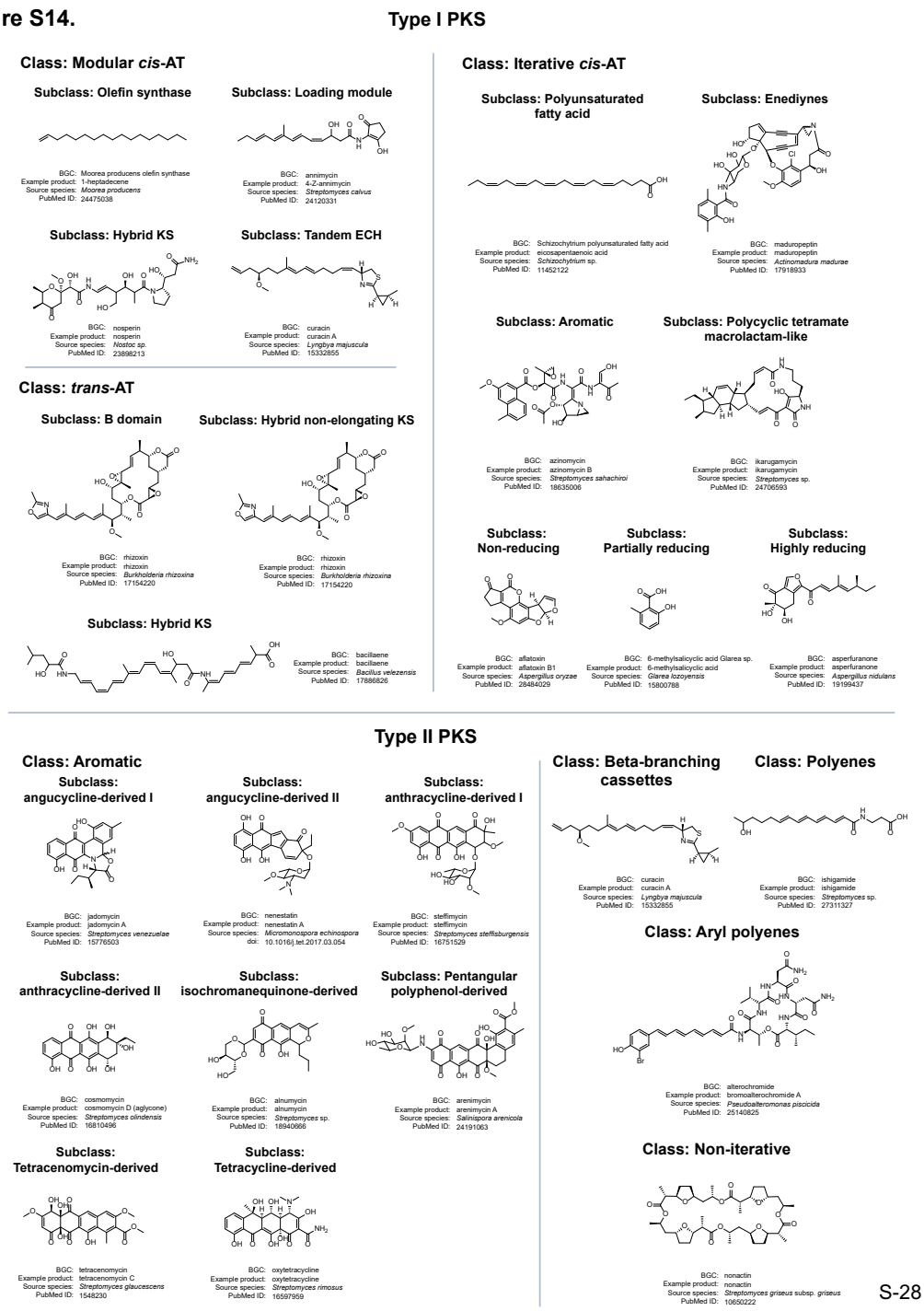


Figure 2.S13. Amplicon detection accuracy.

Figure S13. **Amplicon detection accuracy.** A) Phylogenetic tree of 20 amplicons from Elfeki *et al.* 2018[15] (bolded) and 61 condensing enzyme superfamily sequences from Jiang *et al.* 2008[12]. The 20 amplicons from Elfeki *et al.* 2018[15] (bolded) were detected by NaPDoS2 at 30aa, 100aa, or 200aa minimum alignment lengths (red, orange, or teal circle, respectively) or not detected (not a "hit") at any alignment length (black circle). Grey and yellow brackets include positive control KS sequences as described in Figure S10. B) The conserved domains of all sequences as defined by TREND[16] (http://trend.zhulinlab.org/) from the NCBI conserved domain database are illustrated (gray squares are predicted transmembrane regions). The example "KS" amplicons not detected by NaPDoS2 are off-target amplifications with conserved domains relating to phage proteins. The KS amplicons detected by NaPDoS2 cluster within the PKS and KAS_I_II conserved domain positive control sequences from type I and II PKSs, respectively.

Abbreviations in (A) taxa branches from Jiang *et al.* 2008[12] are as listed in Figure S10; sequences from Elfeki *et al.* 2018[15] are annotated with the minimum amino acid alignment length setting/or "no" for "not a hit", NaPDoS2 classification, SRA dataset, sample information, and NaPDoS2 domain range.
Abbreviations in (B) conserved domains as defined by TREND[16].

Figure S14.



S-28

Figure S14. Example chemical structures for each KS class and subclass in NaPDoS2. Each example includes the compound name, the BGC name in parentheses, the species name for the biological source, and the PubMed ID. For each structure, the relevant KS domain can be found in the BGC tab under the product name on the NaPDoS2 website.

Figure 2.S14. Example chemical structures for each KS class/subclass.

Table 2.S1. Processing times for NaPDoS release (V1) versus NaPDoS2 (V2)

Table S1.

Table S1. Typical NaPDoS pipeline processing times for genomic, PCR amplicon, and metagenomic data sets for the original NaPDoS release (V1) and NaPDoS2 (V2). Initial data upload times are not included, as they vary widely according to internet connection speed. Matches were analyzed using a minimum alignment length of 200 amino acids for all data sets except the 243 nt amplicons, which were screened using a 50 amino acid cutoff value. Dashes indicate missing values for queries that could not be analyzed in NaPDoS version 1 because they exceeded maximum size limitations (50,000 sequences or 30 MB file size). Limits in NaPDoS Version 2 have been increased to 500,000 sequences or 500 MB file size.

Data set	Type	Acession numbers	File size (MB)	Input seqs	V1 sec	V2 sec	Num. domain matches
Salinispora arenicola CNS-205, complete genome	nucleic acid	GCF_000018265.1	5.9	1	106	10	33
Aplysina aerophoba assembed metagenome - random subsample	nucleic acid (assembled contigs)	IMG 3300002222	32.9	1,250	192	13	32
Aplysina aerophoba assembed metagenome - random subsample	nucleic acid (assembled contigs)	IMG 3300002222	46.1	2,500	276	19	38
Aplysina aerophoba assembed metagenome - random subsample	nucleic acid (assembled contigs)	IMG 3300002222	82.5	10,000	-	33	62
Aplysina aerophoba assembed metagenome - random subsample	nucleic acid (assembled contigs)	IMG 3300002222	106.4	20,000	-	42	72
Aplysina aerophoba assembed metagenome - random subsample	nucleic acid (assembled contigs)	IMG 3300002222	141	49,000	-	57	87
Aplysina aerophoba assembed metagenome	nucleic acid (assembled contigs)	IMG 3300002222	211.4	219,427	-	91	94
Salinispora arenicola CNS-205, complete genome	predicted protein	GCF_000018265.1	2.0	4,820	22	6	33
23 draft bacterial genomes (MAGs) - random subsample	predicted protein	PRJNA320446	9.5	25,000	79	4	16
23 draft bacterial genomes (MAGs) - random subsample	predicted protein	PRJNA320446	13.4	35,000	107	6	20
23 draft bacterial genomes (MAGs) - random subsample	predicted protein	PRJNA320446	19.1	50,000	149	7	31
23 draft bacterial genomes (MAGs)	predicted protein	PRJNA320446	28.1	73,530	216	9	49
Aplysina aerophoba assembed metagenome	predicted protein	IMG 3300002222	70.9	365,131	-	41	94
S31 Antarctic soil, KS amplicons - random subsample	nucleic acid amplicon (243 nt)	ERR1527879	1.5	5,000	50	8	3,194
S31 Antarctic soil, KS amplicons	nucleic acid amplicon (243 nt)	ERR1527879	6.5	19,909	198	36	14,024

Table 2.S2. NaPDoS2 database summary.

Class	Subclass		Bacteria	Fungi	Other Eukaryota	Total (across all taxa)
Polyketide Synthases						
type I modular *cis* -AT	olefin synthase		4			4
	loading module		51			51
	hybrid KS		56			56
	tandem ECH		3			3
	no subclass		952			952
		Total	**1066**	**0**	**0**	**1066**
type I iterative *cis* -AT	PUFA (polyunsaturated fatty acids)		18		3	21
	enediynes		9			9
	aromatic		10			10
	PTM-type (polycyclic tetramate macrolactam)		7			7
	non-reducing			31		31
	partially reducing			7		7
	highly reducing			41		41
	no subclass		1			1
		Total	**45**	**79**	**3**	**127**
type I *trans* -AT	beta-branching module		5			5
	hybrid KS		19			19
	hybrid KS0 (non-elongating KS)		7			7
	no subclass		411			411
		Total	**442**	**0**	**0**	**442**
type I Metazoa-type PKS	no subclass				9	9
type I Protist-type PKS	no subclass				9	9
		Total	**0**	**0**	**18**	**18**
		Total	**1553**	**79**	**21**	**1653**
type II aromatic	angucycline-derived I		32			32
	angucycline-derived II		4			4
	anthracycline-derived I		16			16
	anthracycline-derived II		8			8
	isochromanequinone-derived		24			24
	pentangular polyphenol-derived		26			26
	tetracenomycin-derived		6			6
	tetracycline-derived		10			10
	spore pigment		6			6
	unclassified		12			12
		Total	**144**	**0**	**0**	**144**
type II beta-branching cassettes	no subclass		19			19
type II polyenes	KSa, KSb		14			14
type II aryl polyenes	KSa, KSb		13			13
type II non-iterative	no subclass		6			6
type II unclassified	no subclass		5			5
		Total	**57**	**0**	**0**	**57**
		Total	**201**	**0**	**0**	**201**
Fatty acid synthases						
type I FAS	Bacterial-Fungal-type		2	2		4
	Metazoan-type				4	4
	Protist-type				7	7
		Total	**2**	**2**	**11**	**15**
type II FAS	no subclass		8			8
		Total	**8**	**0**	**0**	**8**
		Total	**10**	**2**	**11**	**23**
		Grand Total	**1764**	**81**	**32**	**1877**

Table 2.S3. Accession numbers and dataset references (Excel file).

Table S3. **Accession numbers for negative controls, application use cases, type II aromatic KS sequences, and associated references.** The first tab "Negative Control Sequences" lists the accession numbers for the 697 negative control sequences listed by source: NCBI Conserved Domain family outside, Jiang *et al.* 2008[12] tree, and MIBiG 2.0[5] type III PKS BGCs. The second tab "Application Use Case Accessions" lists the sequence/dataset accession numbers for the application use cases (listed by biological source type, data type, and relevant reference). The third tab "TypeIIAromaticKS_ AccessNum_Refs" lists the GenBank protein ID for alpha and beta sequences, together with literature references used for the biosynthetic annotations in Figure 3 (i.e. poly-beta-keto chain length, starting unit, C-C cyclisation position).

(provided as a Microsoft Excel file).

Separate Microsoft Excel file not included in this dissertation but is accessible on the OSF project page: https://osf.io/uzhcp/

Table 2.S4. *Salinispora* spp. type II KS domains.

Table S4. ***Salinispora* spp. type II KS domains identified by NaPDoS version 1 (NaPDoS1)[1] and NaPDoS2.** 118 *Salinispora* genomes[17] were analyzed using the following default settings: NaPDoS1: HMM 1e-5 cutoff, 200aa minimum alignment length, pathway assignment BLASTP e-value 1e-5 cutoff; NaPDoS2: e-value cutoff 1e-8 and 200aa minimum alignment length. The NaPDoS1 output is limited to the total number of type II KSs detected. The NaPDoS2 output includes the total number of type II KSs ("Total NP2") and their subclassification. KSα and KSβ sequences are grouped together.

Abbreviations: Betabranch= type II beta-branching cassettes; Angl-I= type II aromatic angucycline-derived I; Angl-II= type II aromatic angucycline-derived II; Anth-I= type II aromatic anthracycline-derived I; Isochrom= type II aromatic isochromanequinone-derived; Polyphen= type II aromatic pentangular polyphenol-derived; Unclass= type II aromatic unclassified; Type II-uc= type II unclassified no subclass.

Table S4.

		NaPDoS1	**NaPDoS2**									
						Aromatic						
Species	**# genomes**	Type II	*Total NP2*	Betabranch	Polyene	Angl-I	Angl-II	Anth-I	Isochrom	Polyphen	Unclass	Type II-uc
S. tropica	12	48	108	-	60	-	24	-	-	24	-	-
S. fenicalii	2	12	24	-	12	4	4	-	-	4	-	-
S. cortesiana	1	4	5	-	1	-	2	-	-	2	-	-
S. mooreana	3	8	15	-	7	-	-	-	2	6	-	-
S. oceanensis	12	44	44	-	-	8	-	6	4	26	-	-
S. goodfellowii	1	4	5	-	1	-	2	-	-	2	-	-
S. vitiensis	3	12	28	-	14	-	6	-	-	6	-	2
S. pacifica	23	97	126	-	27	-	46	-	4	46	-	3
S. arenicola	61	134	307	2	173	-	-	4	-	126	2	-

Table 2.S5. Complete list of *Salinispora* spp. KS domains identified by NaPDoS2.

Table S5. **Complete list of *Salinispora* spp. KS domains identified by NaPDoS2.** 118 *Salinispora* genomes[17] were analyzed using default settings (e-value cutoff 1e-8 and 200aa minimum alignment length). Table lists the number of KSs detected and their classification by strain. The summed total number of KSs found in each strain is listed as "NP2 Total". KSα and KSβ sequences are grouped together.

Abbreviations: Betabranch= type II beta-branching cassettes; Angl-I= type II aromatic angucycline-derived I; Angl-II= type II aromatic angucycline-derived II; Anth-I= type II aromatic anthracycline-derived I; Isochrom= type II aromatic isochromanequinone-derived; Polyphen= type II aromatic pentangular polyphenol-derived; Unclass= type II aromatic unclassified; Type II-uc= type II unclassified no subclass; *cis*-AT= type I modular *cis*-AT; cisloading= type I modular *cis*-AT loading module; cisHybridKS= type I modular *cis*-AT hybrid KS; Ened= type I iterative *cis*-AT enediynes; iPKSaromatic= type I iterative *cis*-AT aromatic; PTM= type I iterative *cis*-AT PTM-type (polycyclic tetramate macrolactam); *trans*-AT= type I *trans*-AT no subclass; *trans*-hybridKS0= type I *trans*-AT hybrid KS0 (non-elongating KS).

Table S5.

				NaPDoS2: KS classes																
			FAS	Type II PKS									Type I PKS							
		NP2				Aromatic								*cis*-AT modular		*cis*-AT iterative			*trans*-AT	
Species	*Strain*	*Total*	FAS II	Betabranch	Polyene	AngI-I	AngI-II	Anth-I	Isochrom	Polyphen	Unclass	TypeII-uc	*cis*-AT	cisloading	cisHybridKS	Ened	iPKSaromatic	PTM	*trans*-AT	*trans*-hybridKS0
S. tropica	CNB440	28	2		5		2			2			12	1	2	2				
S. tropica	CNB476	22	2		5		2			2			6	1	3	1				
S. tropica	CNB536	29	3		5		2			2			12	1	3	1				
S. tropica	CNH898	27	4		5		2			2			10	1	2	1				
S. tropica	CNR699	24	2		5		2			2			8	1	2	2				
S. tropica	CNS197	24	2		5		2			2			8	1	2	2				
S. tropica	CNS416	24	2		5		2			2			9	1	2	1				
S. tropica	CNT250	24	2		5		2			2			8	1	2	2				
S. tropica	CNT261	27	3		5		2			2			8	1	5	1				
S. tropica	CNY012	24	3		5		2			2			8	1	2	1				
S. tropica	CNY678	24	2		5		2			2			8	1	2	2				
S. tropica	CNY681	24	2		5		2			2			8	1	2	2				
S. fenicalii	CNT569	30	2		6	2	2			2			2	1	2	2			9	
S. fenicalii	CNR942	29	2		6	2	2			2			2	1	2	1			9	
S. cortesiana	CNY202	13	4		1		2			2			2		1	1				
S. mooreana	CNY646	24	5		5					2			7	2	2	1				
S. mooreana	CNS237	48	4		1					2			34	3	2	2				
S. mooreana	CNT150_DSM45549	13	4		1				2	2			2			2				
S. oceanensis	CNT854	10	4							2			1		1	1	1			
S. oceanensis	CNT584	13	4			4				2			1		1	1				
S. oceanensis	CNT124	13	4			4				2			1		1	1				
S. oceanensis	CNT138_DSM45547	30	4							4			17	1	2	1			1	
S. oceanensis	CNT029	9	4							2			1		1	1				
S. oceanensis	CNY703	9	4							2			1		1	1				
S. oceanensis	CNY673	11	4					2		2			1		1	1				
S. oceanensis	CNT045	21	6							2			4	4	3	1	1			
S. oceanensis	CNS996	23	6					2		2			4	5	3	1				
S. oceanensis	CNT403	16	5					2		2			2	2	2	1				
S. oceanensis	CNS860	15	5						2	2			1	2	2	1				
S. oceanensis	CNS863_DSM45543	16	5						2	2			2	2	2	1				
S. goodfellowii	CNY666	50	2		1		2			2			38	1	2	1			1	
S. vitiensis	CNS055	20	1		4		2			2			9		1	1				
S. vitiensis	CNT148_DSM45548	15	3		5		2			2		1	1		1					
S. vitiensis	CNS801	15	3		5		2			2		1	1		1					
S. pacifica	CNH732	20	2		1		2			2			9		3	1				
S. pacifica	CNQ768	19	2		1		2			2			8		3	1				
S. pacifica	CNR114	20	2		1		2			2			10	1	1	1				
S. pacifica	CNR510	23	2		1		2			2			11	1	3	1				
S. pacifica	CNR894	20	2		1		2			2			9		3	1				
S. pacifica	CNR909	21	3		1		2			2			10		1	1	1			
S. pacifica	CNS103	23	3		1		2			2			11	1	2	1				
S. pacifica	CNT001	23	2		1		2			2			11	1	3	1				
S. pacifica	CNT003	28	3		1		2			2			16	1	1	1			1	
S. pacifica	CNT084	28	4		1		2			2			14	2	1	1	1			
S. pacifica	CNT131	30	2		1		2			2			17	1	3	1			1	
S. pacifica	CNT133	29	4		1		2			2			15	2	1	1	1			
S. pacifica	CNT603	20	2		1		2			2			9		3	1				
S. pacifica	CNT609	26	3		1		2			2		1	13	2	1	1				
S. pacifica	CNT796	21	2		1		2		2	2			10		1	1				
S. pacifica	CNT851	20	2		1		2		2	2			9		1	1				
S. pacifica	CNT855	21	2		2		2			2		1	10		1	1				
S. pacifica	CNY239	20	2		1		2			2			8		4	1				
S. pacifica	CNY330	32	2		1		2			2			23		1	1				
S. pacifica	CNY331	37	2		2		2			2			26	1	1	1				
S. pacifica	CNY363	45	2		2		2			2			33	1	1	1			1	
S. pacifica	CNY498	22	2		1		2			2			10		3	1			1	
S. pacifica	CNS960_DSM45544	28	2		2		2			2		1	14		4	1				
S. arenicola	CNB458	24	2		2					2			13		1	2	2			
S. arenicola	CNB527	29	2		2					4			16		1	2	2			
S. arenicola	CNH643	36	3		4					2			21	1	2	2	1			
S. arenicola	CNH646	21	3		4					2			7		2	2	1			
S. arenicola	CNH713	34	3		4					2			19	1	2	2	1			
S. arenicola	CNH718	33	2		2					2			20	1	1	2	2	1		
S. arenicola	CNH877	29	3		4					2			15		2	2	1			
S. arenicola	CNH905	28	3		4					2			15		1	2	1			
S. arenicola	CNH941	23	3		4					2			7	1	3	2	1			
S. arenicola	CNH962	17	2		2					2			8			2	1			
S. arenicola	CNH963	18	2		2					2			9			2	1			
S. arenicola	CNH964	31	3	1	4					2			7	1	3	2	1		6	1
S. arenicola	CNH996B	26	3		2					2			13	1		2	2	1		
S. arenicola	CNH996	26	3		2					2			13	1		2	2	1		
S. arenicola	CNP105	31	3	1	4					2			7	1	3	2	1		6	1
S. arenicola	CNP193	23	3		4					2			7	1	3	2	1			
S. arenicola	CNQ748	38	2		2					2			24	1	2	2	2	1		
S. arenicola	CNQ884	50	2		2					2			37	2	1	2	2			
S. arenicola	CNR107	22	2		3					2			8		2	2	2	1		
S. arenicola	CNR425	33	2		2					2			20	1	2	2	2			
S. arenicola	CNR921	15	2		2					2			4		1	2	2			
S. arenicola	CNS051	16	2		2					2			5		1	2	2			
S. arenicola	CNS205	33	2		2					2			20	1	1	2	2	1		

Table S5.

				NaPDoS2: KS classes																	
			FAS	Type II PKS										Type I PKS							
		NP2				Aromatic								*cis*-AT modular		*cis*-AT iterative			*trans*-AT		
Species	*Strain*	*Total*	FAS II	Betabranch	Polyene	AngI-I	AngI-II	Anth-I	Isochrom	Polyphen	Unclass	TypeII-uc	*cis*-AT	cisloading	cisHybridKS	Ened	iPKSaromatic	PTM	*trans*-AT	*trans*-hybridKS0	
S. arenicola	CNS243	50	2		2					2			36	2	1	2	2	1			
S. arenicola	CNS296	39	2		2					2			24	2	2	2	2	1			
S. arenicola	CNS299	35	2		2					2			22	2	1	2	2				
S. arenicola	CNS325	35	2		2					2	2		19	2	2	2	2				
S. arenicola	CNS342	32	2		2					2			20	1	1	2	2				
S. arenicola	CNS673	28	2		2					2			15		3	2	2				
S. arenicola	CNS744	27	2		2					4			12		2	2	2	1			
S. arenicola	CNS820	34	2		2					2			20	1	2	2	2	1			
S. arenicola	CNS848	38	3		4					2			20	3	3	2	1				
S. arenicola	CNT005	28	2		2					2			16	1	1	2	2				
S. arenicola	CNT798	28	3		4					2			14		2	2	1				
S. arenicola	CNT799	34	3		4					2			19	1	2	2	1				
S. arenicola	CNT800	30	3		4					2			16		2	2	1				
S. arenicola	CNT849	29	3		4					2			15		2	2	1				
S. arenicola	CNT850	27	3		4					2			13		2	2	1				
S. arenicola	CNT857	30	3		4					2			16		2	2	1				
S. arenicola	CNT859	29	3		4					2			16		1	2	1				
S. arenicola	CNX481	20	2		2					2			8		1	2	2	1			
S. arenicola	CNX482	19	2		2					2			7		1	2	2	1			
S. arenicola	CNX508	28	2		2					2			16		1	2	2	1			
S. arenicola	CNX814	21	2		2			2		2			7		1	2	2	1			
S. arenicola	CNX891	20	2		2			2		2			6		1	2	2	1			
S. arenicola	CNY011	30	3		4					2			16		2	2	1				
S. arenicola	CNY230	40	2		6					2			22	2	2	2	2				
S. arenicola	CNY231	30	2		2					2			18	1	1	2	2				
S. arenicola	CNY234	25	2		2					2			14		1	2	2				
S. arenicola	CNY237	25	2		2					2			13		1	2	2	1			
S. arenicola	CNY244	32	2		2					2			19	1	1	2	2	1			
S. arenicola	CNY256	23	2		2					2			12		1	2	2				
S. arenicola	CNY260	27	2		2					2			15		1	2	2	1			
S. arenicola	CNY280	40	3		8					2			19	2	3	2	1				
S. arenicola	CNY282	23	2		2					2			12		1	2	2				
S. arenicola	CNY486	24	2		2					2			13	1		2	2				
S. arenicola	CNY679	36	3		4					2			21	1	2	2	1				
S. arenicola	CNY685	31	2		2					2			19	1	1	2	2				
S. arenicola	CNY690	30	2		2					2			18	1	1	2	2				
S. arenicola	CNY694	30	2		2					2			18	1	1	2	2				
S. arenicola	CNS991_DSM45545	40	3		4					2			24	2	2	2	1				
	total	*3103*	*313*	*2*	*295*	*12*	*84*	*10*	*10*	*242*	*2*	*5*	*1490*	*92*	*200*	*186*	*104*	*18*	*36*	*2*	

Table 2.S6. KS domains identified in 27 fungal genomes by NaPDoS2.

Table S6. **KS domains identified in 27 fungal genomes by NaPDoS2.** Table lists the total number of KSs detected using default settings (e-value cutoff 1e-8 and 200aa minimum alignment length) and their class and subclass distributions. The summed total number of KSs found in each genome is listed as "Total NP2". The number of protein sequences in each analyzed genome file is listed as "# protein seq".

Abbreviations: bfFASI= type I FAS Bacterial-Fungal-type; FASII= type II FAS; *cis*-AT= type I modular *cis*-AT; cisHybridKS= type I modular *cis*-AT hybrid KS; PR= type I iterative *cis*-AT partially reducing; NR= type I iterative *cis*-AT non-reducing; HR= type I iterative *cis*-AT highly reducing.

Table S6.

					Type I PKS				
					cis-AT modular		*cis*-AT iterative		
Fungal genome	**# protein seq**	*Total NP2*	bfFASI	FASII	*cis*-AT	cisHybridKS	PR	NR	HR
Aspergillus niger ATCC 1015	12,885	50	5	1	1	-	1	8	34
Aspergillus sergii CBS 130017	13,713	39	5	1	-	-	1	12	20
Aspergillus nidulans FGSC A4	9,556	37	5	1	1	-	-	14	16
Penicillium rolfsii F1880	9,955	26	2	1	-	-	-	6	17
Penicillium rubens Wisconsin 54-1255	12,791	25	2	1	-	-	2	3	17
Sclerotinia sclerotiorum 1980 UF-70	14,490	20	1	1	-	-	-	4	14
Leptosphaeria maculans JN3	12,469	17	2	1	-	-	-	4	10
Nectria haematococca MPVI isolate 77-13-4	15,708	15	1	1	-	-	-	2	11
Alternaria alternata SRC1lrK2f	13,466	14	1	1	-	1	-	3	8
Zymoseptoria tritici IPO323	10,941	14	1	1	-	-	-	2	10
Neurospora crassa OR74A	10,812	11	2	1	-	-	-	1	7
Arthrobotrys oligospora TWF154	13,042	9	1	1	-	-	1	-	6
Ustilago maydis 521	6,782	5	2	-	-	-	-	3	-
Allomyces macrogynus ATCC 38327	19,447	4	2	2	-	-	-	-	-
Laccaria bicolor S238N-H82	18,215	3	1	-	1	-	-	1	-
Mucor circinelloides 1006PhL	12,227	3	2	1	-	-	-	-	-
Coprinopsis cinerea okayama 7#130 CC3	13,356	2	1	-	-	-	-	1	-
Pyronema omphalodes CBS 100304	13,367	3	1	1	-	-	-	1	-
Schizophyllum commune H4-8	13,193	3	2	-	-	-	-	1	-
Batrachochytrium dendrobatidis JAM81	8,677	2	1	1	-	-	-	-	-
Debaryomyces hansenii CBS767	6,286	2	1	1	-	-	-	-	-
Neolecta irregularis DAH-3	5,579	2	1	1	-	-	-	-	-
Puccinia graminis tritici CRL 75-36-700-3	15,979	2	2	-	-	-	-	-	-
Saccharomyces cerevisiae S288C	6,002	2	1	1	-	-	-	-	-
Saitoella complicata NRRL Y-17804	7,023	2	1	1	-	-	-	-	-
Schizosaccharomyces pombe 972h-	5,132	2	1	1	-	-	-	-	-
Malassezia globosa CBS 7966	4,286	1	-	-	1	-	-	-	-

Table 2.S7. KS detection using NaPDoS versions 1 and 2.

Table S7. **KS detection using NaPDoS versions 1 and 2.**
MIBiG Fungal PKS: NaPDoS2 detected all of the KS domains in 159 MIBiG 2.0[5] fungal PKS BGCs. NaPDoS version 1 and 2 analyses were run with the default settings.
Wawrik 2005 KS clones: While both NaPDoS versions 1 and 2 detected all 147 type II KS amplicons from Wawrik *et al*. 2005[18], NaPDoS2 could further delineate these sequences into three subclasses. NaPDoS version 1 and 2 analyses were run with the following settings: NaPDoS version 1: HMM 1e-5, 200aa minimum alignment length, pathway assignment: e-value cutoff of 1e-5; NaPDoS2: e-value cutoff 1e-8 and 50aa minimum alignment length. The summed total number of KSs found in each genome is listed as "Total NP1" and "Total NP2" for NaPDoS versions 1 and 2, respectively.

Abbreviations: Iter= Iterative type I; FA= fatty acid synthase; bfFASI= type I FAS Bacterial-Fungal-type; *cis*-AT= type I modular *cis*-AT; PR= type I iterative *cis*-AT partially reducing; NR= type I iterative *cis*-AT non-reducing; HR= type I iterative *cis*-AT highly reducing; Aro-KSα= type II aromatic unclassified KSα; Angl-I-KSα= type II aromatic angucycline-derived I KSα; Polyphen-KSα= type II aromatic pentangular polyphenol-derived KSα.

Table S7.

	NaPDoS1				**NaPDoS2**		Type I PKS				Type II PKS		
							cis-AT modular	*cis*-AT iterative			Aromatic		
Sequence collection	*Total NP1*	Type II	Iter	FA	*Total NP2*	bfFASI	*cis*-AT	PR	NR	HR	Aro-KSα	Angl-I-KSα	Polyphen-KSα
MIBiG Fungal PKS (159 BGCs)	14	-	7	7	182	10	4	7	71	90	-	-	-
Wawrik 2005 KS clones (147 seqs)	147	147	-	-	147	-	-	-	-	-	1	45	101

Table 2.S8. NaPDoS2 analysis of the *Elysia chlorotica* genome.

Table S8. **NaPDoS2 analysis of the *Elysia chlorotica* genome.** NaPDoS2 identified nine KSs from the *Elysia chlorotica* CDSs (fna), protein (faa), and translated CDSs (faa) genomes[19]. Highlighted KSs were identified in Torres *et al*. 2020[20] as being associated with new FAS-like animal PKSs (EcPKS1, EcPKS2) and an FAS (EcFAS). The analyses were run with NaPDoS2 default settings (e-value cutoff 1e-8 and 200aa minimum alignment length).

Table S8.

Match	Query ID	Database match ID	% ID	Align length	E-value	BGC Match	Domain Class	Domain Subclass
EcFAS	RUS90834.1	OryziasFAS_KS01_MetazoaFASI	60	403	5.7e-152	Oryzias latipes FAS	type I FAS	Metazoan-type
EcPKS1	RUS71288.1	HomoFAS_KS01_MetazoaFASI	52	406	9.6e-119	Homo sapiens FAS	type I FAS	Metazoan-type
	RUS77019.1	OryziasFAS_KS01_MetazoaFASI	49	403	2.7e-117	Oryzias latipes FAS	type I FAS	Metazoan-type
EcPKS2	RUS75294.1	OryziasFAS_KS01_MetazoaFASI	49	403	4.7e-117	Oryzias latipes FAS	type I FAS	Metazoan-type
	RUS92164.1	MelopsittacusFAS_KS01_MetazoaFASI	39	410	2.9e-76	Melopsittacus undulatus FAS	type I FAS	Metazoan-type
	RUS75295.1	OryziasFAS_KS01_MetazoaFASI	45	308	6.3e-73	Oryzias latipes FAS	type I FAS	Metazoan-type
	RUS68442.1	AliivibrioAPE_KS03_FASII	60	407	1.3e-137	Aliivibrio fischeri aryl polyene	type II FAS	no subclass
	RUS69056.1	EscherichiaFAS_KS02_FASII	57	420	1.9e-136	Escherichia coli FAS	type II FAS	no subclass
	RUS77541.1	EscherichiaFAS_KS02_FASII	47	421	4.8e-103	Escherichia coli FAS	type II FAS	no subclass

Table 2.S9. Moorea sediment metagenomes analyzed with NaPDoS2.

Table S9. **Moorea sediment metagenomes analyzed with NaPDoS2.** KS domains from 20 marine sediment metagenomes[21] were assigned to 26 different subclasses. NaPDoS2 analyses were run with default settings (e-value cutoff 1e-8 and 200aa minimum alignment length). KSα and KSβ sequences are grouped together.

Abbreviations: MetazoaFASI= type I FAS Metazoan-type; ProtistFASI= type I FAS Protist-type; FASII= type II FAS no subclass; Betabranch= type II beta-branching cassettes; Polyene= type II polyenes; Ape= type II aryl polyenes; Aro= type II aromatic unclassified; Angl= type II aromatic angucycline-derived I & II; Tetcyc= type II aromatic tetracycline-derived; Anth= type II aromatic anthracycline-derived I & II; Isochrom= type II aromatic isochromanequinone-derived; Tetcen= type II aromatic tetracenomycin-derived; Polyphen= type II aromatic pentangular polyphenol-derived; SPKS= type II aromatic spore pigment; *cis*-AT= type I modular *cis*-AT; cisloading= type I modular *cis*-AT loading module; cisHybridKS= type I modular *cis*-AT hybrid KS; cisOLS= type I modular *cis*-AT olefin synthase; cistandemECH= type I modular *cis*-AT tandem ECH; iPKS= type I iterative *cis*-AT no subclass; iPKSPUFA= type I iterative *cis*-AT PUFA (polyunsaturated fatty acids); Ened= type I iterative *cis*-AT enediynes; iPKSaromatic= type I iterative *cis*-AT aromatic; PTM= type I iterative *cis*-AT PTM-type (polycyclic tetramate macrolactam); HR= type I iterative *cis*-AT highly reducing; *trans*-AT= type I *trans*-AT no subclass; *trans*-hybridKS= type I *trans*-AT hybrid KS; *trans*-hybridKS0= type I *trans*-AT hybrid KS0 (non-elongating KS).

Table S9.

				Type II PKS															Type I PKS														
				FAS I					Aromatic										*cis*-AT modular					*cis*-AT iterative							*trans*-AT		
Metagenome	***# contigs***	***N50***	***Total NP2***	MetazoaFASI	ProtistFASI	FASII	MetazoaPKS	Betabranch	Polyene	Ape	Aro	AngI	Tetcyc	Anth	Isochrom	Tetcen	Polyphen	SPKS	*cis*-AT	cisloading	cisHybridKS	cisOLS	cistandemECH	iPKS	iPKSPUFA	Ened	iPKSaromatic	PTM	HR	*trans*-AT	*trans*-hybridKS	*trans*-hybridKS0
MO18_007	152,894	5,918	145	-	-	61	-	-	5	3	1	5	-	-	-	-	-	-	19	3	5	1	-	1	34	-	-	-	-	7	-	-
MO18_010	184,034	3,212	150	-	-	54	-	1	4	5	2	2	-	1	1	1	-	-	17	1	14	-	-	3	34	1	-	-	-	9	-	-
MO18_039	336,958	1,769	58	-	-	13	4	-	1	1	-	1	-	-	1	-	-	-	8	-	17	-	-	-	3	1	-	-	-	8	-	-
MO18_042	378,383	1,596	121	-	-	60	-	1	1	5	1	2	-	-	-	-	-	-	13	3	3	2	-	2	22	-	-	1	-	5	-	-
MO18_071	37,150	2,093	87	-	1	23	-	-	-	2	-	1	-	-	-	-	1	-	24	-	9	3	-	-	5	-	-	1	-	16	-	1
MO18_074	211,343	2,041	312	-	-	149	1	-	5	8	6	6	-	-	-	1	4	1	41	1	28	3	-	8	39	4	-	-	1	6	-	-
MO18_103	186,468	1,851	269	-	1	114	-	1	6	13	3	1	-	-	1	1	6	-	34	3	15	3	1	4	46	3	-	-	-	13	-	-
MO18_106	440,441	1,847	48	-	-	12	-	-	-	-	-	-	-	-	-	-	-	-	19	1	-	6	-	-	7	-	-	-	-	3	-	-
MO18_135	230,831	2,120	270	-	1	91	-	4	1	5	4	5	1	1	-	-	2	-	34	2	35	6	3	2	32	6	-	-	-	35	-	-
MO18_138	104	1,300	0	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
MO18_167	311,846	2,402	706	-	-	272	-	7	5	21	12	13	-	4	1	1	8	1	83	11	48	18	3	8	129	18	2	1	-	39	1	-
MO18_170	234,614	1,716	336	-	-	192	-	1	5	20	6	7	-	1	1	2	2	-	16	1	8	2	-	3	60	6	-	1	-	2	-	-
MO18_180	182,647	1,599	227	-	-	125	-	1	5	12	-	4	-	-	1	-	3	-	13	1	10	4	1	5	29	3	-	-	-	10	-	-
MO18_184	309,098	3,457	51	-	-	11	-	1	-	3	1	1	-	-	-	-	1	-	17	1	-	5	-	-	5	-	-	-	-	5	-	-
MO18_188	259,773	2,959	505	-	-	188	1	5	5	9	4	10	1	-	-	-	7	-	106	4	41	20	3	7	53	8	1	1	-	30	1	-
MO18_192	241,995	1,727	341	-	-	116	2	4	8	8	2	5	-	3	-	-	4	-	60	3	34	10	2	3	49	8	-	-	-	20	-	-
MO18_199	229,239	2,042	200	3	-	77	-	4	2	3	1	2	-	1	1	-	2	-	46	3	2	1	4	3	31	2	-	-	-	12	-	-
MO18_202	242,491	2,136	280	-	1	122	-	2	5	10	5	8	-	-	-	1	3	-	41	3	14	2	3	3	48	2	-	-	-	7	-	-
MO18_231	110,677	2,357	91	-	-	30	-	1	-	5	-	2	-	-	-	-	1	-	22	1	2	-	1	-	20	1	-	-	-	5	-	-
MO18_234	182,155	2,286	170	-	3	62	-	2	-	-	2	4	-	-	-	-	-	-	37	-	13	6	1	1	20	2	-	-	-	16	1	-

Table 2.S10. NaPDoS2 analysis of an eSNaPD v2.0 dataset.

Table S10. NaPDoS2 analysis of KS amplicon eSNaPD v2.0 data. NaPDoS2 detected and classified all 381 sequences from a New Mexico desert soil KS amplicon library (NM_KS_ARRY_LIB01, Owen *et al.* 2013 PNAS[22]). Additionally, NaPDoS2 classified all but one group of the 756 uncharacterized "Novel Clusters 1-60" from the same eSNaPD v2.0 soil amplicon library. NaPDoS2 settings: e-value cutoff 1e-8 and 50aa minimum alignment length (due to the amplicon sequence length). The "% ID" column lists the percent of sequences that were classified by NaPDoS2 (i.e. "Total NP2"/"# seq"; color scale of light grey (low % ID) to dark grey (high % ID)).

Abbreviations: *cis*-AT= type I modular *cis*-AT; cisloading= type I modular *cis*-AT loading module; cisHybridKS= type I modular *cis*-AT hybrid KS; cisOLS= type I modular *cis*-AT olefin synthase; cistandemECH= type I modular *cis*-AT tandem ECH; iPKSaromatic= type I iterative *cis*-AT aromatic; PTM= type I iterative *cis*-AT PTM-type (polycyclic tetramate macrolactam); NR= type I iterative *cis*-AT non-reducing; HR= type I iterative *cis*-AT highly reducing; *trans*-AT= type I *trans*-AT no subclass; *trans*-hybridKS= type I *trans*-AT hybrid KS.

Table S10.

				Type I PKS										
				cis-AT modular					*cis*-AT iterative				*trans*-AT	
% ID	**Sequence set**	**# seq**	*Total NP2*	*cis*-AT	cisloading	cisHybridKS	cisOLS	cistandemECH	iPKSaromatic	PTM	NR	HR	*trans*-AT	*trans*-hybridKS
102%	NM_KS_ARRY_LIB01	381	388	310	7	15	2	-	5	13	-	-	36	-
99%	NovClst1	192	191	2	-	-	-	-	-	-	189	-	-	-
100%	NovClst2	42	42	4	15	-	15	-	-	4	-	4	-	-
100%	NovClst3	38	38	3	-	-	-	-	-	35	-	-	-	-
94%	NovClst4	31	29	-	-	-	-	-	-	-	-	-	29	-
79%	NovClst5	29	23	-	-	22	-	-	-	-	-	-	-	1
108%	NovClst6	25	27	-	-	-	-	-	-	-	27	-	-	-
100%	NovClst7	22	22	-	-	22	-	-	-	-	-	-	-	-
112%	NovClst8	17	19	-	-	18	-	-	-	-	-	-	-	1
86%	NovClst9	14	12	-	-	11	-	-	-	-	-	-	-	1
92%	NovClst10	12	11	8	-	-	-	-	-	-	-	-	3	-
100%	NovClst11	12	12	6	-	-	-	-	-	-	-	-	6	-
100%	NovClst12	12	12	-	-	-	-	-	-	12	-	-	-	-
92%	NovClst13	12	11	-	-	11	-	-	-	-	-	-	-	-
110%	NovClst14	10	11	4	-	-	3	-	-	-	-	-	4	-
100%	NovClst15	10	10	-	-	-	10	-	-	-	-	-	-	-
100%	NovClst16	9	9	-	-	9	-	-	-	-	-	-	-	-
89%	NovClst17	9	8	7	-	-	1	-	-	-	-	-	-	-
111%	NovClst18	9	10	10	-	-	-	-	-	-	-	-	-	-
78%	NovClst19	9	7	5	-	-	-	1	-	-	-	-	1	-
89%	NovClst20	9	8	-	-	7	-	-	-	-	-	-	-	1
89%	NovClst21	9	8	8	-	-	-	-	-	-	-	-	-	-
100%	NovClst22	8	8	-	-	-	-	-	-	-	8	-	-	-
113%	NovClst23	8	9	9	-	-	-	-	-	-	-	-	-	-
113%	NovClst24	8	9	1	-	-	-	-	-	-	-	8	-	-
88%	NovClst25	8	7	-	-	-	-	-	-	-	-	-	7	-
100%	NovClst26	8	8	-	-	8	-	-	-	-	-	-	-	-
100%	NovClst27	8	8	8	-	-	-	-	-	-	-	-	-	-
114%	NovClst28	7	8	-	-	-	-	-	-	-	8	-	-	-
86%	NovClst29	7	6	-	-	6	-	-	-	-	-	-	-	-
100%	NovClst30	7	7	-	-	7	-	-	-	-	-	-	-	-
100%	NovClst31	7	7	4	-	-	3	-	-	-	-	-	-	-
100%	NovClst32	7	7	-	-	7	-	-	-	-	-	-	-	-
86%	NovClst33	7	6	-	-	-	-	-	-	-	-	-	6	-
100%	NovClst34	7	7	-	-	7	-	-	-	-	-	-	-	-
100%	NovClst35	6	6	1	1	-	-	-	-	-	-	-	4	-
67%	NovClst36	6	4	-	-	-	-	-	-	-	-	-	4	-
100%	NovClst37	6	6	-	-	6	-	-	-	-	-	-	-	-
83%	NovClst38	6	5	-	-	5	-	-	-	-	-	-	-	-
100%	NovClst39	6	6	6	-	-	-	-	-	-	-	-	-	-
67%	NovClst40	6	4	-	-	4	-	-	-	-	-	-	-	-
100%	NovClst41	6	6	6	-	-	-	-	-	-	-	-	-	-
0%	NovClst42	6	0	-	-	-	-	-	-	-	-	-	-	-
100%	NovClst43	6	6	1	-	-	5	-	-	-	-	-	-	-
100%	NovClst44	5	5	-	-	-	-	-	-	-	-	-	5	-
100%	NovClst45	5	5	-	-	5	-	-	-	-	-	-	-	-
100%	NovClst46	5	5	5	-	-	-	-	-	-	-	-	-	-
140%	NovClst47	5	7	2	-	-	4	-	-	-	-	1	-	-
60%	NovClst48	5	3	-	-	-	-	-	-	-	-	-	3	-
100%	NovClst49	4	4	-	-	4	-	-	-	-	-	-	-	-
100%	NovClst50	4	4	-	-	4	-	-	-	-	-	-	-	-
100%	NovClst51	4	4	-	-	4	-	-	-	-	-	-	-	-
100%	NovClst52	4	4	-	-	-	-	-	-	-	-	-	4	-
25%	NovClst53	4	1	1	-	-	-	-	-	-	-	-	-	-
100%	NovClst54	4	4	4	-	-	-	-	-	-	-	-	-	-
100%	NovClst55	4	4	-	-	4	-	-	-	-	-	-	-	-
100%	NovClst56	4	4	-	-	-	-	-	-	-	-	-	4	-
100%	NovClst57	4	4	-	-	4	-	-	-	-	-	-	-	-
75%	NovClst58	4	3	3	-	-	-	-	-	-	-	-	-	-
75%	NovClst59	4	3	-	-	3	-	-	-	-	-	-	-	-
175%	NovClst60	4	7	-	-	-	-	-	-	-	7	-	-	-

Table 2.S11. NaPDoS2 analysis of amplicon datasets from Borsetto *et al.* 2019.

Table S11. **NaPDoS2 analysis of 12 type II KS amplicon datasets from Borsetto *et al.* 2019[23].** The total number of amplicon sequences not detected by NaPDoS2 represented 36-95% of the sequences in the amplicon libraries ("# seqs"). Of the KS sequences that were detected by NaPDoS2, 19-93% were classified as type II PKSs and could be assigned to a wide range of type II subclasses while the remainder were classified as type II fatty acid synthases (FASII). NaPDoS2 settings: e-value cutoff 1e-8 and 50aa minimum alignment length (due to the amplicon sequence length).

Abbreviations: FASII= type II FAS no subclass; Betabranch= type II beta-branching cassettes; Polyene-KSα= type II polyenes KSα; Ape-KSα= type II aryl polyenes KSα; Aro-KSα= type II aromatic unclassified KSα; Angl-I-KSα= type II aromatic angucycline-derived I KSα; Angl-II-KSα= type II aromatic angucycline-derived II KSα; Tetcyc-KSα= type II aromatic tetracycline-derived KSα; Anth-I-KSα= type II aromatic anthracycline-derived I KSα; Anth-I-KSβ= type II aromatic anthracycline-derived I KSβ; Anth-II-KSα= type II aromatic anthracycline-derived II KSα; Isochrom-KSα= type II aromatic isochromanequinone-derived KSα; Tetcen-KSα= type II aromatic tetracenomycin-derived KSα; Tetcen-KSβ= type II aromatic tetracenomycin-derived KSβ; Polyphen-KSα= type II aromatic pentangular polyphenol-derived KSα; Polyphen-KSβ= type II aromatic pentangular polyphenol-derived KSβ; SPKS-KSα= type II aromatic spore pigment KSα; *cis*-AT= type I modular *cis*-AT; cisHybridKS= type I modular *cis*-AT hybrid KS.

Table S11.

				Type II PKS																Type I PKS	
							Aromatic													*cis*-AT modular	
Dataset	**# seqs**	***Total NP2***	FASII	Betabranch	Polyene-KSα	Ape-KSα	Aro-KSα	Angl-I-KSα	Angl-II-KSα	Tetcyc-KSα	Anth-I-KSα	Anth-I-KSβ	Anth-II-KSα	Isochrom-KSα	Tetcen-KSα	Tetcen-KSβ	Polyphen-KSα	Polyphen-KSβ	SPKS-KSα	*cis*-AT	cisHybridKS
S31_Antarctica	34,600	12,757	9,593	6	54	2	63	419	163	13	10	1	21	201	75	3	333	20	1,780	-	-
S32_Antarctica	31,118	7,905	6,414	4	50	12	52	369	51	13	42	1	11	134	47	1	343	16	344	1	-
S33_Antarctica	14,703	774	328	-	8	-	18	264	1	1	20	-	-	31	30	2	18	9	44	-	-
S22_AlgeriaB3	166,814	9,345	1,172	1	177	15	1,253	1,003	37	4	2,573	-	76	366	461	-	1,370	-	836	1	-
S23_AlgeriaB3	119,567	19,317	3,155	4	429	38	836	4,648	22	39	4,687	-	106	418	954	-	1,212	-	2,768	1	-
S24_AlgeriaB3	126,771	21,643	3,072	-	373	26	1,354	4,149	27	4	6,824	-	148	414	1,075	-	2,097	-	2,080	-	-
S28_CubaFir	182,418	13,220	3,605	-	308	17	931	2,718	9	10	1,174	-	90	851	48	-	527	-	2,931	1	-
S29_CubaFir	194,450	18,079	2,394	-	146	12	1,279	7,113	8	3	3,331	-	18	1,174	18	-	896	-	1,687	-	-
S30_CubaFir	204,614	30,393	2,701	-	1,056	11	993	7,993	6	8	1,822	-	194	4,820	39	-	973	-	9,768	9	-
S37_Warwick	241,475	115,324	19,560	8	560	30	3,964	15,941	160	20	6,114	-	269	38,472	38	-	8,868	17	21,276	23	4
S38_Warwick	110,892	58,876	7,728	4	321	22	4,390	6,556	124	1	3,158	-	269	7,026	38	-	4,913	11	24,312	1	2
S39_Warwick	135,287	86,200	5,716	1	118	10	2,760	12,604	344	6	2,936	-	600	33,933	60	-	4,122	15	22,974	-	1

Table 2.S12. NaPDoS2 analysis of amplicon sequences from Elfeki *et al.* 2018

Table S12. **NaPDoS2 analysis of 5,000 randomly selected KS amplicon sequences from the Elfeki *et al.* 2018[15] "Chitin" type I** (SRR7206837_H054B_Chitin_KSdomain) **and type II** (SRR7206805_H054B_Chitin_KSalphadomain) **KS datasets run at varying minimum amino acid alignment lengths.** Decreasing the minimum amino acid alignment length increases the number of KSs detected but also increases the likelihood of false positives and misclassifications, as may be evidenced by the detection of type I KSs in the type II dataset. Notably, below an alignment length of 100aa, the number of KSs detected in the type I dataset exceeds the number of sequences analyzed, as can be expected when shorter domain fragment hits are identified from the same longer amplicon sequences.

Table S12.

Type II KSα: Chitin				
Alignment Length	Class	Subclass	NP2 Hits	*Total NP2 hits*
200aa	**-**	-	0	0
150aa	type II aromatic	pentangularpolyphenolKSa	1	1
100aa	type II aromatic	angucycline I KSa	469	1,229
	type II aromatic	isochromanequinone KSa	225	
	type II aromatic	pentangular polyphenol KSa	200	
	type II aromatic	unclassified KSa	146	
	type II aromatic	tetracenomycin KSa	111	
	type II aromatic	anthracycline I KSa	70	
	type I modular cis-AT	no subclass	2	
	type II aromatic	angucycline II KSa	2	
	type II aromatic	spore pigment KSa	1	
	type I trans-AT	hybrid KS	1	
	type I trans-AT	no subclass	1	
	type II aromatic	anthracycline II KSa	1	
50aa	type II aromatic	angucycline I KSa	958	2,999
	type II aromatic	pentangular polyphenol KSa	673	
	type II aromatic	isochromanequinone KSa	570	
	type II aromatic	unclassified KSa	378	
	type II aromatic	tetracenomycin KSa	221	
	type II aromatic	anthracycline I KSa	132	
	type II aromatic	spore pigment KSa	31	
	type I modular cis-AT	no subclass	20	
	type II aromatic	anthracycline II KSa	4	
	type II aromatic	tetracycline KSa	4	
	type II aromatic	angucycline II KSa	4	
	type I modular cis-AT	hybrid KS	2	
	type I trans-AT	hybrid KS	1	
	type I trans-AT	no subclass	1	
30aa	type II aromatic	angucycline I KSa	994	3,099
	type II aromatic	pentangular polyphenol KSa	693	
	type II aromatic	isochromanequinone KSa	584	
	type II aromatic	unclassified KSa	381	
	type II aromatic	tetracenomycin KSa	224	
	type II aromatic	anthracycline I KSa	138	
	type II aromatic	spore pigment KSa	40	
	type I modular cis-AT	no subclass	22	
	type II aromatic	tetracycline KSa	9	
	type II aromatic	anthracycline II KSa	5	
	type II aromatic	angucycline II KSa	5	
	type I modular cis-AT	hybrid KS	2	
	type I trans-AT	hybrid KS	1	
	type I trans-AT	no subclass	1	

Type I KS: Chitin				
Alignment Length	Class	Subclass	NP2 Hits	*Total NP2 hits*
200aa	type I modular cis-AT	no subclass	21	23
	type I iterative cis-AT	aromatic	2	
150aa	type I modular cis-AT	no subclass	385	423
	type I iterative cis-AT	aromatic	24	
	type I trans-AT	no subclass	8	
	type I trans-AT	hybrid KS	6	
100aa	type I modular cis-AT	no subclass	3,406	3,594
	type I trans-AT	no subclass	75	
	type I iterative cis-AT	aromatic	57	
	type I modular cis-AT	hybrid KS	39	
	type I trans-AT	hybrid KS	15	
	type II aromatic	pentangular polyphenol KSa	1	
	type II aromatic	spore pigment KSa	1	
50aa	type I modular cis-AT	no subclass	7,441	7,768
	type I trans-AT	no subclass	139	
	type I modular cis-AT	hybrid KS	85	
	type I iterative cis-AT	aromatic	72	
	type I trans-AT	hybrid KS	17	
	type II aromatic	pentangular polyphenol KSa	5	
	type II aromatic	angucycline I KSa	3	
	type II aromatic	spore pigment KSa	2	
	type II aromatic	anthracycline I KSa	1	
	type II aromatic	isochromanequinone KSa	1	
	type II aromatic	tetracenomycin KSa	1	
	type II aromatic	unclassified KSa	1	
30aa	type I modular cis-AT	no subclass	7,600	7,930
	type I trans-AT	no subclass	140	
	type I modular cis-AT	hybrid KS	86	
	type I iterative cis-AT	aromatic	72	
	type I trans-AT	hybrid KS	17	
	type II aromatic	pentangular polyphenol KSa	5	
	type II aromatic	angucycline I KSa	3	
	type II aromatic	spore pigment KSa	3	
	type II aromatic	anthracycline I KSa	1	
	type II aromatic	isochromanequinone KSa	1	
	type II aromatic	tetracenomycin KSa	1	
	type II aromatic	unclassified KSa	1	

2.10 Acknowledgements

The authors acknowledge James Wang and Ghulam Mustafa for assistance in identifying sequences added to the database. We also thank Dulce Guillén-Matus and Jeong Sang Yi for providing useful feedback on beta-versions of the webtool.

Chapter 2, in full, is a reprint of the materials as it was submitted to the *Journal of Biological Chemistry*. Klau, L.J.; Podell, S.; Creamer, K.E.; Demko, A.M.; Singh, H.W.; Allen, E.E.; Moore, B.S.; Ziemert, N.; Letzel, A.C.; Jensen, P.R., 2022. The dissertation author was one of three equally contributing primary authors of this manuscript.

2.11 References

Agrawal, P., Amir, S., Deepak, Barua, D., and Mohanty, D. (2021) RiPPMiner-Genome: A Web Resource for Automated Prediction of Crosslinked Chemical Structures of RiPPs by Genome Mining. *J Mol Biol* **433**: 166887.

Almeida, H., Tsang, A., and Diallo, A.B. (2019) Supporting supervised learning in fungal Biosynthetic Gene Cluster discovery: new benchmark datasets. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. pp. 1280–1287.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990) Basic local alignment search tool. *J Mol Biol* **215**: 403–10.

Austin, M.B., Saito, T., Bowman, M.E., Haydock, S., Kato, A., Moore, B.S., Kay, R.R., and Noel, J.P. (2006) Biosynthesis of Dictyostelium discoideum differentiation-inducing factor by a hybrid type I fatty acid-type III polyketide synthase. *Nat Chem Biol* **2**: 494–502.

Bachmann, B.O. and Ravel, J. (2009) Chapter 8 Methods for In Silico Prediction of Microbial Polyketide and Nonribosomal Peptide Biosynthetic Pathways from DNA Sequence Data, 1st ed. Elsevier Inc.

Bauman, K.D., Butler, K.S., Moore, B.S., and Chekan, J.R. (2021) Genome mining methods to discover bioactive natural products. *Nat Prod Rep*.

Bauman, K.D., Shende, V. V., Chen, P.Y.-T., Trivella, D.B.B., Gulder, T.A.M., Vellalath, S., Romo, D., and Moore, B.S. (2022) Enzymatic assembly of the salinosporamide γ-lactam-β-lactone anticancer warhead. *Nat Chem Biol*.

Beedessee, G., Hisata, K., Roy, M.C., Satoh, N., and Shoguchi, E. (2015) Multifunctional polyketide synthase genes identified by genomic survey of the symbiotic dinoflagellate, Symbiodinium minutum. *BMC Genomics* **16**: 1–11.

Blin, K., Shaw, S., Kloosterman, A.M., Charlop-Powers, Z., van Wezel, G.P., Medema, M.H., and Weber, T. (2021) antiSMASH 6.0: improving cluster detection and comparison capabilities. *Nucleic Acids Res* 1–7.

Borsetto, C., Amos, G.C.A., Nunes da Rocha, U., Mitchell, A.L., Finn, R.D., Laidi, R.F., Vallin, C., Pearce, D.A., Newsham, K.K., and Wellington, E.M.H. (2019) Microbial community drivers of PK / NRP gene diversity in selected global soils. *Microbiome* 1–11.

Bretschneider, T., Heim, J.B., Heine, D., Winkler, R., Busch, B., Kusebauch, B., Stehle, T., Zocher, G., and Hertweck, C. (2013) Vinylogous chain branching catalysed by a dedicated polyketide synthase module. *Nature* **502**: 124–128.

Buchfink, B., Xie, C., and Huson, D.H. (2015) Fast and sensitive protein alignment using

DIAMOND. *Nat Methods* **12**: 59–60.

Cai, H., Li, Q., Fang, X., Li, J., Curtis, N.E., Altenburger, A., Shibata, T., Feng, M., Maeda, T., Schwartz, J.A., Shigenobu, S., Lundholm, N., Nishiyama, T., Yang, H., Hasebe, M., Li, S., Pierce, S.K., and Wang, J. (2019) Data descriptor: A draft genome assembly of the solar-powered sea slug elysia chlorotica. *Sci Data* **6**: 1–13.

Cao, S., Blodgett, J.A.V., and Clardy, J. (2010) Targeted discovery of polycyclic tetramate macrolactams from an environmental Streptomyces strain. *Org Lett* **12**: 4652–4654.

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

Chen, A., Re, R.N., and Burkart, M.D. (2018) Type II fatty acid and polyketide synthases: deciphering protein-protein and protein-substrate interactions. *Nat Prod Rep* **35**: 1029–1045.

Chen, H. and Du, L. (2016) Iterative polyketide biosynthesis by modular polyketide synthases in bacteria. *Appl Microbiol Biotechnol* **100**: 541–557.

Chen, I.-M.A., Chu, K., Palaniappan, K., Pillay, M., Ratner, A., Huang, J., Huntemann, M., Varghese, N., White, J.R., Seshadri, R., Smirnova, T., Kirton, E., Jungbluth, S.P., Woyke, T., Eloe-Fadrosh, E.A., Ivanova, N.N., and Kyrpides, N.C. (2018) IMG/M v.5.0: an integrated data management and comparative analysis system for microbial genomes and microbiomes. *Nucleic Acids Res* **47**: D666–D677.

Chooi, Y.H. and Tang, Y. (2012) Navigating the fungal polyketide chemical space: From genes to molecules. *J Org Chem* **77**: 9933–9953.

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., and Fischbach, M.A. (2014) Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters. *Cell* **158**: 412–421.

Coates, R.C., Podell, S., Korobeynikov, A., Lapidus, A., Pevzner, P., Sherman, D.H., Allen, E.E., Gerwick, L., and Gerwick, W.H. (2014) Characterization of cyanobacterial hydrocarbon composition and distribution of biosynthetic pathways. *PLoS One* **9**:.

Du, D., Katsuyama, Y., Horiuchi, M., Fushinobu, S., Chen, A., Davis, T.D., Burkart, M.D., and Ohnishi, Y. (2020) Structural basis for selectivity in a highly reducing type II polyketide synthase. *Nat Chem Biol* **16**: 776–782.

Edgar, R.C. (2004) MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**: 1792–1797.

Elfeki, M., Alanjary, M., Green, Stefan J, Ziemert, N., and Murphy, B.T. (2018) Assessing the efficiency of cultivation techniques to recover natural product biosynthetic gene populations from sediment. *ACS Chem Biol*.

Elfeki, M., Alanjary, M., Green, Stefan J., Ziemert, N., and Murphy, B.T. (2018) Assessing the Efficiency of Cultivation Techniques to Recover Natural Product Biosynthetic Gene Populations from Sediment. *ACS Chem Biol* **13**: 2074–2081.

Fawcett, T. (2006) An introduction to ROC analysis. *Pattern Recognit Lett* **27**: 861–874.

Feng, L., Gordon, M.T., Liu, Y., Basso, K.B., and Butcher, R.A. (2021) Mapping the biosynthetic pathway of a hybrid polyketide-nonribosomal peptide in a metazoan. *Nat Commun* **12**:.

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

Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012) CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**: 3150–3152.

Gallo, A., Ferrara, M., and Perrone, G. (2013) Phylogenetic study of polyketide synthases and nonribosomal peptide synthetases involved in the biosynthesis of mycotoxins. *Toxins (Basel)* **5**: 717–742.

Gerlt, J.A. (2017) Genomic Enzymology: Web Tools for Leveraging Protein Family Sequence-Function Space and Genome Context to Discover Novel Functions. *Biochemistry* **56**: 4293–4308.

Goksuluk, D., Korkmaz, S., Zararsiz, G., and Karaagaoglu, A.E. (2016) EasyROC: An interactive web-tool for roc curve analysis using r language environment. *R J* **8**: 213–230.

Goodwin, S., McPherson, J.D., and McCombie, W.R. (2016) Coming of age: Ten years of next-generation sequencing technologies. *Nat Rev Genet* **17**: 333–351.

Grammbitter, G.L.C., Schmalhofer, M., Karimi, K., Shi, Y.M., Schöner, T.A., Tobias, N.J., Morgner, N., Groll, M., and Bode, H.B. (2019) An Uncommon Type II PKS Catalyzes Biosynthesis of Aryl Polyene Pigments. *J Am Chem Soc* **141**: 16615–16623.

Gumerov, V.M. and Zhulin, I.B. (2020) TREND: a platform for exploring protein function in prokaryotes based on phylogenetic, domain architecture and gene neighborhood analyses. *Nucleic Acids Res* **48**: W72–W76.

Hannigan, G.D., Prihoda, D., Palicka, A., Soukup, J., Klempir, O., Rampula, L., Durcak, J., Wurst, M., Kotowski, J., Chang, D., Wang, R., Piizzi, G., Temesi, G., Hazuda, D.J., Woelk, C.H., and Bitton, D.A. (2019) A deep learning genome-mining strategy for biosynthetic gene cluster prediction. *Nucleic Acids Res* **47**:.

Helfrich, E.J.N., Ueoka, R., Dolev, A., Rust, M., Meoded, R.A., Califano, G., Costa, R., Gugger, M., Steinbeck, C., Moreno, P., and Piel, J. (2019) Automated structure prediction of trans-acyltransferase polyketide synthase products. *Nat Chem Biol* **15**:.

Herbst, D.A., Townsend, C.A., and Maier, T. (2018) The architectures of iterative type I PKS and

FAS. *Nat Prod Rep* **35**: 1046–1069.

Hertweck, C. (2015) Decoding and reprogramming complex polyketide assembly lines: Prospects for synthetic biology. *Trends Biochem Sci* **40**: 189–199.

Hertweck, C. (2009) The biosynthetic logic of polyketide diversity. *Angew Chemie - Int Ed* **48**: 4688–4716.

Hillenmeyer, M.E., Vandova, G.A., Berlew, E.E., and Charkoudian, L.K. (2015) Evolution of chemical diversity by coordinated gene swaps in type II polyketide gene clusters. *Proc Natl Acad Sci* **112**: 201511688.

Jenke-Kodama, H., Sandmann, A., Müller, R., and Dittmann, E. (2005) Evolutionary implications of bacterial polyketide synthases. *Mol Biol Evol* **22**: 2027–2039.

Jiang, C., Kim, S.Y., and Suh, D.Y. (2008) Divergent evolution of the thiolase superfamily and chalcone synthase family. *Mol Phylogenet Evol* **49**: 691–701.

Katoh, K. and Standley, D.M. (2013) MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol Biol Evol* **30**: 772–780.

Kautsar, S.A., Blin, K., Shaw, S., Navarro-Muñoz, J.C., Terlouw, B.R., van der Hooft, J.J.J., van Santen, J.A., Tracanna, V., Suarez Duran, H.G., Pascal Andreu, V., Selem-Mojica, N., Alanjary, M., Robinson, S.L., Lund, G., Epstein, S.C., Sisto, A.C., Charkoudian, L.K., Collemare, J., Linington, R.G., Weber, T., and Medema, M.H. (2020) MIBiG 2.0: a repository for biosynthetic gene clusters of known function. *Nucleic Acids Res* **48**: D454–D458.

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C., Thierer, T., Ashton, B., Meintjes, P., and Drummond, A. (2012) Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **28**: 1647–1649.

Kim, J. and Yi, G.S. (2012) PKMiner: A database for exploring type II polyketide synthases. *BMC Microbiol* **12**: 1–12.

Kim, M.C., Machado, H.A., Jang, K.H., Trzoss, L., Jensen, P.R., and Fenical, W. (2018) Integration of Genomic Data with NMR Analysis Enables Assignment of the Full Stereostructure of Neaumycin B, a Potent Inhibitor of Glioblastoma from a Marine-Derived Micromonospora. *J Am Chem Soc* jacs.8b04848.

Komaki, H. and Harayama, S. (2006) Sequence Diversity of Type-II Polyketide Synthase Genes in Streptomyces. *Actinomycetologica* **20**: 42–48.

Kwon, H.J., Smith, W.C., Scharon, A.J., Sung, H.H., Kurth, M.J., and Shen, B. (2002) C-O bond formation by polyketide synthases. *Science (80-)* **297**: 1327–1330.

Lee, W.C., Choi, S., Jang, A., Yeon, J., Hwang, E., and Kim, Y. (2021) Structural basis of the complementary activity of two ketosynthases in aryl polyene biosynthesis. *Sci Rep* **11**: 1–10.

Lefort, V., Longueville, J.E., and Gascuel, O. (2017) SMS: Smart Model Selection in PhyML. *Mol Biol Evol* **34**: 2422–2424.

Lemoine, F., Correia, D., Lefort, V., Doppelt-Azeroual, O., Mareuil, F., Cohen-Boulakia, S., and Gascuel, O. (2019) NGPhylogeny.fr: New generation phylogenetic services for non-specialists. *Nucleic Acids Res* **47**: W260–W265.

Lemoine, F., Domelevo Entfellner, J.B., Wilkinson, E., Correia, D., Dávila Felipe, M., De Oliveira, T., and Gascuel, O. (2018) Renewing Felsenstein's phylogenetic bootstrap in the era of big data. *Nature* **556**: 452–456.

Lohman, J.R., Ma, M., Osipiuk, J., Nocek, B., Kim, Y., Chang, C., Cuff, M., Mack, J., Bigelow, L., Li, H., Endres, M., Babnigg, G., Joachimiak, A., Phillips, G.N., and Shen, B. (2015) Structural and evolutionary relationships of "AT-less" type I polyketide synthase ketosynthases. *Proc Natl Acad Sci U S A* **112**: 12693–12698.

Lu, S., Wang, J., Chitsaz, F., Derbyshire, M.K., Geer, R.C., Gonzales, N.R., Gwadz, M., Hurwitz, D.I., Marchler, G.H., Song, J.S., Thanki, N., Yamashita, R.A., Yang, M., Zhang, D., Zheng, C., Lanczycki, C.J., and Marchler-Bauer, A. (2020) CDD/SPARCLE: the conserved domain database in 2020. *Nucleic Acids Res* **48**: D265–D268.

Marchler-Bauer, A. and Bryant, S.H. (2004) CD-Search: Protein domain annotations on the fly. *Nucleic Acids Res* **32**: 327–331.

Medema, M.H. (2021) The year 2020 in natural product bioinformatics: An overview of the latest tools and databases. *Nat Prod Rep* **38**: 301–306.

Medema, M.H., de Rond, T., and Moore, B.S. (2021) Mining genomes to illuminate the specialized chemistry of life. *Nat Rev Genet*.

Metsä-Ketelä, M., Halo, L., Munukka, E., Hakala, J., Mäntsäla, P., and Ylihonko, K. (2002) Molecular evolution of aromatic polyketides and comparative sequence analysis of polyketide ketosynthase and 16S ribosomal DNA genes from various Streptomyces species. *Appl Environ Microbiol* **68**: 4472–4479.

Metz, J.G., Roessler, P., Facciotti, D., Levering, C., Dittrich, F., Lassner, M., Valentine, R., Lardizabal, K., Domergue, F., Yamada, A., Yazawa, K., Knauf, V., and Browse, J. (2001) Production of polyunsaturated fatty acids by potyketide synthases in both prokaryotes and eukaryotes. *Science (80-)* **293**: 290–293.

Millán-Aguiñaga, N., Chavarria, K.L., Ugalde, J.A., Letzel, A.-C., Rouse, G.W., and Jensen, P.R. (2017) Phylogenomic Insight into *Salinispora* (Bacteria, Actinobacteria) Species Designations. *Sci Rep* **7**: 3564.

Moffitt, M.C. and Neilan, B.A. (2003) Evolutionary affiliations within the superfamily of ketosynthases reflect complex pathway associations. *J Mol Evol* **56**: 446–457.

Mungan, M.D., Alanjary, M., Blin, K., Weber, T., Medema, M.H., and Ziemert, N. (2020) ARTS

2.0: feature updates and expansion of the Antibiotic Resistant Target Seeker for comparative genome mining. *Nucleic Acids Res* **48**: 546–552.

Nakamura, H., Hamer, H.A., Sirasani, G., and Balskus, E.P. (2012) Cylindrocyclophane biosynthesis involves functionalization of an unactivated carbon center. *J Am Chem Soc* **134**: 18518–18521.

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

Nurk, S., Koren, S., and Rhie, A. (2022) The complete sequence of a human genome. *Science (80-)* **376**: 44–53.

Oliver, A., Podell, S., Pinowska, A., Traller, J.C., Smith, S.R., McClure, R., Beliaev, A., Bohutskyi, P., Hill, E.A., Rabines, A., Zheng, H., Allen, L.Z., Kuo, A., Grigoriev, I. V., Allen, A.E., Hazlebeck, D., and Allen, E.E. (2021) Diploid genomic architecture of Nitzschia inconspicua, an elite biomass production diatom. *Sci Rep* **11**: 1–14.

Owen, J.G., Reddy, B.V.B., Ternei, M.A., Charlop-Powers, Z., Calle, P.Y., Kim, J.H., and Brady, S.F. (2013) Mapping gene clusters within arrayed metagenomic libraries to expand the structural diversity of biomedically relevant natural products. *Proc Natl Acad Sci U S A* **110**: 11797–11802.

Price, M.N., Dehal, P.S., and Arkin, A.P. (2010) FastTree 2 - Approximately maximum-likelihood trees for large alignments. *PLoS One* **5**:.

Rebets, Y., Brötz, E., Manderscheid, N., Tokovenko, B., Myronovskyi, M., Metz, P., Petzke, L., and Luzhetskyy, A. (2015) Insights into the pamamycin biosynthesis. *Angew Chemie - Int Ed* **54**: 2280–2284.

Reddy, B.V. ija. B., Milshteyn, A., Charlop-Powers, Z., and Brady, S.F. (2014) eSNaPD: a versatile, web-based bioinformatics platform for surveying and mining natural product biosynthetic diversity from metagenomes. *Chem Biol* **21**: 1023–1033.

Rice, P., Longden, L., and Bleasby, A. (2000) EMBOSS: The European Molecular Biology Open Software Suite. *Trends Genet* **16**: 276–277.

Robinson, S.L., Piel, J., and Sunagawa, S. (2021) A roadmap for metagenomic enzyme discovery. *Nat Prod Rep*.

Schmitt, I. and Lumbsch, H.T. (2009) Ancient horizontal gene transfer from bacteria enhances biosynthetic capabilities of fungi. *PLoS One* **4**: 1–8.

Schöner, T.A., Gassel, S., Osawa, A., Tobias, N.J., Okuno, Y., Sakakibara, Y., Shindo, K., Sandmann, G., and Bode, H.B. (2016) Aryl Polyenes, a Highly Abundant Class of Bacterial Natural Products, Are Functionally Related to Antioxidative Carotenoids. *ChemBioChem* **17**:

247–253.

Schorn, M.A., Verhoeven, S., Ridder, L., Huber, F., Acharya, D.D., Aksenov, A.A., Aleti, G., Moghaddam, J.A., Aron, A.T., Aziz, S., Bauermeister, A., Bauman, K.D., Baunach, M., Beemelmanns, C., Beman, J.M., Berlanga-Clavero, M.V., Blacutt, A.A., Bode, H.B., Boullie, A., Brejnrod, A., Bugni, T.S., Calteau, A., Cao, L., Carrión, V.J., Castelo-Branco, R., Chanana, S., Chase, A.B., Chevrette, M.G., Costa-Lotufo, L. V., Crawford, J.M., Currie, C.R., Cuypers, B., Dang, T., de Rond, T., Demko, A.M., Dittmann, E., Du, C., Drozd, C., Dujardin, J.C., Dutton, R.J., Edlund, A., Fewer, D.P., Garg, N., Gauglitz, J.M., Gentry, E.C., Gerwick, L., Glukhov, E., Gross, H., Gugger, M., Guillén Matus, D.G., Helfrich, E.J.N., Hempel, B.F., Hur, J.S., Iorio, M., Jensen, P.R., Kang, K. Bin, Kaysser, L., Kelleher, N.L., Kim, C.S., Kim, K.H., Koester, I., König, G.M., Leao, T., Lee, S.R., Lee, Y.Y., Li, X., Little, J.C., Maloney, K.N., Männle, D., Martin H, C., McAvoy, A.C., Metcalf, W.W., Mohimani, H., Molina-Santiago, C., Moore, B.S., Mullowney, M.W., Muskat, M., Nothias, L.F., O'Neill, E.C., Parkinson, E.I., Petras, D., Piel, J., Pierce, E.C., Pires, K., Reher, R., Romero, D., Roper, M.C., Rust, M., Saad, H., Saenz, C., Sanchez, L.M., Sørensen, S.J., Sosio, M., Süssmuth, R.D., Sweeney, D., Tahlan, K., Thomson, R.J., Tobias, N.J., Trindade-Silva, A.E., van Wezel, G.P., Wang, M., Weldon, K.C., Zhang, F., Ziemert, N., Duncan, K.R., Crüsemann, M., Rogers, S., Dorrestein, P.C., Medema, M.H., and van der Hooft, J.J.J. (2021) A community resource for paired genomic and metabolomic data mining. *Nat Chem Biol* **17**: 363–368.

Shen, B. (2003) Polyketide biosynthesis beyond the type I, II and III polyketide synthase paradigms. *Curr Opin Chem Biol* **7**: 285–295.

Sigrist, R., Luhavaya, H., McKinnie, S.M.K., Ferreira da Silva, A., Jurberg, I.D., Moore, B.S., and Gonzaga de Oliveira, L. (2020) Nonlinear Biosynthetic Assembly of Alpiniamide by a Hybrid cis/trans-AT PKS-NRPS. *ACS Chem Biol*.

Skinnider, M.A., Johnston, C.W., Gunabalasingam, M., Merwin, N.J., Kieliszek, A.M., MacLellan, R.J., Li, H., Ranieri, M.R.M., Webster, A.L.H., Cao, M.P.T., Pfeifle, A., Spencer, N., To, Q.H., Wallace, D.P., Dejong, C.A., and Magarvey, N.A. (2020) Comprehensive prediction of secondary metabolite structure and biological activity from microbial genome sequences. *Nat Commun* **11**: 1–9.

Tang, Y., Tsai, S.C., and Khosla, C. (2003) Polyketide Chain Length Control by Chain Length Factor. *J Am Chem Soc* **125**: 12708–12709.

Torres, J.P., Lin, Z., Winter, J.M., Krug, P.J., and Schmidt, E.W. (2020) Animal biosynthesis of complex polyketides in a photosynthetic partnership. *Nat Commun* **11**: 1–12.

Villebro, R., Shaw, S., Blin, K., and Weber, T. (2019) Sequence-based classification of type II polyketide synthase biosynthetic gene clusters for antiSMASH. *J Ind Microbiol Biotechnol* 1–7.

Walczak, R.J., Woo, A.J., Strohl, W.R., and Priestley, N.D. (2000) Nonactin biosynthesis: The potential nonactin biosynthesis gene cluster contains type II polyketide synthase-like genes. *FEMS Microbiol Lett* **183**: 171–175.

Walker, P.D., Weir, A.N.M., Willis, C.L., and Crump, M.P. (2021) Polyketide β-branching: Diversity, mechanism and selectivity. *Nat Prod Rep* **38**: 723–756.

Wawrik, B., Kerkhof, L., Zylstra, G.J., and Kukor, J.J. (2005) Identification of unique type II polyketide synthase genes in soil. *Appl Environ Microbiol* **71**: 2232–2238.

Wietz, M., Duncan, K., Patin, N. V., and Jensen, P.R. (2013) Antagonistic Interactions Mediated by Marine Bacteria: The Role of Small Molecules. *J Chem Ecol* **39**: 879–891.

Yan, X., Ge, H., Huang, T., Hindra, Yang, D., Teng, Q., Crnovčić, I., Li, X., Rudolf, J.D., Lohman, J.R., Gansemans, Y., Zhu, X., Huang, Y., Zhao, L.X., Jiang, Y., van Nieuwerburgh, F., Rader, C., Duan, Y., and Shen, B. (2016) Strain prioritization and genome mining for enediyne natural products. *MBio* **7**: 1–12.

Zhang, G., Zhang, W., Saha, S., and Zhang, C. (2015) Recent Advances in Discovery, Biosynthesis and Genome Mining of Medicinally Relevant Polycyclic Tetramate Macrolactams. *Curr Top Med Chem* **16**: 1727–1739.

Ziemert, N., Podell, S., Penn, K., Badger, J.H., Allen, E., and 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.

CHAPTER 3. Genomic comparison of ketosynthases across the tree of life reveals patterns of unique polyketide diversity

3.1 Abstract

Living organisms on Earth share genome sequences that are full of mysterious potential. As the number of sequenced genomes increases, so does our ability to unlock new natural product biosynthetic potential. Specifically, polyketides and fatty acids are important molecules, and their encoding ketosynthase (KS) and fatty-acid synthase (FAS) domains, respectively, perform a phylogenetically conserved condensation reaction, resulting in structurally diverse and often bioactive natural products. To identify where novel ketosynthase and fatty-acid synthase biosynthetic potential exists on Earth, we analyzed ~620,000 genomes from all three domains of life—Bacteria, Archaea, and Eukaryota— using the latest NaPDoS2 webtool. We discovered new distributions of KS domains while abundances in some taxa supported previously known biosynthetic patterns. The identification of 52,000 KS sequences revealed novel exchanges and shared domains between diverse taxa and complex evolutionary patterns. The expansion of our understanding of where KS and FAS enzymatic domain diversity is found across the tree of life illustrates how significant amounts of polyketide and fatty-acid chemical novelty remain to be uncovered. Furthermore, these distribution patterns will aid in the targeted discovery of new natural products to expand our therapeutic drug repertoire.

3.2 Introduction

Organisms across the tree of life from tiny viruses to ginormous blue whales produce organic chemical metabolites that are essential for their living functions. These biological natural products often perform important ecological roles for the organism, such as defense, nutrient

acquisition, reproduction, and communication. Natural products have been the source of many traditional indigenous remedies for thousands of years (Heinrich, 2000), and more recently, the source and inspiration for many pharmaceutical drugs and therapeutic treatments (Newman and Cragg, 2020). Among the many types of natural products, polyketides constitute one of the major classes of natural products, comprising an incredible amount of structurally diverse compounds all originating from simple acyl building blocks. While polyketide natural products could be searched for anywhere in the environment, finding new polyketide scaffolds with interesting structural moieties and even biological activity can be streamlined by taking a targeted approach to avoid rediscovery. Like other natural products, the genes responsible for the biosynthesis of polyketides are encoded in an organism's genome, and thus genome mining by searching for specific genes that together biosynthesize the targeted type of molecule, is a powerful approach to new polyketide natural product discovery (Nivina *et al.*, 2019; Bauman *et al.*, 2021).

Since the first animal genome sequence in1998, initiatives to sequence genomes across the tree of life have made astounding progress (Grigoriev *et al.*, 2014; Leebens-Mack *et al.*, 2019; Hotaling *et al.*, 2021). However, the number of genera described across the tree of life does not yet match the number of publicly accessible genomes in the NCBI genome database, where bacteria comprise >75% of available genome sequences (Medema *et al.*, 2021). The relative ease of sequencing a microbial genome has led to a revolution in the ability to genome mine for bacterial natural products. This has been made possible by the development of sophisticated genome mining tools such as antiSMASH 6.0 (Blin *et al.*, 2021) and PRISM 4 (Skinnider *et al.*, 2020). Genome mining tools identify biosynthetic gene clusters (BGCs) based on the co-localization of core biosynthetic genes and key tailoring enzymes, thus defining a BGC neighborhood predicted to include genes involved in the regulation, biosynthesis, resistance to, and export of the encoded

small molecule. Many studies have elegantly applied these tools to predict the biosynthetic potential in bacterial genomes, uncovering novel patterns of taxa-specific BGC diversity (Cimermancic *et al.*, 2014; Wei *et al.*, 2021; Chen *et al.*, 2022). Recent analyses have shown which species of bacteria should be targeted for novel natural product discovery based on their unique BGC family biosynthetic potential (Gavriilidou *et al.*, 2022). A similar analysis of 1,000 fungal genomes also revealed striking patterns of taxa enriched in BGCs (Robey *et al.*, 2021), and those efforts coupled with advanced "Hex" synthetic biology techniques revealed a systematic way to isolate novel compounds from challenging fungal scaffolds (Harvey *et al.*, 2018). However, it can be challenging to apply these same tools to non-bacterial genomes as BGCs may be missed if they are not fully assembled with the requisite core genes required to be returned as a match (as in fragmented genome assemblies) or if genes of certain pathways are not co-localized in the same genome region or chromosome (as in animals), among other challenges of eukaryotic splice-variant gene calling.

As described in Chapter 2, we have developed an updated NaPDoS2 webtool that detects and classifies ketosynthase (KS) and condensation (C) domains from polyketide synthase (PKS) and non-ribosomal peptide synthetase (NRPS) BGCs (Ziemert *et al.*, 2012). The ability of NaPDoS2 to make broader predictions about PKS and NRPS BGCs based on these relatively short biosynthetic domains makes it ideal for the analysis of genomes beyond bacteria. Our updates of NaPDoS2 included an upgrade for the webtool to now analyze 500,000 sequences/500 MB of data at a time and significant speed improvement with implementation of DIAMOND (Buchfink *et al.*, 2015). Furthermore, we implemented a new polyketide classification scheme where KS queries are assigned to one of the 50 classes/subclasses, indicative of the predicted broader encoding BGC architecture and salient structural features of the compound class. Thus, NaPDoS2 is an ideal

genome mining tool to apply to genomes beyond bacteria and predict specific biosynthetic potential in genomes that are difficult to mine for natural products.

In this study, we harnessed the availability of thousands of genomes across the tree of life to search for polyketide biosynthetic potential using the NaPDoS2 webtool. For the first time, one single tool can be used to predict biosynthetic potential to a level of polyketide class and subclass detail beyond what current tools can predict in genomes that are difficult to analyze. Using NaPDoS2, we analyzed 617,968 bacterial, fungal, animal, algae, plant, protist, plasmid, archaeal, and CPR (candidate phyla radiation) genomes to create a comprehensive dataset of 53,713 ketosynthase domains. These ketosynthase domains were further classified with the NaPDoS2 webtool into specific polyketide and fatty acid classes and subclasses. This facilitated taxa-specific biosynthetic gene distribution predictions and shared KS diversity types across unexpected taxa. Our results revealed which KS types were most similar and where new types of polyketide synthase diversity could be uncovered.

3.3 Methods

3.3.1 Genome dataset selection

To select genomes across the tree of life, I targeted aggregated datasets that included phylogenetic trees of included genomes, when possible, and used a variety of databases to identify and select genome sequences from multiple repositories around the world (**Table 3.1**). My goal was to be comprehensive across the tree of life but select representative genomes when necessary; for example, by only selecting one genome of the same 100 *E. coli* strains.

Bacteria and Archaea. I downloaded all genomes included in the "Web of Life Reference Phylogeny for Microbes" release 1 (WoL, April 5, 2019, https://biocore.github.io/wol/), which is a comprehensive dataset of 10,575 genomes (and metagenome-assembled genomes, MAGs) from Bacteria, Archaea, and CPR (Zhu *et al.*, 2019). This dataset included pre-annotated genomes, taxonomy, and phylogenomic trees of reference genomes across the microbial tree of life where the original goal of the WoL project was to create a reference genome collection and corresponding phylogenetic tree comprised of 381 single-copy marker genes. Custom scripts that I developed were used to append level 1 and 2 metadata to each of the WoL genome sequence; level 1 and 2 generally correspond to Kingdom and Phylum for each sequence (this was not always perfect for some members of the CPR (candidate phyla radiation of bacteria group), some Archaea, and some Eubacteria where taxonomic classification was unclear).

Plasmids. I downloaded the COMPASS database (https://github.com/itsmeludo/COMPASS) of 12,084 completed plasmids from 1,571 species worldwide collected from the past 100 years (Douarre *et al.*, 2020). Reference information for this database included the host species for each plasmid sequence, replication type, and mobilization apparatus type, among others.

Viruses. I collected viral genomes from four different databases including the Reference Viral Database (RVDB, protein database, https://rvdb-prot.pasteur.fr/ , 2021-02, version 21) (Goodacre *et al.*, 2018; Bigot *et al.*, 2020), the Virus Pathogen Resource (ViPR) database (downloaded June 23, 2021) (Pickett *et al.*, 2012), the PATRIC virus and phage database (version 3.6.9) (Davis *et al.*, 2020) as inspired by the analysis in (Dragoš *et al.*, 2021), and the CheckV (version v1.0) curated database of viral genomes (Paez-Espino *et al.*, 2017; Nayfach *et al.*, 2021) clustered into representative viral genome sequences.

Fungi. Because NaPDoS2 does not excise introns and exons, I targeted databases of annotated Eukaryotic genomes with predicted amino acid coding sequences. This is important because introns have been observed to be in the middle of polyketide synthase biosynthetic gene clusters (Harvey *et al.*, 2018), and thus could cause false positives or negatives from NaPDoS2 not realizing it is intronic sequences. For fungi, I downloaded 1,644 fungal genomes from a recent phylogenetic analysis (Li *et al.*, 2021). However, manual inspection of these genomes revealed most were nucleic acid genome sequences, thus I used the metadata list of all genome accessions from the genome collection to download all amino acid genome sequences using the "ncbi-genome-download" script (https://github.com/kblin/ncbi-genome-download); additional individual genomes were downloaded manually.

Plants, Algae, Protists. All amino acid protein genomes from the JGI PhycoCosm repository (149 genomes at time of download September 2021; https://phycocosm.jgi.doe.gov/phycocosm/home) (Grigoriev *et al.*, 2021) and the JGI Phytozome database (version v13, https://phytozome-next.jgi.doe.gov/) (Goodstein *et al.*, 2012) were downloaded. Duplicate genomes were removed based on a comparison with the "animal" genome dataset from NCBI.

Animals. The NCBI Taxonomy browser (https://www.ncbi.nlm.nih.gov/taxonomy) and the NCBI Beta Genomes Datasets tool (https://www.ncbi.nlm.nih.gov/datasets/genomes/) were used to manually inspect the phylogenetic tree at every phylum, class, order, and family level starting at Opisthokonta (~22,632 available amino acid protein genome assemblies) and to download amino acid protein reference genome sequences. Fungi in the Opisthokonta were ignored, but following taxa within Holozoa, Choanozoa, and Metazoa (*Porifera*, *Ctenophora*, *Placazoa*, *Bilateria*, and *Cnidaria*) were targeted. Only "annotated" genomes were targeted

(containing amino acid protein CDS predictions) and when available between assemblies, a reference genome ("RefSeq") genome assembly was selected; at least one genome per genus was downloaded. The "ncbi-genome-download" script (https://github.com/kblin/ncbi-genome-download) was used to download all genome assemblies, resulting in 1,125 protein genomes.

Reference database collections: The NaPDoS2 (Chapter 2) (Ziemert *et al.*, 2012) database (version pksdb_20200830) of 1,877 KS reference sequences was downloaded from https://npdomainseeker.sdsc.edu/napdos2/pathways_v2.html. The MIBiG database (version 2.0) of characterized BGCs FASTA sequences with known products was downloaded and concatenated (Kautsar *et al.*, 2020).

Table 3.1. Number of genomes in each taxa dataset.

Dataset	**Number of Genomes**	**Reference**
Bacteria, Archaea	10,575	(Zhu *et al.*, 2019)
Plasmids	12,084	(Douarre *et al.*, 2020)
Viruses	591,387	(Pickett *et al.*, 2012; Bigot *et al.*, 2020; Davis *et al.*, 2020; Nayfach *et al.*, 2021)
Fungi	1,149	(Li *et al.*, 2021)
PhycoCosm (Green Algae, SAR)	895	(Grigoriev *et al.*, 2021)
Phytozome (plants)	753	(Goodstein *et al.*, 2012)
Animals	1,125	NCBI Datasets - Genomes
Total:	***617,968***	
MIBiG 2.0 (*reference*) - BGCs	2,021	(Kautsar *et al.*, 2020)
NaPDoS2 (*reference*) – KS domains	1,877	Chapter 2, this dissertation (Ziemert *et al.*, 2012)

Before subsequent analysis, all genome datasets were appended with unique identifiable headers based on genome dataset taxa and sub-taxa to facilitate downstream filtering and analyses. A challenge with this dataset is from the time that I started this analysis/downloaded various

genomes to the present, some taxonomic assignments could have changed, or more recent genome assemblies released.

3.3.2 Detection and classification of KS domains with NaPDoS2

All genome datasets were analyzed with NaPDoS2 (Ziemert *et al.*, 2012) (Chapter 2, http://napdos.ucsd.edu/napdos2) with the following settings: uploaded amino acid FASTA sequence file; minimum alignment length of 200 amino acids; E-value 1e-8 (default settings for NaPDoS2). For the bacterial/archaeal genome dataset (Zhu *et al.*, 2019), both the nucleic acid and amino acid contigs of all 10,575 genomes were analyzed, and thus KS domain hits were saved both as nucleic acid and amino acid FASTA sequence files. For each dataset, the KS query hit result tables (.tab file) and trimmed KS domain sequences (.fasta) were saved. NaPDoS2 webtool analyses are limited to 500,000 sequences or 500MB at a time, so in some cases I utilized the provided scripts ("serialize_seqs.pl") on the "File Size Management Tools" NaPDoS2 webpage (https://npdomainseeker.sdsc.edu/napdos2/prefiltering.html) for splitting the sequence datasets into appropriate sized files. Custom scripts and manual annotation were used to annotate the KS query hit FASTA sequence headers with the NaPDoS2 class/subclass identification code. Additionally, custom scripts were used to separate all KS class and subclass hits into individual FASTA sequence files; concatenated datasets of each genome taxa group and each KS class/subclass was used for pertinent analyses. All tabulated result tables were concatenated and summarized using a custom R script in RStudio (version 1.4.1717) (RStudio Team, 2021). RawGraphs (Mauri *et al.*, 2017) was used to visualize the distribution of KS hits across taxa with an alluvial diagram.

3.3.3 Phylogenetic distribution and diversity

For each genome dataset that had corresponding phylogenetic trees, we plotted the abundance and distribution of KS class/subclass hits using iTol (Letunic and Bork, 2019). Briefly, phylogenomic trees for each dataset were used as the internal tree and the distribution (class/subclass) of KS query hits, the abundance of hits per taxa, and the taxonomic classification was plotted. For the type I PKS KS phylogenetic trees, KS domains were clustered into 50% OTUs with UCLUST (Edgar, 2010); phylogenetic trees were calculated using the FastTree (Price *et al.*, 2010) workflow implemented on NGPhylogeny (Lemoine *et al.*, 2019).

Sequence similarity networks were calculated with the EFI-EST Enzyme Similarity Tool (Gerlt *et al.*, 2015; Gerlt, 2017; Zallot *et al.*, 2018, 2019, 2021) and visualized with Cytoscape (version 3.7.2) (Carlin *et al.*, 2017). To calculate the PCOA, KS query hits of each class/subclass from all genomes were divided into phyla within each Kingdome. For each phyla, the KS query hits of each class/subclass were converted into percentages out of 100, and the matrix of percentages was transformed using a Bray-Curtis dissimilarity matrix and plotted using a PCOA in R Studio ((RStudio Team, 2021). Additional figure modifications were performed in Adobe Illustrator (version 24.2).

3.4 Results

We set out to detect and classify polyketide and fatty acid biosynthetic potential across the genome-sequenced tree of life. To do this, we analyzed 617,968 genomes (**Table 3.1**) from multiple taxa, including: Bacteria, Archaea, Plasmids, Viruses, Fungi, Algae, Protists, Plants, and Animals. To give context of where known polyketide and fatty acid biosynthetic potential has been characterized, we analyzed the MIBiG 2.0 reference database (Kautsar *et al.*, 2020) of 2,000 biosynthetic gene clusters (BGCs) and the NaPDoS2 (Chapter 2, this dissertation) database of characterized and classified KS (ketosynthase) domains. We identified and classified 53,713 KS domains, with the most being identified from Bacteria (32,473), Fungi (12,339), and Animals (6,508) (**Figure 3.1**). Of the KS domains detected, most were classified as type I PKS (27,023),

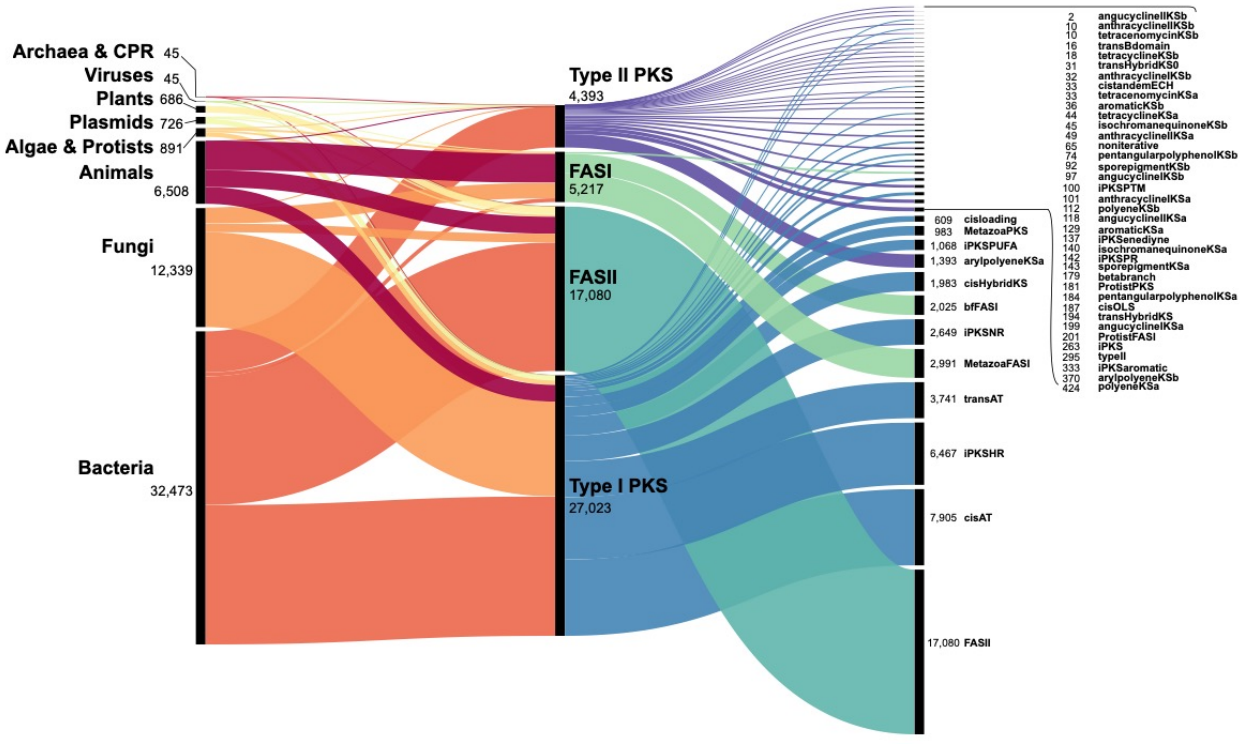


Figure 3.1. Number of KS domains identified in each taxa database, and their subsequent class and subclass type classification by NaPDoS2.

followed type II FAS (17,060) (**Figure 3.1**). Type II PKS biosynthetic potential was the rarest type of KS identified with only 4,393 hits, mostly in Bacterial genomes; however, the type II KSs were the most diverse with many subclasses identified. Of the specific subclasses, KSβ (also referred to as CLF or chain-length factor) were observed less than their cognate KSα domains. This could be due to our sampling scheme of 10,575 representative bacterial genomes where KSβ-containing gene clusters could be predominantly enriched in specific bacterial species or genera and thus that diversity might not be captured. Other rare KS subclasses included type I *trans*-AT beta-branching modules (16) which are KSs that typically co-occur with a B-branching domain responsible for the formation of beta-branches; and type I *trans*-AT hybrid KS0 non-elongating KS domains (31) that follow an NRPS module (**Figure 3.1**). Many of these rarer specific subclass KS domains could serve as hooks that could be used to go back into the genome and find the encoding BGC. The most common KS domain identified was the type II FAS domain with every database taxon type having at least one type II FAS domain (**Figure 3.1**). The animal genomes were most enriched in type I FAS domains, followed by fungi genomes, which contrasted the very small fraction of type I FAS domains identified in the bacterial genomes. The plasmid genome dataset included many of the same types of KS hits as bacteria, which could be a factor of the plasmids carrying KS-encoding BGCs that were bacterial in origin, but this would warrant further investigation. The viral and Archaeal groups contained the least amount of KS hits, which supports previous observations that BGCs are rare in these taxa, or too different from our known databases to accurately identify and characterize (Sharrar *et al.*, 2020; Dragoš *et al.*, 2021).

Our first goal was to understand if the KS diversity that we characterized across the various genomic datasets were similar to one another. We assessed diversity by measuring the dissimilarity between KS class and subclass types at the dataset (**Figure 3.2**) and phyla levels (**Figure 3.3**) and

visualizing these in principal coordinate analysis ordinate analysis (PCoA) plots where objets (here, KS domains) that are closer together have less dissimilarity. We observed that the KS

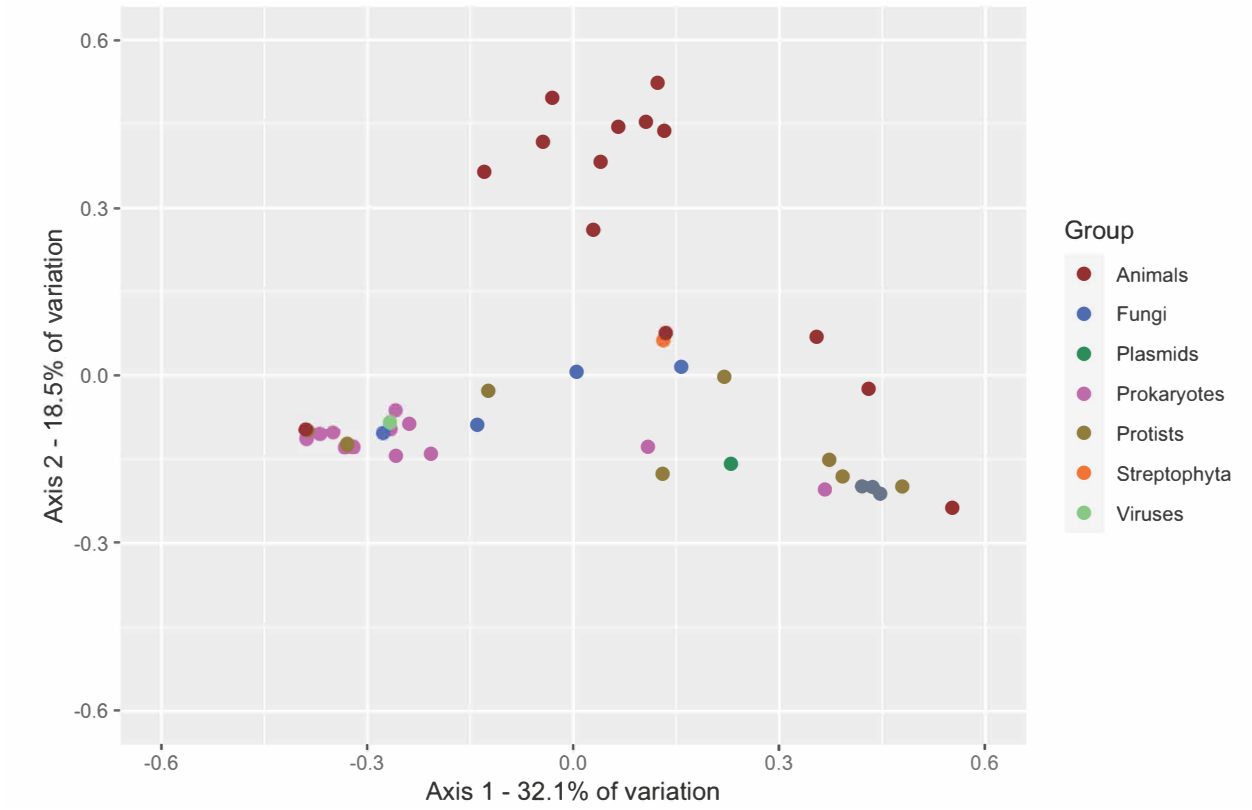


Figure 3.2. Diversity of KS domains, colored by dataset taxa group. Points represent all KSs within a phyla group.

domains identified in the animal genomes were most dissimilar from other KS domains. These could be interesting targets for further phylogenetic comparisons (**Figure 3.2**). The Prokaryotes (defined here as Bacteria and Archaea) formed a cluster with some viral, animal, fungal, and protist KS domains (**Figure 3.2**). However, the plasmid KSs did not cluster with other datasets, most notably the bacterial hosts, indicating the composition of KS diversity was unique in the plasmid dataset (**Figure 3.2**). To investigate this further, we split the KS types further into class/subclass

for each phylum within each of the taxa datasets and performed the same dissimilarity PCoA analysis (**Figure 3.3**). By doing this, we could see which phyla were driving the variation in each

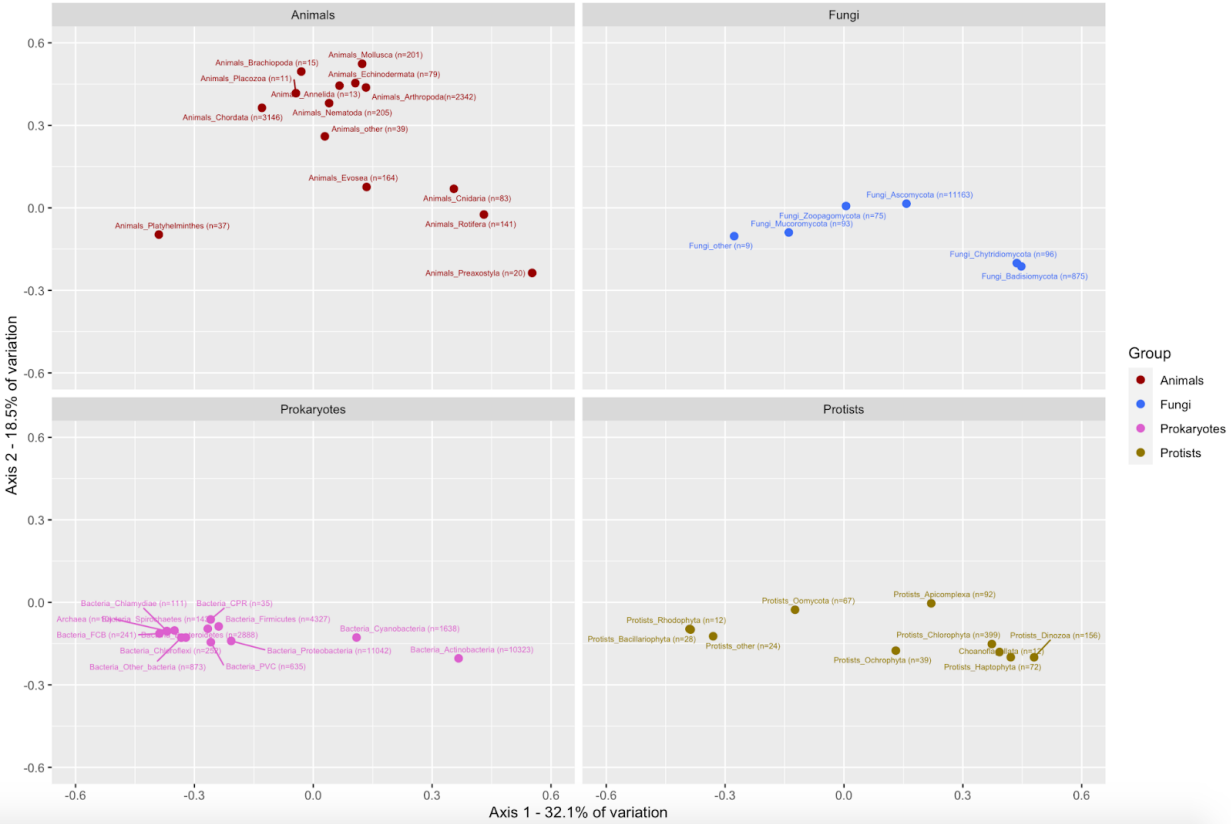


Figure 3.3. Diversity of KS domain class/subclass type, split by Phylum for each dataset taxa group.

dataset. In the animal dataset, we observed that the KS diversity of the Platyhelminthes (common: flatworms) was the most different from other animals, whereas Mollusca (mollusks), Echinodermata (echinoderms), Arthropoda (arthropods), Annelida (segmented worms), and others were more similar to each other (**Figure 3.3**). In fungi, we observed a stark separation between the Ascomycota group and the Basidiomycota and Chytridiomycota, which indicated these fungal groups had very different KS diversity (**Figure 3.3**). In the Prokaryotes (defined here to include

both Bacteria and Archaea), Actinobacteria were the most dissimilar in their KS composition, followed by Cyanobacteria (**Figure 3.3**). This supported our hypothesis that those phyla, long targeted for their specialized metabolism biosynthetic capabilities, could be distinguished based on their polyketide KS diversity alone. Finally, there was wide dissimilarity between protist taxa, including Dinozoa (dinoflagellates), Haptophyta (haptophytes), and Chlorophyta (green algae) compared with Rhodophyta (red algae) and Bacillariophyta (diatoms) (**Figure 3.3**). These differences in KS composition were investigated further with targeted phylogenetic comparisons.

In order to understand the phylogenetic distribution of bacteria PKS classes and subclasses, we next mapped the top 10 most abundant KS types onto the bacterial and archaeal (WoL) phylogenetic tree (Zhu *et al.*, 2019) (**Figure 3.4**). These included type I modular *cis*-AT (ring 1, dark red), type I modular *cis*-loading (ring 3, blue), and type I iterative *cis*-AT aromatic (ring 5, red), which were enriched in the Actinobacteria phylum (**Figure 3.4**). The only taxon with a similar distribution was the Acidobacteria (**Figure 3.4**). However, there also appeared to be a wide diversity and large amount of KS diversity observed in Cyanobacteria, and type II aromatic,

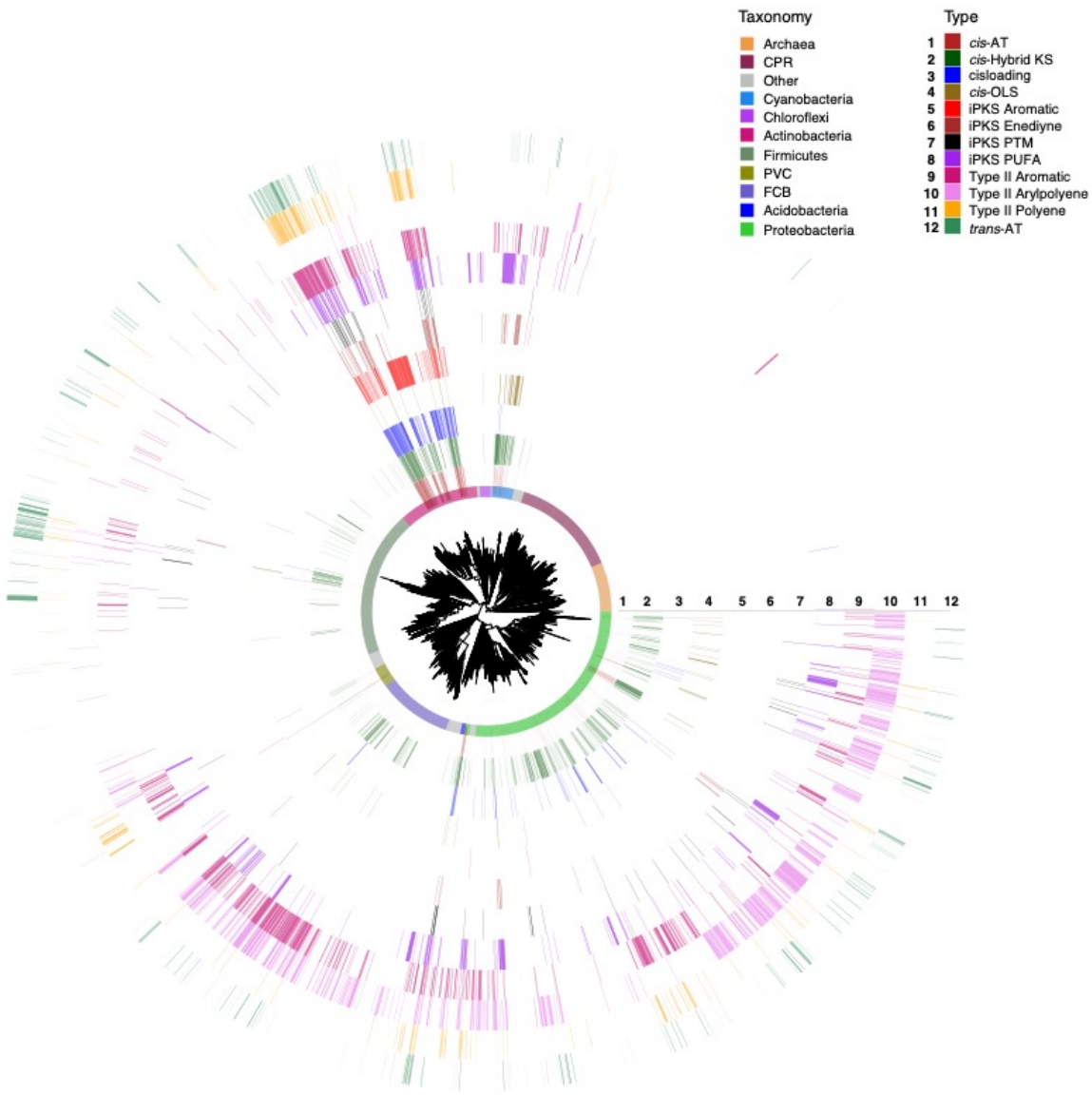


Figure 3.4. Distribution of KS diversity mapped onto the Bacterial and Archaeal phylogenetic tree. The inner ring is colored by taxonomy; and the numbered outer rings are shaded by the abundance of each KS subclass type.

arylpolyene, and polyene biosynthetic potential enriched in FCB (Fibrobacterota, Chlorobiota, and Bacteroidota) and Proteobacteria (**Figure 3.4**).

To investigate this further, we mapped on the less abundant type II and other KS domains onto the same bacterial phylogenetic tree (**Figure 3.5**). In contrast to the type II FAS KS domains

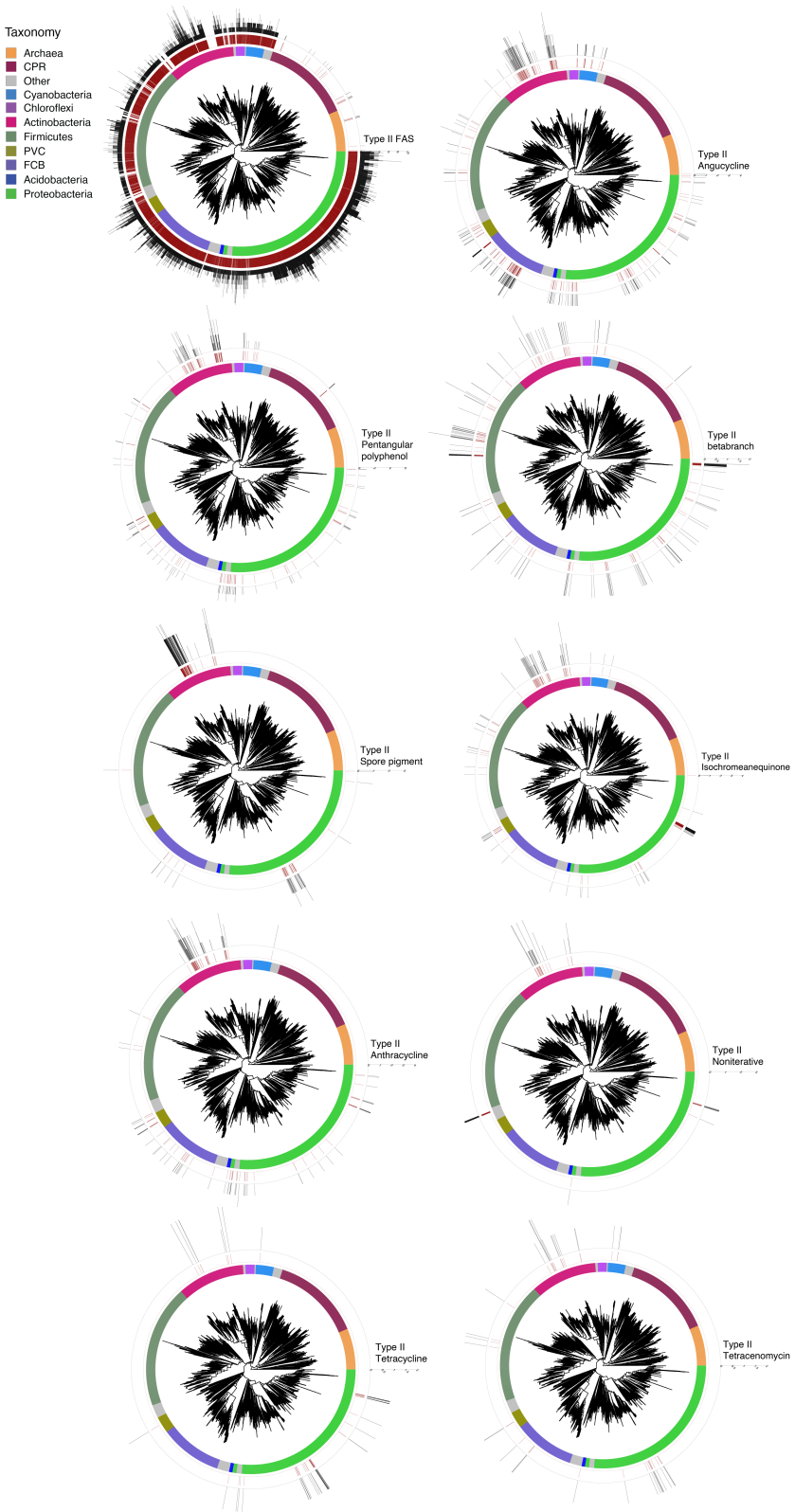


Figure 3.5. Distribution of Type II KS subclass hits across the Bacterial and Archaeal phylogenetic tree. The inner ring is colored by taxonomy; the outer two rings show presence (maroon) and abundance (black bar graphs) of each KS subclass type.

that were observed to be mostly abundant across all taxa except for Archaea, CPR (candidate phyla radiation bacteria), some Actinobacteria, and some "other" phyla, we saw clade-specific patterning of type II angucycline, spore pigment, and noniterative KS domains (**Figure 3.5**). Other type II KS types including tetracycline, tetracenomycin, and isochromanequinone had abundant hits in specific taxa within the larger phyla groups (**Figure 3.5**). In future work, these distribution patterns can be compared to where polyketides of these classes have been described to ascertain which taxa could be targeted for novel compound elucidation.

We next mapped the KS subclass diversity onto the fungal phylogenetic tree (**Figure 3.6**). We hypothesized based on our PCoA analysis (**Figure 3.3**) that we would observe variation in KS diversity between the Ascomycota and Basidiomycota. Our results supported our hypothesis as families within the two groups had drastically different KS subclass types (**Figure 3.6**). For example, in contrast to the Saccharomycotina (including yeasts) which only contained type I and II FAS KS domains, the Pezizomycotina (Ascomycota clade) contained a high diversity of KS types including type I iterative *cis*-AT highly reducing and non-reducing domains (**Figure 3.6**). However, we observed that type I modular *cis*-AT domains were enriched in the Agaricomycotina (mushrooms, bracket fungi, and puffballs; Basidiomycota clade) (**Figure 3.6**). While some of these biosynthetic differences have been reported before at the BGC level, (Harvey *et al.*, 2018; Robey *et al.*, 2021), these differences in diversity and abundances are also readily revealed when looking at KS domains.

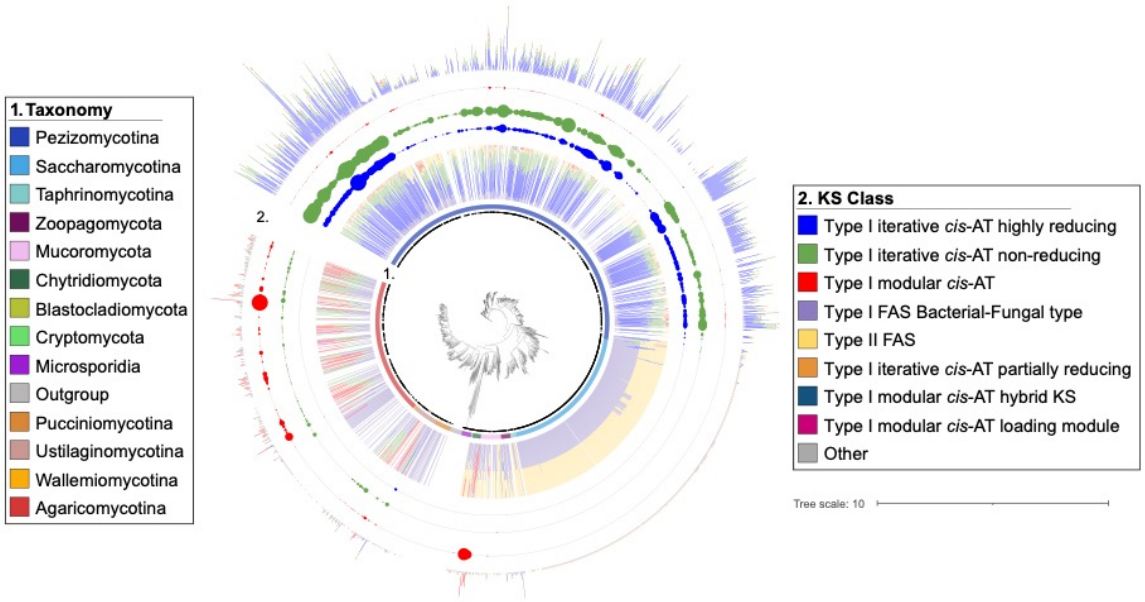


Figure 3.6. Distribution of KS polyketide biosynthetic potential across the fungal phylogeny.

The inner ring is colored by taxonomy; and the outer ring barplots, bubble plots, and bar plots are colored by KS subclass type.

Phylogenetic trees of KS domains of specific classes combined with KSs from reference databases of characterized BGCs can be a powerful method to identify clades with potential biosynthetic novelty. To test for this, we constructed a phylogenetic tree of the type I KS domains, which was the largest KS group across our genome datasets (**Figure 3.7**). In contrast to the clades of KS domains that were specific to certain taxa—for example, large clades of fungal (green) type I iterative *cis*-AT clades (green outer ring, 6'oclock position) and algae/protist (yellow, PhycoCosm dataset) type I *cis*-AT (blue outer ring, 7'oclock position), we observed that the reference MIBiG KS hits were mostly clustered with bacterial KS clades in a variety of subclasses (**Figure 3.7**). This is supported by the observation that the majority of described BGCs in the MIBiG 2.0 database are from bacterial sources, and thus clades without a MIBiG KS hit could be sources of novel KS biosynthetic potential that are unlike BGCs that have been characterized.

There were many clades of various KS subclass types that did not have any MIBiG hits clustered with them (**Figure 3.7**).

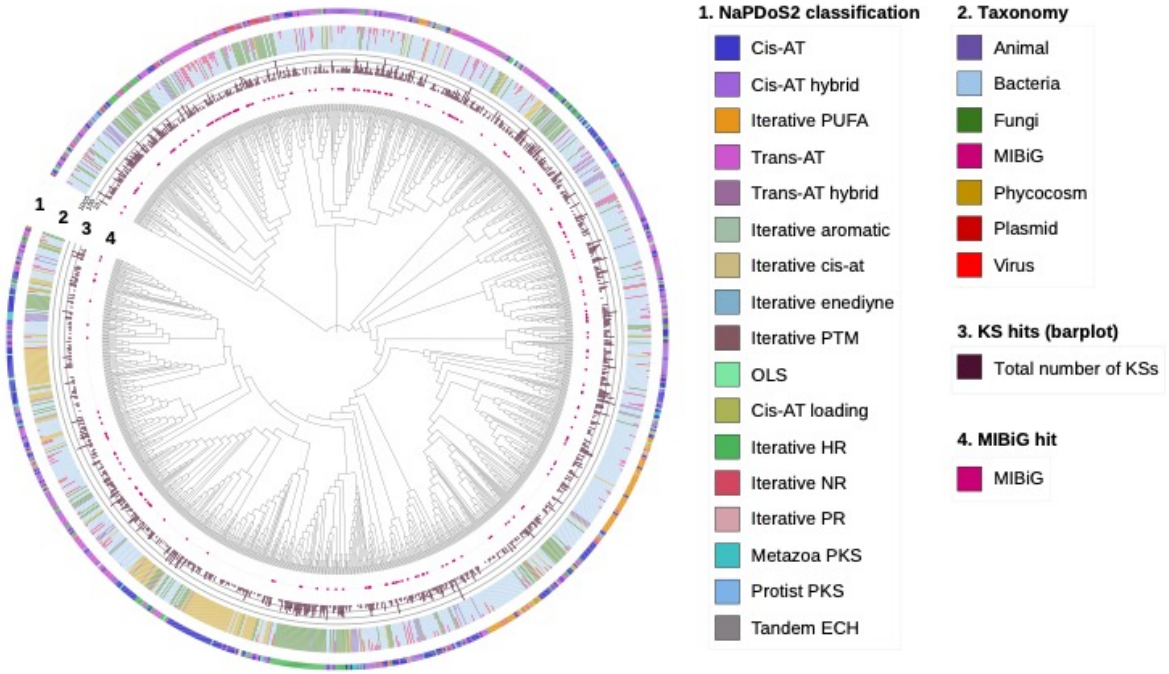


Figure 3.7. Phylogenetic diversity of type I KS domains.

The outer ring 1 is colored by NaPDoS2 subclass type; ring 2 by taxonomic distribution of the 50% clustered OTU clade; ring 3 bar plot of the total number of KS hits; and ring 4 is presence of a MIBiG domain in the OTU cluster.

To look for taxa specific patterns, we constructed phylogenetic trees of the type I KS domains identified in the fungal dataset (**Figure 4.8**). By doing this, we were able to identify specific clades without reference MIBiG hits, thus indicating possible polyketide biosynthetic novelty. We additionally observed that these domains had low percent identities compared to the closest NaPDoS2 database KS hit (**Figure 4.8**). As the NaPDoS2 database contains most of the KS domains from the MIBiG database, it is unsurprising that the domains least similar to the

MIBiG domains have low percent identities to the top NaPDoS2 database match. In this tree, we identified clades where there seems to be biosynthetic novelty in the type I iterative *cis*-AT highly reducing (HR, ring 2 green), type I *cis*-AT (ring 2 blue), type I modular *cis*-AT olefin synthase (ring 2 light blue), type I modular *cis*-AT loading module (ring 2 olive), type I iterative *cis*-AT non-reducing (NR, ring 2 red), type I modular *cis*-AT hybrid KS (ring 2 purple), and type I *trans*-AT (ring 2 seafoam) (**Figure 4.8**). There were also large clades of MIBiG 2.0 KS domains, which could represent KSs from similar pathways that have been discovered or rediscovered in multiple

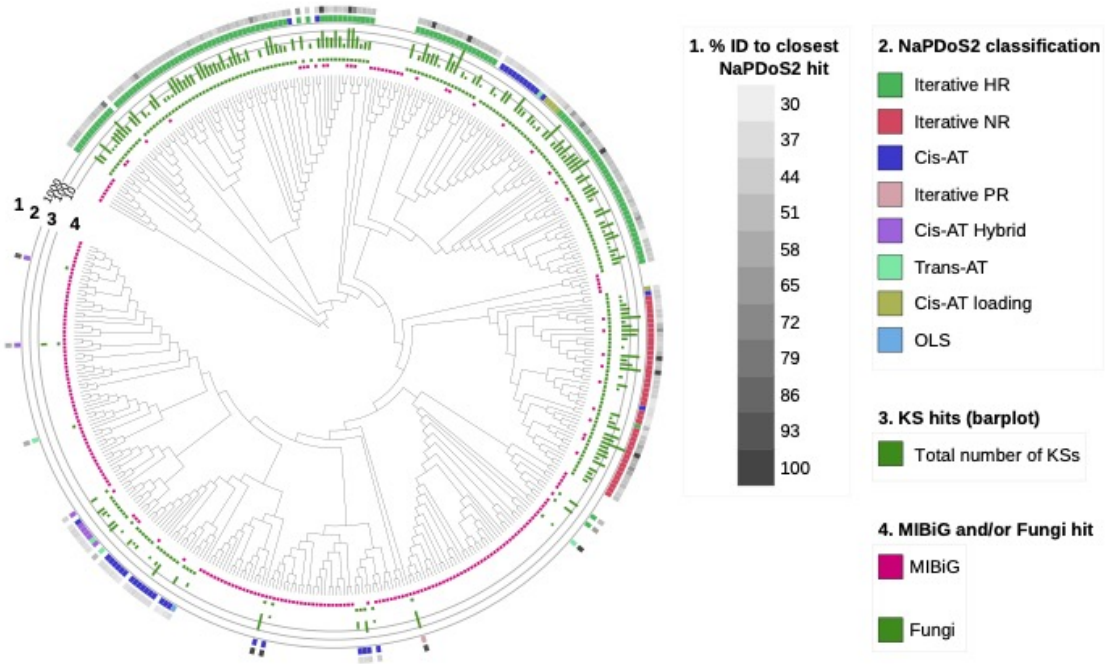


Figure 3.8. Phylogeny of type I fungal KS domains.

Ring 1 indicates percent identity to the top NaPDoS2 database match; ring 2 indicates NaPDoS2 KS class/subclass; ring 3 is an abundance bar plot of KSs in each branch; and ring 4 indicates if the node contains a MIBiG hit.

organisms. Yet, our sampling scheme of representative fungal genomes might not have included all species or strains with the domain (**Figure 4.8**).

Next, from our PCoA analysis which indicated that many animal KS domains were divergent from other KSs, we constructed a phylogenetic tree of the KS domains identified in the algae, protist, and animal datasets (**Figure 3.9**). From this phylogeny, we discovered large clades without any known MIBiG hits, especially from the PhycoCosm (algae, protists) dataset belonging to the type I modular *cis*-AT class (**Figure 3.9**). These domains could belong to the *cis*-AT enriched genomes of Dinoflagellates and Haptophytes which has been previously reported (Verma

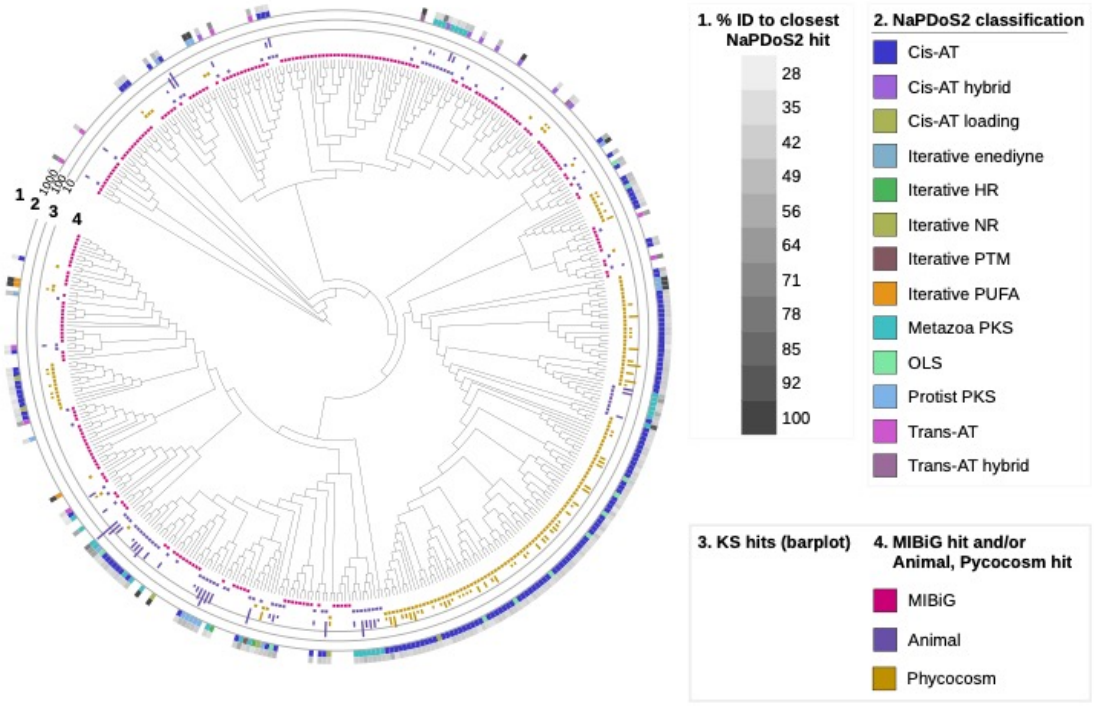


Figure 3.9. Phylogenetic diversity of KS domains from the PhycoCosm and Animal datasets.

Ring 1 indicates percent identity to the top NaPDoS2 database match; ring 2 is colored by NaPDoS2 classification; ring 3 is a bar plot indicating KS total amount per branch, and ring 4 is the database the node is from.

et al., 2019; Chekan *et al.*, 2020). Additionally, as MIBiG is devoid of BGCs from animals, many KS domain detected in the animal genome dataset including the classes of type I Metazoan-type PKS, type I Protist-type PKS and even some type I *trans*-AT and *cis*-AT domains were unique from known pathways (**Figure 3.9**). These could be targets for future investigation into the broader KS-encoding BGC environment. It is striking that the entire left part of the tree (3 to 6 o'clock) is represented by KS domains from the PhycoCosm and animal datasets that have no similarity to MIBiG KSs and low identity to NaPDoS2 database hits (**Figure 3.9**).

One of the main goals of this study was to not only investigate the distribution and diversity of polyketide KS biosynthetic potential across the genome-sequenced tree of life, but to also ask if we could observe if specific KS domains were shared across taxa or if there were unusual similarities across taxa. To do this, we constructed a sequence similarity network of all the type II KS domains and colored the nodes by dataset. While these networks proved very difficult to calculate and visualize (>1-10 million edges in the small network files), we observed preliminary groupings of type II KS domains that supported the NaPDoS2 classification in a taxa specific pattern in some cases (**Figure 3.10**). For example, type II KSα domains from the viral dataset and the PVC superphylum (Planctomycetota, Verrucomicrobiota, and Chlamydiota) each formed a distinct cluster, along with doublets from the Archaea and CPR groups (**Figure 3.10**). There were also smaller clusters that contained KSs from multiple taxa, including a cluster containing KS domains from Cyanobacteria, Archaea, and an Actinobacteria (**Figure 3.10**). The large cluster of many connected nodes needs to be further separated but includes many of the similar type II aromatic KS subclasses (**Figure 3.10**). Of note is that the reference database nodes are mostly clustered with the large cluster, and many of the clusters of taxa with unique type II KSs do not cluster with a known KS domain. Further investigation into what the broader genomic BGC

neighborhood looks like for each of these clusters, especially in Archaea, CPR, and viruses, will

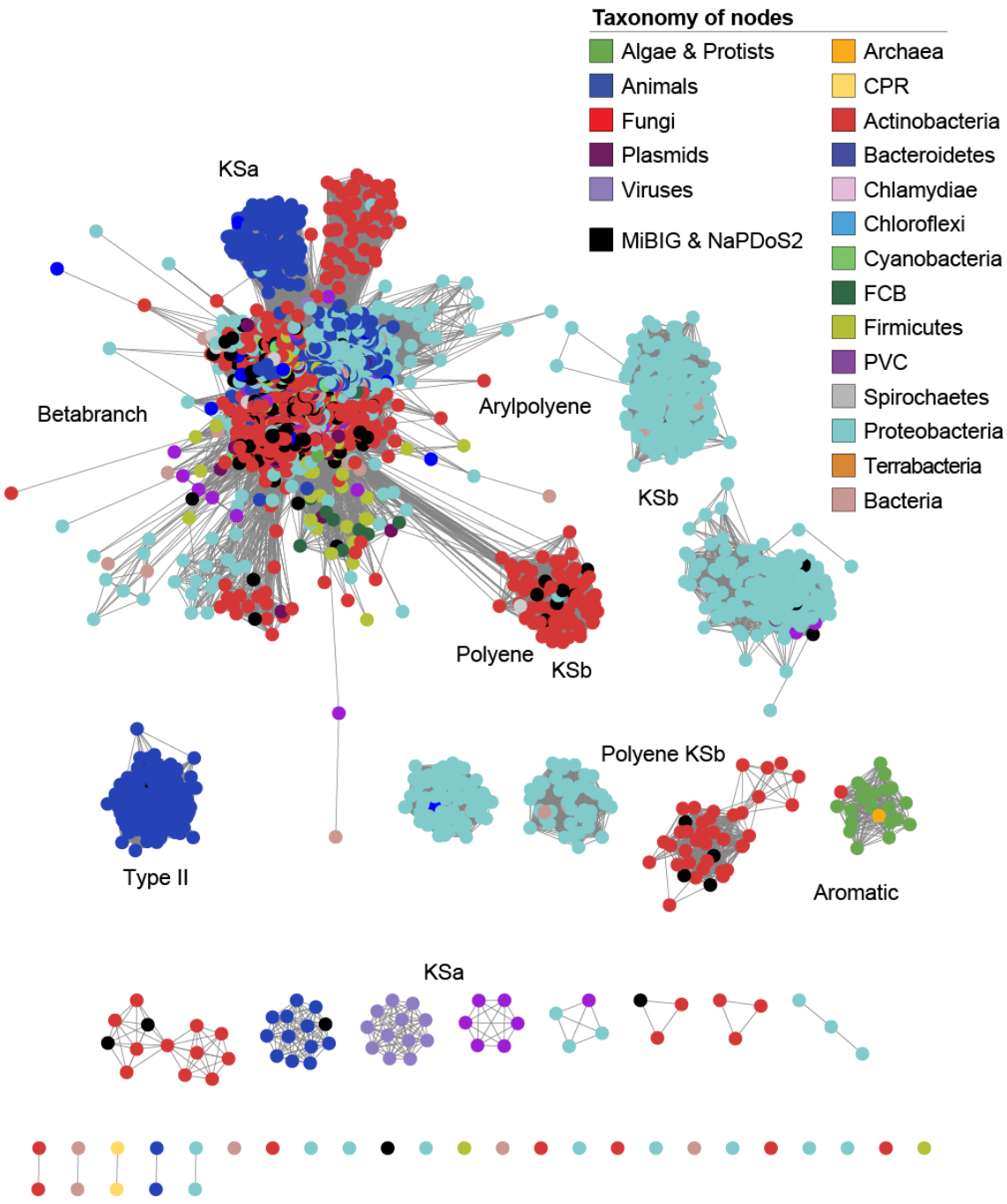


Figure 3.10. Sequence similarity of all type II KS domains colored by taxonomic dataset.

Nodes (n=4,927) are clustered into 90% representative OTUs and connected by edges if >20% shared identity.

be needed.

Finally, to determine how well similar KSs within a specific KS subclass type could reveal similar BGCs in divergent taxa, we constructed a sequence similarity network of all type I iterative *cis*-AT PTM-type KS domains (**Figure 3.11**). We identified 118 PTM-type KS domains belonging to 3 datasets—the bacteria, fungal, and MIBiG reference dataset (**Figure 3.11**). Upon clustering the sequences together connected at >52% identity, all reference MIBiG KS domains clustered with the largest cluster including domains from a wide variety of mostly bacterial taxa (**Figure 3.11**). This could indicate that many of the connected PTM-type KS nodes share similar biosynthetic neighborhoods, but further investigation would be needed, especially for the domains in the PVC and Cyanobacteria. The large cluster also contained one connection to a fungal *Aspergillus* PTM-type KS domain, which could indicate a shared evolutionary history of the domain. Further phylogenetic comparison would help resolve the specific relatedness of the domains (**Figure 3.11**). Other smaller clusters of PTM-type KS domains from Actinobacteria and Proteobacteria indicates similar biosynthetic potential across these taxa (**Figure 3.11**). This

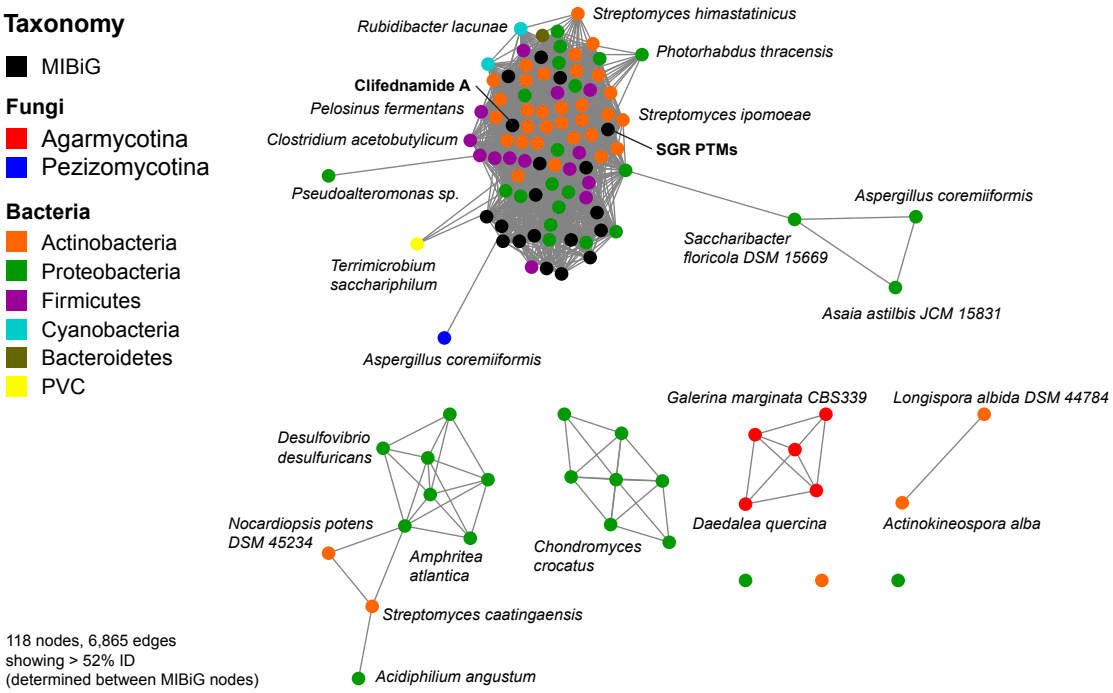


Figure 3.11. Sequence similarity network of type I iterative *cis*-AT PTM-type KS domains colored by taxonomy.

Nodes (n=118) are connected by edges if they share >52% identity as determined between known MIBiG reference KS domain nodes.

contrasts with the identification of a fungal-only cluster—the KSs in this clade could be targets for fungal-specific PTM-type BGCs that perhaps have divergent organization from the characterized reference PTM BGCs and other domain PTM-type KSs (**Figure 3.11**).

3.5 Discussion

In this study, we describe the polyketide and fatty acid biosynthetic potential across the genome-sequenced tree of life at a scale that has not been captured before. The updates to the NaPDoS2 webtool facilitated this analysis as there is no other bioinformatic tool that allows the

detection and specific classification of KS domains in both bacterial and animal genomes at the same time. Thus, we were able to create a dataset of 53,713 KS domains from across the tree of life that could be analyzed.

While there have been detailed studies of biosynthetic diversity at the BGC level across the bacterial kingdom (Cimermancic *et al.*, 2014; Wei *et al.*, 2021; Chen *et al.*, 2022; Gavriilidou *et al.*, 2022), these studies have used tools like antiSMASH (Blin *et al.*, 2021) and PRISM (Skinnider *et al.*, 2020) which do not classify polyketide synthase genes with the level of detail provided by NaPDoS2. For example, type II and type III PKS BGCs are classified by antiSMASH together as "PKS other". While these previous analyses can serve as important comparisons for our analysis to build upon, with our KS dataset we can perform additional analyses such as building sequence similarity networks and complete phylogenetic trees across all taxa instead of just one. Compared to a recent analysis reporting the bacterial distribution of seven subclasses of aromatic polyketide synthase genes, we observed similar distributions in our analysis (Chen *et al.*, 2022). However, we could identify additional type II and type II aromatic biosynthetic potential, adding important information to the conclusions of their study. We have built extensively upon a prior study (Cimermancic *et al.*, 2014), where PKSs were discovered to be the least common BGC when using the ClusterFinder tool to genome mine; and we have expanded upon their identification of aryl polyene PKSs across all genome-sequenced life. In comparison to the recently published analysis of bacterial genomes using the PRISM4 (Skinnider *et al.*, 2020) BGC identification tool, our results supported their findings of Actinobacteria being enriched in many of the polyketide pathways, however other than PKS and aryl polyene, the type I and II polyketide biosynthetic potential seems to have been aggregated in "PKS", "PKS-NRPS", and "others" (Wei *et al.*, 2021).

In a recent analysis of BGC diversity in 1,000 fungal genomes, highly-reducing and non-reducing type I iterative *cis*-AT PKS BGC diversity was enriched in the Ascomycota clade whereas the Basidomycota clade was less enriched in those types and overall contained fewer BGCs (Robey *et al.*, 2021). Our detailed KS analyses support this finding and contribute to our understanding of the biosynthetic differences between two clades of fungi. One consideration of using NaPDoS2 to assess fungal polyketide biosynthetic potential is the question of introns and exons as it does not excise introns and predict splicing variants. It has been previously reported that, on average, fungal genes contain 2.5 introns per gene and genes involved in specialized metabolism can contain 5-9 introns each, which includes PKS BGCs (DeNicola, 2018). Introns can split KS domains, and thus make them difficult to detect. In our tests comparing NaPDoS2 analyses of amino acid versus nucleic acid fungal genome assemblies, we observed more KS hits in the protein than the nucleotide genomes, which corresponds to introns perhaps breaking up KS domains and thus the KS domains not meeting the NaPDoS2 match threshold (data not shown). Nonetheless, it is exciting that NaPDoS2, in combination with other bioinformatic BGC genome-mining tools, can continue to help identify polyketide biosynthetic potential with high accuracy and classification specificity.

We believe that NaPDoS2 is the only genome-mining tool that can be used to analyze animal protein genomes. This allows for the rapid identification of animal KSs as recently reported in birds, fish, (Ganley and Derbyshire, 2020) nematodes (Feng *et al.*, 2021), Sacoglossans (sea slugs) (Torres and Schmidt, 2019; Torres *et al.*, 2020), and Echinoderms (sea urchin) (Li *et al.*, 2022). Reports that animal polyketide-like type I FAS domains and metazoan specific PKS-type KS domains have unique evolutionary histories indicate there are still evolutionary dynamics between and within specific KS subclass types that remain poorly understood (Nivina *et al.*, 2019).

Using phylogenetics to predict functional novelty has facilitated the discovery of new patterns of BGC horizontal gene transfer events and even new types of BGCs that were previously missed in larger clusters, as was the case for a large new clade of type I iterative *cis*-AT BGCs. (Grininger, 2020; Wang *et al.*, 2020; Nivina *et al.*, 2021). However, a goal of those analyses and our analysis is to use biosynthetic potential predictions to link to realized chemical diversity. For example, if a new polyketide BGC was predicted in an unknown taxon, we would want to know if perhaps that taxon produced new natural products. We tested this with our current dataset of KSs from across the genome-sequenced tree of life, first focusing on the type I iterative *cis*-AT PTM-like domains (**Figure 3.11**). By searching for the PTM molecule clifednamide A (Cao *et al.*, 2010) in mass-spectrometry (MS) standard reference databases, we acquired a reference MS spectrum (Wang *et al.*, 2016). We then searched the MS spectrum against the entire GNPS metabolomics repository of millions of MS spectra from thousands of datasets using the MASST tool (Wang *et al.*, 2019). We discovered only 5 hits to the clifednamide A molecule—all of which were MS spectra collected from five different metabolomic experiments where *Streptomyces* bacteria were extracted. This means that the PTM clifednamide A molecule and its close derivatives (which PTM molecules share a core similar structure, and thus our MASST search of similar molecular fragmentation patterns likely would have matched other PTMs, though this is not a given) have only been chemically realized in *Streptomyces*. This means that from our PTM-type KS network analysis, perhaps targeting non-Streptomyces KS hits for future elucidation would be the best method to avoid rediscovery of well-known PTM molecules.

In the future, we believe this dataset of 53,713 KS domains will be a useful reference for KS-amplicon studies and metagenomic datasets. First, this collection of KSs could serve as a reference dataset for creating class and subclass specific amplicon primers. Additionally, this large

reference dataset of KSs with known genomic origins and complete taxonomy can serve as a reference to map both class/subclass function (as characterized by NaPDoS2) and putative taxonomic assignment for both KS-amplicon and metagenomic studies where the contigs might not have assigned taxonomy, akin to 16S rRNA databases used for assigning taxonomy to 16S rRNA amplicon sequencing. We are additionally exploring if this could be an add-on feature to the NaPDoS2 webtool, where users would select this collection of KS sequences to map their query hits to, build phylogenies with, and thus drive hypothesis formation of taxonomic and functional polyketide biosynthetic diversity in a wide variety of applications. The analyses presented in this paper show where targeted efforts for novel polyketide chemistry could focus, especially in taxa that have not been explored before, thus illustrating the power of evolutionarily conserved domains as powerful search hooks for novel specialized metabolite chemistry in any type of sequencing data.

3.6 Acknowledgements

I would like to thank authors of the NaPDoS2 version 1 webtool as our updated webtool to complete this analysis would not have been possible without the foundation of the tool. Additionally, I would like to thank members of the Jensen Laboratory for helpful feedback on this project; and acknowledge other feedback from conference and seminar attendees when this work has been presented (Marine Microbial Chemical Communication seminar series; the Gordon Research Conference on Marine Natural Products, & others) that helped refine our questions and analyses. I thank the authors of the genomes we used for making their data publicly accessible so that we could use it for this study. This work was supported by the National Science Foundation

Graduate Research Fellowship Program under Grant no. DGE-1650112 to KEC and DGE-2038238 to HWS.

Chapter 3 is coauthored with Hans W. Singh, Dr. Sheila Podell, Dr. Leesa J, Klau, and Dr. Paul R. Jensen. The dissertation author was the primary co-investigator and co-author of this chapter with Hans W. Singh.

3.7 References

Bauman, K.D., Butler, K.S., Moore, B.S., and Chekan, J.R. (2021) Genome mining methods to discover bioactive natural products. *Nat Prod Rep*.

Bigot, T., Temmam, S., Pérot, P., and Eloit, M. (2020) RVDB-prot, a reference viral protein database and its HMM profiles. *F1000Research* **8**: 1–13.

Blin, K., Shaw, S., Kloosterman, A.M., Charlop-Powers, Z., van Wezel, G.P., Medema, M.H., and Weber, T. (2021) antiSMASH 6.0: improving cluster detection and comparison capabilities. *Nucleic Acids Res* 1–7.

Buchfink, B., Xie, C., and Huson, D.H. (2015) Fast and sensitive protein alignment using DIAMOND. *Nat Methods* **12**: 59–60.

Cao, S., Blodgett, J.A.V., and Clardy, J. (2010) Targeted discovery of polycyclic tetramate macrolactams from an environmental Streptomyces strain. *Org Lett* **12**: 4652–4654.

Carlin, D.E., Demchak, B., Pratt, D., Sage, E., and Ideker, T. (2017) Network propagation in the cytoscape cyberinfrastructure. 1–9.

Chekan, J.R., Fallon, T.R., and Moore, B.S. (2020) Biosynthesis of marine toxins. *Curr Opin Chem Biol* **59**: 119–129.

Chen, S., Zhang, C., and Zhang, L. (2022) Investigation of the Molecular Landscape of Bacterial Aromatic Polyketides by Global Analysis of Type II Polyketide Synthases. *Angew Chemie Int Ed*.

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., and Fischbach, M.A. (2014) Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters. *Cell* **158**: 412–421.

Davis, J.J., Wattam, A.R., Aziz, R.K., Brettin, T., Butler, R., Butler, R.M., Chlenski, P., Conrad, N., Dickerman, A., Dietrich, E.M., Gabbard, J.L., Gerdes, S., Guard, A., Kenyon, R.W., MacHi, D., Mao, C., Murphy-Olson, D., Nguyen, M., Nordberg, E.K., Olsen, G.J., Olson, R.D., Overbeek, J.C., Overbeek, R., Parrello, B., Pusch, G.D., Shukla, M., Thomas, C., Vanoeffelen, M., Vonstein, V., Warren, A.S., Xia, F., Xie, D., Yoo, H., and Stevens, R. (2020) The PATRIC Bioinformatics Resource Center: Expanding data and analysis capabilities. *Nucleic Acids Res* **48**: D606–D612.

DeNicola, A. (2018) Engineering expanded spliceosome function in *Saccharomyces cerevisiae.*

Douarre, P.E., Mallet, L., Radomski, N., Felten, A., and Mistou, M.Y. (2020) Analysis of COMPASS, a New Comprehensive Plasmid Database Revealed Prevalence of Multireplicon and Extensive Diversity of IncF Plasmids. *Front Microbiol* **11**:.

Dragoš, A., Andersen, A.J.C., Lozano-Andrade, C.N., Kempen, P.J., Kovács, Á.T., and Strube, M.L. (2021) Phages carry interbacterial weapons encoded by biosynthetic gene clusters. *Curr Biol* **31**: 3479-3489.e5.

Edgar, R.C. (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**: 2460–2461.

Feng, L., Gordon, M.T., Liu, Y., Basso, K.B., and Butcher, R.A. (2021) Mapping the biosynthetic pathway of a hybrid polyketide-nonribosomal peptide in a metazoan. *Nat Commun* **12**:.

Ganley, J.G. and Derbyshire, E.R. (2020) Linking genes to molecules in eukaryotic sources: An endeavor to expand our biosynthetic repertoire. *Molecules* **25**:.

Gavriilidou, A., Kautsar, S.A., Zaburannyi, N., Krug, D., Müller, R., Medema, M.H., and Ziemert, N. (2022) Compendium of secondary metabolite biosynthetic diversity encoded in bacterial genomes. *Nat Microbiol* **7**: 726–735.

Gerlt, J.A. (2017) Genomic Enzymology: Web Tools for Leveraging Protein Family Sequence-Function Space and Genome Context to Discover Novel Functions. *Biochemistry* **56**: 4293–4308.

Gerlt, J.A., Bouvier, J.T., Davidson, D.B., Imker, H.J., Sadkhin, B., Slater, D.R., and Whalen, K.L. (2015) Enzyme function initiative-enzyme similarity tool (EFI-EST): A web tool for generating protein sequence similarity networks. *Biochim Biophys Acta - Proteins Proteomics* **1854**: 1019–1037.

Goodacre, N., Aljanahi, A., Nandakumar, S., Mikailov, M., and Khan, A.S. (2018) A Reference Viral Database (RVDB) To Enhance Bioinformatics Analysis of High-Throughput Sequencing for Novel Virus Detection. *mSphere* **3**:.

Goodstein, D.M., Shu, S., Howson, R., Neupane, R., Hayes, R.D., Fazo, J., Mitros, T., Dirks, W., Hellsten, U., Putnam, N., and Rokhsar, D.S. (2012) Phytozome: A comparative platform for green plant genomics. *Nucleic Acids Res* **40**: 1178–1186.

Grigoriev, I. V., Hayes, R.D., Calhoun, S., Kamel, B., Wang, A., Ahrendt, S., Dusheyko, S., Nikitin, R., Mondo, S.J., Salamov, A., Shabalov, I., and Kuo, A. (2021) PhycoCosm, a comparative algal genomics resource. *Nucleic Acids Res* **49**: D1004–D1011.

Grigoriev, I. V., Nikitin, R., Haridas, S., Kuo, A., Ohm, R., Otillar, R., Riley, R., Salamov, A., Zhao, X., Korzeniewski, F., Smirnova, T., Nordberg, H., Dubchak, I., and Shabalov, I. (2014) MycoCosm portal: Gearing up for 1000 fungal genomes. *Nucleic Acids Res* **42**: 699–704.

Grininger, M. (2020) The role of the iterative modules in polyketide synthase evolution. *Proc Natl Acad Sci U S A* **117**: 8680–8682.

Harvey, C.J.B., Tang, M., Schlecht, U., Horecka, J., Fischer, C.R., Lin, H.C., Li, J., Naughton, B., Cherry, J., Miranda, M., Li, Y.F., Chu, A.M., Hennessy, J.R., Vandova, G.A., Inglis, D., Aiyar, R.S., Steinmetz, L.M., Davis, R.W., Medema, M.H., Sattely, E., Khosla, C., Onge, R.P.S., Tang, Y., and Hillenmeyer, M.E. (2018) HEx: A heterologous expression platform for the discovery of fungal natural products. *Sci Adv* **4**:.

Heinrich, M. (2000) Ethnobotany and its role in drug development. *Phyther Res* **14**: 479–488.

Hotaling, S., Kelley, J.L., and Frandsen, P.B. (2021) Toward a genome sequence for every animal: Where are we now? *Proc Natl Acad Sci U S A* **118**: 1–8.

Kautsar, S.A., Blin, K., Shaw, S., Navarro-Muñoz, J.C., Terlouw, B.R., van der Hooft, J.J.J., van Santen, J.A., Tracanna, V., Suarez Duran, H.G., Pascal Andreu, V., Selem-Mojica, N., Alanjary, M., Robinson, S.L., Lund, G., Epstein, S.C., Sisto, A.C., Charkoudian, L.K., Collemare, J., Linington, R.G., Weber, T., and Medema, M.H. (2020) MIBiG 2.0: a repository for biosynthetic gene clusters of known function. *Nucleic Acids Res* **48**: D454–D458.

Leebens-Mack, J.H., Barker, M.S., Carpenter, E.J., Deyholos, M.K., Gitzendanner, M.A., Graham, S.W., Grosse, I., Li, Z., Melkonian, M., Mirarab, S., Porsch, M., Quint, M., Rensing, S.A., Soltis, D.E., Soltis, P.S., Stevenson, D.W., Ullrich, K.K., Wickett, N.J., DeGironimo, L., Edger, P.P., Jordon-Thaden, I.E., Joya, S., Liu, T., Melkonian, B., Miles, N.W., Pokorny, L., Quigley, C., Thomas, P., Villarreal, J.C., Augustin, M.M., Barrett, M.D., Baucom, R.S., Beerling, D.J., Benstein, R.M., Biffin, E., Brockington, S.F., Burge, D.O., Burris, J.N., Burris, K.P., Burtet-Sarramegna, V., Caicedo, A.L., Cannon, S.B., Çebi, Z., Chang, Y., Chater, C., Cheeseman, J.M., Chen, T., Clarke, N.D., Clayton, H., Covshoff, S., Crandall-Stotler, B.J., Cross, H., dePamphilis, C.W., Der, J.P., Determann, R., Dickson, R.C., Di Stilio, V.S., Ellis, S., Fast, E., Feja, N., Field, K.J., Filatov, D.A., Finnegan, P.M., Floyd, S.K., Fogliani, B., García, N., Gâteblé, G., Godden, G.T., Goh, F. (Qi Y., Greiner, S., Harkess, A., Heaney, J.M., Helliwell, K.E., Heyduk, K., Hibberd, J.M., Hodel, R.G.J., Hollingsworth, P.M., Johnson, M.T.J., Jost, R., Joyce, B., Kapralov, M. V., Kazamia, E., Kellogg, E.A., Koch, M.A., Von Konrat, M., Könyves, K., Kutchan, T.M., Lam, V., Larsson, A., Leitch, A.R., Lentz, R., Li, F.W., Lowe, A.J., Ludwig, M., Manos, P.S., Mavrodiev, E., McCormick, M.K., McKain, M., McLellan, T., McNeal, J.R., Miller, R.E., Nelson, M.N., Peng, Y., Ralph, P., Real, D., Riggins, C.W., Ruhsam, M., Sage, R.F., Sakai, A.K., Scascitella, M., Schilling, E.E., Schlösser, E.M., Sederoff, H., Servick, S., Sessa, E.B., Shaw, A.J., Shaw, S.W., Sigel, E.M., Skema, C., Smith, A.G., Smithson, A., Stewart, C.N., Stinchcombe, J.R., Szövényi, P., Tate,

J.A., Tiebel, H., Trapnell, D., Villegente, M., Wang, C.N., Weller, S.G., Wenzel, M., Weststrand, S., Westwood, J.H., Whigham, D.F., Wu, S., Wulff, A.S., Yang, Y., Zhu, D., Zhuang, C., Zuidof, J., Chase, M.W., Pires, J.C., Rothfels, C.J., Yu, J., Chen, C., Chen, L., Cheng, S., Li, J., Li, R., Li, X., Lu, H., Ou, Y., Sun, X., Tan, X., Tang, J., Tian, Z., Wang, F., Wang, J., Wei, X., Xu, X., Yan, Z., Yang, F., Zhong, X., Zhou, F., Zhu, Y., Zhang, Y., Ayyampalayam, S., Barkman, T.J., Nguyen, N. phuong, Matasci, N., Nelson, D.R., Sayyari, E., Wafula, E.K., Walls, R.L., Warnow, T., An, H., Arrigo, N., Baniaga, A.E., Galuska, S., Jorgensen, S.A., Kidder, T.I., Kong, H., Lu-Irving, P., Marx, H.E., Qi, X., Reardon, C.R., Sutherland, B.L., Tiley, G.P., Welles, S.R., Yu, R., Zhan, S., Gramzow, L., Theißen, G., and Wong, G.K.S. (2019) One thousand plant transcriptomes and the phylogenomics of green plants. *Nature* **574**: 679–685.

Lemoine, F., Correia, D., Lefort, V., Doppelt-Azeroual, O., Mareuil, F., Cohen-Boulakia, S., and Gascuel, O. (2019) NGPhylogeny.fr: New generation phylogenetic services for non-specialists. *Nucleic Acids Res* **47**: W260–W265.

Letunic, I. and Bork, P. (2019) Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res* **47**: W256–W259.

Li, F., Lin, Z., Torres, J.P., Hill, E.A., Li, D., Townsend, C.A., and Schmidt, E.W. (2022) Sea Urchin Polyketide Synthase SpPks1 Produces the Naphthalene Precursor to Echinoderm Pigments.

Li, Y., Steenwyk, J.L., Chang, Y., Wang, Y., James, T.Y., Stajich, J.E., Spatafora, J.W., Groenewald, M., Dunn, C.W., Hittinger, C.T., Shen, X.X., and Rokas, A. (2021) A genome-scale phylogeny of the kingdom Fungi. *Curr Biol* **31**: 1653-1665.e5.

Mauri, M., Elli, T., Caviglia, G., Uboldi, G., and Azzi, M. (2017) RAWGraphs: A visualisation platform to create open outputs. *ACM Int Conf Proceeding Ser* **Part F1313**:

Medema, M.H., de Rond, T., and Moore, B.S. (2021) Mining genomes to illuminate the specialized chemistry of life. *Nat Rev Genet*.

Nayfach, S., Camargo, A.P., Schulz, F., Eloe-Fadrosh, E., Roux, S., and Kyrpides, N.C. (2021) CheckV assesses the quality and completeness of metagenome-assembled viral genomes. *Nat Biotechnol* **39**: 578–585.

Newman, D.J. and Cragg, G.M. (2020) Natural Products as Sources of New Drugs over the Nearly Four Decades from 01/1981 to 09/2019. *J Nat Prod* **83**: 770–803.

Nivina, A., Herrera Paredes, S., Fraser, H.B., and Khosla, C. (2021) GRINS: Genetic elements that recode assembly-line polyketide synthases and accelerate their diversification. *Proc Natl Acad Sci* **118**: e2100751118.

Nivina, A., Yuet, K.P., Hsu, J., and Khosla, C. (2019) Evolution and Diversity of Assembly-Line Polyketide Synthases. *Chem Rev* **119**: 12524–12547.

Paez-Espino, D., Chen, I.M.A., Palaniappan, K., Ratner, A., Chu, K., Szeto, E., Pillay, M., Huang,

J., Markowitz, V.M., Nielsen, T., Huntemann, M., Reddy, T.B.K., Pavlopoulos, G.A., Sullivan, M.B., Campbell, B.J., Chen, F., McMahon, K., Hallam, S.J., Denef, V., Cavicchioli, R., Caffrey, S.M., Streit, W.R., Webster, J., Handley, K.M., Salekdeh, G.H., Tsesmetzis, N., Setubal, J.C., Pope, P.B., Liu, W.T., Rivers, A.R., Ivanova, N.N., and Kyrpides, N.C. (2017) IMG/VR: A database of cultured and uncultured DNA viruses and retroviruses. *Nucleic Acids Res* **45**: D457–D465.

Pickett, B.E., Sadat, E.L., Zhang, Y., Noronha, J.M., Squires, R.B., Hunt, V., Liu, M., Kumar, S., Zaremba, S., Gu, Z., Zhou, L., Larson, C.N., Dietrich, J., Klem, E.B., and Scheuermann, R.H. (2012) ViPR: An open bioinformatics database and analysis resource for virology research. *Nucleic Acids Res* **40**: 593–598.

Price, M.N., Dehal, P.S., and Arkin, A.P. (2010) FastTree 2 - Approximately maximum-likelihood trees for large alignments. *PLoS One* **5**:.

Robey, M.T., Caesar, L.K., Drott, M.T., Keller, N.P., and Kelleher, N.L. (2021) An Interpreted Atlas of Biosynthetic Gene Clusters from 1000 Fungal Genomes. *PNAS* **118**: 2020.09.21.307157.

RStudio Team (2021) RStudio: Integrated Development Environment for R. *RStudio*.

Sharrar, A.M., Crits-Christoph, A., Méheust, R., Diamond, S., Starr, E.P., and Banfield, J.F. (2020) Bacterial Secondary Metabolite Biosynthetic Potential in Soil Varies with Phylum, Depth, and Vegetation Type. *MBio* **11**: 1–17.

Skinnider, M.A., Johnston, C.W., Gunabalasingam, M., Merwin, N.J., Kieliszek, A.M., MacLellan, R.J., Li, H., Ranieri, M.R.M., Webster, A.L.H., Cao, M.P.T., Pfeifle, A., Spencer, N., To, Q.H., Wallace, D.P., Dejong, C.A., and Magarvey, N.A. (2020) Comprehensive prediction of secondary metabolite structure and biological activity from microbial genome sequences. *Nat Commun* **11**: 1–9.

Torres, J.P., Lin, Z., Winter, J.M., Krug, P.J., and Schmidt, E.W. (2020) Animal biosynthesis of complex polyketides in a photosynthetic partnership. *Nat Commun* **11**: 1–12.

Torres, J.P. and Schmidt, E.W. (2019) The biosynthetic diversity of the animal world. *J Biol Chem* **294**: 17684–17692.

Verma, A., Barua, A., Ruvindy, R., Savela, H., Ajani, P.A., and Murray, S.A. (2019) The genetic basis of toxin biosynthesis in dinoflagellates. *Microorganisms* **7**: 1–29.

Wang, A.M., Jarmusch, A.K., Vargas, F., Aksenov, A.A., Gauglitz, J.M., Weldon, K., Petras, D., Silva, R., Quinn, R., Alexey, V., Hooft, J.J.J. Van Der, Mauricio, A., Rodríguez, C., Felix, L., Aceves, C.M., Panitchpakdi, M., Brown, E., Di, F., Sikora, N., Elijah, E.O., Labarta-bajo, L., and Gentry, E.C. (2019) MASST: A Web-based Basic Mass Spectrometry Search Tool for Molecules to Search Public Data.

Wang, B., Guo, F., Huang, C., and Zhao, H. (2020) Unraveling the iterative type I polyketide synthases hidden in Streptomyces . *Proc Natl Acad Sci* **117**: 201917664.

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., Crüsemann, M., Boudreau, P.D., Esquenazi, E., Sandoval-Calderón, 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., Boya P, C.A., 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., Rodríguez, 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., Müller, R., Waters, K.M., Shi, W., Liu, X., Zhang, L., Knight, R., Jensen, P.R., Palsson, B.Ø., Pogliano, K., Linington, R.G., Gutiérrez, M., Lopes, N.P., Gerwick, W.H., Moore, B.S., Dorrestein, P.C., and Bandeira, N. (2016) Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nat Biotechnol* **34**: 828–837.

Wei, B., Du, A., Zhou, Z., Lai, C., Yu, W., Yu, J., Yu, Y., Chen, J., Zhang, H., Xu, X., and Wang, H. (2021) An atlas of bacterial secondary metabolite biosynthesis gene clusters. *Environ Microbiol* **00**:

Zallot, R., Oberg, N., and Gerlt, J.A. (2021) Discovery of new enzymatic functions and metabolic pathways using genomic enzymology web tools. *Curr Opin Biotechnol* **69**: 77–90.

Zallot, R., Oberg, N., and Gerlt, J.A. (2019) The EFI Web Resource for Genomic Enzymology Tools: Leveraging Protein, Genome, and Metagenome Databases to Discover Novel Enzymes and Metabolic Pathways. *Biochemistry*.

Zallot, R., Oberg, N.O., and Gerlt, J.A. (2018) 'Democratized' genomic enzymology web tools for functional assignment. *Curr Opin Chem Biol* **47**: 77–85.

Zhu, Q., Mai, U., Pfeiffer, W., Janssen, S., Asnicar, F., Sanders, J.G., Belda-Ferre, P., Al-Ghalith, G.A., Kopylova, E., McDonald, D., Kosciolek, T., Yin, J.B., Huang, S., Salam, N., Jiao, J.Y., Wu, Z., Xu, Z.Z., Cantrell, K., Yang, Y., Sayyari, E., Rabiee, M., Morton, J.T., Podell, S., Knights, D., Li, W.J., Huttenhower, C., Segata, N., Smarr, L., Mirarab, S., and Knight, R. (2019) Phylogenomics of 10,575 genomes reveals evolutionary proximity between domains Bacteria and Archaea. *Nat Commun* **10**:.

Ziemert, N., Podell, S., Penn, K., Badger, J.H., Allen, E., and 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.

CHAPTER 4. Phylogenetic analysis of the salinipostin γ-butyrolactone gene cluster uncovers new potential for bacterial signaling-molecule diversity

4.1 Introduction to Chapter 4

Building from the idea that bioinformatic tools can be used to discover new biosynthetic potential, chapter 4 describes the discovery of the incredible diversity and distribution of the *Salinispora* salinipostin biosynthetic gene cluster. Salinipostins A-K are γ-butyrolactone compounds with rare bicyclic phoshostriesters that were first identified from *Salinispora* in anti-malarial screening assays (Schulze *et al.*, 2015). A subsequent transcriptomic analysis linked a biosynthetic gene cluster (BGC) to salinipostin biosynthesis (named *spt*) which was further confirmed with a knockout of one of the key biosynthetic genes, *spt9*, which abolished salinipostin production (Amos *et al.*, 2017). The complete description of the salinipostin biosynthetic pathway was recently described by Dr. Yuta Kudo, and the work in this chapter was inspired by his initial analysis of the *spt* BGC (Kudo *et al.*, 2020). Dr. Kudo had observed that the salinipostin BGC seemed to be distributed across the Actinobacteria phylum, and I set out to describe the diversity and distribution of *spt*, asking what other microbes have the potential to produce salinipostin-like molecules. With a fully characterized BGC, I could use the complete pathway and each of the genes in the *spt* BGC as search hooks to find gene clusters with similar gene organization. Additionally, with the biochemical knowledge that the essential gene *spt9* is an *afsA* homolog, which is required to produce the A-factor signaling molecule in *Streptomyces* bacteria (Khokhlov *et al.*, 1967; Kato *et al.*, 2007), I could compare similar gene cluster organization to hypothesize if they have the genes required to produce salinipostin-like molecules, which might then act as signaling molecules. Additional molecules have been linked to the *spt* gene cluster, including Sal-GBL1, Sal-GBL2, (Kudo *et al.*, 2020) and the salinilactone A-H (Schlawis *et al.*, 2018, 2020) which share structural similarities to other γ-butyrolactone molecules. I performed comparative

analyses of the *spt* BGC as part of the description of salinilactones D-H, however, we could not discern differences between the *spt* BGCs that could account for the varied salinilactone production among multiple *Salinispora* species (Schlawis *et al.*, 2020).

Overall, the ability to use biosynthetic knowledge about an entire BGC to identify similar gene clusters is a powerful approach to understand their diversity and distribution. By carefully studying BGCs in this context, we can observe new patterns of evolution, as seen in the *Salinispora* genus where the *spt* BGC seems to have been horizontally acquired in a location-specific manner between *Salinispora arenicola* and *Salinispora tropica*. As a follow-up to my methods described in this chapter, Appendix A details a similar approach where the gene organization of a BGC was not only used to inform the biosynthetic prediction of the new *Salinispora pacifica* molecule pacificamide, but also a comparative BGC analysis illustrated how rare the BGC and its gene organization was across all bacteria (Castro-Falcón *et al.*, 2022).

Chapter 4 (Section 4.3), in full, is a reprint of the material as it appears in *Microbial Genomics* 7(5), Creamer, K.E.; Kudo, Y; Moore, B.S.; Jensen, P.R., 2021. "Phylogenetic analysis of the salinipostin γ-butyrolactone gene cluster uncovers new potential for bacterial signalling-molecule diversity." The dissertation author was the primary investigator and author of this paper.

4.2 Chapter 4 Introduction References

Amos, G.C.A., Awakawa, T., Tuttle, R.N., Letzel, A.-C., Kim, M.C., Kudo, Y., Fenical, W., Moore, B.S., and Jensen, P.R. (2017) Comparative Transcriptomics as a Guide to Natural Product Discovery and Biosynthetic Gene Cluster Functionality. *PNAS* **114**: E11121–E11130.

Castro-Falcón, G., Creamer, K.E., Chase, A.B., Kim, M.C., Sweeney, D., Glukhov, E., Fenical, W., and Jensen, P.R. (2022) Structure and Candidate Biosynthetic Gene Cluster of a Manumycin-Type Metabolite from Salinispora pacifica. *J Nat Prod* **85**: 980–986.

Kato, J.-Y., Funa, N., Watanabe, H., Ohnishi, Y., and Horinouchi, S. (2007) Biosynthesis of y-butyrolactone autoregulators that switch on secondary metabolism and morphological development in *Streptomyces*. *Proc Natl Acad Sci* **104**: 2378–2383.

Khokhlov, A.S., Tovarova, I.I., Borisova, L.N., Pliner, S.A., Shevchenko, L.N., Kornitskaia, E.I., Ivkina, N.S., and Rapoport, I.A. (1967) A-faktor, obespechivaiushchii biosintez streptomitsina mutantnym shtammom *Actinomyces streptomycini*. *Dokl Akad Nauk SSSR* **177**: 232–235.

Kudo, Y., Awakawa, T., Du, Y.-L., Jordan, P.A., Creamer, K.E., Jensen, P.R., Linington, R.G., Ryan, K.S., and Moore, B.S. (2020) Expansion of Gamma-Butyrolactone Signaling Molecule Biosynthesis to Phosphotriester Natural Products. *ACS Chem Biol* **15**: 3253–3261.

Schlawis, C., Harig, T., Ehlers, S., Guillen-Matus, D.G., Creamer, K.E., Jensen, P.R., and Schulz, S. (2020) Extending the Salinilactone Family. *ChemBioChem* **21**: 1629 –1632.

Schlawis, C., Kern, S., Kudo, Y., Grunenberg, J., Moore, B., and Schulz, S. (2018) Structural Elucidation of Trace Components Combining GC/MS, GC/IR, DFT-Calculation and Synthesis - Salinilactones, Unprecedented Bicyclic Lactones from *Salinispora* Bacteria. *Angew Chemie Int Ed* **57**: 14921–14925.

Schulze, C.J., Navarro, G., Ebert, D., DeRisi, J., and Linington, R.G. (2015) Salinipostins A-K, long-chain bicyclic phosphotriesters as a potent and selective antimalarial chemotype. *J Org Chem* **80**: 1312–1320.

4.3 Reprint of: "Phylogenetic analysis of the salinipostin γ-butyrolactone gene cluster uncovers new potential for bacterial signalling-molecule diversity"

MICROBIAL GENOMICS

RESEARCH ARTICLE

Creamer *et al.*, *Microbial Genomics* 2021;7:000568

DOI 10.1099/mgen.0.000568

Phylogenetic analysis of the salinipostin γ-butyrolactone gene cluster uncovers new potential for bacterial signalling-molecule diversity

Kaitlin E. Creamer[1], Yuta Kudo[1]†, Bradley S. Moore[1,2] and Paul R. Jensen[1,*]

Abstract

Bacteria communicate by small-molecule chemicals that facilitate intra- and inter-species interactions. These extracellular signalling molecules mediate diverse processes including virulence, bioluminescence, biofilm formation, motility and specialized metabolism. The signalling molecules produced by members of the phylum Actinobacteria generally comprise γ-butyrolactones, γ-butenolides and furans. The best-known actinomycete γ-butyrolactone is A-factor, which triggers specialized metabolism and morphological differentiation in the genus *Streptomyces*. Salinipostins A–K are unique γ-butyrolactone molecules with rare phosphotriester moieties that were recently characterized from the marine actinomycete genus *Salinispora*. The production of these compounds has been linked to the nine-gene biosynthetic gene cluster (BGC) *spt*. Critical to salinipostin assembly is the γ-butyrolactone synthase encoded by *spt9*. Here, we report the surprising distribution of *spt9* homologues across 12 bacterial phyla, the majority of which are not known to produce γ-butyrolactones. Further analyses uncovered a large group of *spt*-like gene clusters outside of the genus *Salinispora*, suggesting the production of new salinipostin-like diversity. These gene clusters show evidence of horizontal transfer and location-specific recombination among *Salinispora* strains. The results suggest that γ-butyrolactone production may be more widespread than previously recognized. The identification of new γ-butyrolactone BGCs is the first step towards understanding the regulatory roles of the encoded small molecules in Actinobacteria.

DATA SUMMARY

All sequences analysed in this paper were retrieved from publicly accessible databases including the Joint Genome Institute (JGI) Integrated Microbial Genomes & Microbiomes system (IMG)/MER and the National Center for Biotechnology Information (NCBI) databases, with all sequence accession information included in the supplementary dataset S1 (available via Open Science Framework: https://osf.io/4g3mn/). PCR sequences produced as part of this work can be accessed at NCBI GenBank (accession numbers MW321490–MW321495) and are also listed in Table S1 (available with the online version of this article). Additionally, all sequence alignment and tree files used for the phylogenetic analyses are available through the Open Science Framework: https://osf.io/4g3mn/ with DOI 10.17605/OSF.IO/4G3MN. Supplementary material can be found on Figshare at 10.6084/m9.figshare.14325233.

INTRODUCTION

Bacteria use chemical signalling molecules to regulate gene expression in a population-dependent manner. This process, known as quorum sensing, controls group behaviours including swarming, bioluminescence, virulence, biofilm formation, cell competence, DNA uptake, public-goods production and specialized metabolism. In many

Received 28 January 2021; Accepted 24 March 2021; Published 12 May 2021

Author affiliations: [1]Center for Marine Biotechnology and Biomedicine, Scripps Institution of Oceanography, University of California San Diego, La Jolla, CA, USA; [2]Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California San Diego, La Jolla, CA, USA.

***Correspondence:** Paul R. Jensen, pjensen@ucsd.edu

Keywords: actinomycetes; bacterial signalling molecules; biosynthetic gene clusters; γ-butyrolactone; salinipostin; *Salinispora*.

Abbreviations: AHL, acyl-homoserine lactone; BGC, biosynthetic gene cluster; GI, genomic island; IMG, Integrated Microbial Genomes; JGI, Joint Genome Institute; NCBI, National Center for Biotechnology Information; NRPS, non-ribosomal peptide synthetase; PKS, polyketide synthase.

†Present address: Frontier Research Institute for Interdisciplinary Sciences, Japan Graduate School of Agricultural Science, Tohoku University, Sendai, Miyagi, Japan.

The GenBank/EMBL/DDBJ accession numbers for the PCR sequences of the *Salinispora* isolates are MW321490–MW321495.

Data statement: All supporting data, code and protocols have been provided within the article or through supplementary data files. One supplementary table, seven supplementary figures and other supplementary material are available with the online version of this article, and via the Open Science Framework (https://osf.io/4g3mn/) and Figshare (10.6084/m9.figshare.14325233).

Gram-negative bacteria, quorum sensing is mediated by acyl-homoserine lactone (AHL) autoinducers and their cognate receptors [1]. In some Gram-positive bacteria, autoinducing peptides and their respective transmembrane two-component histidine sensor kinases control similar group behaviours [2]. Among Actinobacteria, γ-butyrolactone signalling molecules regulate morphological development and specialized metabolite production. Given the importance of Actinobacteria for the production of antibiotics and other useful compounds, the discovery of new signalling molecules could facilitate the discovery of new natural products from the large number of 'cryptic' gene clusters detected in actinomycete genome sequences.

To date, the types of signalling molecules known to be produced by Actinobacteria include γ-butyrolactones [3–25], γ-butenolides [26–30], furans [31, 32], PI factor [33] and *N*-methylphenylalanyl-dehydrobutyrine diketopiperazine [34] (Fig. 1). Most of these were discovered from members of the genus *Streptomyces*. Sometimes referred to as actinobacterial 'hormones', signalling molecules are commonly

Impact Statement

Signalling molecules orchestrate a wide variety of bacterial behaviours. Among Actinobacteria, γ-butyrolactones mediate morphological changes and regulate specialized metabolism. Despite their importance, few γ-butyrolactones have been linked to their cognate biosynthetic gene clusters (BGCs). A new series of γ-butyrolactones called the salinipostins was recently identified from the marine actinomycete genus *Salinispora* and linked to the *spt* BGC. Here, we report the detection of *spt*-like gene clusters in diverse bacterial families not known for the production of this class of compounds. This finding expands the taxonomic range of bacteria that may employ this class of compounds and provides opportunities to discover new compounds associated with chemical communication.

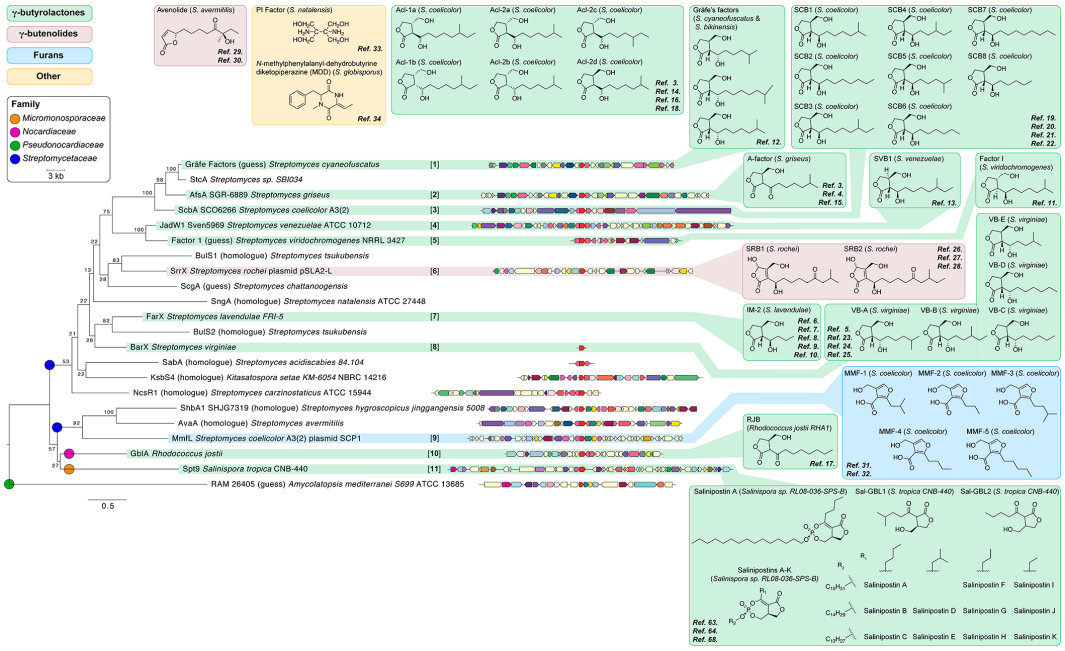


Fig. 1. Actinobacterial AfsA homologue phylogeny, gene neighbourhoods and small molecule signalling products. Maximum-likelihood phylogeny of 22 AfsA homologues created with RAxML using a LG+I+G+F ProtTest model; branches are labelled with bootstrap support (500 replicates). Scale bar represents the mean number of amino acid substitutions per site. Coloured circles indicate the actinobacterial family. Gene neighbourhoods are drawn 5′ to 3′ when genome sequences were available and aligned by AfsA homologue (red); other genes are coloured by COG function. Coloured boxes delineate γ-butyrolactones (green), γ-butenolides (red), furans (blue) and others (yellow). Compounds mapped to the tree have been experimentally linked to their respective gene cluster (references are indicated in the insets). Those not mapped to the tree have not been linked to AfsA-containing gene clusters. Bracketed numbers are used in subsequent figures to refer to specific AfsA homologues and their associated signalling molecule products.

produced in low amounts and have proven difficult to isolate and characterize. Many of these molecules not only induce the production of specialized metabolites, but also regulate bacterial morphogenesis and control complex regulatory systems [15]. The first bacterial signalling molecule discovered was A-factor (autoregulatory factor, 2-isocapryloyl-3*R*-hydroxy methyl-γ-butyrolactone) from the actinomycete *Streptomyces griseus*. It was shown to trigger sporulation and the production of the antibiotic streptomycin [3]. A-factor biosynthesis requires a γ-butyrolactone synthase and a reductase encoded by the genes *afsA* and *bprA*, respectively [15], and its elucidation revitalized the search to link signalling molecules to their biosynthetic genes [16, 32, 35–41]. Most of the biosynthetically characterized γ-butyrolactones, γ-butenolides and furans have been linked to *afsA* gene homologues via sequence similarity, biochemical verification or A-factor receptor binding assays. However, many *afsA* gene homologues observed in *Streptomyces*, *Kitasatospora* and *Amycolatopsis* genomes have yet to be linked to a small molecule [42–51]. Likewise, the Acl series of γ-butyrolactones reported from *Streptomyces coelicolor* has not been linked to an *afsA* homologue [3, 14, 18] (Fig. 1).

While most A-factor-like molecules have been identified from the genus *Streptomyces*, it remains possible that other actinobacterial taxa produce related signalling molecules. This includes the obligate marine actinomycete genus *Salinispora*, which comprises nine species: *Salinispora tropica*, *Salinispora arenicola*, *Salinispora oceanensis*, *Salinispora mooreana*, *Salinispora cortesiana*, *Salinispora fenicalii*, *Salinispora goodfellowii*, *Salinispora vitiensis* and *Salinispora pacifica* [52, 53] isolated from marine sediments [54–57], seaweeds [55] and sponges [58, 59]. This genus has proven to be a prolific source of specialized metabolites [60] including the proteasome inhibitor salinosporamide A [61], which is currently in phase III clinical trials as an anticancer agent. Whole-genome sequencing of 118 *Salinispora* strains revealed 176 distinct biosynthetic gene clusters (BGCs), of which only 25 had been linked to their products [62]. In a subsequent study, a majority of *Salinispora* BGCs were shown to be transcriptionally active under standard cultivation conditions, suggesting that many of their small molecule products were being missed using traditional detection and isolation techniques [63]. Given that little is known about the regulation of specialized metabolism in this genus, it remains possible that signalling molecules play a role in the regulation of BGC expression.

Recently, a series of compounds known as salinipostins A–K with rare bicyclic phosphotriesters were identified from a *Salinispora* sp. RL08-036-SPS-B [64]. While these compounds were identified based on anti-malarial activity against *Plasmodium falciparum*, the γ-butyrolactone part of the salinipostin structure is reminiscent of *Streptomyces* A-factor [64, 65]. Salinipostin biosynthesis was linked to the *spt* gene cluster via a knockout of the *afsA* homologue *spt9* in *Salinispora tropica* CNB-440, which resulted in the elimination of salinipostin production [63]. Subsequently, eight volatile bicyclic lactones, salinilactones A–H, were isolated and characterized from *Salinispora arenicola* CNS-205 [66, 67]. Two γ-butyrolactones, Sal-GBL1 and Sal-GBL2, were also recently characterized from multiple *Salinispora* strains [68]. The Sal-GBLs, salinilactones A–H and salinipostins A–K all share a bicyclic lactone motif and are proposed to originate from the same *spt* BGC [63, 66–68] (Fig. S1).

In this study, we set out to determine the distribution of Spt9 butyrolactone synthase homologues among sequenced bacteria and the diversity of BGCs in which they reside. We uncovered that salinipostin-like BGCs are widely distributed outside of the genus *Salinispora* and exhibit gene rearrangements and unusual gene fusions relevant to γ-butyrolactone biosynthesis. Finally, the evolutionary history of the salinipostin BGC indicates that it was horizontally transferred between *Salinispora* species at a location where they are known to co-occur.

METHODS

Identification and distribution of Spt9 homologues

The Pfam function of Spt9 was identified using the National Center for Biotechnology Information (NCBI) Conserved Domain Database prediction tool [69]. AnnoTree [70] was then used to determine the taxonomic distribution of the Spt9/AfsA Pfam03756 'A-factor biosynthesis hotdog domain-containing protein'. The top 500 Spt9 homologues were identified using the *Salinispora tropica* CNB-440 319 amino acid Spt9 sequence as a BLASTP (2.6.0+) [71] query against the Joint Genome Institute (JGI) Integrated Microbial Genomes and Microbiomes system (IMG)/MER sequence database (publicly available genomic sequence data integrated with JGI sequence data, all_img_core 2019) with an *E* value and sequence identity cut-off of 1×10^{-5} and >25%, respectively. Gene neighbourhoods were evaluated 20 kb upstream and downstream of all top Spt9 homologues. Sequences were grouped into actinobacterial family or gammaproteobacterial class. Also included were 22 previously characterized AfsA homologues, including 9 linked to the production of 34 γ-butyrolactone molecules, 1 linked to the production of two γ-butenolide molecules and 1 linked to the production of five furan molecules.

To identify Spt9 homologues within the genus *Salinispora*, protein–protein BLASTP with an *E* value cut-off of 1×10^{-5} was used to search all public *Salinispora* genomes. PCR was used to confirm the integrity of the split *spt* BGC in *Salinispora arenicola* CNS-296 and the presence of a hypothetical gene in *Salinispora pacifica* CNS-143. PCR was performed by aliquoting 90 ng genomic DNA into a PCR mixture consisting of 2× Phusion green hot start II high-fidelity PCR master mix (1.5 mM $MgCl_2$, 200 µM each dNTP, 0.4 U Phusion enzyme; Thermo Scientific), 3% DMSO and 0.5 µM of each forward and reverse primer [primer pair A – 6F (5′-ATCGAACGTGTC ATCGAATGGC-3′), 6dntransR (5′-CGTAGCCGAGGA AAGAAGCATC-3′); primer pair B – 6F, 6dntrans-IGR_R (5′-TCGTTCATCAGAGGTCCCCTTC-3′); primer pair C – 6F, 7R (5′-GATCAGATAGAGCATGGCGAGC-3′)]. PCR conditions were as follows: primer pair A (6F, 6dntransR),

30 s of initial denaturation at 98 °C, followed by 30 cycles of denaturation at 98 °C for 5 s, annealing at 66 °C for 20 s and extension at 72 °C for 30–50 s, followed by a final extension for 7 min at 72 °C; primer pair B (6F, 6dntrans-IGR_R) same as the previous but with annealing at 65.6 °C for 20 s and extension at 72 °C for 69 s; primer pair C (6F, 7R), same as the previous but with annealing at 65.7–66 °C for 20 s and extension at 72 °C for 35–50 s, followed by a final extension for 5–7 min at 72 °C. The resulting products were visualized in a 0.8 % agarose gel run in 1× TAE (Tris-acetate-EDTA buffer) at 95–97 V for 30–60 min; then, excised, purified, Sanger sequenced in forward and reverse directions (Eton Bioscience), trimmed, and mapped to their respective genomes in Geneious v8.1.9 [72].

Identification of salinipostin-like BGCs

ClusterScout [73] searches were performed to identify salinipostin-like BGCs in sequenced genomes using the following Pfam functions: Spt1 Pfam00391, Pfam01326; Spt2 Pfam00501; Spt4 Pfam00550; Spt5 Pfam07993; Spt6 Pfam00334; Spt7 Pfam01040; Spt8 Pfam00296; Spt9 Pfam03756 (Fig. S1). It should be noted that antiSMASH v4 and v5 [74, 75] do not fully identify *spt1–2* in the salinipostin butyrolactone BGC; thus, other methods were used to find *spt*-like BGCs. Independent ClusterScout searches were run with a minimum requirement of either 3, 4 or 5 Pfam matches, a maximum distance of <10 000 bp between each Pfam match, and a minimum distance of 1 bp from the scaffold edge. The boundaries of each match were extended by a maximum of 10 000 bp to help identify full biosynthetic operons. For some searches, the Spt9 Pfam was defined as essential. MultiGeneBlast [76] was also used to query the contiguous *Salinispora tropica* CNB-440 salinipostin *spt1–9* gene cluster against the NCBI GenBank Bacteria BCT subdivision database. Finally, the STRING v11 database [77] was queried using Spt1–9 to identify significant protein–protein interactions, gene neighbourhoods and gene co-occurrences within 5090 organisms. Biosynthetic clusters retrieved from each ClusterScout, MultiGeneBlast and STRING search were manually inspected for *spt* Pfams and gene organization.

Phylogenetic distribution of Spt9 homologues and salinipostin-like BGCs

A maximum-likelihood amino acid phylogeny was generated from the top 403 Spt9 homologues and an additional 22 experimentally characterized AfsA homologues. The sequences were aligned with MUSCLE [78] within the Mesquite system for phylogenetic computing [79] and analysed using ProtTest 3.4.2 [80] to determine an amino acid model for tree calculations. RAxML [81] was used to create a tree using ML+rapid bootstraps with 500 replicates. A second phylogeny was generated for the Spt9 homologues observed in salinipostin-like BGCs using the same parameters. The topologies of these trees and branch support were confirmed using PhyML [82] with SMS Smart Model Selection (AIC model selection; BIONJ tree searching, NNI tree improvement and an aLRT SH-like fast likelihood method) [83].

To test whether the *Salinispora* salinipostin BGC was acquired as an intact gene cluster, Spt1–9 protein sequences from 116 *Salinispora* genomes were aligned with MUSCLE [78] and PhyML [82] was used to calculate a phylogenetic tree for each protein with automatic SMS Smart Model Selection (AIC model selection; BIONJ tree searching, NNI tree improvement and an aLRT SH-like fast likelihood-based method) [83]. The nine Spt1–9 protein trees were compared for congruency with a concatenated *Salinispora* species tree created using the following 11 single-copy protein sequences: DnaA, GyrB1, GyrB2, PyrH, RecA, Pgi, TrpB, AtpD, SucC, RpoB and TopA, as previously reported [84]. Spt1–9 protein sequences were also concatenated (3758 total amino acid characters), aligned with MUSCLE [78], and a maximum-likelihood tree calculated using SMS and PhyML with the previously described parameters.

FigTree [85] and the Interactive Tree of Life (iTOL v4) [86] were used to visualize phylogenetic trees. Actinobacterial families were assigned using a recently proposed phylogeny [87]. Fused genes consisting of functional domains from two Spt proteins were identified using Geneious v8.1.9 [72].

RESULTS

Taxonomic distribution of Spt9 homologues

The γ-butyrolactone synthase AfsA is critical for the biosynthesis of the *Streptomyces* signalling molecule A-factor [15]. The identification of the *afsA* homologue *spt9* in the salinipostin (*spt*) BGC and its essential role in catalysing the γ-butyrolactone ring formation in salinipostin biosynthesis [63] led us to explore the distribution of Spt9 homologues among sequenced bacterial genomes. We first identified that Spt9 belongs to the Pfam03756 'A-factor biosynthesis hotdog domain-containing protein' family. It contains two AfsA-like hotdog fold superfamily domains and is distantly related to the FabA and FabZ *β*-hydroxyacyl-ACP dehydratases associated with fatty-acid biosynthesis in *Escherichia coli* [88]. Using AnnoTree [70], 1230 Spt9 Pfam03756 hits were identified out of the 27 000 reference genomes in the Genome Taxonomy Database (Fig. 2). Surprisingly, these sequences were distributed among 12 bacterial phyla, the majority of which are not known for the production of γ-butyrolactone signalling molecules. The phylum Actinobacteria contained 74% of the hits, Proteobacteria had 21% of the hits and the remainder were scattered across 10 additional phyla. Noticeably, 25% (911) of the 3579 Actinobacteria in the reference Genome Taxonomy Database contained the AfsA Pfam03756 compared to only 3% (256) of the 8882 Proteobacteria. These results inspired a more detailed analysis of Spt9 homologues among bacterial genome sequences.

Spt9 phylogeny and gene environment

We next conducted a BLASTP search to identify Spt9 homologues among the ~70 000+ bacterial genomes in the 2019 JGI IMG BLAST database [89]. The top 500 matches shared at least 25% amino acid identity with Spt9. After removing duplicate *Salinispora* sequences, 403 Spt9 homologues were further

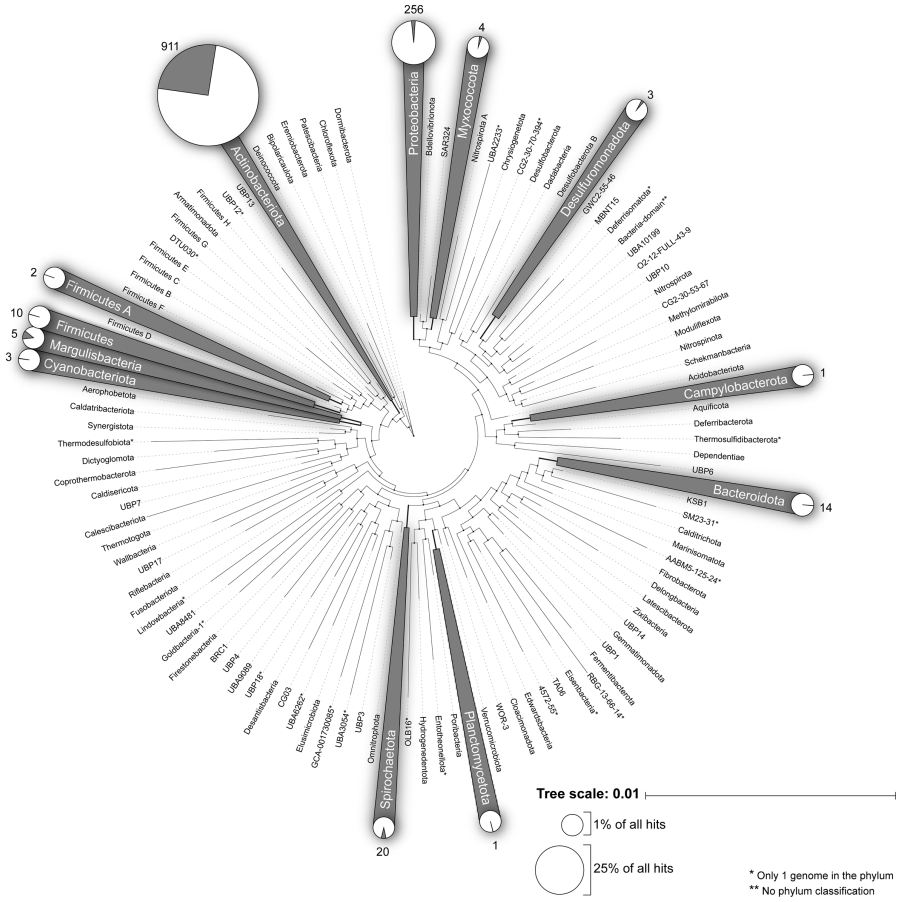


Fig. 2. Distribution of Spt9 Pfam homologues across 27 000 bacterial genomes. Shaded taxa contain Spt9/AfsA Pfam03756 (A-factor biosynthesis hotdog domain-containing protein) homologues as determined using AnnoTree. The phylogeny is from the Genome Taxonomy Database. Scale bar indicates the mean number of amino acid substitutions per site. Pie charts show the proportion of genomes in each taxon with a Spt9 homologue, with the total number of hits indicated. Pie chart sizing is proportional to the percentage of hits out of the 1230 detected across all taxa.

analysed. We additionally included 22 AfsA homologues that have been bioinformatically or experimentally linked to the production of a diverse array of γ-butyrolactones, γ-butenolides and furans (Fig. 1). A maximum-likelihood phylogeny generated using these Spt9 homologues was incongruent with the established taxonomic relationships of the strains in which the sequences were detected (Figs 3 and S2). One prominent example includes the *Salinispora* Spt9 sequences, which are sister to a homologue in *Streptomyces phaeofaciens*. These sequences fall within a larger clade comprising diverse members of the *Gammaproteobacteria* (*Citrobacter koseri* and *Dickeya* sp.) and Actinobacteria (*Cellulomonas cellasea* and *Rhodococcus* sp.) as opposed to forming a clade with the family *Micromonosporaceae* to which *Salinispora* belongs. The 22 experimentally characterized AfsA homologues are restricted to one large clade in the phylogeny and distinct from the northern end of the tree, which contains the *Salinispora* Spt9 sequences (Fig. 3).

The large number of Spt9 homologues suggests considerable potential for the discovery of new γ-butyrolactone-synthase-mediated chemical diversity (Fig. 3). New biosynthetic routes are supported by the diverse gene environments in which these Spt9 homologues are observed. For example, antiSMASH 5 [75] analyses revealed that some Spt9 homologues are close to ketosynthase- and thiolase-encoding genes, suggesting

Fig. 3. Phylogeny and gene environments of AfsA and Spt9 homologues. Condensed maximum-likelihood phylogeny of the top Spt9 homologues (403, black) and experimentally characterized AfsA homologues (22, red) linked to known molecules. The tree was calculated with a WAG+I+G+F ProtTest model with 500 replicates in RAxML; branches are labelled with bootstrap support. Scale bar represents the mean number of amino acid substitutions per site. Taxonomically coherent clades are collapsed with the number of sequences indicated in parentheses. The *Pseudomonas* sp. RIT357 Spt9 homologue was used as an outgroup. Gene neighbourhoods are drawn 5′ to 3′ and aligned with the Spt9 homologue (red); genes are coloured by COG function as annotated by JGI IMG/MER. Shaded rectangles indicate actinobacterial family or gammaproteobacterial class (see the key) with circles proportional to the number of sequences in each familial clade. Representative chemical structures are shown [γ-butyrolactones – salinipostin A from *Salinispora tropica* CNB-440, A-factor from *Streptomyces griseus*; furan – methylenomycin furan MMF-1 from *Streptomyces coelicolor* A3(2)] and bracketed numbers correspond to the AfsA homologues and their associated compounds in Fig. 1. Stars indicate salinipostin-like BGCs.

they are part of larger polyketide synthase (PKS) gene clusters. Among *Salinispora* strains, six Spt9 homologues were observed outside of the *spt* BGC. Two of these were observed in *Salinispora oceanensis* strains CNT-124 and CNT-584, each of which contains a *spt9* gene in a type II PKS BGC in addition to the *spt* BGC. The other four were observed in *Salinispora fenicalii* strains CNT-569 and CNR-942, which have *spt9* gene homologues in both a type II PKS BGC and a butyrolactone non-ribosomal peptide synthetase (NRPS) BGC while lacking the *spt* BGC.

Despite incongruence with the species phylogeny, the gene environments surrounding some Spt9 homologues are conserved. For example, many *Streptomyces* and *Kitasatospora* species in the centre of the tree have a *bprA* homologue (dark pink) next to the *spt9* homologues as required for A-factor biosynthesis in *Streptomyces griseus* (Fig. 3). The three sequences that share the highest similarity to Spt9 in *Salinispora* (observed in *Streptomyces phaeofaciens*, *Dickeya* sp. and *Citrobacter koseri*) also contain *spt4* acyl carrier protein gene homologues (pale blue). Below the *Salinispora* Spt9 clade, conservation of two 3-oxoacyl (acyl-carrier-protein) synthases (light green), an acyl carrier protein (light blue, *spt4* homologues), a hydrolase (pale yellow) and a 3-oxoacyl-(acyl-carrier-protein) reductase (light blue) is observed across taxonomically diverse *Streptomycetaceae*, *Nocardiopsaceae*, *Nocardiaceae*, *Thermomonosporaceae*, *Pseudonocardiaceae* and *Streptosporangiaceae* strains. At the bottom of the tree, the Spt9 homologue in the gammaproteobacterial outgroup *Pseudomonas* sp. RIT357 shares conserved genes with other diverse families of bacteria including *Gammaproteobacteria*, *Nocardiaceae*, *Streptomycetaceae* and *Pseudonocardiaceae* with a putative hydrolase of the haloacid dehydrogenase (HAD) superfamily (light yellow), a cytochrome P450 (light teal) and a MFS (major facilitator superfamily) protein transporter (light yellow). While evidence of gene conservation around Spt9 homologues suggests some functional similarities in the *spt9* gene neighbourhoods across diverse bacterial families, the overall diversity of gene environments illustrates the potential for new routes of γ-butyrolactone, γ-butenolide and furan production among bacteria not known to produce these molecules.

Targeted search for *spt*-like BGCs

A number of *spt9* gene neighbourhoods outside of *Salinispora* caught our attention due to similarities with the salinipostin BGC (Fig. 3). To search more thoroughly for *spt*-like BGCs among sequenced genomes, we searched for *spt*-like BGCs using ClusterScout [73], MultiGeneBlast [76] and STRING v11 [77]. These efforts led to the identification of 91 *spt*-like BGCs spanning six actinomycete families within the genera *Nocardia*, *Gordonia*, *Tsukamurella*, *Mycobacterium*, *Dietzia, Streptomyces*, *Kitasatospora, Rhodococcus* and *Kutzneria* (Fig. 4). All of these BGCs possess *spt1, spt5*, *spt6*, *spt7* and *spt9* homologues, with *spt9* towards the 3′ end of the cluster as seen in *Salinispora*. Notably, none of these *spt*-like BGCs contain the flavin-dependent oxidoreductase *spt8*, whose role is unknown in salinipostin biosynthesis, and none of the *spt*-like BGCs have been linked to the small molecules they encode.

A maximum-likelihood phylogeny of the Spt9 homologues observed in *spt*-like BGCs clearly delineates them from the AfsA homologues linked to γ-butyrolactone, γ-butenolide and furan biosynthesis (Fig. 4). Compared with the nine gene salinipostin BGC identified in *Salinispora* (Fig. S1), we observed gene reorganizations and fusions in other bacteria (Fig. 4). Most notably, *spt2* and *spt3* are fused across the large clade bracketed by *Nocardia* and *Dietzia timorensis*, as well as two *Kutzneria* species and five *Streptomyces* species at the most southern part of the tree. This fusion is conserved across most BGCs except for those observed in *Salinispora*, four *Streptomyces* species and *Rhodococcus rhodnii* NRRL B-16535. A second gene fusion is observed between *spt6* and *spt9* in *D. timorensis*. Alignment of the *spt2*, *spt3*, *spt6* and *spt9* fused and individual genes reveals maintenance of the functional domains (Fig. S3). These gene fusions, also known as Rosetta gene fusions [90, 91], suggest a functional interaction between the encoded proteins in the biosynthesis of salinipostin-like γ-butyrolactone molecules. The gene fusions could have arisen from the single-domain *spt2*, *spt3*, *spt6* and *spt9* genes in the *Salinispora spt* BGC and, thus, suggest some selective advantage for these co-localized biosynthetic genes to become fused.

Many of the *spt*-like BGCs differed in gene order compared to that observed in *Salinispora*, while others contained extra genes in the cluster including a nitroreductase (dark pink in Fig. 4). As noted, the flavin-dependent oxidoreductase *spt8* (Fig. S1) was unique to the *Salinispora spt* BGC, yet has no proposed function in salinipostin [68] or salinilactone [66, 67] biosynthesis. Similarly, some *Streptomyces spt*-like BGCs did not contain *spt2* (AMP-ligase) and *spt4* (acyl carrier protein) homologues, which are proposed to help load and carry the R_2 aliphatic sidechain during salinipostin biosynthesis, respectively [68]. These observations suggest additional structural diversity remains to be discovered. Furthermore, two BGCs found in *Kitasatospora cheerisanensis* and *Frankia* sp. contained *spt9* and *spt2* homologues next to a type I PKS, suggesting a potential role in PKS BGC regulation as observed in methylenomycin biosynthesis [31]. We observed that some of the 91 *spt*-like BGCs occur in different genomic locations within the same genus, which could support BGC migration or horizontal gene transfer. However, there is also evidence of vertical inheritance based on gene conservation in some strains. To investigate this further, we focused on the genus *Salinispora*, where BGC migration and transfer events have been previously reported [62].

The *spt* BGC in the genus *Salinispora*

The salinipostin BGC (*spt1–9*) is highly conserved within the genus *Salinispora* [62]. Notably, only 5 of 118 strains with available genome sequences lack the *spt* BGC. These belong to the recently described species *Salinispora fenicalii*, *Salinispora goodfellowii* and *Salinispora vitiensis* (Fig. S4a). At the species level, the *spt* BGC is commonly observed in the same

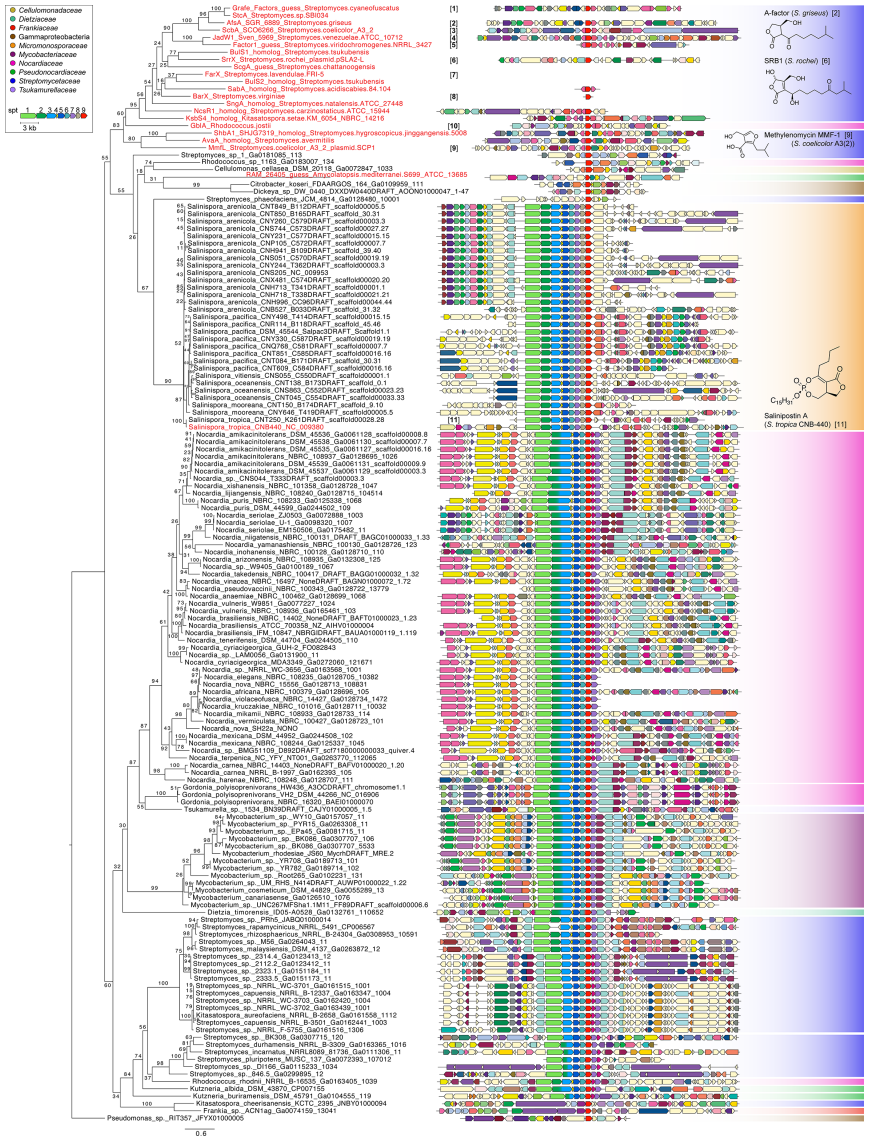


Fig. 4. Phylogeny of Spt9 homologues within salinipostin-like BGCs. The maximum-likelihood phylogeny was calculated with a WAG+I+G+F ProtTest model with 500 replicates in RAxML; branches are labelled with bootstrap support. The branch length scale bar represents the mean number of amino acid substitutions per site. Gene neighbourhoods are drawn 5′ to 3′ and aligned with the Spt9 homologue (red). The names of AfsA homologues linked to the production of specific compounds are coloured in red. Gene fusions are shown with approximate transition points and neighbouring genes are coloured by COG function as annotated by JGI IMG/MER. Coloured rectangles indicate actinobacterial family or gammaproteobacterial class (see the key). Representative chemical structures are shown [γ-butyrolactones – A-factor from *Streptomyces griseus*, salinipostin A from *Salinispora tropica* CNB-440; γ-butenolide – SRB-1 from *Streptomyces rochei*; furan – methylenomycin furan MMF-1 from *Streptomyces coelicolor* A3(2)] and bracketed numbers correspond to the AfsA homologues and associated signalling molecule products in Fig. 1.

genomic environment (Fig. S5) within previously defined genomic islands (GIs) [62, 92]. For example, it occurs in GI 20 in *Salinispora tropica* [62] and GI 15 [62] in most *Salinispora arenicola* and *Salinispora pacifica* strains with notable conservation of upstream and downstream regions.

Variations in the *spt* BGC are observed in three *Salinispora* strains (Fig. S5). In *Salinispora arenicola* CNS-296, *spt1–6* and *spt7–9* are split onto different contigs and flanked by transposases. Targeted PCR amplifications of *spt6* and the neighbouring transposase confirmed that the BGC is indeed split (Fig. S6a). Attempts to amplify a region between *spt7* and the downstream hypothetical gene resulted in multiple PCR products likely due to multiple copies of the hypothetical gene in the *Salinispora arenicola* CNS-296 genome and were, thus, uninformative. However, primers spanning *spt6–7* resulted in a product that was 3.2 kb larger than what was observed from a contiguous BGC in *Salinispora arenicola* CNQ-884 (Fig. S6b). Sequences obtained from the ends of this amplicon mapped poorly to *spt6* (67% identity) and better to *spt7* (99% identity), providing additional support for an insertion between *spt6–7* in the *Salinispora arenicola* CNS-296 *spt* BGC. It remains to be determined whether *Salinispora arenicola* CNS-296 produces salinipostins. In contrast, the detection of *spt* genes on multiple contigs in *Salinispora arenicola* CNT-088 and *Salinispora pacifica* CNS-143 is likely due to poor genome assembly [93]. The hypothetical gene annotated between *spt6–7* in *Salinispora pacifica* CNS-143 is also likely an error given that PCR products spanning *spt6–7* in this strain and *Salinispora arenicola* CNQ-884, where the *spt* BGC is contiguous, yielded amplicons of the same size and with the same conserved *spt6–7* domains (Fig. S6b).

While conservation of the *spt* BGC within *Salinispora* supports vertical inheritance, the Spt9 phylogeny (Figs S4b and S5) reveals incongruencies with the established *Salinispora* species phylogeny [52, 62] (Fig. S4a) that are consistent with horizontal gene transfer. In one example, Spt9 sequences from *Salinispora vitiensis* CNS-055 and *Salinispora cortesiana* CNY-202 occur within the *Salinispora oceanensis* clade as opposed to outside of it (Figs S4 and S5). A more pronounced example is the placement of *Salinispora tropica* Spt9 sequences within the larger *Salinispora arenicola* clade, suggesting the former acquired the sequences from the latter. This transfer or recombination event appears to have involved the entire BGC, since all *Salinispora tropica* Spt1–8 sequences share a similar phylogeny (Fig. S4b). As predicted, a concatenated phylogeny of Spt1–9 (Fig. 5a) is incongruent with the *Salinispora* species phylogeny (Fig. S4a) and supports a horizontal exchange of the BGC between *Salinispora arenicola* and *Salinispora tropica*. Mapping the geographical origin of these strains onto the tree reveals that all 12 *Salinispora tropica* and the 6 most closely related *Salinispora arenicola* sequences all originated from ocean sediments collected in the Bahamas and the Yucatán (Fig. 5b). A closer examination reveals that the two species co-occur at four of the five collection sites (Fig. S7). This geographical proximity would provide opportunities for BGC horizontal gene transfer to occur.

DISCUSSION

Specialized metabolites that function as signalling molecules regulate important functional traits in bacteria. However, only a small number of bacterial signalling molecules have been identified to date. This may be because they are small in size, generally produced in low yields, and often lack activity in the bioassays commonly used to guide small molecule discovery. γ-Butyrolactones represent an important class of signalling molecules produced by Actinobacteria (Fig. 1). The salinipostins, salinilactones, Sal-GBL1 and Sal-GBL2 were recently reported from the marine actinomycete genus *Salinispora* [64, 66–68], and bear structural similarities to previously characterized actinomycete γ-butyrolactones.

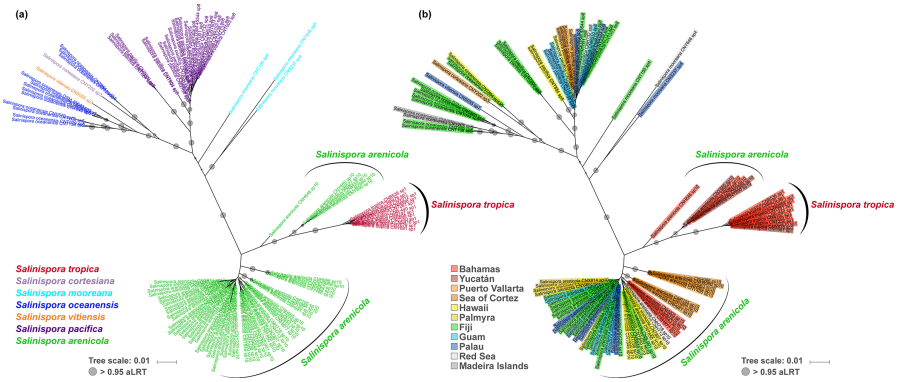


Fig. 5. Concatenated Spt1–9 phylogeny. (a) Coloured by *Salinispora* species. (b) Coloured by *Salinispora* strain isolation location. The maximum-likelihood tree was calculated in PhyML with a Smart Model Selection HIVb+G+I+F model and midpoint-rooting; branches have proportional circles representing aLRT branch support. Scale bar represents the mean number of amino acid substitutions per site.

Linkage between the biosynthesis of these compounds and the γ-butyrolactone synthase Spt9 encoded by the salinipostin *spt* BGC led us to more broadly explore the potential for signalling molecule production by assessing the distribution of this protein among sequenced bacterial genomes. Surprisingly, we detected Spt9 homologues across 12 diverse bacterial phyla, many of which are not known to produce γ-butyrolactones (Fig. 2). Despite the unexpectedly wide distribution of Spt9 homologues, only 285 of the ~25 500 currently described microbial natural products in The Natural Products Atlas [94] contain a γ-butyrolactone moiety. Of these, only 14 have been linked to their respective BGCs in the Minimum Information about a Biosynthetic Gene cluster (MIBiG) repository [95] and only four of these, including the salinipostins in *Salinispora*, A-factor in *Streptomyces griseus*, SCB1-3 in *Streptomyces coelicolor* A3(2) and lactonamycin in *Streptomyces rishiriensis*, contain an Spt9 homologue. Thus, opportunities remain to identify the products of Spt9-containing BGCs and to establish formal links between these compounds and their biosynthetic origins. Our results suggest that the production of γ-butyrolactones and related compounds may be more common than previously recognized.

A phylogenetic tree of the top 403 Spt9 homologues, including experimentally characterized AfsA homologues, showed that the associated γ-butyrolactone, γ-butenolide and furan signalling molecules are restricted to a clade that is distinct from the majority of uncharacterized Spt9 sequences (Fig. 3). The Spt9 tree also showed major incongruencies with recognized actinobacterial and gammaproteobacterial classification, suggesting extensive horizontal gene transfer. The genomic environments around the Spt9 homologues were diverse, suggesting the potential production of considerable chemical diversity. It remains to be seen whether all of these Spt9 homologues catalyse γ-butyrolactone synthase-like reactions, especially when they are distantly related to experimentally characterized AfsA homologues. Heterologous expression to determine whether these Spt9 homologues perform the canonical AfsA condensation reaction that assembles a fatty acid ester (*β*-ketoacyl-DHAP ester) intermediate is a next step towards establishing their functionality [15].

Surprisingly, we discovered a large clade of Spt9 homologues in the genus *Nocardia* that occurred in operons with similar structure to the *Salinispora spt* BGC (starred in Fig. 3). To date, no small molecules isolated from *Nocardia* spp. have been linked to these BGCs. Differences between the *Salinispora spt* BGC and the 91 *spt*-like BGCs observed both in *Nocardia* and other genera (Fig. 4) include the absence of *spt8* (a flavin-dependent oxidoreductase) and gene organization, with *spt7* occurring after *spt9*. Several *spt*-like BGCs also have an additional nitroreductase gene between *spt9* and *spt7* (Fig. 4), suggesting the production of a γ-butyrolactone with a reduced nitrogen or nitro functional group. Other *spt*-like BGCs lack the AMP-ligase *spt2* and the acyl carrier protein *spt4*, suggesting the products may lack the extended aliphatic sidechain observed in the salinipostins. These variations further support the production of new chemical diversity and provide opportunities to link structural changes to BGC evolution.

Also of note are the *spt2–spt3* and *spt6–spt9* gene fusions observed in the genera *Nocardia*, *Gordonia*, *Tsukamurella*, *Mycobacterium*, *Dietzia* and *Streptomyces*. Both pairs of fused genes appear functional based on the maintenance of conserved functional domains (Fig. S3). *D. timorensis* is the only strain with both *spt2–spt3* and *spt6–spt9* fusions, and is sister to the large clade containing the other gene fusions. Protein fusions can arise when clustered genes are co-transcribed and co-translated, providing evidence of functional interaction and, perhaps, a selective advantage over individual proteins [90, 91, 96]. The gene fusions observed in the *spt*-like BGCs appear to represent the evolution of more complex, multifunctional proteins in these strains. The *spt2–spt3* and *spt6–spt9* gene fusions are similar to recently described multi-domain enzyme fusions in the desferrioxamine (*des*) BGC, where they are hypothesized to contribute to chemical diversification [96]. Two additional fusions involving an AfsA/Spt9 homologue were observed within *trans*-AT PKS modules associated with gladiofungin and gladiostatin biosynthesis, where AfsA functions for unprecedented offloading and butenolide formation [97, 98]. Identifying unusual gene fusions such as these represents an exciting new avenue for genome-mining-driven natural product discovery [96, 99].

Conservation of the *spt* BGC in the 116 *Salinispora* genomes examined here suggests it was present in a common ancestor of the genus. Yet, the *Salinispora spt* phylogeny is incongruent with the established species phylogeny (Figs 5a and S4). These incongruencies are likely due to horizontal gene transfer events, which have been identified as important avenues of BGC transfer and diversification, especially in Actinobacteria [100]. The most apparent of these events is represented by the clustering of the *Salinispora tropica* Spt9 sequences within the *Salinispora arenicola* clade. All 12 *Salinispora tropica* Spt1–9 sequences share the same evolutionary history (Fig. S4), providing evidence that the horizontal gene transfer affected the entire *spt* BGC (Fig. 5a). Co-localization of *Salinispora tropica* and *Salinispora arenicola* in Bahamian and Yucatán sediments provides spatial opportunities for these exchange events to occur (Figs 5b and S7). BGC exchange is well documented in the genus *Salinispora* and has been linked to gene gain, loss, duplication and divergence in lineage-specific patterns [62, 84, 101]. It remains unknown whether the *Salinispora arenicola spt* locus that appears to have replaced the ancestral version or been acquired *de novo* in *Salinispora tropica* provides a selective advantage or affects the compounds produced. The apparent interspecies exchange of *spt* adds to growing evidence that this process occurs between both closely and distantly related bacteria, as seen in the granaticin, coronafacoyl phytotoxins, tunicamycin, foxicin, antimycin, streptomycin and bicyclomycin BGCs [100]. Overall, acquisition of the *spt* BGC by *Salinispora tropica* highlights the importance of understanding the functional roles of its products and the effects of these exchange events on population and species-level dynamics.

The *spt* BGC has been linked to the production of both the salinipostins A–K [63, 64, 68], Sal-GBL1 and Sal-GBL2 [68], and the salinilactones A–H [66, 67], which share structural similarities to the A-factor family of γ-butyrolactone signalling molecules. Additionally, *Salinispora arenicola* and *Salinispora pacifica* produce two additional AHLs that have yet to be linked to their biosynthetic origins [102]. AHLs are the most common class of autoinducer signalling molecules produced by Gram-negative bacteria; thus, *Salinispora* appears to employ both γ-butyrolactone and homoserine lactone signalling molecules. While further studies are needed to understand the ecological functions of the *spt* products, there is evidence that lactone signalling molecules affect microbial community organization and function [103], and can elicit specialized metabolite production [40, 104, 105]. Thus, the small molecule products of *spt* BGCs may regulate the expression of other biosynthetic pathways in *Salinispora*. In support of this, *spt9* was detected within *Salinispora* PKS and NRPS gene clusters. This is reminiscent of the methylenomycin BGC in *Streptomyces coelicolor* A3(2), where methylenomycin furan (MMF) signalling molecules induce methylenomycin production [106]. Additionally, we identified an *spt*-like BGC neighbouring the recently identified cyphomycin PKS BGC in a Brazilian *Streptomyces* sp. ISID311 isolated from the fungus-growing ant *Cyphomyrmex* sp. [107]. None of the genes in this *spt*-like BGC have been linked to cyphomycin biosynthesis [107], which suggests they encode a different small molecule that may have a regulatory role. The recently reported total synthesis of salinipostin [108] and the identification of molecules from orphan *spt*-like BGCs could support future studies to explore the roles of signalling compounds in regulating actinomycete specialized metabolism.

Our results reveal unexplored biosynthetic potential related to γ-butyrolactone signalling molecules in bacteria. The γ-butyrolactone synthase *spt9* is broadly distributed among diverse bacteria and observed in a wide range of gene environments suggesting the potential for unrealized chemical diversity. Experimentally characterized γ-butyrolactone, γ-butenolide and furan BGCs are largely restricted to the genus *Streptomyces*, yet *spt*-like BGCs are observed among bacterial genera that are not widely recognized for the production of signalling molecules. Evidence of gene fusions and gene gain/loss in the newly described *spt*-like BGCs suggest that new chemical diversity awaits discovery within this unusual class of compounds.

Funding information
This research was supported by the National Institutes of Health (grant no. 5R01GM085770 to P.R.J. and B.S.M.), the National Science Foundation Graduate Research Fellowship Program (grant no. DGE-1650112 to K.E.C.), and the Japan Society for Promotion of Science (JSPS Overseas Research Fellowship to Y.K.). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation nor the other funding providers.

Acknowledgements
We thank our colleagues Dr Leesa J. Klau and Dr Henrique R. Machado for advice on analyses and figure design.

Conflicts of interest
The authors declare that there are no conflicts of interest.

References

1. Papenfort K, Bassler BL. Quorum sensing signal–response systems in Gram-negative bacteria. *Nat Rev Microbiol* 2016;14:576–588.
2. Novick RP, Geisinger E. Quorum sensing in staphylococci. *Annu Rev Genet* 2008;42:541–564.
3. Khokhlov AS, TovarovaBLN, Pliner SA, Shevchenko LN, Kornitskaia EI *et al*. A-faktor, obespechivaiushchii biosintez streptomitsina mutantnym shtammom *Actinomyces streptomycini*. *Dokl Akad Nauk SSSR* 1967;177:232–235.
4. Ando N, Matsumori N, Sakuda S, Beppu T, Horinouchi S. Involvement of AfsA in A-factor biosynthesis as a key enzyme. *J Antibiot* 1997;50:847–852.
5. Lee YJ, Kitani S, Nihira T. Null mutation analysis of an afsA-family gene, barX, that is involved in biosynthesis of the γ-butyrolactone autoregulator in *Streptomyces virginiae*. *Microbiology* 2010;156:206–210.
6. Sato K, Nihira T, Sakuda S, Yanagimoto M, Yamada Y. Isolation and structure of a new butyrolactone autoregulator from *Streptomyces* sp. FRI-5. *J Ferment Bioeng* 1989;68:170–173.
7. Hashimoto K, Nihira T, Sakuda S, Yamada Y. IM-2, a butyrolactone autoregulator, induces production of several nucleoside antibiotics in *Streptomyces* sp. FRI-5. *J Ferment Bioeng* 1992;73:449–455.
8. Kitani S, Yamada Y, Nihira T. Gene replacement analysis of the butyrolactone autoregulator receptor (FarA) reveals that FarA acts as a novel regulator in secondary metabolism of *Streptomyces lavendulae* FRI-5. *J Bacteriol* 2001;183:4357–4363.
9. Kitani S, Iida A, Izumi T, Maeda A, Yamada Y *et al*. Identification of genes involved in the butyrolactone autoregulator cascade that modulates secondary metabolism in *Streptomyces lavendulae* FRI-5. *Gene* 2008;425:9–16.
10. Waki M, Nihira T, Yamada Y. Cloning and characterization of the gene (*farA*) encoding the receptor for an extracellular regulatory factor (IM-2) from *Streptomyces* sp. strain FRI-5. *J Bacteriol* 1997;179:5131–5137.
11. Gräfe U, Schade W, Eritt I, Fleck WF, Radics L. A new inducer of anthracycline biosynthesis from *Streptomyces viridochromogenes*. *J Antibiot* 1982;35:1722–1723.
12. Gräfe U, Reinhardt G, Schade W, Eritt I, Fleck WF *et al*. Interspecific inducers of cytodifferentiation and anthracycline biosynthesis from *Streptomyces bikinensis* and *S. cyaneofuscatus*. *Biotechnol Lett* 1983;5:591–596.
13. Zou Z, Du D, Zhang Y, Zhang J, Niu G *et al*. A γ-butyrolactone-sensing activator/repressor, JadR3, controls a regulatory mini-network for jadomycin biosynthesis. *Mol Microbiol* 2014;94:490–505.
14. Joo H-S, Yang Y-H, Lee C-S, Kim J-H, Kim B-G. Fragmentation study on butanolides with tandem mass spectrometry and its application for the screening of ScbR-captured quorum sensing molecules in *Streptomyces coelicolor* A3(2). *Rapid Commun Mass Spectrom* 2007;21:764–770.
15. Kato J, Funa N, Watanabe H, Ohnishi Y, Horinouchi S. Biosynthesis of y-butyrolactone autoregulators that switch on secondary metabolism and morphological development in *Streptomyces*. *Proc Natl Acad Sci USA* 2007;104:2378–2383.
16. Efremenkova OV. A-factor-like autoregulators. *Russ J Bioorganic Chem* 2016;42:457–472.
17. Ceniceros A, Dijkhuizen L, Petrusma M. Molecular characterization of a *Rhodococcus jostii* RHA1 γ-butyrolactone(-like) signalling molecule and its main biosynthesis gene *gblA*. *Sci Rep* 2017;7:17743.
18. Onoprienko VV, Anisova LN, Blinova IN, Efremenkova OV, Koz'min YP. Bioregulators of *Streptomyces coelicolor* A3(2). *In VII Sovetsko-Indiĭskiĭ Simpozium Po Khimii Prirodnykh Soedineniĭ*. Tbilisi; 1983. pp. 111–112.

19. Takano E, Nihira T, Hara Y, Jones JJ, Gershater CJL *et al*. Purification and structural determination of SCB1, a gamma-butyrolactone that elicits antibiotic production in *Streptomyces coelicolor* A3(2). *J Biol Chem* 2000;275:11010–11016.
20. Hsiao N-H, Söding J, Linke D, Lange C, Hertweck C *et al*. ScbA from *Streptomyces coelicolor* A3(2) has homology to fatty acid synthases and is able to synthesize γ-butyrolactones. *Microbiology* 2007;153:1394–1404.
21. Hsiao NH, Nakayama S, Merlo ME, de Vries M, Bunet R *et al*. Analysis of two additional signaling molecules in *Streptomyces coelicolor* and the development of a butyrolactone-specific reporter system. *Chem Biol* 2009;16:951–960.
22. Sidda JD, Poon V, Song L, Wang W, Yang K *et al*. Overproduction and identification of butyrolactones SCB1-8 in the antibiotic production superhost: *Streptomyces* M1152. *Org Biomol Chem* 2016;14:6390–6393.
23. Yamada Y, Sugamura K, Kondo K, Yanagimoto M, Okada H. The structure of inducing factors for virginiamycin production in *Streptomyces virginiae*. *J Antibiot* 1987;40:496–504.
24. Kondo K, Higuchi Y, Sakuda S, Nihira T, Yamada Y. New virginiae butanolides from *Streptomyces virginiae*. *J Antibiot* 1989;42:1873–1876.
25. Kawachi R, Akashi T, Kamitani Y, Sy A, Wangchaisoonthorn U *et al*. Identification of an AfsA homologue (BarX) from *Streptomyces virginiae* as a pleiotropic regulator controlling autoregulator biosynthesis, virginiamycin biosynthesis and virginiamycin M1 resistance. *Mol Microbiol* 2000;36:302–313.
26. Arakawa K, Mochizuki S, Yamada K, Noma T, Kinashi H. γ-Butyrolactone autoregulator-receptor system involved in lankacidin and lankamycin production and morphological differentiation in *Streptomyces rochei*. *Microbiology* 2007;153:1817–1827.
27. Arakawa K, Tsuda N, Taniguchi A, Kinashi H. The butenolide signaling molecules SRB1 and SRB2 induce lankacidin and lankamycin production in *Streptomyces rochei*. *Chembiochem* 2012;13:1447–1457.
28. Yamamoto S, He Y, Arakawa K, Kinashi H. γ-Butyrolactone-dependent expression of the *Streptomyces* antibiotic regulatory protein gene *srrY* plays a central role in the regulatory cascade leading to lankacidin and lankamycin production in *Streptomyces rochei*. *J Bacteriol* 2008;190:1308–1316.
29. Kitani S, Miyamoto KT, Takamatsu S, Herawati E, Iguchi H *et al*. Avenolide, a *Streptomyces* hormone controlling antibiotic production in *Streptomyces avermitilis*. *Proc Natl Acad Sci USA* 2011;108:16410–16415.
30. Zhu J, Sun D, Liu W, Chen Z, Li J *et al*. AvaR2, a pseudo γ-butyrolactone receptor homologue from *Streptomyces avermitilis*, is a pleiotropic repressor of avermectin and avenolide biosynthesis and cell growth. *Mol Microbiol* 2016;102:562–578.
31. Corre C, Song L, O'Rourke S, Chater KF, Challis GL. 2-Alkyl-4-hydroxymethylfuran-3-carboxylic acids, antibiotic production inducers discovered by *Streptomyces coelicolor* genome mining. *Proc Natl Acad Sci USA* 2008;105:17510–17515.
32. Sidda JD, Corre C. Gamma-butyrolactone and furan signaling systems in *Streptomyces*. *Methods Enzymol* 2012;517:71–87.
33. Recio E, Colinas A, Rumbero A, Aparicio JF, Martín JF. PI factor, a novel type quorum-sensing inducer elicits pimaricin production in *Streptomyces natalensis*. *J Biol Chem* 2004;279:41586–41593.
34. Matselyukh B, Mohammadipanah F, Laatsch H, Rohr J, Efremenkova O *et al*. N-methylphenylalanyl-dehydrobutyrine diketopiperazine, an A-factor mimic that restores antibiotic biosynthesis and morphogenesis in *Streptomyces globisporus* 1912-B2 and *Streptomyces griseus* 1439. *J Antibiot* 2015;68:9–14.
35. Horinouchi S. A microbial hormone, A-factor, as a master switch for morphological differentiation and secondary metabolism in *Streptomyces griseus*. *Front Biosci* 2002;7:d2045-57.
36. Takano E. γ-Butyrolactones: *Streptomyces* signalling molecules regulating antibiotic production and differentiation. *Curr Opin Microbiol* 2006;9:287–294.
37. Nishida H, Ohnishi Y, Beppu T, Horinouchi S. Evolution of γ-butyrolactone synthases and receptors in *Streptomyces*. *Environ Microbiol* 2007;9:1986–1994.
38. Willey JM, Gaskell AA. Morphogenetic signaling molecules of the streptomycetes. *Chem Rev* 2011;111:174–187.
39. Polkade AV, Mantri SS, Patwekar UJ, Jangid K. Quorum sensing: An under-explored phenomenon in the phylum *Actinobacteria*. *Front Microbiol* 2016;7:131.
40. Niu G, Chater KF, Tian Y, Zhang J, Tan H. Specialised metabolites regulating antibiotic biosynthesis in *Streptomyces* spp. *FEMS Microbiol Rev* 2016;40:554–573.
41. Daniel-Ivad M, Pimentel-Elardo S, Nodwell JR. Control of specialized metabolism by signaling and transcriptional regulation: opportunities for new platforms for drug discovery? *Annu Rev Microbiol* 2018;72:25–48.
42. Lee KM, Lee C-K, Choi S-U, Park H-R, Kitani S *et al*. Cloning and in vivo functional analysis by disruption of a gene encoding the γ-butyrolactone autoregulator receptor from *Streptomyces natalensis*. *Arch Microbiol* 2005;184:249–257.
43. Healy FG, Eaton KP, Limsirichai P, Aldrich JF, Plowman AK *et al*. Characterization of γ-butyrolactone autoregulatory signaling gene homologs in the angucyclinone polyketide WS5995B producer *Streptomyces acidiscabies*. *J Bacteriol* 2009;191:4786–4797.
44. Choi S-U, Lee C-K, Hwang Y-I, Kinoshita H, Nihira T. Cloning and functional analysis by gene disruption of a gene encoding a γ-butyrolactone autoregulator receptor from *Kitasatospora setae*. *J Bacteriol* 2004;186:3423–3430.
45. Ichikawa N, Oguchi A, Ikeda H, Ishikawa J, Kitani S *et al*. Genome sequence of *Kitasatospora setae* NBRC 14216T: an evolutionary snapshot of the family *Streptomycetaceae*. *DNA Res* 2010;17:393–406.
46. Aroonsri A, Kitani S, Hashimoto J, Kosone I, Izumikawa M *et al*. Pleiotropic control of secondary metabolism and morphological development by KsbC, a butyrolactone autoregulator receptor homologue in *Kitasatospora setae*. *Appl Environ Microbiol* 2012;78:8015–8024.
47. Intra B, Euanorasetr J, Nihira T, Panbangred W. Characterization of a gamma-butyrolactone synthetase gene homologue (*stcA*) involved in bafilomycin production and aerial mycelium formation in *Streptomyces* sp. SBI034. *Appl Microbiol Biotechnol* 2016;100:2749–2760.
48. Salehi-Najafabadi Z, Barreiro C, Rodríguez-García A, Cruz A, López GE *et al*. The gamma-butyrolactone receptors BulR1 and BulR2 of *Streptomyces tsukubaensis*: tacrolimus (FK506) and butyrolactone synthetases production control. *Appl Microbiol Biotechnol* 2014;98:4919–4936.
49. Tan G-Y, Peng Y, Lu C, Bai L, Zhong J-J. Engineering validamycin production by tandem deletion of γ-butyrolactone receptor genes in *Streptomyces hygroscopicus* 5008. *Metab Eng* 2015;28:74–81.
50. Choi S-U, Lee C-K, Hwang Y-I, Kinosita H, Nihira T. γ-Butyrolactone autoregulators and receptor proteins in non-*Streptomyces* actinomycetes producing commercially important secondary metabolites. *Arch Microbiol* 2003;180:303–307.
51. Du Y-L, Shen X-L, Yu P, Bai L-Q, Li Y-Q. Gamma-butyrolactone regulatory system of *Streptomyces chattanoogensis* links nutrient utilization, metabolism, and development. *Appl Environ Microbiol* 2011;77:8415–8426.
52. Millán-Aguiñaga N, Chavarria KL, Ugalde JA, Letzel A-C, Rouse GW *et al*. Phylogenomic insight into *Salinispora* (bacteria, actinobacteria) species designations. *Sci Rep* 2017;7:3564.
53. Román-Ponce B, Millán-Aguiñaga N, Guillen-Matus D, Chase AB, Ginigini JGM *et al*. Six novel species of the obligate marine actinobacterium *Salinispora*, *Salinispora cortesiana* sp. nov., *Salinispora fenicalii* sp. nov., *Salinispora goodfellowii* sp. nov., *Salinispora mooreana* sp. nov., *Salinispora oceanensis* sp. nov. and *Salinispora vitiensis* sp. nov., and emended description of the genus *Salinispora*. *Int J Syst Evol Microbiol* 2020;70:4668–4682.

54. Mincer TJ, Jensen PR, Kauffman CA, Fenical W. Widespread and persistent populations of a major new marine actinomycete taxon in ocean sediments. *Appl Environ Microbiol* 2002;68:5005–5011.
55. Jensen PR, Gontang E, Mafnas C, Mincer TJ, Fenical W. Culturable marine actinomycete diversity from tropical Pacific Ocean sediments. *Environ Microbiol* 2005;7:1039–1048.
56. Mincer TJ, Fenical W, Jensen PR. Culture-dependent and culture-independent diversity within the obligate marine actinomycete genus *Salinispora*. *Appl Environ Microbiol* 2005;71:7019–7028.
57. Maldonado LA, Fenical W, Jensen PR, Kauffman CA, Mincer TJ *et al*. *Salinispora arenicola* gen. nov., sp. nov. and *Salinispora tropica* sp. nov., obligate marine actinomycetes belonging to the family *Micromonosporaceae*. *Int J Syst Evol Microbiol* 2005;55:1759–1766.
58. Kim TK, Garson MJ, Fuerst JA. Marine actinomycetes related to the '*Salinospora*' group from the Great Barrier Reef sponge *Pseudoceratina clavata*. *Environ Microbiol* 2005;7:509–518.
59. Vidgen ME, Hooper JNA, Fuerst JA. Diversity and distribution of the bioactive actinobacterial genus *Salinispora* from sponges along the Great Barrier Reef. *Antonie Van Leeuwenhoek* 2012;101:603–618.
60. Jensen PR, Moore BS, Fenical W. The marine actinomycete genus *Salinispora*: a model organism for secondary metabolite discovery. *Nat Prod Rep* 2015;32:738–751.
61. Feling RH, Buchanan GO, Mincer TJ, Kauffman CA, Jensen PR *et al*. 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* 2003;42:355–357.
62. Letzel A-C, Li J, Amos GCA, Millán-Aguiñaga N, Ginigini J *et al*. Genomic insights into specialized metabolism in the marine actinomycete *Salinispora*. *Environ Microbiol* 2017;19:3660–3673.
63. Amos GCA, Awakawa T, Tuttle RN, Letzel A-C, Kim MC *et al*. Comparative transcriptomics as a guide to natural product discovery and biosynthetic gene cluster functionality. *Proc Natl Acad Sci USA* 2017;114:E11121–E11130.
64. Schulze CJ, Navarro G, Ebert D, DeRisi J, Linington RG. Salinipostins A-K, long-chain bicyclic phosphotriesters as a potent and selective antimalarial chemotype. *J Org Chem* 2015;80:1312–1320.
65. Yoo E, Schulze CJ, Stokes BH, Onguka O, Yeo T *et al*. The antimalarial natural product salinipostin A identifies essential α/β serine hydrolases involved in lipid metabolism in *P. falciparum* parasites. *Cell Chem Biol* 2020;27:143–157.
66. Schlawis C, Kern S, Kudo Y, Grunenberg J, Moore BS *et al*. Structural elucidation of trace components combining GC/MS, GC/IR, DFT-calculation and synthesis – salinilactones, unprecedented bicyclic lactones from *Salinispora* bacteria. *Angew Chem Int Ed Engl* 2018;57:14921–14925.
67. Schlawis C, Harig T, Ehlers S, Guillen-Matus DG, Creamer KE *et al*. Extending the salinilactone family. *Chembiochem* 2020;21:1629–1632.
68. Kudo Y, Awakawa T, Du Y-L, Jordan PA, Creamer KE *et al*. Expansion of gamma-butyrolactone signaling molecule biosynthesis to phosphotriester natural products. *ACS Chem Biol* 2020;15:3253–3261.
69. Marchler-Bauer A, Bo Y, Han L, He J, Lanczycki CJ *et al*. CDD/SPARCLE: functional classification of proteins via subfamily domain architectures. *Nucleic Acids Res* 2017;45:D200–D203.
70. Mendler K, Chen H, Parks DH, Lobb B, Hug LA *et al*. AnnoTree: visualization and exploration of a functionally annotated microbial tree of life. *Nucleic Acids Res* 2019;47:4442–4448.
71. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z *et al*. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 1997;25:3389–3402.
72. Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M *et al*. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* 2012;28:1647–1649.
73. Hadjithomas M, Chen I-MA, Chu K, Huang J, Ratner A *et al*. IMG-ABC: new features for bacterial secondary metabolism analysis and targeted biosynthetic gene cluster discovery in thousands of microbial genomes. *Nucleic Acids Res* 2017;45:D560–D565.
74. Blin K, Wolf T, Chevrette MG, Lu X, Schwalen CJ *et al*. antiSMASH 4.0 – improvements in chemistry prediction and gene cluster boundary identification. *Nucleic Acids Res* 2017;45:W36–W41.
75. Blin K, Shaw S, Steinke K, Villebro R, Ziemert N *et al*. antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline. *Nucleic Acids Res* 2019;47:W81–W87.
76. Medema MH, Takano E, Breitling R. Detecting sequence homology at the gene cluster level with MultiGeneBlast. *Mol Biol Evol* 2013;30:1218–1223.
77. Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S *et al*. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res* 2019;47:D607–D613.
78. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 2004;32:1792–1797.
79. Maddison WP, Maddison DR. Mesquite: a modular system for evolutionary analysis, 3.40; 2018. http://www.mesquiteproject.org
80. Darriba D, Taboada GL, Doallo R, Posada D. ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics* 2011;27:1164–1165.
81. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 2014;30:1312–1313.
82. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W *et al*. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol* 2010;59:307–321.
83. Lefort V, Longueville JE, Gascuel O. SMS: smart model selection in PhyML. *Mol Biol Evol* 2017;34:2422–2424.
84. Ziemert N, Lechner A, Wietz M, Millán-Aguiñaga N, Chavarria KL *et al*. Diversity and evolution of secondary metabolism in the marine actinomycete genus *Salinispora*. *Proc Natl Acad Sci USA* 2014;111:E1130–1139.
85. Rambaut A. FigTree v1.4.3; 2016. https://github.com/rambaut/figtree/releases
86. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res* 2019;47:W256–W259.
87. Nouioui I, Carro L, García-López M, Meier-Kolthoff JP, Woyke T *et al*. Genome-based taxonomic classification of the phylum Actinobacteria. *Front Microbiol* 2018;9:2007.
88. Dillon SC, Bateman A. The hotdog fold: wrapping up a superfamily of thioesterases and dehydratases. *BMC Bioinformatics* 2004;5:109.
89. Chen I-MA, Chu K, Palaniappan K, Pillay M, Ratner A *et al*. IMG/M v.5.0: an integrated data management and comparative analysis system for microbial genomes and microbiomes. *Nucleic Acids Res* 2018;47:666–677.
90. Enright AJ, Iliopoulos I, Kyrpides NC, Ouzounis CA. Protein interaction maps for complete genomes based on gene fusion events. *Nature* 1999;402:86–90.
91. Marcotte EM, Pellegrini M, Ng HL, Rice DW, Yeates TO *et al*. Detecting protein function and protein–protein interactions from genome sequences. *Science* 1999;285:751–753.
92. Penn K, Jenkins C, Nett M, Udwary DW, Gontang EA *et al*. Genomic islands link secondary metabolism to functional adaptation in marine actinobacteria. *ISME J* 2009;3:1193–1203.
93. Ziemert N, Podell S, Penn K, Badger JH, Allen E *et al*. The natural product domain seeker NaPDoS: a phylogeny based bioinformatic tool to classify secondary metabolite gene diversity. *PLoS One* 2012;7:e34064.
94. Van Santen JA, Jacob G, Singh AL, Aniebok V, Balunas MJ *et al*. The Natural Products Atlas: an open access knowledge

base for microbial natural products discovery. *ACS Cent Sci* 2019;5:1824–1833.

95. Kautsar SA, Blin K, Shaw S, Navarro-Muñoz JC, Terlouw BR *et al*. MIBiG 2.0: a repository for biosynthetic gene clusters of known function. *Nucleic Acids Res* 2020;48:D454–D458.
96. Chevrette MG, Gutiérrez-García K, Selem-Mojica N, Aguilar-Martínez C, Yañez-Olvera A *et al*. Evolutionary dynamics of natural product biosynthesis in bacteria. *Nat Prod Rep* 2020;37:566–599.
97. Niehs SP, Kumpfmüller J, Dose B, Little RF, Ishida K *et al*. Insect-associated bacteria assemble the antifungal butenolide gladiofungin by non-canonical polyketide chain termination. *Angew Chem Int Ed Engl* 2020;59:23122–23126.
98. Nakou IT, Jenner M, Dashti Y, Romero-Canelón I, Masschelein J *et al*. Genomics-driven discovery of a novel glutarimide antibiotic from *Burkholderia gladioli* reveals an unusual polyketide synthase chain release mechanism. *Angew Chem Int Ed Engl* 2020;59:23145–23153.
99. de Rond T, Asay JE, Moore BS. Co-occurrence of enzyme domains guides the discovery of an oxazolone synthetase. *bioRxiv* 2020:147165.
100. Park CJ, Smith JT, Andam CP. Horizontal gene transfer and genome evolution in the phylum Actinobacteria. *Horizontal Gene Transfer*. Cham: Springer; 2019. pp. 155–174.
101. Bruns H, Crüsemann M, Letzel A-C, Alanjary M, McInerney JO *et al*. Function-related replacement of bacterial siderophore pathways. *ISME J* 2018;12:320–329.
102. Bose U, Ortori CA, Sarmad S, Barrett DA, Hewavitharana AK *et al*. Production of *N*-acyl homoserine lactones by the sponge-associated marine actinobacteria *Salinispora arenicola* and *Salinispora pacifica*. *FEMS Microbiol Lett* 2017;364:fnx002.
103. McBride SG, Strickland MS. Quorum sensing modulates microbial efficiency by regulating bacterial investment in nutrient acquisition enzymes. *Soil Biol Biochem* 2019;136:107514.
104. Patteson JB, Lescallette AR, Li B. Discovery and biosynthesis of azabicyclene, a conserved nonribosomal peptide in *Pseudomonas aeruginosa*. *Org Lett* 2019;21:4955–4959.
105. Okada BK, Seyedsayamdost MR. Antibiotic dialogues: induction of silent biosynthetic gene clusters by exogenous small molecules. *FEMS Microbiol Rev* 2017;41:19–33.
106. Alberti F, Leng DJ, Wilkening I, Song L, Tosin M *et al*. Triggering the expression of a silent gene cluster from genetically intractable bacteria results in scleric acid discovery. *Chem Sci* 2019;10:453–463.
107. Chevrette MG, Carlson CM, Ortega HE, Thomas C, Ananiev GE *et al*. The antimicrobial potential of *Streptomyces* from insect microbiomes. *Nat Commun* 2019;10:516.
108. Okamura H, Fujioka T, Mori N, Taniguchi T, Monde K *et al*. First enantioselective synthesis of salinipostin A, a marine cyclic enolphosphotriester isolated from *Salinispora* sp. *Tetrahedron Lett* 2019;60:150917.

Supporting Information

Phylogenetic analysis of the salinipostin γ-butyrolactone gene cluster uncovers new potential for bacterial signaling-molecule diversity

Kaitlin E. Creamer[a], Yuta Kudo[a], Bradley S. Moore[b,c], Paul R. Jensen[a]#

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

[b] Center for Oceans and Human Health, Scripps Institution of Oceanography, University of California San Diego, La Jolla, California, USA

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

Figure S1.

Salinipostin (*spt*) BGC as described in Amos *et al.* 2017. *PNAS.*

Spt1: Phosphoenolpyruvate synthase **(pfam00391; pfam01326)**
Spt2: AMP-ligase **(pfam00501)**
Spt3: Nucleotidyltransferase
Spt4: Acyl carrier protein **(pfam00550)**
Spt5: Thiolreductase **(pfam07993)**
Spt6: Nucleoside diphosphate kinase **(pfam00334)**
Spt7: UbiA-type prenyltransferase **(pfam01040)**
Spt8: Flavin-dependent oxidoreductase, luciferase family **(pfam00296)**
Spt9: y-butyrolactone synthase, AfsA **(pfam03756)**

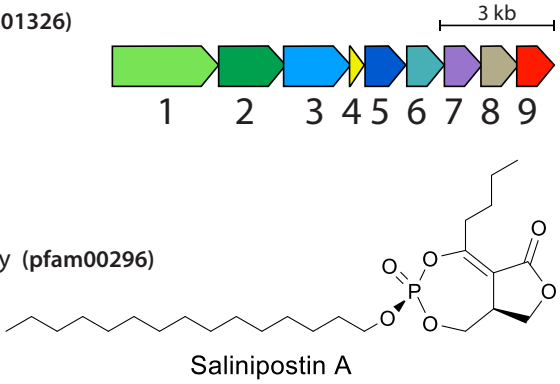


Figure 4.S1. Organization and functional annotation of the *Salinispora* salinipostin *spt* gene cluster. Pfam characterizations for the nine *spt* genes (1-9) are given in parentheses. The structure of salinipostin A is shown.

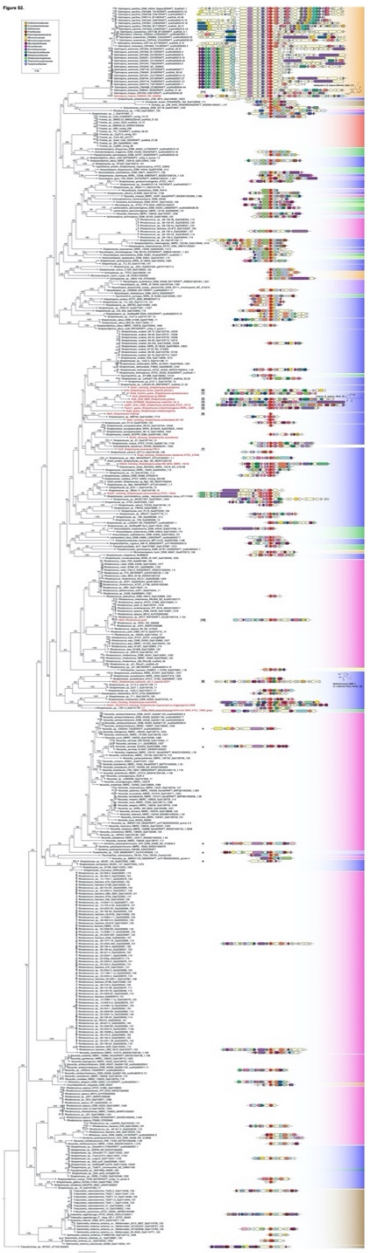


Figure 4.S2. Expanded phylogenetic tree of the top 403 Spt9 homologs (black) and 22 experimentally characterized AfsA homologs (red).

The RaxML maximum likelihood tree was calculated with a WAG+I+G+F ProtTest model with 500 replicates; branches are labeled with bootstrap support. Gene neighborhoods are drawn 5' to 3' for one representative taxa in each monophyletic clade and aligned with the Spt9 homolog (red); genes are colored by their COG function as annotated by JGI IMG/MER. Shaded rectangles indicate Actinobacterial family or Gammaproteobacterial class (see legend). Representative chemical structures are shown (g-butyrolactones: salinipostin A from *Salinispora tropica* CNB-440, A-factor from *Streptomyces griseus*; furan: methylenomycin MMF-1 from *Streptomyces coelicolor* A3(2)) and bracketed numbers correspond to the AfsA homologs and their associated compounds in Figure 4.1. Stars indicate salinipostin-like BGCs.

Figure S3.

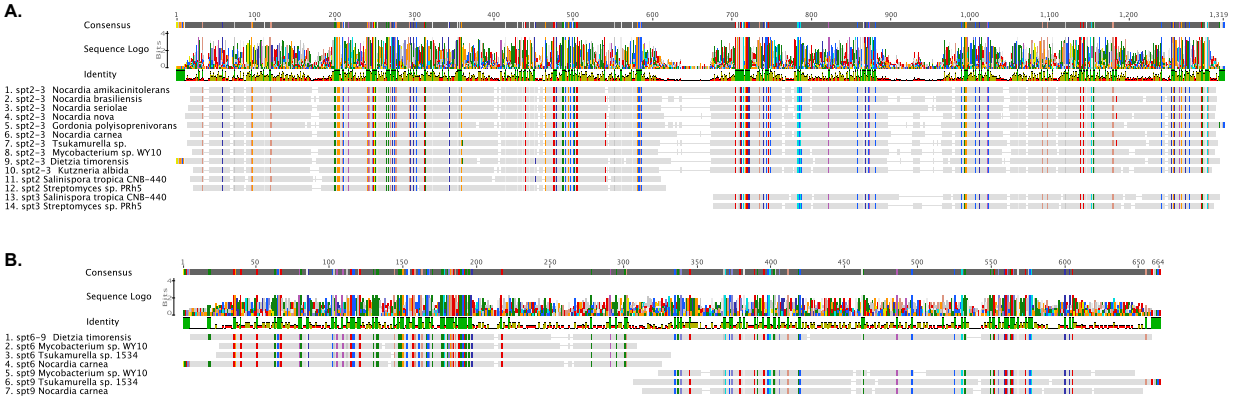


Figure 4.S3. Alignment of fused and individual Spt sequences.

A) Fused Spt2-3 protein sequences aligned with individual Spt2 and Spt 3 sequences from *Salinispora tropica* CNB-440 and *Streptomyces sp.* PRh5. Conserved regions are in green on the identity graph and colored by amino acid residue. The fused Spt2-3 proteins have conserved residues in the Spt2 and Spt3 functional domains (as predicted by the NCBI Conserved Domain Database tool, E-value cutoff 0.1).

B) Fused Spt6-9 protein sequence from *Dietzia timorensis* ID05-A0528 aligned with individual Spt6 and Spt9 sequences from *Mycobacterium sp.* WY10, *Tsukamurella sp.* 1534, and *Nocardia carnea* NRRL B-1997. Conserved regions are in green on the identity graph and colored by amino acid residue. The fused Spt6-9 protein in *Dietzia timorensis* ID05-A0528 includes both functional domains and conserved sequence similarity to the individual Spt6 and Spt9 proteins (as predicted by the NCBI Conserved Domain Database tool, E-value cutoff 0.01).

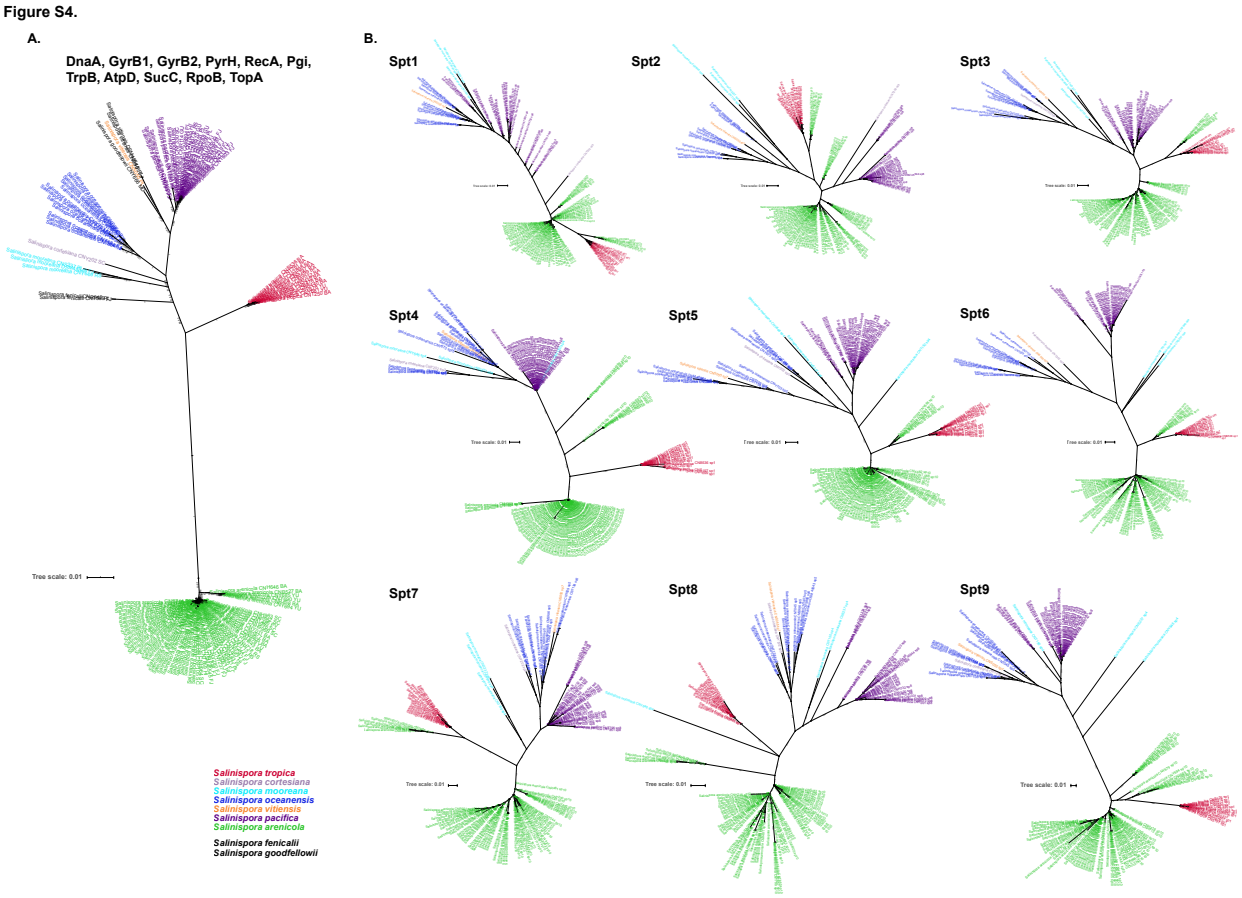


Figure 4.S4. *Salinispora* species and Spt1-9 phylogenies.

A) Maximum likelihood phylogenetic tree of 11 single-copy, concatenated proteins (DnaA, GyrB1, GyrB2, PyrH, RecA, Pgi, TrpB, AtpD, SucC, RpoB, and TopA) from 118 *Salinispora* genomes as reported in Ziemert *et al.* 2014. The PhyML tree was midpoint rooted and calculated with a Smart Model Selection AIC HIVb+G+I+F amino acid model; branches are labeled with aLRT support. Taxa lacking *spt* are in black and those with the *spt* BGC are colored by species as indicated in the legend.

B) Individual maximum likelihood phylogenetic trees for Spt1-9 amino acid sequences from 116 *Salinispora* strains. The PhyML trees were midpoint-rooted and calculated with a Smart Model Selection for each salinipostin protein: Spt1 (HIVb+G+I+F), Spt2 (Flu+G+I+F), Spt3 (HIVb+G+I+F), Spt4 (JTT+G), Spt5 (HIVb+G+I+F), Spt6 (HIVb+G+F), Spt7 (HIVb+G+I+F), Spt8 (HIVw+G+I+F), Spt9 (JTT+G+I+F); branches are labeled with their aLRT support. Taxa are colored by species as indicated in the legend.

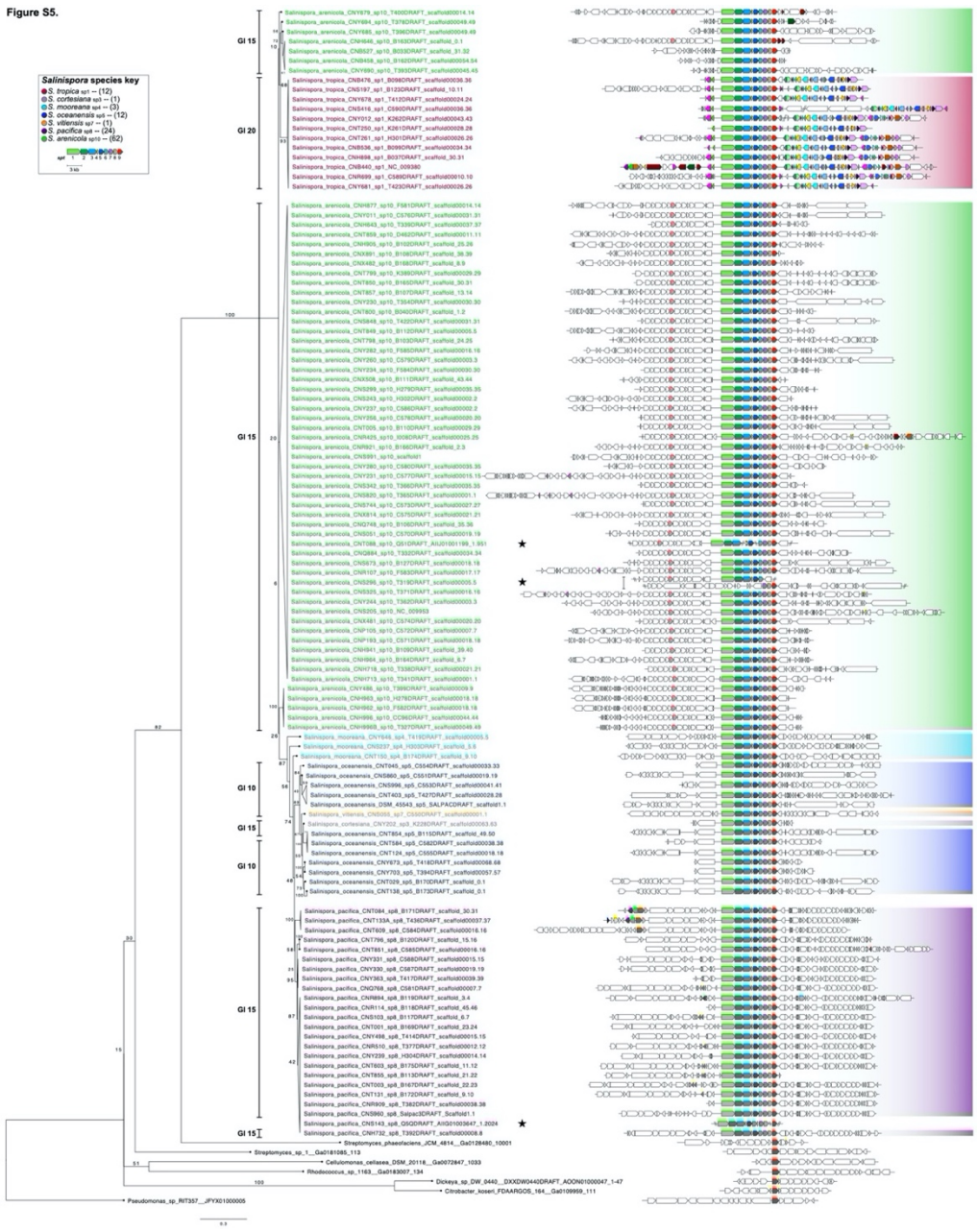


Figure 4.S5. *Salinispora* Spt9 phylogeny and gene cluster neighborhoods.

Maximum likelihood tree for Spt9 amino acid sequences from 116 Salinispora strains with the spt BGC was calculated with a JTT+I+G ProtTest model with 500 replicates in RaxML; branches are labeled with bootstrap support. Spt9 homologs in non-Salinispora bacteria were used as outgroups. Brackets on the left indicate genomic island (GI) locations of the spt BGC. Taxa are colored by species (see key). Gene neighborhoods are drawn 5' to 3' and aligned with the Spt9 homolog (red). Neighboring genes are colored as per MultiGeneBlast only if they share homology to genes in the Salinispora tropica CNB-440 spt BGC contig; white indicates no homology to genes in the S. tropica CNB-440 spt BGC contig. Stars indicate BGCs in which variations from the canonical spt BGC were observed.

Figure S6.

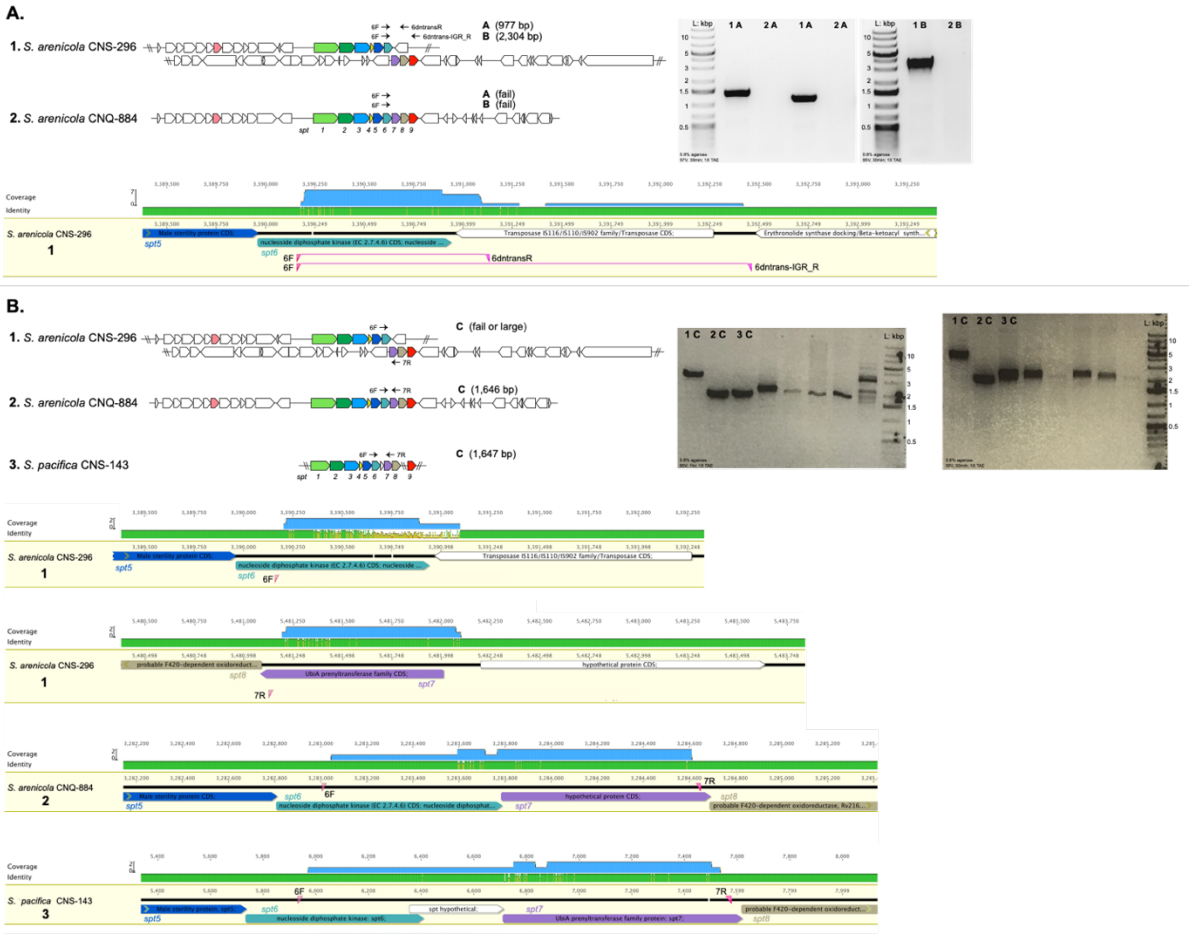


Figure 4.S6. Targeted PCR of *Salinispora spt* BGC variants.

A) The *spt* BGC in *S. arenicola* CNS-296 was observed on two contigs. Primer sets A (6F, 6dntransR) and B (6F, 6dntrans-IGR_R), which span *spt6* and the neighboring transposase gene, successfully amplified ~1400bp and ~3500bp products (gel image, lanes 1A and 1B, respectively), from *S. arenicola* CNS-296 (expected product size in parentheses). These primers failed to yield a product from *S. arenicola* CNQ-884, which possesses a contiguous *spt* BGC (gel image, lanes 2A and 2B, respectively). The light blue coverage graph indicates coverage of 8 PCR product sequences mapped to the reference genome and the identity graph shows mean pairwise identity over all mapped sequences (green: 100% identity; yellow: between 30%-99% identity; red: below 30% identity).

B) Primer pair C targeting *spt6-7* amplified an ~5kb product in *S. arenicola* CNS-296 (gel images, lane 1C). Sequencing from the 5' and 3' ends (900bp each) mapped back to *spt6* (poorly, 67% identity) and *spt7* (99% identity) indicating an ~3.2kb insertion between *spt6* and *spt7*. Primer pair C amplified similarly sized contiguous *spt6-7* products (~1,600bp) in both *S. arenicola* CNQ-884 (gel images, lane 2C) and *S. pacifica* CNS-143 (gel images, lane 3C). The sequences contained identical *spt6-7* sequence length and conserved domains indicating that the hypothetical gene between *spt6* and *spt7* in *S. pacifica* CNS-143 might be a genome assembly or annotation artifact. Coverage and identity graphs of mapped sequences is the same as described above.

Figure S7.

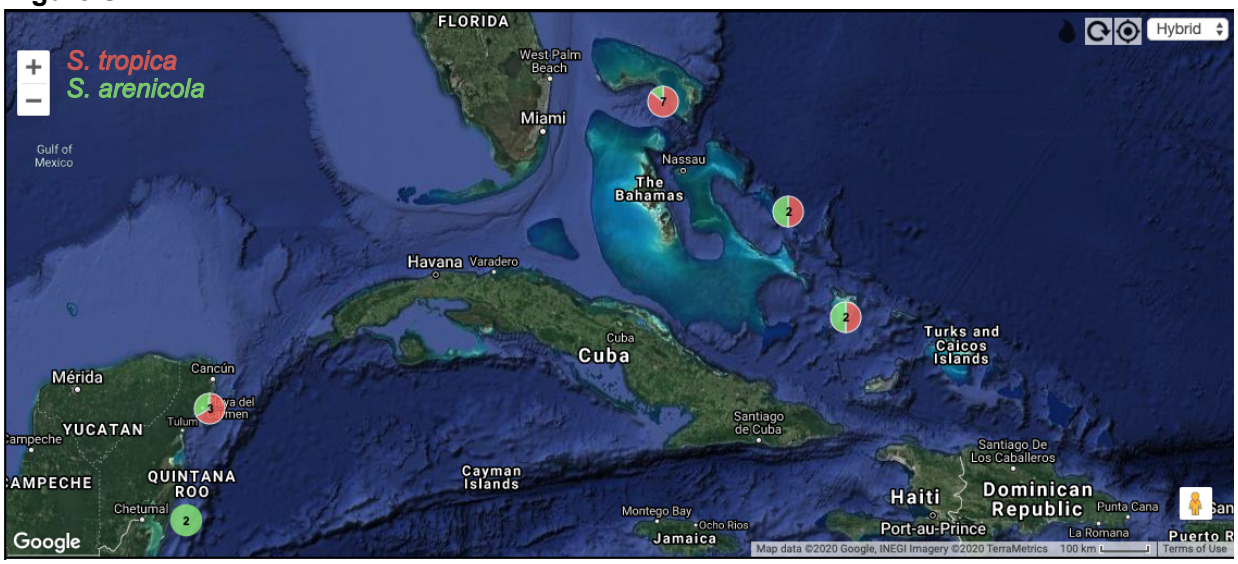


Figure 4.S7. Co-occurrence of *S. tropica* and *S. arenicola* strains.

Circles indicate geographic areas within the Bahamas and the Yucatán from which the 18 *Salinispora tropica* and *S. arenicola* strains that are most closely related in the Spt1- 9 phylogeny (Figure 4.S5) were isolated. The total number of strains are indicated with pie charts showing the proportion of each species; two strains are not shown as exact collection location coordinates in the Bahamas are not known for *S. tropica* strains CNT- 250 and CNT-261.

Table 4.S1. NCBI GenBank Accession numbers for the *spt6-7* PCR products as described in Figure 4.S6.

Sequence Name	NCBI GenBank Accession	Description	Corresp. Figure
SaCNS296_spt6-transp_1A	MW321490	PCR-amplified region from spt6 to transposase using primers 6F and 6dntransR, sequenced from both sides, partial CDS.	Figure S6-A, 1A
SaCNS296_spt6-transp_1B	MW321491	PCR-amplified region from spt6 to transposase using primers 6F and 6dntransR-IGR_R, sequenced from 6dntransR-IGR_R side; partial CDS.	Figure S6-A, 1B
SaCNS296_spt6_6F-7R_1C	MW321492	PCR-amplified spt6-7 using primers 6F and 7R sequenced from spt6 side; partial CDS.	Figure S6-B, 1C
SaCNS296_spt6_7R-6F_1C	MW321493	PCR-amplified spt6-7 using primers 6F and 7R sequenced from spt7 side; partial CDS.	Figure S6-B, 1C
SaCNQ884_spt6-7_6F-7R_2C	MW321494	PCR-amplified spt6-7 using primers 6F and 7R sequenced from both sides; partial CDS.	Figure S6-B, 2C
SpCNS143_spt6-7_6F-7R_3C	MW321495	PCR-amplified spt6-spt hypothetical gene-spt7 using primers 6F and 7R sequenced from both sides; partial CDS.	Figure S6-B, 3C

Supplemental Datasets.

Dataset S1. List of all sequence datasets with relevant accession information used in this paper's analyses.

Sheet/Tab 1: **Fig1_AnnoTree_hits_genes**: Number of Spt9 (pfam PF03766) genome hits identified by AnnoTree across Phyla, Class, Order, Family, Genus, and Species; the proportion of all hits and the number of genomes in each clade were used to draw and scale the pie charts of Figure 2. Also listed is the Gene ID, GTDB ID, and protein sequence of all genome hits.

Sheet/Tab 2: **403_spt9homolog_IMGgeneinfo**: List of all 403 Spt9 homolog sequence information (including locus tag, gene product name, genome ID, genome name, Genbank accession, amino acid sequence length, scaffold ID, scaffold external accession, scaffold length, scaffold GC%, and Pfam). Corresponds to Figure 3 and Figure S2.

Sheet/Tab 3: **22_known_afsAhomologs**: List of 22 characterized AfsA homologs belonging in the gamma-butyrolactone, furan, gamma-butenolide, and 'other' class of signaling molecules; corresponds to Figure 1. Information includes gene locus tag, gene symbol/name, molecule linked to gene if known, genome strain, NCBI protein accession, JGI Gene ID, JGI genome ID, and gene product name.

Sheet/Tab 4: **152_spt-likeBGCs**: List of all 152 Spt9 homologs identified within Spt-like BGCs and additional homologs using for building the tree in Figure 4. Gene information includes JGI gene ID, locus tag, gene product name, JGI genome ID,

genome name, start coordinate, end coordinate, strand, DNA sequence length, amino acid length, scaffold ID, scaffold external accession, scaffold GC %.

Sheet/Tab 5: **11_MLST_Salinispora**: List of 119 *Salinispora* species strain DnaA, GyrB1, GyrB2, Pgi, TrpB, SucC, RecA, PyrH, TopA, AtpD, RpoD JGI gene ID accessions for Figure S4-A.

Sheet/Tab 6: **Spt1_Salinispora_116**: List of all 116 Spt1 *Salinispora* gene sequence information for Figure 5 and S4-B (including JGI gene ID, locus tag, gene product name, JGI genome ID, genome name, start coordinate, end coordinate, strand, DNA sequence length, amino acid sequence length, scaffold ID, scaffold external accession name, and pfam).

Sheet/Tab 7: **Spt2_Salinispora_116**: List of all 116 Spt2 *Salinispora* gene sequence information for Figure 5 and S4-B (included information is the same as Spt1).

Sheet/Tab 8: **Spt3_Salinispora_116**: List of all 116 Spt3 *Salinispora* gene sequence information for Figure 5 and S4-B (including JGI gene ID, locus tag, gene product name, JGI genome ID, genome name, start coordinate, end coordinate, strand, DNA sequence length, amino acid sequence length, scaffold ID, and scaffold external accession name).

Sheet/Tab 9: **Spt4_Salinispora_116**: List of all 116 Spt4 *Salinispora* gene sequence information for Figure 5 and S4-B (included information is the same as Spt1).

Sheet/Tab 10: **Spt5_Salinispora_116**: List of all 116 Spt5 *Salinispora* gene sequence information for Figure 5 and S4-B (included information is the same as Spt1).

Sheet/Tab 11: **Spt6_Salinispora_116**: List of all 116 Spt6 *Salinispora* gene sequence information for Figure 5 and S4-B (included information is the same as Spt1).

Sheet/Tab 12: **Spt7_Salinispora_116**: List of all 116 Spt7 *Salinispora* gene sequence information for Figure 5 and S4-B (included information is the same as Spt1).

Sheet/Tab 13: **Spt8_Salinispora_116**: List of all 116 Spt8 *Salinispora* gene sequence information for Figure 5 and S4-B (included information is the same as Spt1).

Sheet/Tab 14: **Spt9_Salinispora_116**: List of all 116 Spt9 *Salinispora* gene sequence information for Figure 5 and S4-B (included information is the same as Spt1).

4.4 Acknowledgements

Chapter 4 (Section 4.3), in full, is a reprint of the material as it appears in *Microbial Genomics* 7(5), Creamer, K.E.; Kudo, Y; Moore, B.S.; Jensen, P.R., 2021. The dissertation author was the primary investigator and author of this paper.

CHAPTER 5. Characterization of micro- and macroscale genomic and biosynthetic gene cluster diversity in the marine actinomycete genus *Salinispora*

5.1 Abstract

Marine sediments contain some of the highest predicted abundances of bacterial cells on Earth. Even individual sand grains can harbor over 100,000 bacterial cells, and thus competition for resources and complex community interactions likely affect the population structure of individual bacterial species. However, we do not understand how the natural products produced by community members influence community structure on a close spatial scale. The marine obligate actinobacterial genus *Salinispora* is a model system for studying these patterns across two different spatial scales. On the "macroscale", *Salinispora* have been isolated from marine seaweed, coral, and sediments from worldwide locations and there is some understanding of their distributions at this scale. Yet we do not know the fine scale spatial distribution of *Salinispora* in marine sediments. To study the spatial effects on chemical innovation at the population scale in *Salinispora*, we selectively isolated 176 and whole-genome sequenced 99 "microscale" *Salinispora* strains from a 1-meter^2 marine sediment quadrant in Fiji. Three species of *Salinispora—S. arenicola, S. pacifica*, and *S. oceanensis*— were isolated from 16 sub-quadrants, and comparative genomics revealed that the microscale genomes accounted for most of the currently known *S. arenicola* species diversity and were not clonal. The microscale *Salinispora* had diverse morphologies with *S. arenicola* showing signs of intraspecies diversification, which was supported by the diverse assemblage of biosynthetic gene clusters detected. We captured new examples of BGCs previously rare in the *Salinispora* genus and unique BGCs found only in the microscale *Salinispora* genomes, thus contributing to our understanding of how intraspecies divergence and microscale population dynamics could be contributing to chemical diversity. Overall, we effectively doubled the collection of whole-genome sequenced *Salinispora*, which will

facilitate future evolutionary investigations into patterns driving chemical and bacterial speciation along with new bioactive compound discovery.

5.2 Introduction

Bacteria have been discovered in all environments on earth and perform important ecosystem functions from biogeochemical cycling, climate regulation, symbioses, and plant, agricultural, animal, and human health (van der Meij *et al.*, 2017; Cavicchioli *et al.*, 2019; Flemming and Wuertz, 2019). In terms of abundance, estimates for the total number of microbial cells on earth are astronomical, as microbial cells are believed to exist in places we've yet to uncover them and likely exist in complex biofilms miles deep below the surface of the earth (Flemming and Wuertz, 2019). Recent estimates predicted that there are 0.2, 1 and 1.3 x 10^{26} bacterial cells in animal hosts, plant hosts, and freshwater systems, respectively (Lloyd *et al.*, 2018). In contrast, marine sediments, soil, and terrestrial subsurface are predicted to have more than 2,000 times the amount of cells (2900, 2560, and 2500 x 10^{26}) respectively, making marine sediments one of the most microbially rich ecosystems (Lloyd *et al.*, 2018). These estimates were recently expanded upon with the incredible discovery that a single marine sediment sand grain can harbor as many as 100,000 individual bacterial cells belonging to several thousand bacterial species (Probandt *et al.*, 2018)—multiply that by the amount of marine sediment worldwide and the number of bacteria is astounding. The bacterial diversity and abundance on a single marine sediment sand grain begs the question at what scale important biogeochemical functions are occurring, and at what levels we can accurately capture complex bacterial community composition, especially with culture-dependent methods.

These findings have important implications for investigating the evolutionary diveristy of bacteria. Culture-dependent methods from marine sediments may only capture 3% of the bacterial community, however, the isolates from a community that can be cultured in the lab can include rare community members that culture-independent approaches miss (Demko *et al.*, 2021). Additionally, to understand bacterial evolutionary dynamics, whole-genomes (Undabarrena *et al.*, 2021), re-sequencing approaches mapped to reference genome strains (Harden *et al.*, 2015; Creamer *et al.*, 2017; Hamdallah *et al.*, 2018), or genome-resolved metagenomic analyses (Chen *et al.*, 2020) are necessary. In marine sediments, the marine obligate genus *Salinispora* can serve as a model system for investigating these dynamics, especially through the lens of specialized metabolite production.

The bacterial taxon *Salinispora* was the first widespread obligate marine actinomycete genus to be described (Mincer *et al.*, 2002; Maldonado *et al.*, 2005). To date, hundreds of *Salinispora* strains have been isolated from marine sediments (Mincer *et al.*, 2002, 2005; Jensen *et al.*, 2005), marine seaweed (Jensen *et al.*, 2005), and marine sponges (Kim *et al.*, 2005; Vidgen *et al.*, 2012). *Salinispora* are slow-growing, aerobic, Gram-positive actinomycetes that have branching, filamentous substrate mycelia and form black non-motile spores after extended growth in agar and liquid culture media (Maldonado *et al.*, 2005). Next-generation sequencing technology was used to sequence 118 *Salinispora* genomes (Millán-Aguiñaga *et al.*, 2017) and subsequent whole-genome average nucleotide identity (ANI) analyses revealed that the *Salinispora* genus includes nine species: *Salinispora tropica, S. arenicola, S. pacifica, S.mooreana, S. cortesiana, S. fenicalii, S. vitiensis, S. goodfellowii,* and *S. oceanensis* (Román-Ponce *et al.*, 2020).

In-depth phylogenomic comparisons of *Salinispora* species isolated from worldwide ocean sediments uncovered that *Salinispora* have an incredible specialized metabolite biosynthetic

potential (Udwary *et al.*, 2007; Penn *et al.*, 2009; Ziemert *et al.*, 2014; Letzel *et al.*, 2017). Early isolation and characterization efforts found that *Salinispora* had promising bioactivities typically associated with other prolific Actinomycetia (formerly Actinobacteria) (Salam *et al.*, 2020) metabolite producers in the genus *Streptomyces* as *Salinispora* chemical extracts contained significant cancer cell cytotoxicity, antibiotic, and antifungal properties (Mincer *et al.*, 2002). *Salinispora* species produce specialized metabolites in species-specific patterns (Jensen *et al.*, 2007) including molecules like the proteasome inhibitor salinosporamide A in *S. tropica* (Feling *et al.*, 2003) and the RNA-polymerase-targeting antibiotic rifamycin in *S. arenicola* (Kim *et al.*, 2006), among many other exciting small molecules produced by different *Salinispora* strains (Jensen *et al.*, 2015). Complementing continued chemical work to isolate new compounds, *Salinispora* genome sequences have facilitated an *in-silico* genome-mining approach to predict specialized BGC operons. Each *Salinispora* strain harbors 15-30 BGCs (Ziemert *et al.*, 2014; Letzel *et al.*, 2017; Chase *et al.*, 2021). These BGCs may be silent (the BGC is not expressed and associated molecule is not observed) or cryptic (a molecule is observed but not linked to a BGC) (Ziemert *et al.*, 2014; Letzel *et al.*, 2017).

Salinispora BGCs are largely clustered in genomic islands (Penn and Jensen, 2012). The presence of core BGCs common to individual *Salinispora* species and the existence of unique BGC singletons/doublets in some strains indicates that their small molecule products are a major factor in the evolution and speciation of *Salinispora* (Fischbach *et al.*, 2008; Penn *et al.*, 2009; Ziemert *et al.*, 2014; Jensen, 2016; Letzel *et al.*, 2017). Genomic evidence shows that BGCs persist through vertical inheritance; for example, 1) some BGCs are only found within certain *Salinispora* species sharing a conserved genomic neighborhood, and 2) some BGCs were present in the most recent common ancestor of *Salinispora* with moderate differences across species and strains likely

due to the effects of natural selection and divergence over time (Letzel *et al.*, 2017; Chase *et al.*, 2021). However, there is also striking evidence that many of the rare BGCs in *Salinispora* have been subject to migration, replacement, degradation, exchange, and horizontal transfer as there are 1) BGCs that are unique to strains and species indicating recent acquisition, 2) different BGCs that are present in the same genome environment appear to produce compounds with similar ecological roles, and 3) a small number of BGCs show evidence of genomic island migration and degradation indicating that they are under selection and contributors to adaptive genomic flux (Penn *et al.*, 2009; Bruns *et al.*, 2017; Letzel *et al.*, 2017). The salinipostin BGC would be one such example and was the subject of Chapter 4 of this thesis (Creamer *et al.*, 2021).

While the previous evolutionary analyses of *Salinispora* and its BGCs have focused on 118 strains isolated from worldwide "macroscale" marine sediment samples, the goal of this chapter was to isolate, genome-sequence, and characterize the biosynthetic potential of a new "micro-scale" collection of *Salinispora* genomes. These "microscale" *Salinispora* strains were isolated from a 1 m^2 sediment quadrant, split into 16 sub-quadrants, adjacent to a coral reef in Fiji. To the best of our knowledge, there have been no prior studies that have used comparative genomics to assess biosynthetic potential at these spatial scales. Additionally, we report the putative presence of plasmids in *Salinispora*, which could be a mechanism by which BGCs and other gene content are shared between and across species. The creation of this microscale *Salinispora* genome collection will allow us to evaluate if there is evidence of BGC exchange between *Salinispora* strains mediated by close proximity.

5.3 Methods

5.3.1 Sampling and selective isolation of *Salinispora*.

All marine sediment samples were collected as previously described (Demko, 2021). Briefly, samples were collected around Nacula Island, Fiji via SCUBA in June of 2017 by Alyssa M. Demko and Paul R. Jensen (**Figure 5.1**). At sampling site 4, the fine scale sampling of a 1 m^2 quadrant, divided into 16 even segments (4 x 4), was performed next to a reef area. Surface sediment from each of the 16 sections was collected into Whirl-pak (Nasco) bags denoting their location within the quadrat. Site 4, named "Off Honeymoon Island", characteristics included: 10.36m depth; Latitude 16° 53.578 S; Longitude 177° 23.076' (E); coarse calcareous sediment, *Halimede* rubble, next to small reef. Sediment samples were stored in a cooler following collection, frozen upon returning to shore (20°C), and kept frozen until processing in San Diego, CA.

To selectively culture *Salinispora*, 1g chunks of frozen sediment from each of the 16 sub-quadrant Whirl-pak bags were placed in sterile petri dishes and dried over the weekend in a laminar flow hood. Sterile sponges wetted with sterile seawater were used to stamp sediment onto two different types of agar plates: A1 (10g/L starch; 4g/L yeast extract; 2g/L peptone; 22g/L Instant Ocean; 16g/L agar; 1L diH_2O; autoclaved and cycloheximide added for a final concentration of 100μg/mL) and SWA (22g/L Instant Ocean, 16/L agar; 1L diH_2O; autoclaved and cycloheximide added for a final concentration of 100μg/mL) (**Figure 5.1**). The stamping was performed in a spiral pattern where no stamp covered the previous stamp, as to dilute from a single stamp (**Figure 5.1**). Per sub-quadrant sample, 4 plates of each A1 and SWA media were stamped (128 total plates). Plates were incubated at room temperature and monitored for over 2 months for signs of

actinomycete-like growth (**Figure 5.1**). Single colonies were selectively isolated by re-streaking onto clean plates of the same media type. All isolates were tracked to know what sub-quadrant and media type it was isolated from. The first round resulted in over >600 isolates; 30 from A1 media and ~600 from SWA media. Isolates went through two rounds of purification to make sure they were single cultures. Obligate salt-water growth assays were performed using split-petri plates with one side containing A1 with full strength seawater and the other side with A1 prepared with deionized water (**Figure 5.1**). This helped to selectively target orange colonies that could only grow in the presence of saltwater, which indicated that they could be *Salinispora*.

5.3.2 Putative species identification with colony PCR.

Bacterial isolates that had been passaged at least 2 times were analyzed with colony PCR to putatively identify the species with 16S rRNA gene primers (**Figure 5.1**). Briefly, a single colony was suspended in 10µL of sterile DMSO. PCR amplification reactions were assembled as follows: 12.5µL GoTaq Green Mastermix; 1.25µL each FC127 (5'-AGAGTTTGATCCTGGCTCAG-3') and RC1492 (5'- TACGGCTACCTTGTTACGACTT-3') primer (each stock at 10µM); 1µL DMSO, 8µL ddH_2O, and 1µL of colony DNA/DMSO mixture. Reactions were run with the following settings: initial denaturation at 95°C for 3 min followed by 35 cycles of 95°C for 45 s, 60°C for 45 s, and 72°C for 120s; followed by a 5 min extension at 72°C. Amplification products were analyzed by gel electrophoresis on a 0.8% agarose TAE gel, run at 95V for 30minutes in 1X TAE; with a 1 kb GeneRuler Plus ladder; and 6x loading dye (New England Biosciences) and SYBR Green I nucleic acid gel stain (Life Technologies) added to each sample. Successful PCR products with a length of ~1500bp were cleaned and purified with the Qiagen PCR cleanup kit. From 225 total reactions that were sent to Eton Bioscience (San Diego,

CA) for sequencing from both primer directions, 91% were successful, capturing the entire ~1,400bp targeted 16S rRNA region. Each pair of sequences were trimmed, aligned with Geneious (Kearse *et al.*, 2012), manually inspected, and the consensus sequence was run with NCBI BLAST.

To assess the diversity of the 16S rRNA identified isolates, we calculated a phylogenetic tree of all *Salinispora* 16S sequences using MUSCLE (Edgar, 2004) for alignment, jModelTest 2.1.4 for model selection (Darriba *et al.*, 2012), RAxML (Stamatakis, 2014) for tree calculation,

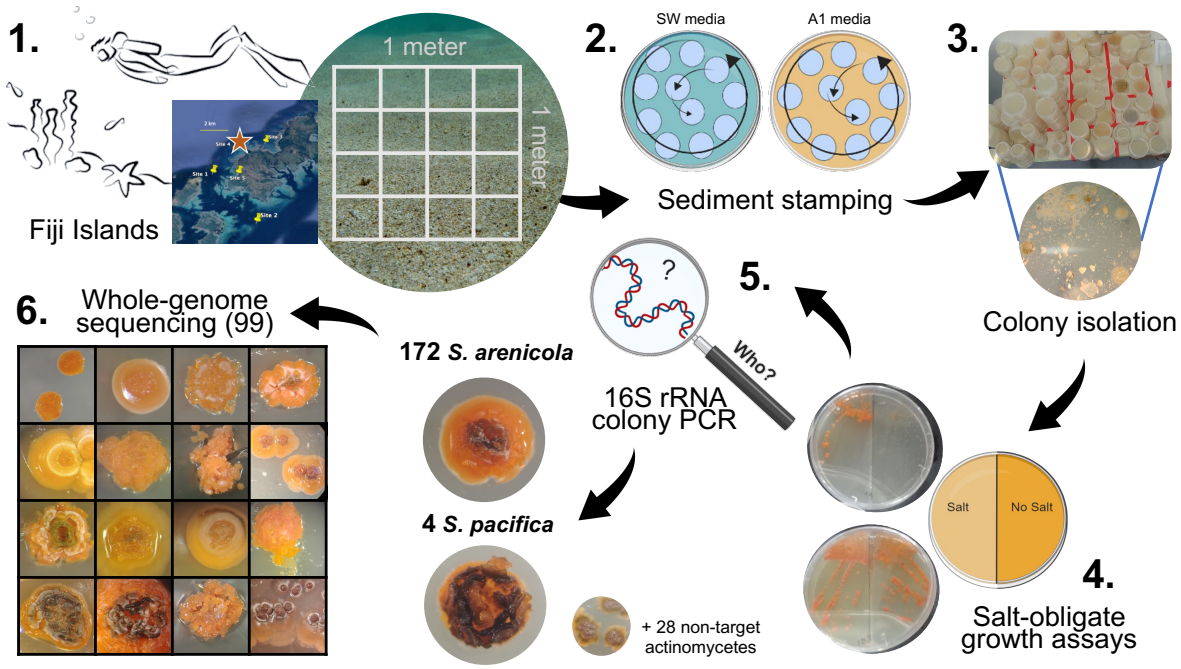


Figure 5.1. Sampling and selective isolation of microscale *Salinispora* workflow.

and FigTree (Rambaut, 2016) for visualization. Additionally, manual inspection of alignments of all microscale *Salinispora* 16S rRNA sequences in Geneious (Kearse *et al.*, 2012) was performed to identify SNP (single nucleotide polymorphisms).

5.3.3 Cultivation and extraction for whole-genome sequencing, plasmids, and metabolomics.

We selected 5-6 microscale *Salinispora* strains from each of the 16 sub-quadrants to culture in larger volumes for DNA and plasmid extractions, cryo-preservation, and metabolomic analyses. Briefly, each microscale *Salinispora* strain was grown (inoculated from the original 2mL frozen cryovial) in 60mL of A1 75% seawater liquid media (10 g/L soluble starch (Affymetrix), 2 g/L peptone (Fischer Scientific), 4 g/L yeast extract (Fischer Scientific), 22 g/L instant ocean mix (Marineland) in 1L DI water; 14g/L of agar added for solid plate media) at 28°C at 230rpm with glass beads until they reached exponential cell density. Cultures were checked daily to watch for signs of contamination; cultures averaged 1-2 weeks to reach exponential phase (cell biomass assessed by eye and color of culture). At each point where we interacted with the culture, purity plates were used to ensure nothing became contaminated. When the cells were ready to harvest, 18mL of dense culture was mixed with 6-6.5mL of 50% glycerol/seawater for cryovial preservation; 4 x 1mL cultures were aliquoted and centrifuged as cell pellets for DNA extraction; 8 x 2mL aliquots were centrifuged into cell pellets for plasmid DNA extraction; and the rest of the culture, up to 14mL was saved for metabolomic analyses. Supernatants were removed from all cell pellets; and at the end of the workflow, another purity plate of the culture was created, thus ensuring no contamination during the aliquoting. All DNA, plasmid, and cryovial samples were saved at -80°C and the metabolomic samples at -20°C.

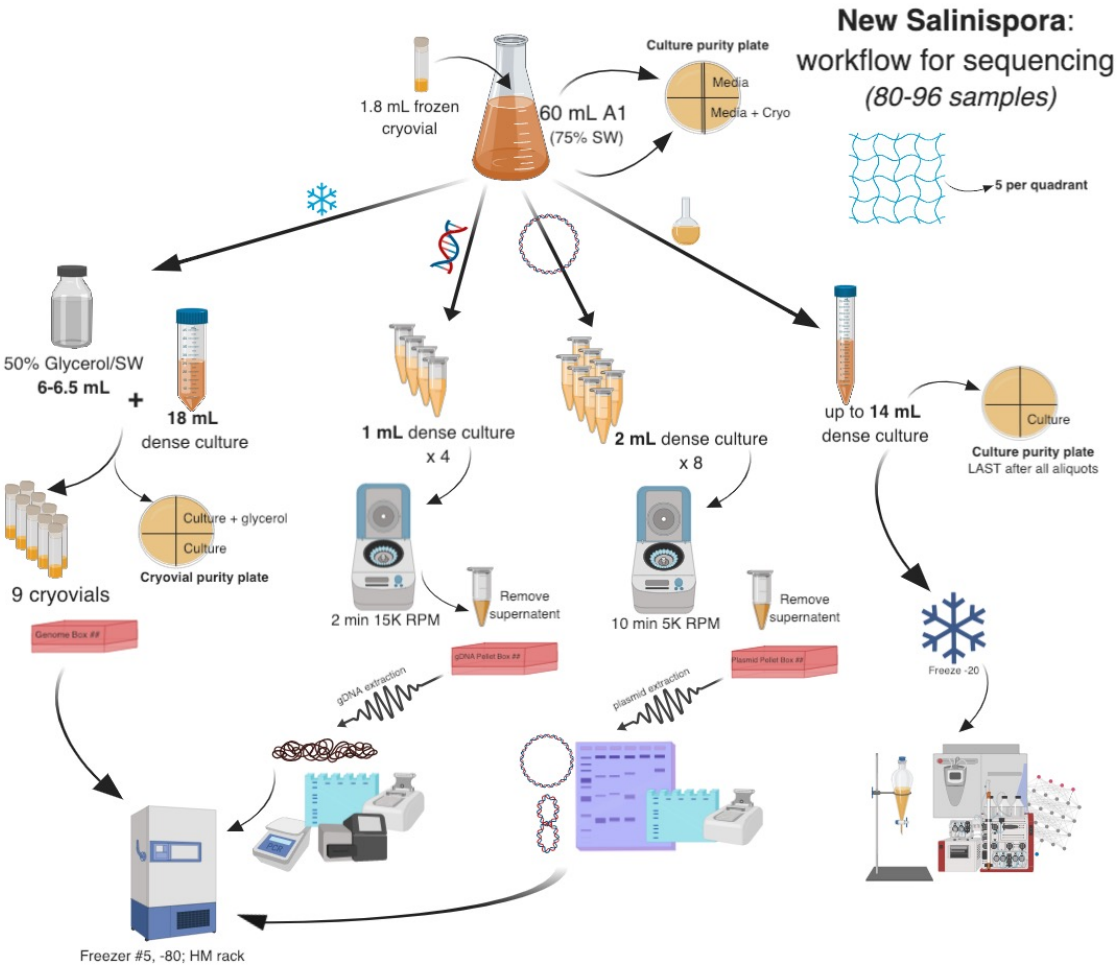


Figure 5.2. Workflow for growing and saving microscale *Salinispora* cultures for subsequent DNA, plasmid, cryovial, and metabolomic work.

The protocol developed for plasmid DNA extraction and visualization is explained in detail in the Chapter 5 Appendix. For genomic DNA extractions, we used the Wizard Genomic DNA Purification Kit (Promega) with modifications for Gram-positive cells. Briefly, this included the addition of freshly prepared 10mg/mL lysozyme (Sigma Aldrich), and following the recommended instructions by the manufacturer with the maximum time recommended for each step. Wide-bore tips were used to prevent shearing of gDNA. We doubled the amount of DNA pellet ethanol

washes in the protocol and used Lo-Bind DNA tubes (Eppendorf) for DNA precipitation. All gDNA samples were quantified with a NanoDrop 1000 spectrophotometer (Thermo Fisher Scientific) and a Qubit 3.0 fluorometer (Thermo Fisher Scientific) according to the manufacturer's instructions. The gDNA quality was analyzed by running 3µL of gDNA on a 1% agarose TAE gel for 60 minutes at 86V alongside a GeneRuler 1kb Plus DNA ladder (Thermo Fisher Scientific). Finally, a 16S rRNA PCR of each gDNA sample was performed as described above to confirm each strain matched the sequence from the original colony PCR of the starting cryovial. Metabolomic samples were not analyzed as part of the work described in this chapter.

5.3.4 Whole-genome sequencing, assembly, and annotation.

Genomic DNA samples were sent to two sequencing cores for sequencing: 1) Illumina MiSeq (UC Davis) for PE300, 700bp inserts; resulting in ~17.7 million reads with ~11.7% PhiX spike-in, with an overall Q30 > 60%; 2) Illumina NovaSeq 6000 (UCSD IGM) for S4 PE150, 400bp inserts prepared with a Nextera XT library kit (Illumina) and a pre-pooling MiSeq check run; resulting in ~3 billion reads with an overall Q30 >90% for each genome sample (5-27 million reads per genome; other reads were for co-sequenced metagenome samples).

Bioinformatic analyses were performed on the Triton Shared Computing Cluster (TSCC) at the San Diego Supercomputer Center (SDSC) on a 756 GB RAM, 36 Intel (R) Xeon (R) Gold 6240 CPU @ 2.60GHzCPU server. Each raw sequence dataset from the MiSeq and NovaSeq sequencers were assembled separately as they had drastically different number of reads per sample. Whole-genome assembly and annotation was performed using the bactopia version 2.0.3 (Petit and Read, 2020) which is a Nextflow-enabled (version 22.04.0) nf-core workflow that wraps many assembly and annotation tools. Briefly, the versions of the tools used for each genome assembly

step are as listed: Raw reads QC and trimming (bbduk: 38.93 (run with options --minlength 35; --trimq 10 (sourceforge.net/projects/bbmap/)); fastq-scan: 0.4.4 (https://github.com/rpetit3/fastq-scan); fastqc: 0.11.9 (https://github.com/s-andrews/FastQC); lighter: 1.1.2; mash: '2.3'; mccortex: 0.0.3-610-g400c0e3; sourmash: 4.2.2); genome assembly (shovill: 1.1.0, which included tools: seqtk, bbmap, flash: 1.2.11, spades.py: 3.15.3 (run with --isolate option) (https://github.com/tseemann/shovill), bwa: 0.7.17-r1188, samtools: '1.12', pilon: '1.24' (Walker *et al.*, 2014)); genome assembly QC (checkm: 1.1.3 (Parks *et al.*, 2015); quast: 5.0.2 (Gurevich *et al.*, 2013)); and genome annotation (prokka: 1.14.6 (--compliant; --centre JensenSIO) (Seemann, 2014)). Three different assembly methods were compared, including spades.py: 3.15.3; skesa: 2.4.0 (Souvorov *et al.*, 2018); and velvetg: 1.2.10 (Zerbino and Birney, 2008); however, the SPades (Bankevich *et al.*, 2012) assembly resulted in the most contiguous assembly (longest contigs) with the least number of contigs. We assessed this by comparing assembly settings and coverage settings and visualizing the resulting assemblies with Bandage (Wick *et al.*, 2015). For the MiSeq samples, no coverage correction was needed as genome assemblies averaged ~48x (median depth ~20x) when run at 100x maximum coverage of reads. For the NovaSeq samples that contained significantly more reads per sample, we tested 100x, 200x, and 600x coverage for assemblies, finding that the 600x coverage resulted in the most contiguous and complete genome assembly. We further investigated the high coverage assembly method where SNPs and indels were only corrected with Pilon (Walker *et al.*, 2014) if the coverage of reads was 0.25% of the coverage on contigs >10,000bp; the SPades assembly used both R1, R2, and the additional dataset of overlapping merged reads assembled with Flash (Magoč and Salzberg, 2011); with more coverage, we did not see error correction rate increase, only an increase in the longest contig and

average N50 value; and finally, manual inspection of the genome assemblies revealed that the 600x coverage assemblies contained less orphaned nodes/contigs and dead ends.

5.3.5 Comparative genomics, phylogenomics, and measurement of microscale *Salinispora* biosynthetic potential.

Within the bactopia (Petit and Read, 2020) workflow, fastani (version 1.32) (Jain *et al.*, 2018) was used to calculate ANI values between all 99 microscale and 118 macroscale *Salinispora* strains (and together, n=217 genomes). ANI results were parsed and visualized as heatmaps, dendrograms, bar charts, and genome clusters using the packages bactaxR (https://github.com/lmc297/bactaxR) (Carroll *et al.*, 2020) and custom scripts utilizing packages reshape2, ComplexHeatmap, gplots, and ggtree (Yu *et al.*, 2017; Yu, 2020) in RStudio (RStudio Team, 2021). A phylogenetic tree of conserved single-copy core genes for the microscale (99 genomes), macroscale (118 genomes), and combined (217 genomes) *Salinispora* genomes was calculated using PhyloPhlAn 2.0 (Segata *et al.*, 2013; Asnicar *et al.*, 2020). PhyloPhlAn was run with settings --diversity medium and mapping to the "phylophlan" database of conserved single-copy genes, resulting in 314 marker genes in the microscale genome dataset: 327 marker genes in the macroscale dataset, and 324 marker genes in all 217 *Salinispora* genomes. Subsequent phylogenetic trees of all concatenated marker genes were calculated with a RAxML with PROTCATLG model of evolution with 100 bootstraps (Stamatakis, 2014). Phylogenetic trees were visualized with FigTree (Rambaut, 2016); other figure creation was performed in Adobe Illustrator.

Biosynthetic potential of the 99 microscale *Salinispora* genomes and 118 macroscale *Salinispora* genomes was measured with antiSMASH 6.0 (with all options on, including --taxon

bacteria, --cb-general, cb-knownclusters, --cb-subclusters, --cc-mibig, --asf, --rre, --pfam2go, --tigrfam, --smcog-trees, --clusterhmmer, --fullhmmer, --genefinding-tool prodigal) (Blin *et al.*, 2021) and NaPDoS2 (Chapter 2, this dissertation) (Ziemert *et al.*, 2012). BiG-SCAPE and CORASON (Navarro-Muñoz *et al.*, 2019) were used to calculate similarity between biosynthetic gene clusters including gene cluster families (GCFs) and gene cluster family clans. BiG-SCAPE was run with the following settings: Pfam 35.0 database; --include_singletons only when the comparison to the MIBiG 2.0 (Kautsar *et al.*, 2020) option (--mibig) was toggled off (and vice versa when the –mibig option was toggled on, singletons were excluded); --cutoffs 0.1 0.3 0.5 0.75 1.0; and --mode auto. BiG-SCAPE networks were visualized in Cytoscape (Carlin *et al.*, 2017). Custom R scripts were used to parse out KS classification results for each genome from NaPDoS2.

5.4 Results

We set out to selectively cultivate *Salinispora* on a microscale, here defined as 16 (4x4) sub-quadrants from a $1m^2$ quadrant from marine sediments adjacent to a coral reef in Fiji. By using selective cultivation techniques (**Figure 5.1**), we cultured >600 isolates. Obligate saltwater-growth assays to target *Salinispora* and 16S rRNA colony PCR identification of targeted isolates resulted in 172 *Salinispora arenicola*, 4 *S. pacifica*, and 28 non-target actinomycetes from the genera *Micromonospora* and *Verrucosispora* (**Figure 5.1**). We observed that across our isolates, there was a high amount of morphological diversity (**Figure 5.1**). This included *Salinispora* strains that grew like orange juice with no clumps to those that grew with large clumps of cells in liquid media. On solid agar, we observed that all of the *Salinispora* strains produced the characteristic dark brown to black spores. We also observed many strains producing fuzzy white growth that likely

represents very short aerial hyphae (**Figure 5.1**). This was confirmed to not be contamination and was reproducible in many strains where white fuzzy hyphae would form first before the dark spores, while in other colonies, only dark spores would form. There were other *Salinispora* colonies that had more popcorn-like texture compared to the circular concentric growth seen in many strains (**Figure 5.1**). Some microscale *Salinispora* colonies were observed to grow only on the surface with a small amount of growth into the solid media, however there were many strains that grew mostly into the agar, spreading orange hyphae deep into the agar where only the tops of the colonies with black and white-fuzzy growth at the surface (**Figure 5.1**). Other microscale *Salinispora* strains grew as very small colonies on solid agar plates; some strains were significantly less orange where they would grow to be pale orange and then sporulate, which was contrasted by colonies that grew to be a deep orange color before sporulation. We believe this is the first description of the white-fuzzy morphology that co-occurs with the dark brown/black spore patterning in *Salinispora*. We were amazed to observe such a wide morphological diversity from the *Salinispora* strains isolated in 16 sub-quadrants from a $1m^2$ plot, and this led us to believe that the strains were likely not clonal and instead were genetically diverse (**Figure 5.1**).

To assess the diversity based on 16S rRNA sequence of the new microscale *Salinispora* isolates, we constructed a phylogenetic tree of all 118 macroscale (the previously reported *Salinispora* genomes from (Millán-Aguiñaga *et al.*, 2017)) and 102 microscale *Salinispora* 16S rRNA colony PCR sequences (**Figure 5.3**).

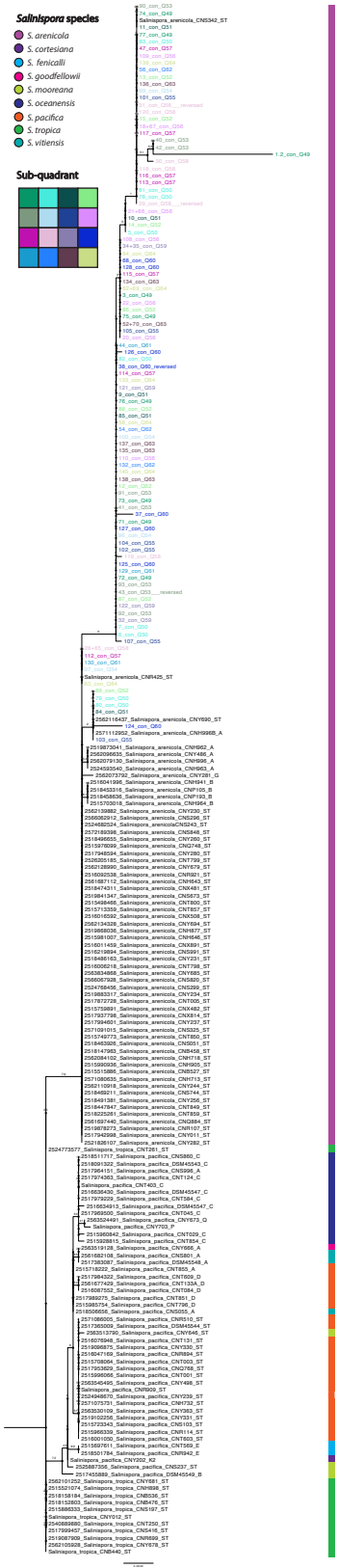


Figure 5.3. Phylogenetic tree of 102 microscale and 118 macroscale *Salinispora* 16S rRNA sequences. Microscale taxa are colored by sub-quadrant isolation location and the bar on the right denotes *Salinispora* species.

When I colored the microscale *Salinispora* strains by quadrant isolation location, I observed that 1) all of the microscale 16S rRNA sequence formed a large group outside the top clade of the tree and 2) most were identical if not very closely related with very short branch lengths (**Figure 5.3**). In the microscale clades, there were 4 macroscale *S. arenicola* of "ST" and "SA" 16S rRNA sequence type sequences, thus we could tentatively classify all microscale *Salinispora* strains as *S. arenicola* "ST" type, with *S. arenicola* CNZ-922 as sequence type "A" (**Figure 5.3**). We did not observe any 16S rRNA phylogenetic grouping of strains isolated from the same sub-quadrant location (**Figure 5.3**). *Salinispora* species can share 99% similarity in the 16S rRNA gene, so even if our complete PCR products were 1,500bp long, there would still be not enough significant base pairs in an alignment to resolve phylogenetic diversity. We thus manually inspected the alignment of all 16S rRNA sequences to identify single nucleotide polymorphisms (SNPs) which can be used in some cases in *Salinispora* to differentiate between species. We discovered that all 102 of the sequences were identical, apart from three strains with 2 total SNPs (**Figure 5.4**). Based on the closest 16S rRNA *Salinispora* sequence type, we could putatively identify that the three strains with SNPs were most closely related to a *S. pacifica* 16S rRNA reference sequence (**Figure 5.4**). Thus, across the microscale quadrant, we had isolated 98 *S. arenicola* strains and 3 *S. pacifica* strains with high morphological diversity (**Figure 5.1**). It is possible that the varying SNPs in one of the putative *S. pacifica* strains could indicate another species, as the genus of *Salinispora* was recently updated with six additional species (Román-Ponce *et al.*, 2020), but this differentiation can only be resolved with whole-genome sequencing.

Strain #	Different **SNP**s										
S. pacifica ref. seq.	A	A	C	…	A	…	G	…	A	C	G
CNZ-865	A	A	C	…	**G**	…	G	…	A	C	G
CNZ-875	A	A	C	…	**G**	…	**T**	…	A	C	G
CNZ-966	A	A	C	…	**G**	…	**T**	…	A	C	G

Figure 5.4. SNP analysis of all 102 microscale *Salinispora* 16S rRNA colony PCR sequences revealed 2 SNPs (bolded) between three strains which were most closely related to *S. pacifica* via 16S rRNA sequence alignment.

We whole-genome sequenced 99 microscale *Salinispora* genomes using two different short-reads sequencing platforms (MiSeq PE300 and NovaSeq PE150). On average, the microscale *Salinispora* genome assemblies contained: 88 contigs with an average length of 5,635,033 bp (base pairs); N50 185,576 bp with the largest contig length of 470,884 bp; 69.62% GC content; 5,059 genes with 3 rRNA and 64 tRNA. Estimates with CheckM (Parks *et al.*, 2015) indicated on average the genome assemblies were 99-100% complete with 314 of the marker genes for *Actinomycetales* present and on average 0.19% contamination. These metrics are similar to, and for some genomes, better than the original macroscale *Salinispora* genome assemblies (Millán-Aguiñaga *et al.*, 2017). Upon calculation of the average nucleotide identity between all 99 microscale *Salinispora*, we observed that most of the genomes were the same *S. arenicola* species with >95% ANI values; however, there were three genomes belonging to the *S. pacifica* (1) and *S. oceanensis* (2) species with 87-88% ANI (**Figure 5.5**). Of the genomes with >95% ANI, there also seems to be an intra-species division as seen by the two peaks split at 99% ANI (**Figure 5.5**).

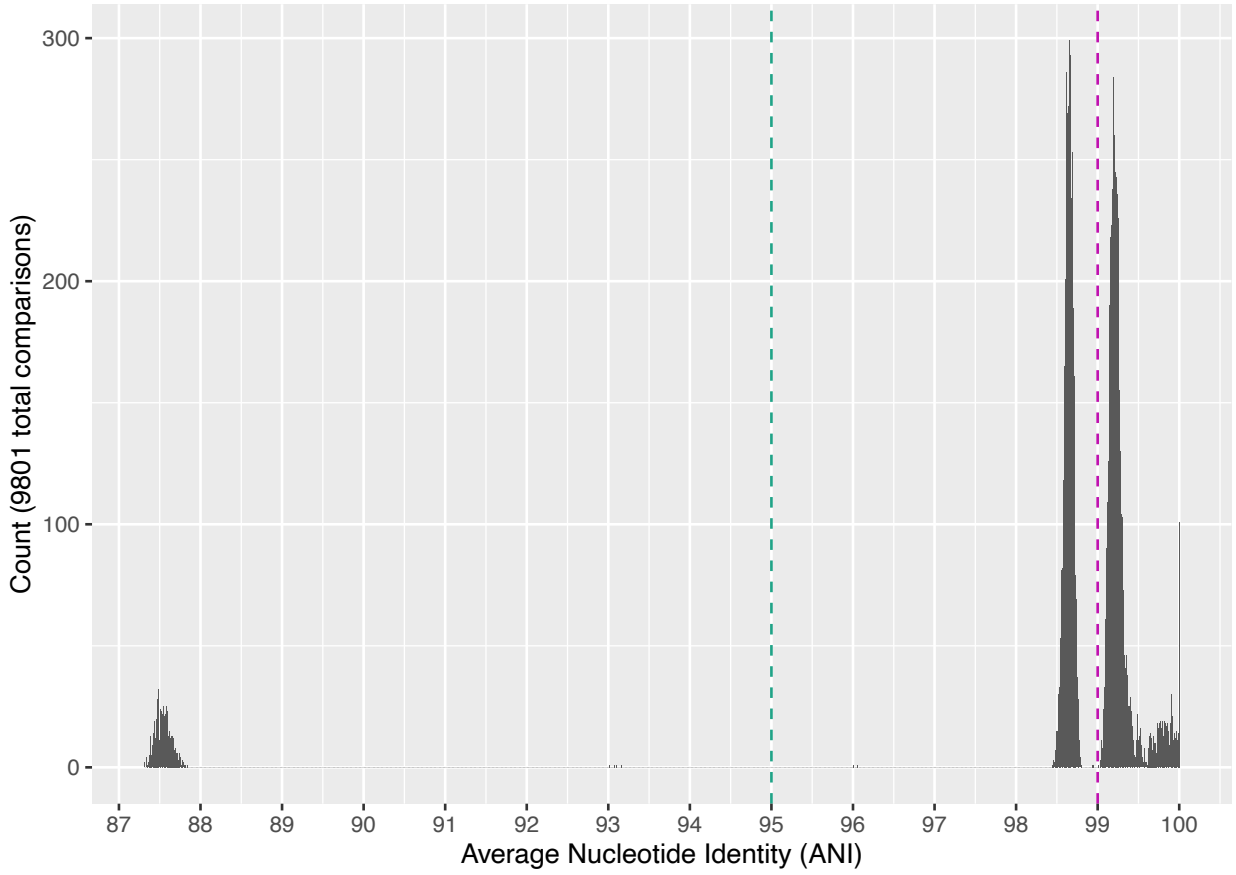


Figure 5.5. Average nucleotide identity (ANI) for the new 99 microscale *Salinispora* genomes. Lines denoting 95% (teal) and 99% (pink) ANI are drawn.

To investigate the relatedness of the 99 microscale *Salinispora* genomes further, we calculated an ANI heatmap with a cutoff at 95% ANI for species designation (**Figure 5.6**). This confirmed that we indeed had isolated 3 different species from the microscale 1m^2 quadrant— 96 *S. arenicola,* 2 *S. oceanensis,* and 1 *S. pacifica* (**Figure 5.6**). Within the *S. arenicola,* we observed there were two main groups that were split around 96-98% ANI, which indicates that there could be diversification within the *S. arenicola* on a sub-quadrant spatial scale (**Figure 5.6**). This type of pattern would not have been observed in the macroscale genome dataset. However, when we mapped on the sub-quadrant isolation location of the microscale genomes, there was no apparent

pattern of *S. arenicola* strains isolated from the same sub-quadrant being more related to one another than strains in other quadrants (**Figure 5.6**). When we plotted the 99% ANI values in a heatmap, there were clear groups of *S. arenicola* strains that were more closely related to each other in the 99 microscale dataset (**Figure 5.7**). The two ANI groups were split into 3 groups and many subgroups (**Figure 5.7**). This could indicate that there is a heterogenous, diverging population of *S. arenicola* in the Fiji marine sediment community, which has further implications for considering the *Salinispora* species diversity on a worldwide scale.

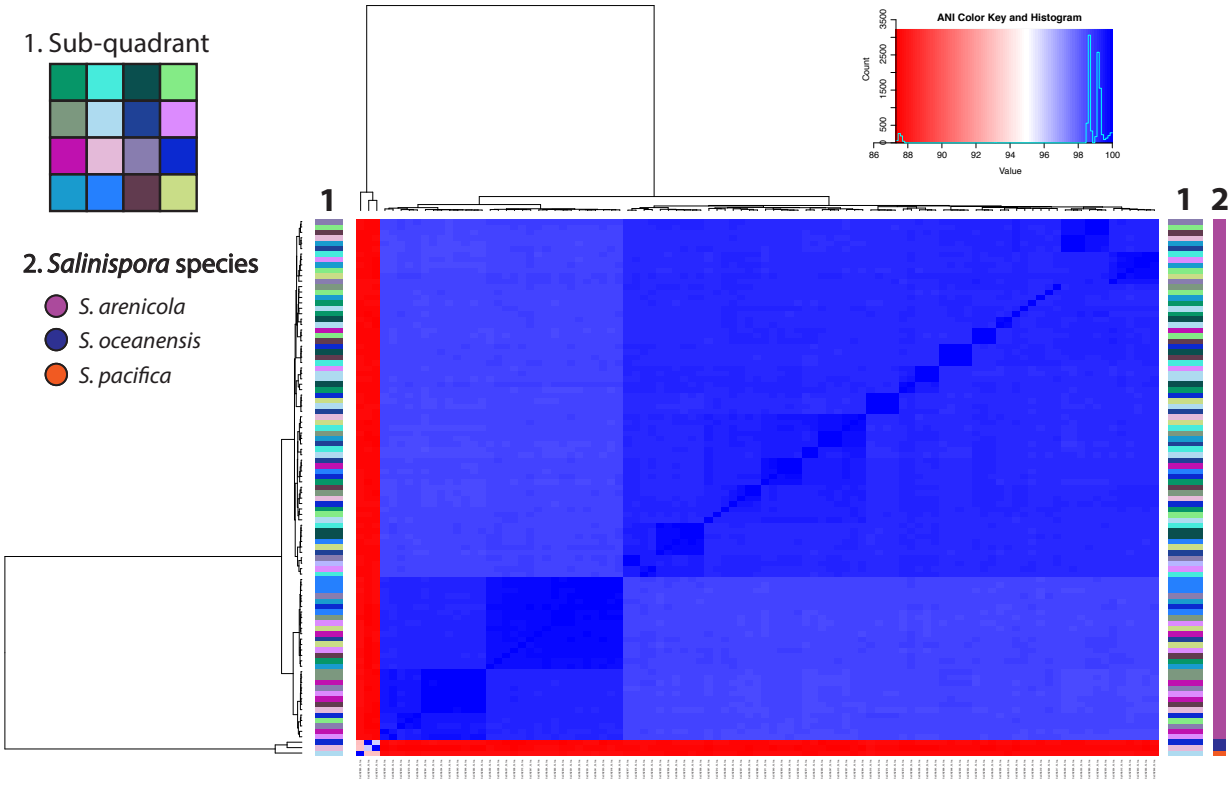


Figure 5.6. ANI heatmap of the 99 new microscale *Salinispora* genomes, with the cutoff colored at 95% ANI. Colored bars indicate 1) sub-quadrant isolation location and 2) Salinispora species.

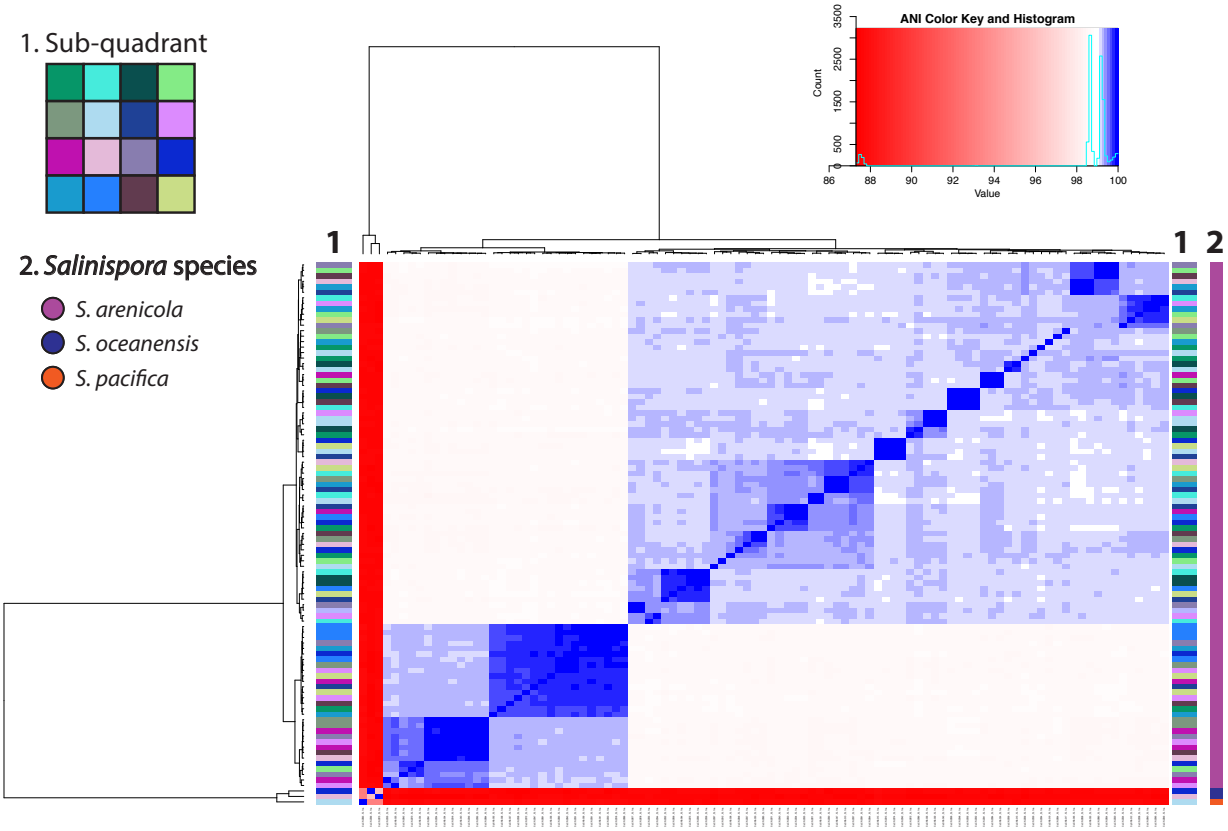


Figure 5.7. ANI heatmap of the 99 new microscale *Salinispora* genomes, with the cutoff colored at 99% ANI. Bar colors indicate 1) sub-quadrant isolation location and 2) *Salinispora* species.

To summarize the new microscale *Salinispora* genomic diversity and confirm the species designations, we constructed an ANI dendrogram which confirmed the identification of three *Salinispora* species based on 95% ANI values (**Figure 5.8**). If the macroscale *Salinispora* genomes were grouped based on 99% ANI, we observe two subclades within *S. oceanensis* and *S. arenicola* (**Figure 5.8**). This is striking because to the best of our knowledge, no intra-species divergence has been observed at this scale from marine sediment. Contrary to the hypothesis that all of the

isolated microscale strains would be clonal *S. arenicola*, it appears as if there are at least two major *S. arenicola* populations within the 1m^2 sediment quadrant.

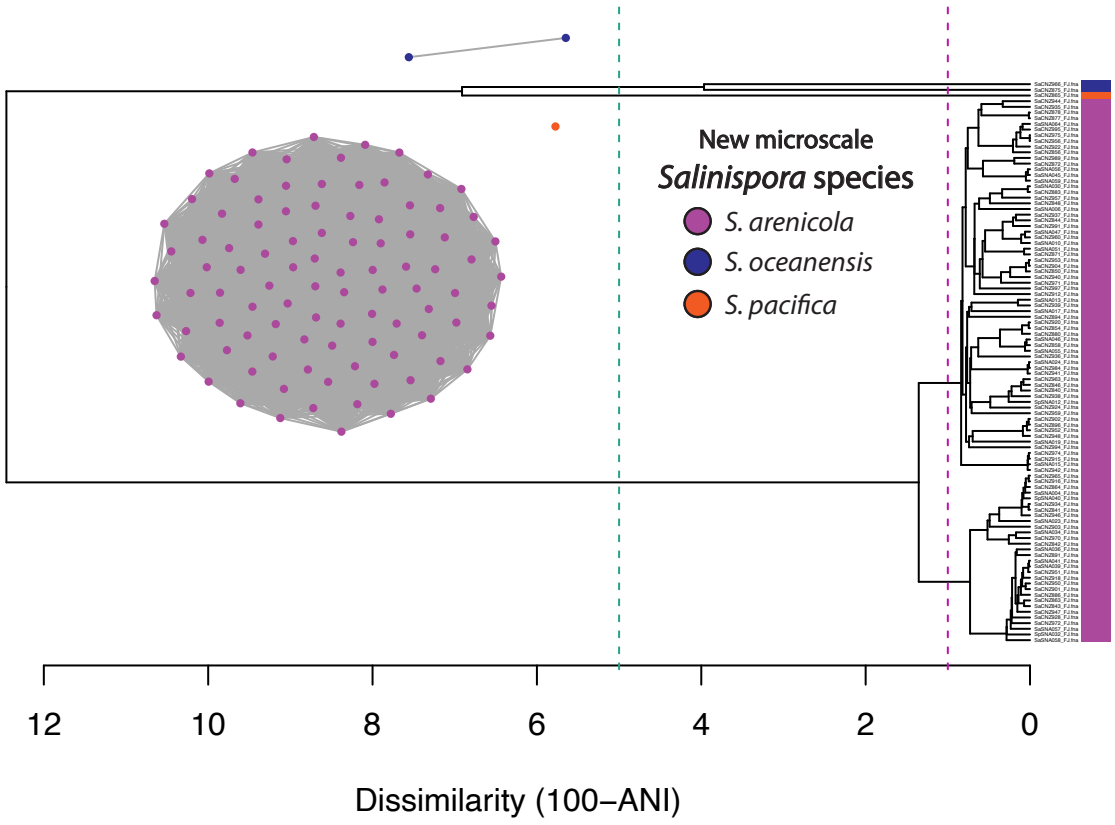


Figure 5.8. ANI dendrogram of the new 99 microscale *Salinispora* genomes. The teal line indicates 95% ANI; the pink line indicates 99% ANI. ANI clusters are colored by *Salinispora* species.

Our next aim was to understand what was driving the diversification within the microscale *S. arenicola* clade and compare the genomes to the current macroscale *Salinispora* genomes which were isolated from locations worldwide. An ANI analysis of all 217 *Salinispora* genomes showed significant divergence below 95%, as predicted for the nine named *Salinispora* species and a large number of intra-species genomes sharing 98-99% ANI (**Figure 5.9**). In comparison with the 99 microscale *Salinispora* dataset, it is clear that the six species are captured from other worldwide

isolation locations thus contributing to the larger number of genomes with lower comparative ANI values (**Figure 5.9**).

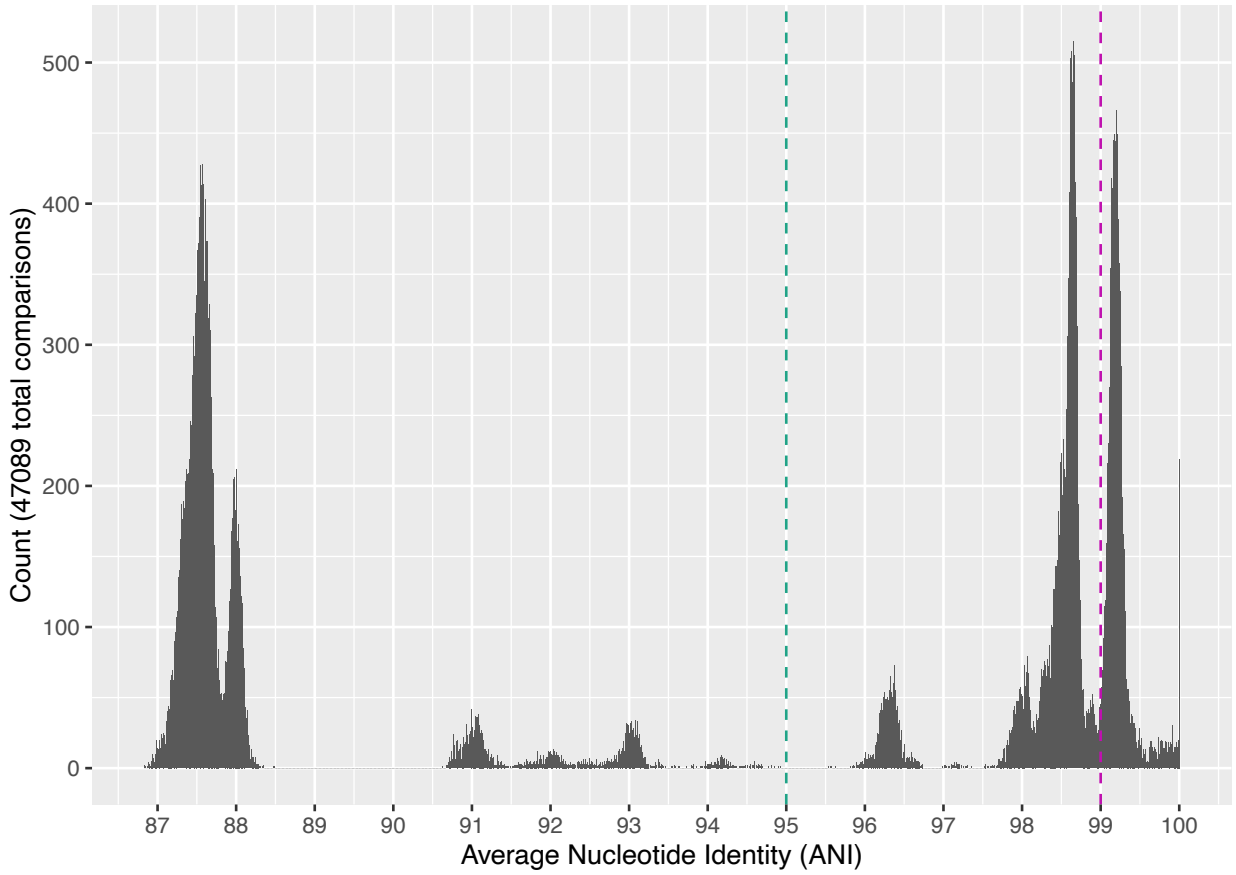


Figure 5.9. Average nucleotide identity (ANI) for all 217 micro- and macroscale *Salinispora* genomes. Lines denoting 95% (teal) and 99% (pink) ANI are drawn.

When all *Salinispora* were compared in a 95% ANI heat plot, we discovered that the new microscale *Salinispora* genomes were broadly distributed among the previously described *S. arenicola* diversity (**Figure 5.10**). Microscale *S. arenicola* strains formed clades with multiple macroscale *S. arenicola* strains, with only two groups at the bottom of the *S. arenicola* clade not containing any new microscale *S. arenicola* genomes (**Figure 5.10**). Additionally, the 2 *S. oceanensis* and 1 *S. pacifica* microscale genomes were closely related to other members of their respective species (**Figure 5.10**). Interestingly, the two microscale *S. oceanensis* genomes were

divergent from one another and each were more similar to other macroscale *S. oceanensis* instead of each other, despite being isolated from the same microscale location (**Figure 5.10**). We investigate this further by plotting the 99% ANI heatmap where the intra-species groups with both micro- and macroscale *Salinispora* strains could be better observed (**Figure 5.11**). It appears as if the two large intra-species *S. arenicola* groups identified from the microscale dataset drive the diversification of those two groups whereas many of the other macroscale *S. arenicola* strains fall within smaller groups or are very small components of the larger 2 groups (**Figure 5.11**). This could indicate that at the microscale spatial scale of sampling, we have detected two distinct yet co-occurring *S. arenicola* populations that were also observed in the macroscale genomes. The final designation of all 217 *Salinispora* genomes was determined from an ANI dendrogram and cluster analysis (**Figure 5.12**). With the 99 new microscale *Salinispora* genome dataset, we have increased the microdiversity within *S. arenicola* and two other *Salinispora* species (**Figure 5.12**). Of note, our analysis indicated that *S. mooreana* should be split into 2 different species as one of the three genomes had < 95% ANI compared to the other two genomes within the species (**Figure 5.12**) (light green clade at the top of the dendrogram; 95% ANI teal line splits the branching clade). While this could be due to differences in kmer values for ANI calculations, the *S. mooreana* CNY-646 strain was previously noted to not cluster with the other 2 members of its species in ANI-AF calculations (Millán-Aguiñaga *et al.*, 2017). This should be further investigated as it seems this strain could be a new *Salinispora* species based on the most recent 95% ANI cutoff (**Figure 5.12**).

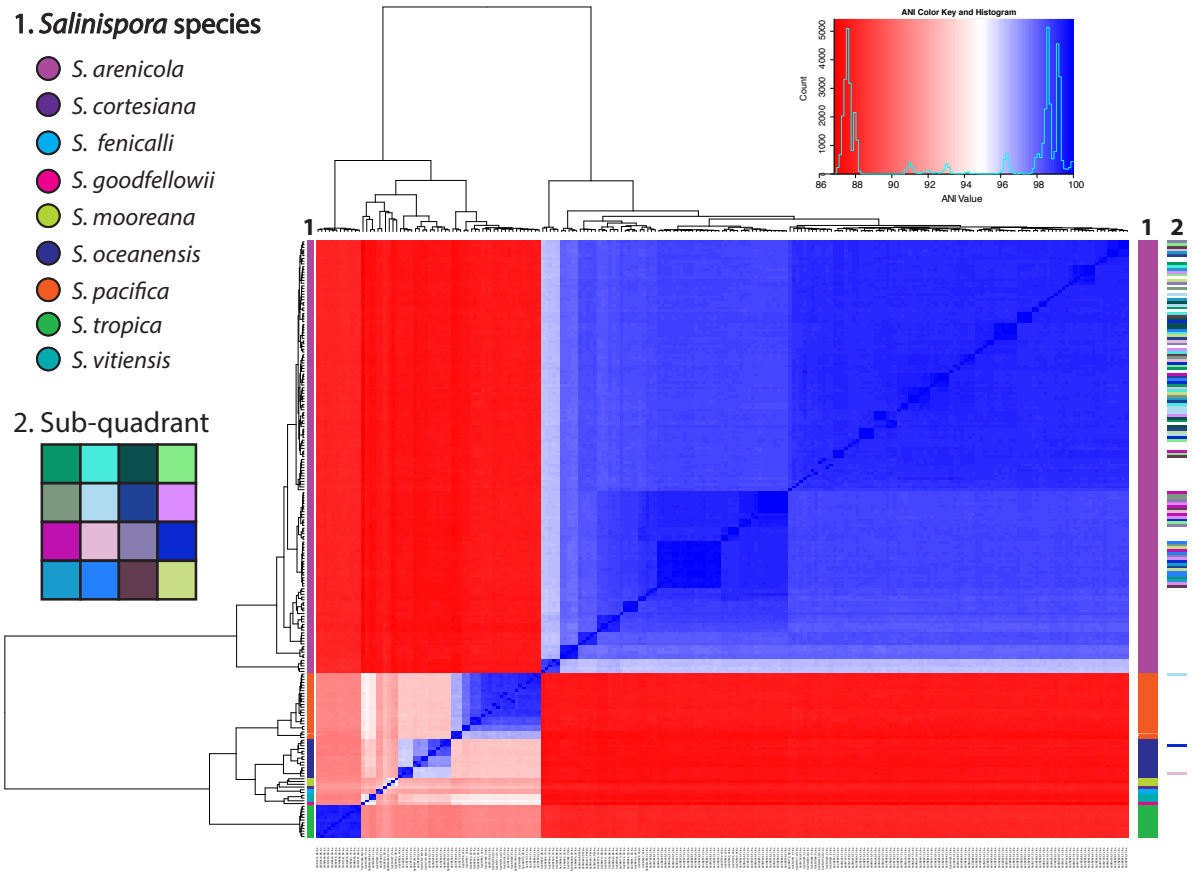


Figure 5.10. ANI heatmap of all 217 micro- and macroscale *Salinispora* genomes, with the cutoff colored at 95% ANI. Bar colors indicate 1) sub-quadrant isolation location and 2) *Salinispora* species.

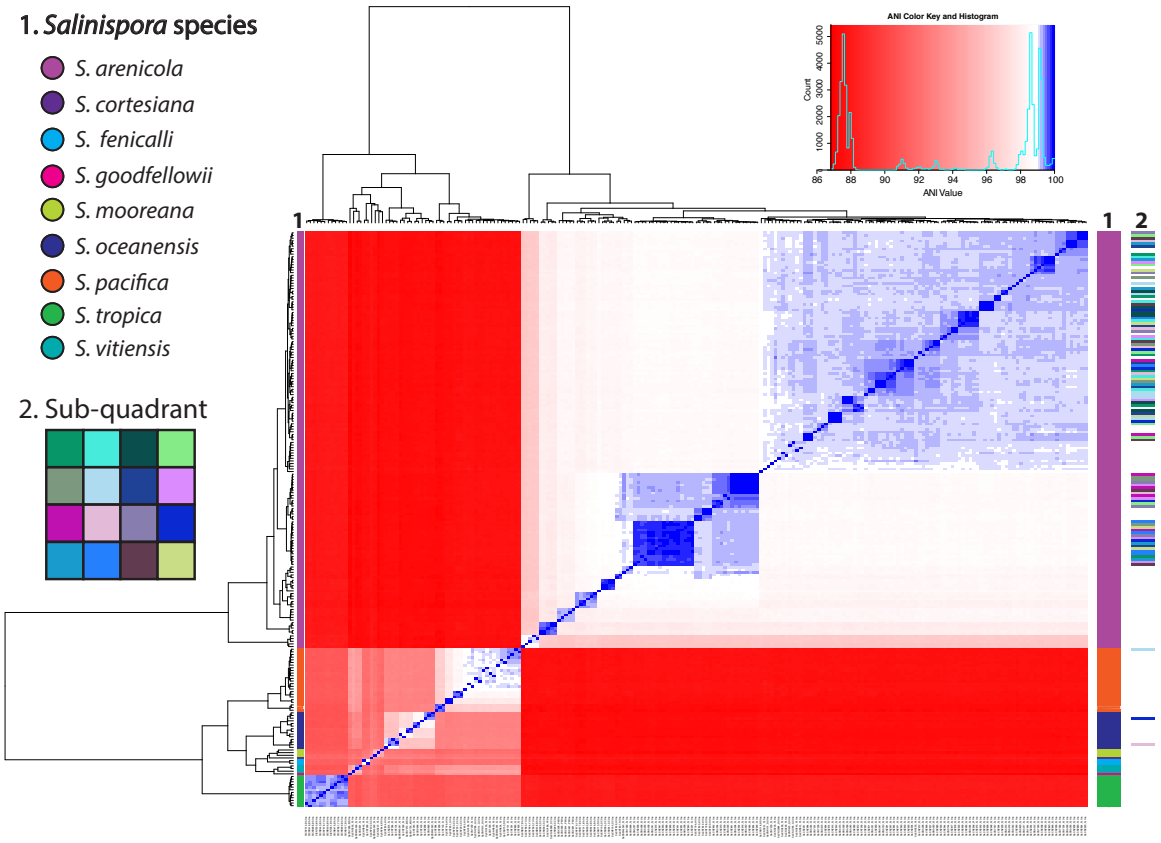


Figure 5.11. ANI heatmap of all 217 micro- and macroscale *Salinispora* genomes, with the cutoff colored at 99% ANI. Bar colors indicate 1) sub-quadrant isolation location and 2) *Salinispora* species.

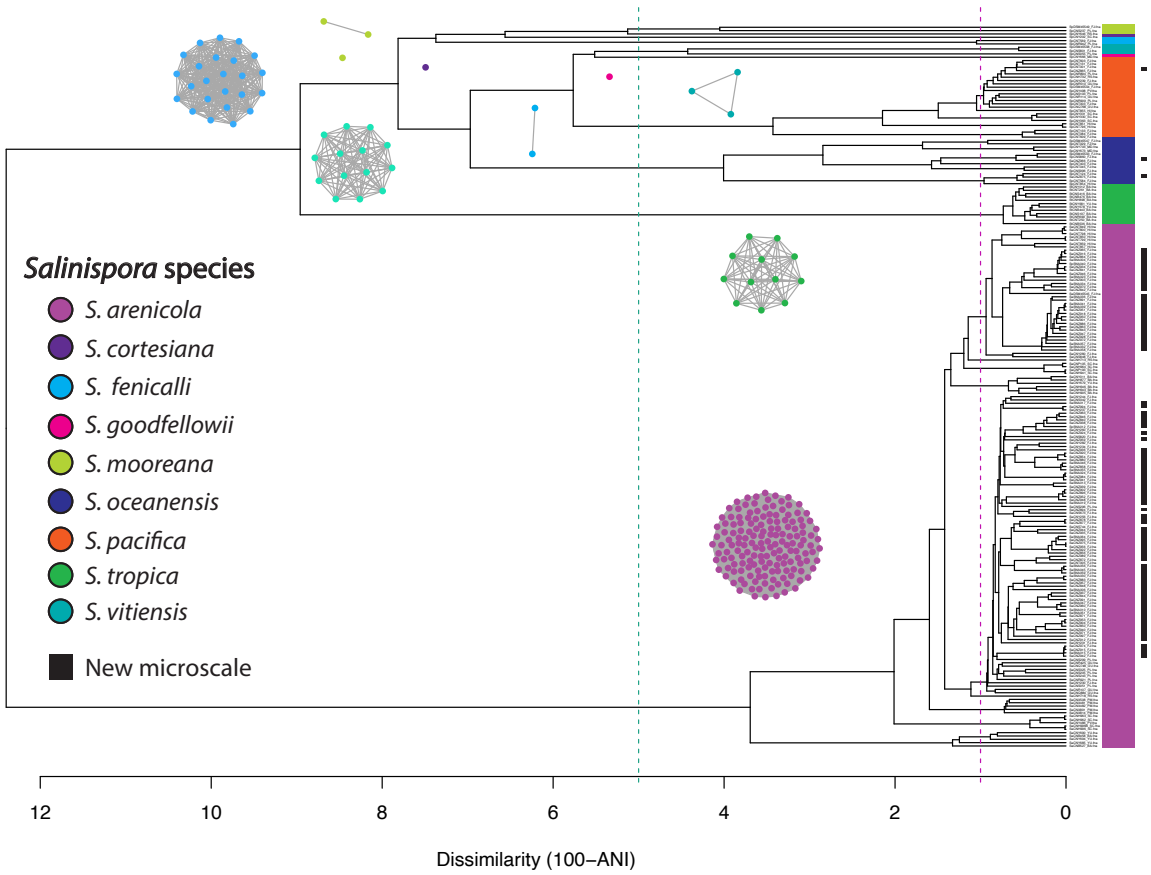


Figure 5.12. ANI dendrogram of all 217 micro- and macroscale *Salinispora* genomes. The teal line indicates 95% ANI; the pink line indicates 99% ANI. ANI clusters are colored by *Salinispora* species; black boxes mark new microscale strains.

While ANI comparisons can inform species-boundary predictions, we additionally constructed a phylogenetic tree comprised of 324 single-copy conserved genes from all 217 *Salinispora* genomes (**Figure 5.13**). Again, we observed that the microscale *S. arenicola* were spread throughout the *S. arenicola* clade with other macroscale *S. arenicola* genomes, and there was no apparent pattern based on sub-quadrant isolation location (**Figure 5.13**). However, we did observe that many of the macroscale *S. arenicola* genomes that claded with microscale strains were isolated from Fiji (**Figure 5.13**). This is fascinating, as while the isolation locations are likely not the same, the strains were isolated years apart, thus closer genomic comparisons could be performed to look for any evolutionary adaptations over time. There were also macroscale *S.*

arenicola strains isolated from Palau, the Red Sea, Guam, Palmyra, among others that claded next to microscale strains (**Figure 5.13**). This observation could support the hypothesis that all *S. arenicola* subspecies types are everywhere, and the environment and marine sediment community selects for what is present in the population. Closer analysis of the specific genomic differences between the strains isolated from different locations could help elucidate these patterns. It is apparent however that there is a location-dependent factor driving diversification within *S. arenicola* as most macroscale strains isolated from the Bahamas, Yucatán, Puerto Vallarta, and the Sea of Cortez formed a macroscale-only group at the base of the phylogenetic tree (**Figure 5.13**). The microscale *S. pacifica* CNZ-865 strain was observed to clade next to other *S. pacifica* strains that were also isolated from Fiji and the same pattern was observed for the two different *S. oceanensis* strains, though one was sister to a strain isolated from Hawaii (**Figure 5.13**).

I next wanted to investigate if the biosynthetic potential of the new microscale *Salinispora* genomes could be linked to any of the intra-species diversification observed and if we captured any new *Salinispora* biosynthetic diversity. To do this, I analyzed all 99 genomes with antiSMASH 6.0 (Blin *et al.*, 2021) and compared the BGC relatedness with BiG-SCAPE (Navarro-Muñoz *et al.*, 2019). I uncovered 3,023 total BGCs from all 99 genomes, with most belonging to the "others" BGC class (930), followed by RiPPs (463), NRPS (450), and type 1 PKS (366). On average, the microscale genomes had 27-32 BGCs per genome. The BGCS were clustered into gene cluster families based on similarity distance scores, resulting in a total of 145 GCFs with an average of 21 BGCs per GCF. In order to ascertain the relatedness of the BGCs, I built a BGC similarity network with characterized BGCs from the MIBiG 2.0 reference database (Kautsar *et al.*, 2020) (**Figure 5.14**).

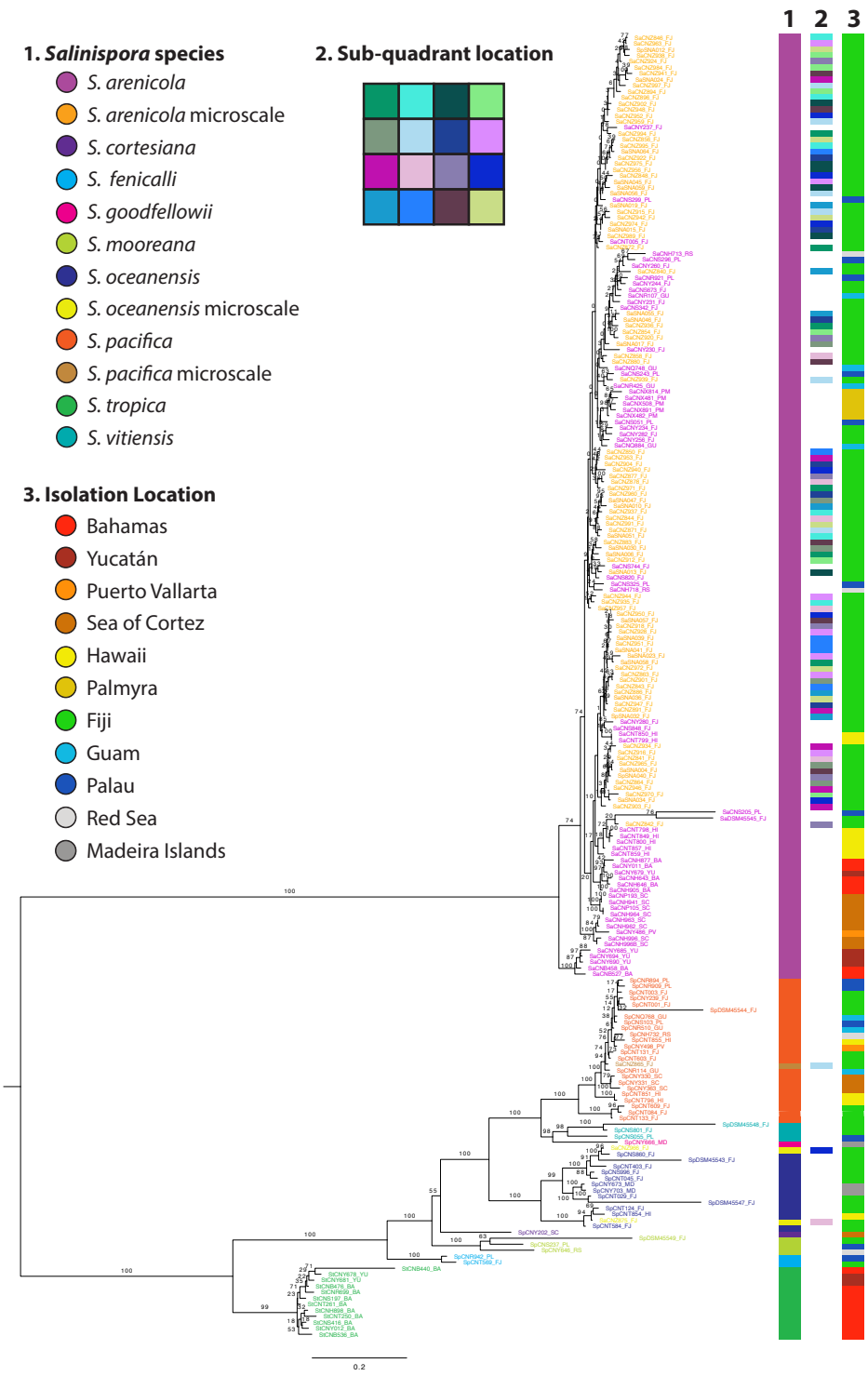


Figure 5.13. Phylogenetic tree of 324 conserved single-copy genes from all 217 micro- and macroscale *Salinispora* genomes.

Maximum likelihood tree was calculated with a PROTCATLG model of evolution with 100 bootstraps in RAxML. Colored bars indicate 1) *Salinispora* species, 2) microscale sub-quadrant location, and 3) geographical isolation location.

Only 11 clusters in the BGC network had known products associated with them, including: cyclomarin, calichaeamicin, alkyl-O-dihydrogeranyl-methyoxyhydroquinones, lymphostin, rifamycin, staurosporine, griseusin, retimycin, salinichelins, lomaiviticin, and thiolactomycin (**Figure 5.14**). Many of the clusters did not have known products associated, which suggests there is a large amount of biosynthetic potential still to be realized from these genomes (**Figure 5.14**). Most of the BGC clusters were microscale *S. arenicola* specific, and there were also *S. pacifica* and *S. oceanensis* microscale species-specific clusters (**Figure 5.14**). What is striking is that from a $1m^2$ plot, I was able to capture a large biosynthetic potential including rare BGCs like retimycin, salinichelin, and thiolactomycin.

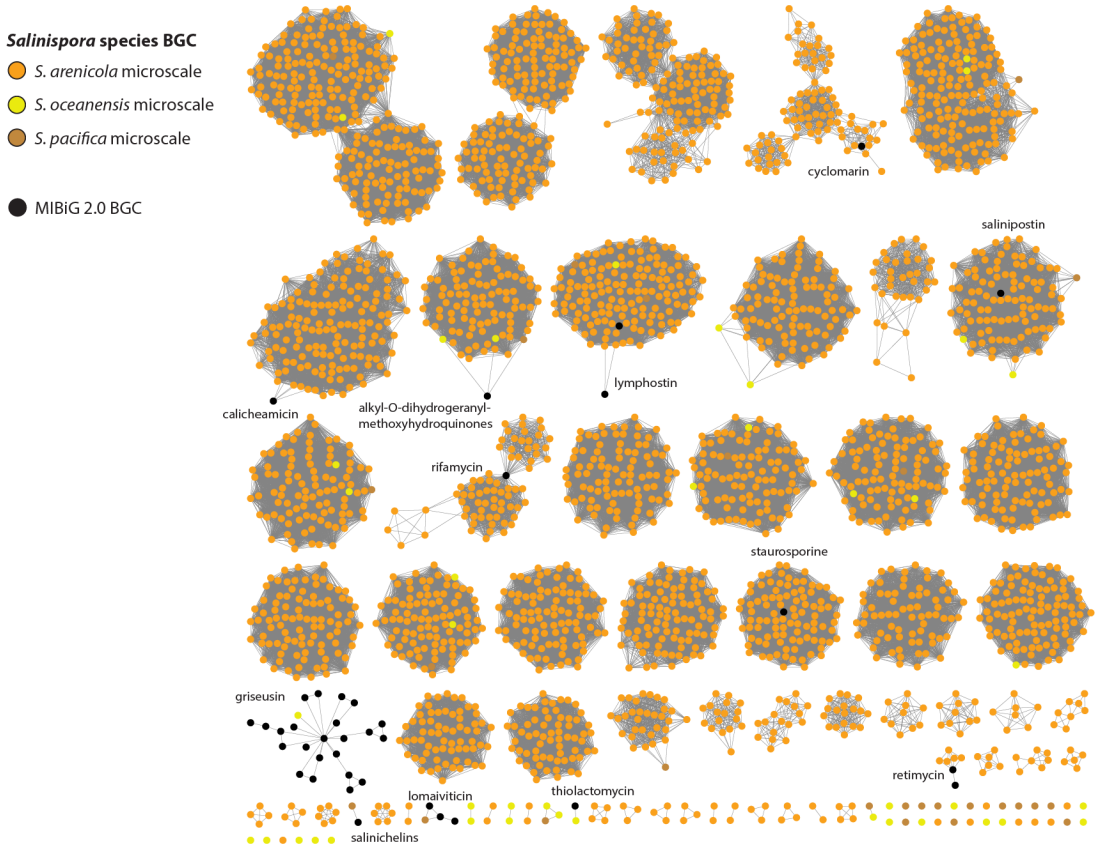


Figure 5.14. Network of BGCs from the new 99 microscale *Salinispora* genomes, as identified by antiSMASH 6.0 and BiG-SCAPE (c=0.3).

Black nodes with labels are characterized BGCs from the MIBiG 2.0 database.

Next, I hypothesized that some of the biosynthetic potential observed in the microscale *S. arenicola* genomes would differ from the macroscale and could help drive separation of the clade groups seen in our ANI and phylogenetic analyses. I identified the BGCs (n=6,425; 417 GCFs) and their similarity from all 217 *Salinispora* strains (**Figure 5.15**). In the BGC network, 23 clusters could be linked to a putative BGC compound from the MIBiG 2.0 (Kautsar *et al.*, 2020) database. The microscale *S. arenicola* genomes captured new BGC diversity, including an additional five retimycin BGCs where there was only one previously predicted in the macroscale *Salinispora* genomes. Additionally, one more thiolactomycin BGC was captured in the microscale *S. oceanensis* genomes, which not only increases the BGC diversity of that compound class but illustrates a BGC that one of the microscale *S. oceanensis* has but the other does not (**Figure 5.15**). Overall, the microscale *S. arenicola* genomes increased BGC diversity by capturing additional BGCs linked to known molecules and detecting new BGC clusters with unknown products. Furthermore, many diverging BGCs that could contain modifications with regards to gene content or SNPs within the BGC thus contributing to the overall chemical diversity. Finally, there were many species-specific BGC clusters as previously reported (Letzel *et al.*, 2017; Chase *et al.*, 2021).

To explore the polyketide chemical diversity of the microscale *Salinispora*, I analyzed all 99 genomes with the NaPDoS2 webtool (described in Chapter 2 of this dissertation) and uncovered four uniquely distributed type II beta-branching type BGCs (**Table 5.1**), among many other types of polyketide KSs. Across all 99 genomes, there was commonly shared KS biosynthetic potential including type I modular *cis*-AT and *cis*-hybrid domains, however there were also some unique KS domains. In the microscale *S. pacifica*, I identified type II angucycline KSα and KSβ domains, whereas one of the *S. oceanensis* strains contained two type II angucycline KSα and KSβ domains.

However, I was interested in the type II beta-branching type of KS domain found in only four of the microscale strains (**Table 5.1**) as that type of KS domain has only been observed in two macroscale *S. arenicola* genomes (Chapter 2, this dissertation). It appears that the type II beta-branching BGC is highly conserved between the four microscale *S. arenicola* that contain the BGC, and the beta-branching domains are found in a "super-cluster" BGC with other *trans*-AT PKS, NRPS, and lanthipeptide class II biosynthetic machinery (**Figure 5.16**).

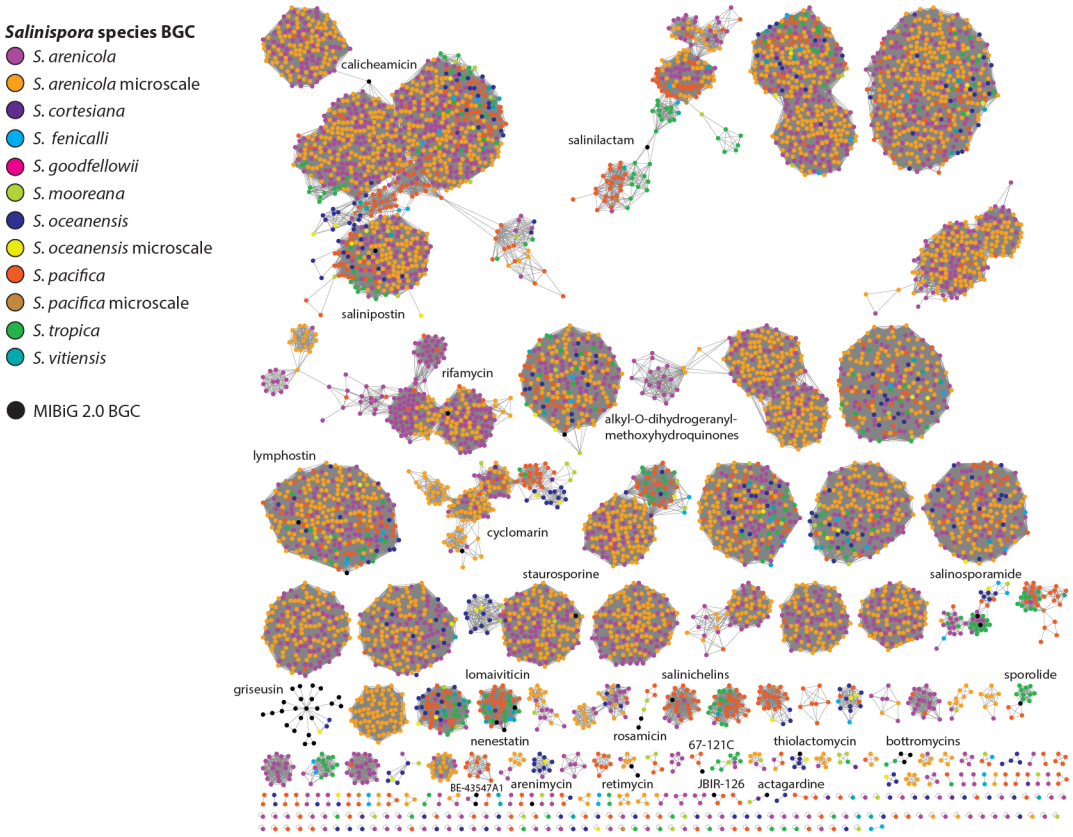


Figure 5.15. Network of BGCs from all 217 micro- and macroscale *Salinispora* genomes, as identified by antiSMASH 6.0 and BiG-SCAPE (n=6,425 BGCs; clustering cutoff c=0.3).

Black nodes with labels are characterized BGCs from the MIBiG 2.0 database.

Table 5.1. Polyketide KS biosynthetic potential predicted with the NaPDoS2 webtool for all 99 new microscale *Salinispora* genomes (run with default NaPDoS2 parameters: 1e-8, min. alignment length 200aa)

				NaPDoS2: KS classes																		
				FAS	Type II PKS											Type I PKS						
								Aromatic								*cis* -AT modular			cis-AT iterative			*trans* -AT	
Species	*Strain*	*Quad*	*NP2 total*	FASII	Betabranch	Polyene	Angu-I KSa	Angu-I KSb	Angu-II KSa	Angu-II KSb	Polyphen KSa	Polyphen KSb	*cis* -AT	cisloading	cisHybridKS	Ened	iPKSaromatic	PTM	*trans* -AT	*trans* -HybridKS0
S. pacifica	CNZ865	54	29	3		2			1	1	2	2	12	1	3	1			1	
S. oceanensis	CNZ875	58	17	5		1	2	2			2	2	1		1	1				
S. oceanensis	CNZ966	60	14	5							1	1	2	2	2	1				
S. arenicola	CNZ840	61	37	2		2					1	1	23	1	2	2	2	1		
S. arenicola	CNZ841	58	34	3		4					1	1	19	1	2	2	1			
S. arenicola	CNZ842	59	40	4		5					2	2	21	1	2	2	1			
S. arenicola	CNZ843	62	32	2		3							21	1	2	2	1			
S. arenicola	CNZ844	58	35	3		3					2	2	18	1	1	2	2	1		
S. arenicola	CNZ846	50	35	2		1							23	2	2	2	2	1		
S. arenicola	CNZ848	60	26	1		2					1	1	14		2	2	2	1		
S. arenicola	CNZ850	62	34	2		3					1	1	19	1	2	2	2	1		
S. arenicola	CNZ854	52	47	2		1					1	1	34	2	1	2	2		1	
S. arenicola	CNZ856	64	37	3		3					3	3	18	1	1	2	2	1		
S. arenicola	CNZ858	58	47	2		2					1	1	33	2	1	2	2		1	
S. arenicola	CNZ863	56	32	2		3							21	1	2	2	1			
S. arenicola	CNZ864	53	34	3		4					1	1	19	1	2	2	1			
S. arenicola	CNZ871	54	33	1		2							22	2	1	2	2	1		
S. arenicola	CNZ872	49	35	2		2					1	1	22	2	1	2	2			
S. arenicola	CNZ877	59	25	2		2					2	2	10		2	2	2	1		
S. arenicola	CNZ878	58	21	1		1					1	1	10		2	2	2	1		
S. arenicola	CNZ880	63	52	3		3					2	2	34	2	1	2	2		1	
S. arenicola	CNZ883	63	28	1		1							18	1	2	2	2	1		
S. arenicola	CNZ886	61	39	4		5					2	2	20	1	2	2	1			
S. arenicola	CNZ891	57	32	2		3							20	1	2	3	1			
S. arenicola	CNZ894	52	29	2		2					1	1	18	1	1	1	2			
S. arenicola	CNZ896	50	35	3		3					2	2	19	1	1	2	2			
S. arenicola	CNZ901	53	32	2		3							21	1	2	2	1			
S. arenicola	CNZ902	51	32	2		2					1	1	19	1	1	2	2		1	
S. arenicola	CNZ903	57	36	3		4					1	1	21	1	2	2	1			
S. arenicola	CNZ904	55	37	3		3					2	2	19	1	2	2	2	1		
S. arenicola	CNZ912	52	33	2		1							22	2	1	2	2	1		
S. arenicola	CNZ915	54	35	1	1	2					1	1	15		2	2	2	1	6	1
S. arenicola	CNZ916	56	35	3		4					1	1	19	1	2	3	1			
S. arenicola	CNZ918	59	34	3		5					1	1	19	1	2	1	1			
S. arenicola	CNZ920	59	48	2		1					1	1	34	2	1	2	3		1	
S. arenicola	CNZ922	55	38	3		3					3	3	20	1	1	2	1	1		
S. arenicola	CNZ924	59	31	1		1							20	1	2	3	2	1		
S. arenicola	CNZ928	56	35	3		4					1	1	20	1	2	2	1			
S. arenicola	CNZ934	57	38	4		5					2	2	19	1	2	1	1	1		
S. arenicola	CNZ935	50	40	2		2					2	2	25	2	1	2	2			
S. arenicola	CNZ936	49	24	1		1							15	1	1	2	3			
S. arenicola	CNZ937	50	33	2		2					1	1	20	1	2	2	1	1		
S. arenicola	CNZ938	52	36	2		2					1	1	21	1	2	2	3	1		
S. arenicola	CNZ939	54	32	2		2					1	1	20	1	1	2	2			
S. arenicola	CNZ940	60	41	3		3					2	2	22	2	2	2	1	2		
S. arenicola	CNZ941	63	35	2		1							24	2	2	2	2			
S. arenicola	CNZ942	64	36	1	1	2					1	1	16		2	2	2	1	6	1
S. arenicola	CNZ944	56	34	2		2					2	2	18	1	1	3	2	1		
S. arenicola	CNZ946	57	34	3		4					1	1	19	1	2	2	1			
S. arenicola	CNZ947	55	33	3		4					1	1	19	1	2	1	1			
S. arenicola	CNZ948	63	33	2		2					1	1	20	1	1	3	2			
S. arenicola	CNZ950	60	37	4		5					2	2	19	1	2	1	1			
S. arenicola	CNZ951	62	31	3		3							19	1	2	2	1			
S. arenicola	CNZ952	60	31	2		2					1	1	19	1	1	2	2			
S. arenicola	CNZ953	57	36	2		3					2	2	19	1	2	2	2	1		
S. arenicola	CNZ956	51	31	1		1					1	1	18	1	2	3	2	1		
S. arenicola	CNZ957	58	30	2		2					1	1	18	1	1	1	2	1		
S. arenicola	CNZ959	54	32	2		2					1	1	19	1	1	2	2	1		
S. arenicola	CNZ960	55	41	3		3					2	2	22	2	2	2	2	1		
S. arenicola	CNZ963	58	30	1		1							19	1	2	3	2	1		
S. arenicola	CNZ965	53	37	5		4					2	2	19	1	2	1	1			
S. arenicola	CNZ970	52	34	3		4					1	1	20	1	2	1	1			
S. arenicola	CNZ971	49	37	2		2					1	1	21	2	2	3	2	1		
S. arenicola	CNZ972	64	38	4		5					2	2	20	1	2	1	1			
S. arenicola	CNZ974	60	41	2	1	2					1	1	19	1	2	2	2	1	6	1
S. arenicola	CNZ975	51	29	1		1					1	1	18	1	1	2	2	1		
S. arenicola	CNZ984	52	36	2		2					1	1	21	2	2	2	2	1		
S. arenicola	CNZ989	51	31	2		2					1	1	19	1	1	2	2			
S. arenicola	CNZ991	64	35	2		2					1	1	21	2	1	2	2	1		
S. arenicola	CNZ994	49	33	2		2					1	1	19	1	1	2	2	2		
S. arenicola	CNZ995	50	36	3		3					3	3	18	1	1	2	2			
S. arenicola	CNZ997	54	30	1		1							19	1	1	3	2	1	1	
S. arenicola	SNA004	63	33	3		4					1	1	19	1	2	1	1			
S. arenicola	SNA006	49	32	3		2					1	1	18	1	2	2	2			
S. arenicola	SNA010	61	33	2		2					1	1	19	1	2	2	2	1		
S. arenicola	SNA012	64	27	1		2					1	1	15		2	2	2	1		
S. arenicola	SNA013	51	28	1		2					1	1	17	1	1	2	2			
S. arenicola	SNA015	55	41	2	1	2					1	1	19	1	2	2	2	1	6	1
S. arenicola	SNA017	53	35	3		3					2	2	18	1	1	2	2	1		
S. arenicola	SNA019	61	29	1		1							19	1	2	3	2			
S. arenicola	SNA023	56	35	3		4					1	1	21	1	2	1	1			
S. arenicola	SNA024	57	35	2		2					1	1	20	2	2	2	2	1		
S. arenicola	SNA030	53	33	2		2					1	1	18	1	2	3	2	1		
S. arenicola	SNA032	61	35	3		4					1	1	19	1	2	3	1			
S. arenicola	SNA034	60	33	3		4					1	1	19	1	2	1	1			
S. arenicola	SNA036	64	35	3		4					1	1	19	1	2	3	1			
S. arenicola	SNA039	62	33	3		4					1	1	19	1	2	1	1			
S. arenicola	SNA040	59	34	3		4					1	1	20	1	2	1	1			
S. arenicola	SNA041	62	40	4		4					1	1	23	2	2	2	1			
S. arenicola	SNA045	56	25	1		3					1	1	14		1	2	2			
S. arenicola	SNA046	55	46	2		1					1	1	33	2	1	2	2		1	
S. arenicola	SNA047	53	36	3		3					2	2	18	1	2	2	2	1		
S. arenicola	SNA051	50	34	1		3							22	2	1	2	2	1		
S. arenicola	SNA055	61	46	2		1					1	1	33	2	1	2	2		1	
S. arenicola	SNA056	54	31	2		2					1	1	18	1	1	3	2			
S. arenicola	SNA057	63	35	3		4					1	1	20	1	2	2	1			
S. arenicola	SNA058	49	33	3		4					1	1	19	1	2	1	1			
S. arenicola	SNA059	51	30	2		2					1	1	18	1	1	2	2			
S. arenicola	SNA064	62	39	3		2					2	2	22	2	1	2	2	1		
	total		*3376*	*235*	*4*	*255*	*2*	*2*	*1*	*1*	*110*	*110*	*1943*	*113*	*162*	*194*	*162*	*45*	*33*	*4*

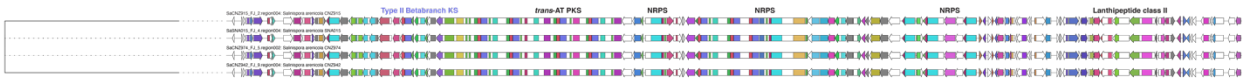


Figure 5.16. BGC alignment of the type II beta-branching containing BGC that is unique to four microscale *Salinispora* strains. Genes contain colored biosynthetic domains.

In addition to the type II beta-branching PKS BGC (**Figure 5.16**), which was a unique BGC to four microscale strains, I also noticed another unique type I *cis*-hybrid polyketide BGC GCF in the same four strains (**Figure 5.17**). These two BGCs could contribute to unique diversity of the four specific microscale strains in comparison to the other microscale strains. In support of that hypothesis, I discovered that these four strains always form their own clade in both 95% and 99% ANI analyses (**Figure 5.8, Figure 5.12**) and they form their own clade in the phylogenetic tree of all *Salinispora* (**Figure 5.13**). These four strains were not isolated from the same sub-quadrant, instead from four different sub-quadrants. It is possible that this four-member microscale clade is distinct from other micro- and macroscale *Salinispora* due to the acquisition of at least these two unique BGCs (**Figure 5.16-5.17**). Further investigation into all microscale BGCs will look for BGCs with similar patterns of rare distribution, which could be evidence of chemical diversification within the genus.

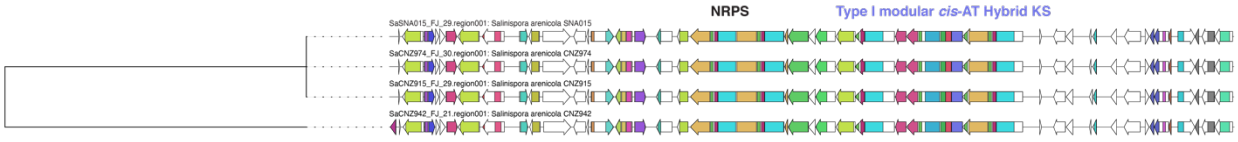


Figure 5.17. BGC alignment of the type I modular *cis*-AT hybrid type BGC that is unique to four microscale *Salinispora* strains. Genes contain colored biosynthetic domains.

Finally, one of the main goals of this project was to look for BGC horizontal transfer events. This could happen between *Salinispora* species, between *Salinispora* and other species, or diversification could be happening at the BGC level instead (the topic of Chapter 6 of this dissertation). However, one aspect of BGC diversification that has not been explored in *Salinispora* is the mechanism by which genes or BGCs could be transferred. To that end, we developed methods to both extract and visualize native plasmids in *Salinispora* (methods described in detail in Chapter 5 Appendix). We extracted plasmids from all 99 microscale *Salinispora* genomes and observed that most seemed to have putative plasmids (**Figure 5.18**). We used pulsed-field gel electrophoresis to characterize and estimate the size of the plasmids. From what we could tell, many of the plasmids were >20kb in length, with others around ~50kb or even > 50kb (**Figure 5.18**). It proved difficult to sequence the plasmids, especially from my attempted methods of using short read sequencing platforms, so future work will be needed to fully confirm the plasmids we saw with conventional gel electrophoresis and PFGE are indeed plasmids. This is further discussed in the Chapter 5 Appendix.

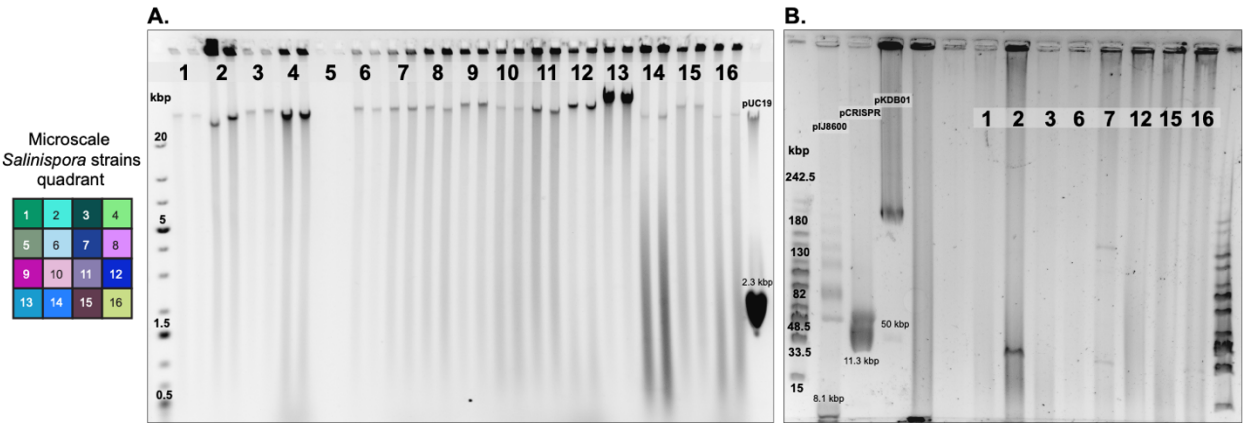


Figure 5.18. Plasmids extracted from microscale *Salinispora* strains numbered by quadrant location.

(A) Conventional gel electrophoresis of microscale *Salinispora* plasmids with a 1kbp ladder and a 2.3kbp plasmid control (pUC19).

(B) PFGE of microscale *Salinispora* plasmids with a midrange PFGE ladder and three plasmid controls: 8.1kbp (pIJ8600), 11.3kbp (pCRISPR-Cas9), and 50kbp (pKDB01).

5.5 Discussion

Previous to the work conducted in this chapter, 118 macroscale *Salinispora* strains isolated from the Bahamas, Yucatán, Puerto Vallarta, the Sea of Cortez, Hawaii, Palmyra, Fiji, Guam, Palau, the Red Sea, and the Madeira Islands were whole-genome sequenced (Millán-Aguiñaga *et al.*, 2017). This collection of macroscale *Salinispora* genomes has not only proven lucrative for novel natural product chemical discovery (Jensen *et al.*, 2015), but has become a model system for evaluating the bacterial speciation and BGC diversification at the global (macroscale) level (Penn *et al.*, 2009; Jensen, 2010; Ziemert *et al.*, 2014; Millán-Aguiñaga *et al.*, 2017). Early speciation analyses using specific 16S rRNA and *gyrB* gene sequencing revealed that *S. arenicola* was globally distributed and that the canonical operational taxonomic unit criteria is likely underestimating global *Salinispora* species diversity (Jensen and Mafnas, 2006). Whole-genome analyses expanded that finding, confirming that the large clade that includes the *S. pacifica* type strain actually includes six additional species (Millán-Aguiñaga *et al.*, 2017; Román-Ponce *et al.*, 2020). Culture-independent experiments selectively targeting *Salinispora* indicate that when physical and chemical nucleic acid lysis methods are used on marine sediment samples, *Salinispora* diversity is the same as culture-dependent methods (Mincer *et al.*, 2005). However, in a recent 16S rRNA sequencing analysis of a Belize marine sediment microbial community, *Salinispora* was identified to be a rare community member (Tuttle *et al.*, 2019). Both culture-dependent and culture-independent biogeographic analyses of *Salinispora* strains have contributed to hypotheses regarding *Salinispora* worldwide macroscale distribution, however, only the culture-based studies provide continued genomic evidence and experimental capacity for answering

questions about *Salinispora* BGC diversity, distribution, and evolution across the ocean's biogeographic expanse.

To that end, the goal of this chapter was for the first time to investigate the "microscale" diversity of *Salinispora* that were selectively cultured from a $1m^2$ quadrant from marine sediment adjacent to a coral reef in Fiji. Over the course of a few months, we isolated 172 new *Salinispora arenicola*, 4 *S. pacifica,* and 28 non-target orange actinomycetes from 16 sub-quadrant sediment samples within the $1m^2$ square quadrant. All *Salinispora* isolates were confirmed to require seawater for growth, a characteristic feature of *Salinispora* (Jensen *et al.*, 2005; Bucarey *et al.*, 2012). The new microscale strains had a wide diversity of morphologies, including pale orange to deep orange coloration, white fluffy spore-parts in addition to the dark brown and back spores, smooth circular, bumpy popcorn-like textures, and surface and subsurface growth on solid culturing media **(Figure 5.1**). A SNP analysis revealed that all strains shared 99% 16S rRNA sequence similarity and only 2 SNPs were discovered between 3 putative *S. pacifica* strains. Upon whole genome sequencing 5-6 strains per the 16 sub-quadrants, we identified 96 *S. arenicola*, two *S. oceanensis,* and one *S. pacifica* strain across the microscale quadrant. Comparative genomic analyses revealed that we captured most of the currently known *S. arenicola* genomic diversity where microscale strains grouped with macroscale *S. arenicola* similarly isolated from Fiji and other locations worldwide rather than other microscale strains isolated from the same sub-quadrant. We report the first evidence of subspecies diversification within *S. arenicola*, as all 217 *Salinispora* strains showed distinct 95-99% ANI groups. One hypothesis to explore would be if sub-species recombination is occurring between specific sub-populations of *S. arenicola*. With the microscale genome dataset, we now have the means to test this. Additionally, the similarity between some microscale *S. arenicola* strains and other *S. arenicola* strains isolated from non-Fiji

locations provides evidence that *S. arenicola* is cosmopolitan in marine sediments and perhaps specific populations are abundant or environments select for specific sub-species or contribute to the intraspecific diversity captured across the species.

A comprehensive BGC analysis of the 118 macroscale *Salinispora* noted that the number of BGCs shared between *Salinispora* did not decrease over increasing geographic distance at both the genus and species level in *Salinispora*, indicating that no biogeographic barriers have prevented BGC flow between *Salinispora* (Letzel *et al.*, 2017). The question of how individual bacterial species are globally distributed and how their diversity differs on a macroscale and microscale continues to fascinate microbiologists. From our 99 new microscale *Salinispora* genomes, we uncovered 3,023 BGCs. All 99 new strains contained the salinipostin BGC (described in detail in Chapter 4 of this dissertation), indicating that all have the potential to produce salinipostin and SAL-GBLs (Kudo *et al.*, 2020; Creamer *et al.*, 2021). To understand if we captured any new biosynthetic potential compared to the macroscale *Salinispora*, we compared BGCs from all 217 strains and discovered that the new microscale genomes both had unique BGCs and GCFs (gene cluster families of BGCs), shared BGCs, and even captured additional rare BGCs. One BGC cluster that was specific to the microscale *Salinispora* strains included 2 domains involved in rifamycin biosynthesis, thus it might be a fragmented of the rifamycin BGC, which has previously been challenging to assemble (Chase *et al.*, 2021). Two GCFs were unique to four of the same *Salinispora* microscale strains, and one of the BGC types included rare type II beta-branching KS domains, which have only been previously seen in 2 out of the 118 macroscale *Salinispora* genomes (the other unique GCF in these four strains was identified by NaPDoS2 as a type I modular *cis*-AT hybrid KS containing polyketide BGC). Plus, 45 of the new microscale *Salinispora* genomes contained PTM-like KS domains, which were previously only observed in a

few *Salinispora* strains. This discovery is exciting because it provides a greater number of *Salinispora* isolates in culture that could be grown to search for PTM molecules (Blodgett *et al.*, 2010; Cao *et al.*, 2010; Qi *et al.*, 2021). One of our microscale *S. oceanensis* strains contained the thiolactomycin BGC, thus increasing the number of strains known to have this BGC from 5 to 6 (Tang *et al.*, 2015). Similarly, the microscale *S. oceanensis* CNZ-966 was discovered to contain the salinosporamide BGC, which is another example of this biomedically relevant and interesting BGC occurring outside of *S. tropica* (Chase *et al.*, 2021; Bauman *et al.*, 2022). A great example of a previously rare BGC that the microscale genome dataset detected is the retimycin BGC, which was previously only observed one *S. arenicola* strain (Duncan *et al.*, 2015; Letzel *et al.*, 2017). We discovered five microscale *S. arenicola* strains with the retimycin BGC, which would be fantastic targets to see if any structural variants of the quinomycin-type depsipeptide retimycin A compound are produced. While it was a risk that all of the microscale *S. arenicola* could have had the exact same BGC biosynthetic potential, instead we find a high amount of across strains, highlighting the importance of studying closely related strains and the diversity they encode. Plus, even if some of the BGCs were already described, many of the BGCs could have gene differences or SNPs that could account for new products and increasing the number of cultured isolates with related BGCs will assist future compound discovery as we know different strains may have different BGC expression patterns. We predict that this could be a powerful way to look for structural variants of known compounds like cyclomarin, rifamycin, and lymphostin if any of the microscale BGCs have gene differences, especially since these strains were isolated years after the original isolates.

In all, the microscale *Salinispora* genome dataset greatly expands our understanding of the diversity within closely related microscale and macroscale *Salinispora* populations, and their

encoding biosynthetic potential diversity. These strains can serve as exciting models to explore if there are additional patterns of new evolutionary chemical innovation on a spatial scale that has not been explored before.

5.6 Acknowledgements

I am grateful to the Republic of Fiji for allowing the marine sediment sample collections and I acknowledge Alyssa Demko and Paul Jensen for collecting the sediment samples with other field team members (Deanna Beatty, Samantha Mascuch, Julia Kubanek, Mark Hay, Samson Viulu and Joape Ginigini). I acknowledge the support of members of the Jensen laboratory for inspiration and helpful feedback on this project, including Natalie Millán-Aguiñaga, Krystle Chavarria, Dulce Guillén-Matus, and Alex B. Chase. I want to thank Robert A. Petit III for troubleshooting bug help with their bactopia workflow; and Jorge C. Navarro-Muñoz for troubleshooting bug-squashing with their BiG-SCAPE workflow; both were quick to assist with posted Github issues.

This research was supported by NIH ICBG Grant U19-TW00740, NIH Grant GM085770, and the National Science Foundation Graduate Research Fellowship under Grant No. DGE-1650112 (to K.E.C). Some of the whole-genome sequencing was carried out at the DNA Technologies and Expression Analysis Cores at the UC Davis Genome Center, supported by NIH Shared Instrumentation Grant 1S10OD010786-01; I acknowledge Vanessa K. Rashbrook for technical assistance. This publication also includes genome-sequencing data generated at the UC San Diego IGM Genomics Center utilizing an Illumina NovaSeq 6000 that was purchased with funding from a National Institutes of Health SIG grant (#S10 OD026929). In addition, the

NovaSeq sequencing data was supported by a mini-grant awarded to A.M.D. and K.E.C. from Illumina, Inc (San Diego, CA); I acknowledge both Christina M. Czerwinski (Illumina) and Kristen Jepsen (UCSD IGM) for the opportunity and technical assistance. I acknowledge the Triton Shared Computing Cluster (University of California, San Diego) hosted at the San Diego Supercomputer Center (doi: 10.57873/T34W2R) for hosting our computational node that was instrumental to this work, and for technical assistance for the large analyses.

Chapter 5 is coauthored with Victoria Vasilat, David Vereau-Gorbitz, Alyssa M. Demko, and Paul R. Jensen. The dissertation author was the primary investigator and author of this chapter.

5.7 Chapter 5 Appendix. Characterization of the macro- and microscale *Salinispora* plasmid "mobilome"

5.7.1 Abstract

Previous work in the genus *Streptomyces* has shown that conjugative double-stranded DNA transfer is mediated by actinobacterial plasmids. While there have been over 118 *Salinispora* genomes sequenced, there has been no investigation into whether *Salinispora* harbor native plasmids. In this chapter appendix, we report methods to successfully extract and visualize plasmid DNA from *Salinispora* strains. We apply this technique to all 99 new microscale *Salinispora* genome sequences (in addition to many of the original 118 macroscale *Salinispora* strains during development of the protocols) and find that many strains harbor different sized plasmids. Attempts to sequence the plasmids proved to be challenging, and we suggest that future work with long-reads or Hi-C sequencing approaches will be needed to fully characterize these *Salinispora* plasmids. The characterization of these novel *Salinispora* plasmids will facilitate future investigation to see if the plasmids harbor BGCs that are being exchanged, in addition to proving useful for molecular genetic experiments.

5.7.2 Introduction

Horizontal gene transfer of DNA between bacteria can occur in four ways: 1) conjugation: which requires cell to cell contact via cell surface pili or adhesins through which DNA is transferred from the donor cell to the recipient cell; 2) transformation: the uptake, integration, and functional expression of naked fragments of extracellular DNA; 3) transduction: which requires specialized or generalized bacteriophages that transfer bacterial DNA from a previously infected

donor cell to the recipient cell; and 4) Gene transfer agents (GTAs): bacteriophage-like particles that carry random pieces of the producing cell's genome where the GTA particles are released through cell lysis and spread to a recipient cell (Von Wintersdorff *et al.*, 2016). A previous study of the macroscale *Salinispora* CRISPR-Cas systems identified 24 strains that contained evidence of the common *Streptomyces* SV1 (Stuttard, 1983) prophage, with 6% of the *Salinispora* spacers matching SV1-related sequences indicating that the phage preys on *Salinispora* (Wietz *et al.*, 2014). While no *Salinispora* specific phage has been identified to date, it is possible that transduction is an active form of DNA acquisition in *Salinispora*. I searched all 118 macroscale *Salinispora* genomes for any evidence of transformation competence related genes (Johnsborg *et al.*, 2007; Mell and Redfield, 2014) and found none, indicating that *Salinispora* is likely not naturally competent. This indicates that *Salinispora* would need to be in a recently described *Streptomyces* wall-less state (Ramijan *et al.*, 2018) for transformation of extracellular DNA to take place.

Conjugation is one of the most widespread mechanisms of HGT in Actinomycetia (formerly Actinobacteria) where DNA is translocated across two bacterial cell membranes mediated by two types of mobile genetic elements: conjugative plasmids and actinomycete integrative and conjugative elements (AICES) (Evelien and Henk, 2008; Ghinet *et al.*, 2011; Bordeleau *et al.*, 2012). Many AICE (integrative conjugative plasmid) sequence regions have been identified in *Salinispora*, however, as AICES can excise from the chromosome to form circular covalently closed molecules, they can only disseminate via conjugation (Ghinet *et al.*, 2011). The order *Actinomycetales* has unique conjugative machinery composed to the essential FtsK-homolog protein TraB, which is reminiscent of the machinery that allows segregation of chromosomal DNA during bacterial cell division and sporulation (Bordeleau *et al.*, 2012). TraB is a single plasmid-

encoded protein that directs the transfer from plasmid-carrying donor to the recipient Actinomycete. Recent fluorescent microscopy studies have shown that inter-mycelial plasmid spreading in *Streptomyces* requires both the hexameric pore TraB and a Spd protein complex (Reuther *et al.*, 2006; Thoma *et al.*, 2015, 2016; Thoma and Muth, 2016). I have been unable to find any TraB homologs or functional conserved TraB domains in the 118 macroscale *Salinispora*. Interestingly, experimental conjugation with *Salinispora* is only successful when the conjugating donor strain contains the "helper" plasmid pUZ8002 which has a *Streptomyces traB*-like transfer machinery on it, as alluded to in (Eustáquio *et al.*, 2009; Bucarey *et al.*, 2012; Zhang *et al.*, 2018) and observed in my own experimental tests. Thus, perhaps like *Streptomyces,* which have a natural plasmid-encoded *traB*, it is possible that the *Salinispora traB* is also plasmid-encoded.

Plasmids are self-replicating genetic elements capable of mobilization between different hosts and are widely recognized as mediators of HGT events that contribute to evolutionary patterns in microbial populations (Hü *et al.*, 2017). Plasmids typically carry conserved functions of DNA replication and mobilization in addition to accessory genes that contribute to their host's phenotypic diversity. These plasmid-encoded characteristics include virulence factors, resistance to antibiotics, production of antimicrobials, degradation of xenobiotics, and functions necessary for bacteria-host interactions (Dib *et al.*, 2015). Plasmids in individual Actinomycetia (formerly Actinobacteria) are reported, many with important accessory genes like BGCs. The report of a linear SCP1 (365kbp) plasmid in *Streptomyces coelicolor* A(3)2 that contains the signaling molecule methylenomycin BGC (*S. coelicolor* also contains a circular SCP2 (31kbp) plasmid) was one of the first surprising actinobacterial plasmid discoveries (Kinashi *et al.*, 1987; Corre *et al.*, 2008). More recently, metagenomic studies have uncovered further evidence that plasmids are able to harbor important, functional BGCs in the case of the NRPS renieramycin 33kbp BGC found in

a sponge bacterial symbiont; in this case, the BGC was the entire plasmid (Tianero *et al.*, 2019). In 2012, a study coined the term "plasmidome" and defined it as the overall plasmid population of a sample (Kav *et al.*, 2012). While metagenomic approaches to study the plasmidome or the synonymous "mobilome" (total amount of mobilizable genetic elements in a population) can reveal complex plasmid communities and functional adaptations in microbial communities in the bovine rumen (Kav *et al.*, 2012) and ground water environment (Kothari *et al.*, 2019), culture-dependent approaches can allow targeted characterization of important clinical and environmental bacterial plasmids.

5.7.3 Methods

Culturing for plasmid extractions: *Salinispora* and other plasmid containing cultures (*Escherichia coli*) for controls were cultured as previously described in Chapter 5. Briefly, *Salinispora* strains were cultured in A1 75% seawater liquid media (10 g/L soluble starch (Affymetrix), 2 g/L peptone (Fischer Scientific), 4 g/L yeast extract (Fischer Scientific), 22 g/L instant ocean mix (Marineland) in 1L DI water) at 28°C at 230rpm until they reached exponential cell density. *E. coli* cultures were grown in LB media with appropriate antibiotics for stability of the plasmids for 16-18h at 37°C at 200rpm. Plasmid DNA was isolated from *Salinispora* as described below. Control plasmids of known sizes were purified from *E. coli* cultures using the QIAprep spin Miniprep Kit (Qiagen) and the ZymoPURE EndoZero Plasmid Midiprep (Zymo Research).

Genomic DNA from *Salinispora* was extracted as follows (Millán-Aguiñaga *et al.*, 2017): The 50mL *Salinispora* cell pellet was re-suspended in TE Buffer (pH 8.0, 10mM Tris-HCl, 1.0mM

EDTA; Teknova, Inc.) to an OD_{600} of 1.0; lysozyme (100mg/mL; end concentration 3mg/mL; Sigma-Aldrich) and RNAse A (100mg/mL; end concentration 100ug/mL; Qiagen) was added and the mixture was incubated at least 80 min at 37°C. After incubation, 10% SDS (Sigma-Aldrich) and Proteinase K (20mg/mL; Qiagen) was added and the entire mixture was incubated overnight at 55°C. The next day, 1.5 mL of 5M NaCl (Ambion) and 1 mL 10% CTAB/NaCl (Sigma-Aldrich) solution was added and incubated for 10 min at 65°C followed by an ice bath for 10 min. Chemical DNA extraction was performed by adding 4mL phenol/chloroform/isoamyl alcohol (25:24:1 ratio saturated with 10 mM Tris, pH 8.0, 1 mM EDTA; Sigma-Aldrich), centrifuging for 10 min at 10,000 rpm, 4°C, followed by transferring the upper aqueous layer to 4 mL chloroform (Sigma-Aldrich). The samples were centrifuged for 10 min at 10,000 rpm, 4°C and the upper aqueous layer was transferred using wide-bore tips to a new tube; DNA was precipitated by adding 0.6 volumes of molecular grade isopropanol (Sigma-Aldrich), followed by centrifugation for 10 min at 10,000 rpm, 4°C. The precipitated DNA was rinsed with ice-cold 70% molecular grade EtOH (Sigma-Aldrich), and centrifuged for 10,000 rpm for 2 min, after which the EtOH was removed and samples air dried; the genomic DNA was dissolved overnight in TE buffer at 4°C. The gDNA purity was confirmed by measuring the 260 nm/280 nm and 260 nm/230 nm absorbance ratios using a NanoDrop 1000 spectrophotometer (Thermo Fisher Scientific), and the concentration of the DNA was measured using both the NanoDrop 1000 spectrophotometer (Thermo Fisher Scientific) and a Qubit 3.0 fluorometer (Thermo Fisher Scientific), according to the manufacturer's instructions. The gDNA molecular weight quality was analyzed by running 60 ng of the gDNA on a 1% agarose gel for 90 minutes at 90V alongside a GeneRuler 1kb Plus DNA ladder (Thermo Fisher Scientific).

5.7.4 *Salinispora* plasmid extraction protocol

This protocol was adapted from a plasmid extraction protocol from Dr. Henrique Machado.

A. Buffer preparation:

Buffer P1 (store in the fridge, 4°C)

Tris – 0.606g

Na_2EDTA x $2H_2O$ – 0.37g

RNAse A (10mg/mL) – 1g (200µL of 100mg/mL Qiagen)

d H_2O – up to 100mL

- Adjust to pH 8.0 (by adding 0.1M HCl); filter sterilize with 0.2uM filter.

Buffer P2

SDS (10% w/v) sterile in H_2O – 10mL

NaOH – 1.17g

dH_2O – up to 100mL

Buffer P3

Potassium acetate – 58.88g

dH_2O – up to 200mL

- Adjust pH to 5.5 by adding acetic acid

B. Plasmid extraction and purification:

1. Collect 2mL *Salinispora* culture in a 2mL microcentrifuge tube (best yield was 2mL of mid-dense culture).
2. Centrifuge for 10min, 5000 rpm, at 4°C; discard the supernatant.
3. Add 300µL P1 Buffer; vortex.

4. Add 300µL P2 Buffer; incubate 4 minutes at room temperature; shake gently by inverting 5 times in the tube rack.
5. Add 300µL P3 Buffer; incubate 5 minutes at room temperature; shake gently by inverting 5 times in the tube rack.
6. Centrifuge for 20min, 13,000 rpm, at 4°C.
7. Place supernatant in a new 2mL microcentrifuge tube.
8. Repeat steps 6 and 7 by spinning the supernatant and transferring again—you do not want any residual cell material. I recommend the use of wide bore tips and slow pipetting to not shear plasmid DNA.
9. Add 640µL isopropanol (0.7 volume of the supernatant); shake gently.
10. Centrifuge for 20min, 13,000 rpm, at 4°C; discard the supernatant (you can carefully pour it off as long as the pellet is not dislodged).
11. Add 550µL ice cold 70% Ethanol.
12. Centrifuge for 10min, 13,000 rpm, at 4°C; discard the supernatant using a pipette tip; the pellet is your plasmid DNA.
13. Dry the pellet for 12 minutes at 37°C in an incubator with the tube caps open.
14. Add 50-100µL EB buffer to resuspend, let sit overnight at 4°C or for at least 20minutes.
15. Store at -20°C, and for long-term storage, -80°C.

C. Quantify plasmid amounts

The plasmid purity should be confirmed by measuring the 260 nm/280 nm and 260 nm/230 nm absorbance ratios using a NanoDrop 1000 spectrophotometer (Thermo Fisher Scientific).

The concentration of the plasmid DNA should be measured using both the NanoDrop 1000 spectrophotometer (Thermo Fisher Scientific) and a Qubit 3.0 fluorometer (Thermo Fisher Scientific), according to the manufacturer's instructions. We recommend measuring both the Qubit dsDNA BR and Qubit RNA BR assays, as it is likely the NanoDrop measurements will not match the Qubit dsDNA measurements, and thus an additional RNAse A cleanup and re-purification step might be needed.

D. RNAse A treatment & plasmid re-purification

1. Add 2.5µL RNAse A (Qiagen 100mg/mL) per 100µL of plasmid. Incubate at least 1 hour at 37°C.
2. Add 5x volume Buffer PB (Qiagen, or homemade 5M guanidine hydrochloride (Gu-HCl), 30% isopropanol) to the plasmid sample and mix.
3. Apply mixture to a silica-binding column (Epoch life sciences, Qiagen QIAquick, or similar).
4. Spin at max RPM for 60 seconds. Discard flow through. Add 750µL Buffer PE (Qiagen, or diluted homemade 5X (80mM NaCL, 8mM Tirs-HCl, pH 7.5) diluted to 1X with Ethanol (final volume 80%).
5. Repeat spin, discard flow through, and repeat wash with PE buffer. Replace column in collection tube and do one final spin for 60seconds.
6. Elute plasmid DNA by adding 70°C Buffer EB (Qiagen, homemade 10mM Tirs-HCl, pH 8.0 or ddH_2O), letting sit for at least 5 minutes, followed by a final spin for 60 seconds.

5.7.5 Conventional gel electrophoresis visualization of *Salinispora* plasmids

Plasmid DNA samples should be fully dissolved in EB buffer or similar. Prepare a TAE gel with 0.5% agarose in a large gel electrophoresis rig (long). To prepare samples, mix 7µL of plasmid DNA with 7µL Sybr Green I (diluted to 1:100 in water; Thermo Fisher Scientific), and 2.3µL 6X loading dye (NEB blue or purple). Pre-chill 1X TAE buffer at 4°C; use to fill gel box and load samples into gel wells. Once ready, the gel should be run at 34V for 6 hours and 5 minutes in the dark with ice packs covering the top, side, and underneath the gel rig (these should be changed if they melt too quickly). Paper towels should be used to prevent water from pooling near electrical wires. Visualize using a gel imager and UV light. We recommend the use of the GeneRuler 1kb Plus DNA ladder (Thermo Fisher Scientific) on either side of the samples on the gel. Our samples were also compared to a control plasmid pUC19 sample (2.3kb in size) and gDNA sample.

5.7.6 PFGE visualization of *Salinispora* plasmids

Below, we describe the PFGE (pulsed-field gel electrophoresis) methods to visualize *Salinispora* plasmids. PFGE is a technique used to separate larger sizes of DNA molecules that can't be accurately sized by conventional gel electrophoresis. PFGE is different than conventional gel electrophoresis because the electric pulses come from multiple angles around the gel, not just north/south poles, and low voltages are run for an extended period with switching electric patterns to create mobility and separation of nucleic acids.

These protocols were written and tested by David Vereau-Gorbitz; they were adapted from the BioRad CHEF-DR III Pulsed Field Electrophoresis Systems instruction manual and applications guide.

A. Prepare TBE (Tris-Borate-EDTA) buffer

TBE is a buffer used in pulsed-field electrophoresis for its greater buffering capacity compared to TAE (Tris-acetate EDTA), for longer electrophoresis runs and for resolving specific sizes of nucleic acids. DNA fragment run faster in TAE than in TBE due to the lower ionic strength of TAE, however this means that TAE cannot be run/used more than 24-36 hours whereas TBE can be run for several days without a change in the mobility of the DNA fragments.

Recipe for 1L of 10X TBE:

7.44 g Disodium EDTA (final concentration 0.02 M)

55 g Boric Acid (final concentration 0.9 M)

108 g Tris Base (final concentration 0.9 M)

Diluted in Milli-Q water

TBE Buffer Procedure

1. Weigh 7.44 g of EDTA salt and transfer to a 100 mL beaker.
2. Measure 35 ml of Milli-Q water and add to the beaker, stir without heating.
3. Adjust the pH to 8.0 by adding NaOH, first add 7-8 pellets of solid NaOH and when the pH gets close to the desired measurement change to adding drops of 1.0 M NaOH solution. The EDTA salt will not dissolve until pH for the solution is basic.
4. Adjust the volume of the solution to 40 mL.
5. Add 800 mL Milli-Q water to an 1800 mL beaker.
6. Add the 40 mL EDTA solution to 800 mL Milli-Q water.

7. Measure 108 g of Tris Base and add to the beaker, stir until dissolved.
8. Measure 55 g of Boric acid and add to the beaker, stir until dissolved.
9. Adjust the volume of the solution to 1 L with Milli-Q water, store in a 1L glass screw-cap bottle at room temp.

To make a working stock of TBE (1x) for PFGE or conventional gel electrophoresis, measure 100mL of 10x buffer in a 1L measuring cylinder, and add Milli-Q water to a total volume of 1 L.

B. Procedure for a PFGE run

Equipment: BioRad CHEF-DR III Pulsed field Electrophoresis Systems; attached to a BioRad Cooling Module.

1. Turn a water bath on to 55°C.
2. Attach all the cords to the PFGE power module and cooling module as seen in the instruction manual setup.
    1. Before starting a run: make sure the pump is on and the flow rate is correct. If you haven't used the PFGE rig in the past month, it is recommended to turn the system and pump on to make sure all the tubes are pumping properly and nothing is clogged. Also, check to make sure that the PFGE cooling module is operational; it should cool to 14°C from room temperature in around 15-20 minutes while the flow system pumps at a 0.75 L/min rate (a setting of 70-80 on the PFGE variable speed pump).

3. Prepare your gel by dissolving 1% (g/mL) agarose into 0.5x TBE (must be SeaKem Gold Agarose (Lonza), BioRad Certified Megabase Agarose, or similar). The casting stand on the rig is for a 100 mL gel, and you need to make a little bit more gel to seal the ladders. Measure 1.1 g of agarose into a 250 mL Erlenmeyer flask. Measure and pour 110 mL of TBE buffer.
4. Agarose will not dissolve at room temperature; heat the solution in the lab microwave until it starts to boil then stop the microwave and carefully remove it with heat resistant gloves and stir vigorously. Rinse and repeat until all of the agarose has dissolved.
5. Once the agarose has dissolved completely, transfer it to the water bath for 15 minutes. Once the solution has reached 55°C, pour 100 mL into the gel casting stand on a levelled surface, with the well comb. Return the rest of the solution to the water bath.
6. The gel will take at least 1 hour to completely set so plan accordingly, you can let the gel solidify for 45 minutes and then carefully transfer it to a level surface on the fridge for the last 15 minutes.
7. Once the gel has set, carefully remove the comb and load your samples. For agarose plug samples, after loading the sample, seal the well with molten agarose from water bath and wait for it to solidify. The most important thing is to watch out for and avoid air bubbles inside the wells. For liquid samples, see step 10.
8. Turn on the PFGE power module, fill the electrophoresis chamber with 2 liters of 1X TBE buffer and turn on the pump and cooling module on.
9. Carefully lower the gel into the electrophoresis rig, if the gel floats it's because of the temperature difference, chill the gel in the fridge for 5 minutes. Make sure to fix the gel in place to ensure that there is no air trapped beneath it.

10. Once the gel is stable in the rig and the water is flowing & chilling, load the liquid samples that have been mixed with loading dye.
11. Turn to the power module, press BLOCK, the left display should show the number “1” to indicate you are changing parameters for block 1. The CHEF-DR III system can run 3 independent blocks consecutively, should you need more than one block.
12. Enter the parameters for initial and final switch time, run time, Volts/cm, and included angle for block 1.
13. Press BLOCK again to view the parameters for blocks 2 and 3, a run time of “0” will disable a block from being run.
14. When you are ready to begin, press the PAUSE/START RUN button, once the program is running, you cannot edit any of the run parameters. To manually end a run, hold down the PAUSE/START RUN button for 4 seconds and wait until you hear two beeps.
15. During the run, condensation will occur along the chilled tygon tubing; in order to avoid accidents, set paper towels around the tubing where condensation may cause accumulation of water. Check to ensure no water will touch electrical wires or outlets.
16. Remember to turn off the water bath, and thoroughly wash any equipment you used for the preparation of the gel. TBE is a hazardous waste; dispose of it in the correct liquid and solid waste containers according to EH&S guidelines.
17. Once the run has started, touching the tubing, changing the pump settings, or even worrying too much in the vicinity of the electrophoresis rig will raise the temperature of the buffer. This can alter the conditions of the run, so it is advised against.
18. Once a run has ended, turn everything off, carefully remove the gel and transfer it to the staining station, if all your samples are over 10kb in TBE buffer, you can afford to leave

the sample idle for a few hours as the samples are too heavy to diffuse into the buffer from the gel.

19. The Jensen laboratory is equipped with SYBR green I staining solution, to prepare the staining solution dilute SYBR green I in 0.5x TBE buffer in 1:10000.
20. Transfer the gel into the container with staining solution for 30 minutes. Remember to shake occasionally. SYBR green I does not require a post stain wash, so you can transfer your gel to the imaging station right away.

Other equipment:

Agarose (SeaKem Gold agarose, Lonza; Certified Megabase agarose, BioRad; or similar)

Ladder:

15-291 kb ladder (NEB midrange marker)

48.5-1000 kb ladder (NEB lambda ladder)

225-1100 kb ladder (2 yeast chromosome markers from NEB)

5.7.7 Results

To date, there have been no native *Salinispora* plasmids described. However, during the large-volume genomic DNA extractions and subsequent gel electrophoresis visualization of the 118 macroscale *Salinispora*, four *Salinispora* strains showed evidence of harboring plasmids (Dr. Natalie Millán-Aguiñaga, unpublished data). The goal of this chapter 5 appendix in tandem with the genome-sequencing of 99 new microscale *Salinispora* strains was to determine if any *Salinispora* have native plasmids and if so, attempt to characterize and compare the macroscale and microscale *Salinispora* mobilome. We optimized a plasmid extraction protocol that uses

alkaline lysis and plasmid precipitation that achieves high yields with test plasmids in *E. coli* and *Salinispora* (described in the methods above). I performed parallel large volume gDNA extractions (the same that (Millán-Aguiñaga *et al.*, 2017) used for all macroscale *Salinispora*) and plasmid extractions on three *Salinispora* strains in which plasmids were previously observed by Dr. Natalie Millán-Aguiñaga. When visualized by low voltage gel electrophoresis, I confirmed that these three strains have plasmids that were faintly seen in the gDNA preps but clearly present in the plasmid extractions (data not shown). I extracted an additional 13 macroscale strains and 16 microscale *Salinispora* strains (one from each sub-quadrant, **Figure 5.19**) and discovered that most *Salinispora*, but not all, have evidence of plasmids (**Figure 5.18**). I expanded this survey to all 99 newly isolated microscale *Salinispora* and again observed that most had evidence of plasmids (**Figure 5.20-5.25**). Some of the strains had very low amounts of extracted plasmid DNA and others had high amounts of contaminating co-purified RNA (low molecular weight black blobs, for example in **Figure 5.21)**, which can affect sequencing downstream and plasmid quantification estimates when only using the Nanodrop spectrophotometer. To this aim, I developed an RNA-cleanup protocol that removes significant amounts, but not all, of the RNA in the plasmid preparations (data not shown; protocol described in methods above). This RNA removal method could be combined with downstream DNA cleaning and concentration protocols.

In order to determine if plasmids were present after extraction, I optimized a 13-16-hour low voltage gel electrophoresis protocol which showed the separation of plasmids into multiple bands (Wang and Lai, 1995). Plasmids are expected to form multiple bands on a gel as the preparations may include nicked open-circular plasmid DNA (with one strand cut), linear plasmid DNA (with both strands cut at one site), supercoiled/covalently closed-circular plasmid DNA (where the DNA is fully intact with both strands uncut), supercoiled denatured DNA (which can

occur after excessive alkaline lysis where both strands are uncut but are not correctly paired resulted in a compacted plasmid form), and relaxed circular DNA (fully intact and relaxed with no supercoils) conformations that will each separate differently due to their varying degrees of mobility. Many of the microscale *Salinispora* plasmids were larger than the 2.3kbp control pUC19 plasmid, including plasmids from quadrants 2, 4, 5, 8, 10, and 11 in **Figure 5.20** which each had the signature 4 bands of plasmids of different conformations. **Figure 5.21** shows an example of a larger plasmids ~20kbp or larger, from quadrant 16 with other samples with a decent amount of RNA co-contamination as dark bands at the bottom of the gel. In **Figure 5.22**, two plasmid samples from quadrants 6 and 9 each showed the typical 4 plasmid banding pattern of plasmids estimated to be at least >2.3kbp in size. In contrast, plasmid samples from quadrant 1 in **Figure 5.23** appeared to be smaller (2.3-5kbp) in size. In **Figure 5.24**, we observed a plasmid with a very high molecular weight over 20kbp with two bands at the top of the gel >20kbp from quadrant 5; and one other plasmid sample from quadrant 12 that seemed to have a size at least larger than the 2.3kbp plasmid control. Finally, **Figure 5.25** showed two plasmid bands high on the gel indicating sizes 20kbp or larger in a sample from quadrant 16. Overall, we observed plasmid banding patterns using conventional gel electrophoresis in ~15/99 samples, however there could be larger plasmids that could not be visualized and separated using this method. To better determine the size the extracted plasmids, I turned to pulsed-field electrophoresis (PFGE).

PFGE is a useful technique for resolving super-coiled and circular chromosome-sized DNAs via an alternating low-voltage electric field between spatially distinct pairs of electrodes that force small to large sized DNA to reorient and move at different speeds during extended period in a chilled buffer. David Vereau-Gorbitz, an undergraduate student under my mentorship, optimized two PFGE runs for a mid-range ladder (15kbp to 250kbp) and smaller-size ladder (8kbp-

50kbp). Upon running 8 of the plasmid extracts from the 16 microscale *Salinispora*, we observed multiple plasmid conformations in a couple of the samples with one plasmid around 50kbp in size and another ~60kbp or larger (**Figure 5.18**). The PFGE method can make it difficult to interpret the size of the plasmids due to multiple conformations, and thus we attempted sequencing the plasmids to fully size and assemble the plasmid contents.

In a pilot sequencing run, I combined nine strains of *Salinispora* plasmid DNA in a "mock" plasmidome for short-read MiSeq Illumina sequencing; however, efforts so far to assemble any plasmids from the sequencing data using the plasmid assembly tool PLACNETw (Vielva *et al.*, 2017) have not been successful. While Illumina sequencing can help provide accuracy and sequencing depth, it has been discussed in the literature that it is very difficult to assemble plasmids from short-read sequencing due to frequent repeat sequences, and in our case, the possibility that there might be BGCs with repeat regions could further complicate assembly (Arredondo-Alonso *et al.*, 2017). There are other tools that could be used to assemble and annotate the *Salinispora* plasmids including PLACNETw (Vielva *et al.*, 2017), plasmidSPAdes (Bankevich *et al.*, 2012) (now part of SPAdes), Recycler (Rozov *et al.*, 2017), Circlator (Hunt *et al.*, 2015), MIplasmids (Arredondo-Alonso *et al.*, 2018), PlaScope (Royer *et al.*, 2019), PLASmapper (Dong *et al.*, 2004), canu (Koren *et al.*, 2017), among others that have been developed specifically for MinION sequencing (George *et al.*, 2018) and the assembly of plasmids from whole genome assembly data. In a trial long-read sequencing attempt, 5 plasmid samples with distinctive 4 banding patterns or high bands from conventional gel electrophoresis were sent to Primordium Labs (Arcadia, CA) which specializes in sequencing plasmids. They were able to successfully sequence one of the five samples (**Figure 5.26**), resulting in a circular 9,468bp sequence with 4 predicted coding sequences as assigned by pLannotate (McGuffie and Barrick, 2021). This specific microscale *Salinispora*

plasmid is from quadrant 10, lane number 10 in gel **Figure 5.20** with the plasmid-distinctive 4 banding pattern. In communication with the company, they indicated that challenges to sequencing these native plasmids include the high amount of co-extracted RNA and thus lower concentrations of the plasmid DNA, along with many DNA fragments which could be from the plasmid or contaminating gDNA. The coding sequences predicted on the plasmid did not seem to be typical plasmid genes; however, when the plasmid was BLASTn against the microscale collection of *Salinispora* genomes, there were multiple hits <2.5kbp alignments and less to the genomes (including the strain the plasmid was isolated from), which could indicate this this plasmid is a small extrachromosomal element with a couple of chromosome-encoded genes, or perhaps part of a small fragment of gDNA (**Figure 5.26**). Future efforts to sequence *Salinispora* plasmids should use Nanopore MinION long-read sequencing or Hi-C methods to sequence the macroscale and microscale *Salinispora* plasmids as both technologies has been successfully applied in the sequencing of plasmids (Lemon *et al.*, 2017; Li *et al.*, 2018; Taylor *et al.*, 2019; Bickhart *et al.*, 2022), and careful cleanup and concentration o the samples will hopefully enable better sequencing results.

Figure 5.19. Sub-quadrant isolation location color and number key.

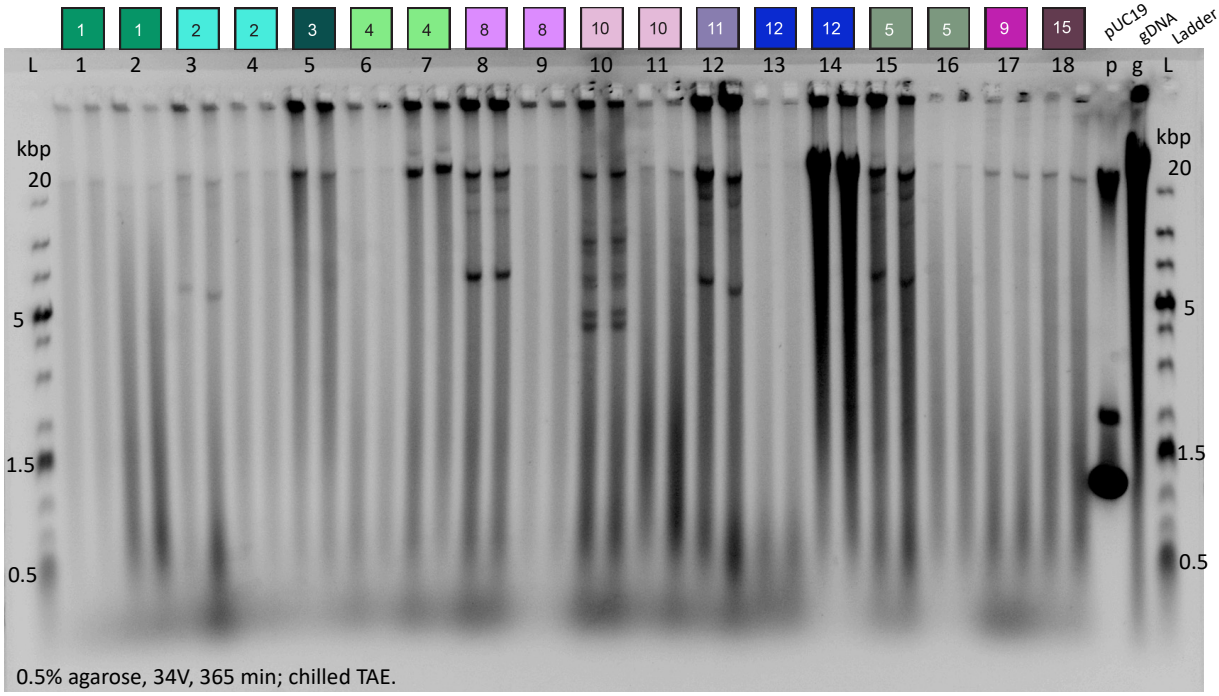


Figure 5.20. Conventional gel electrophoresis of microscale *Salinispora* plasmids (n=2 extractions per strain) numbered by quadrant location. Gel includes 1kbp ladder (L); a pUC19 2.3kbp plasmid and *Salinispora* gDNA for comparative controls.

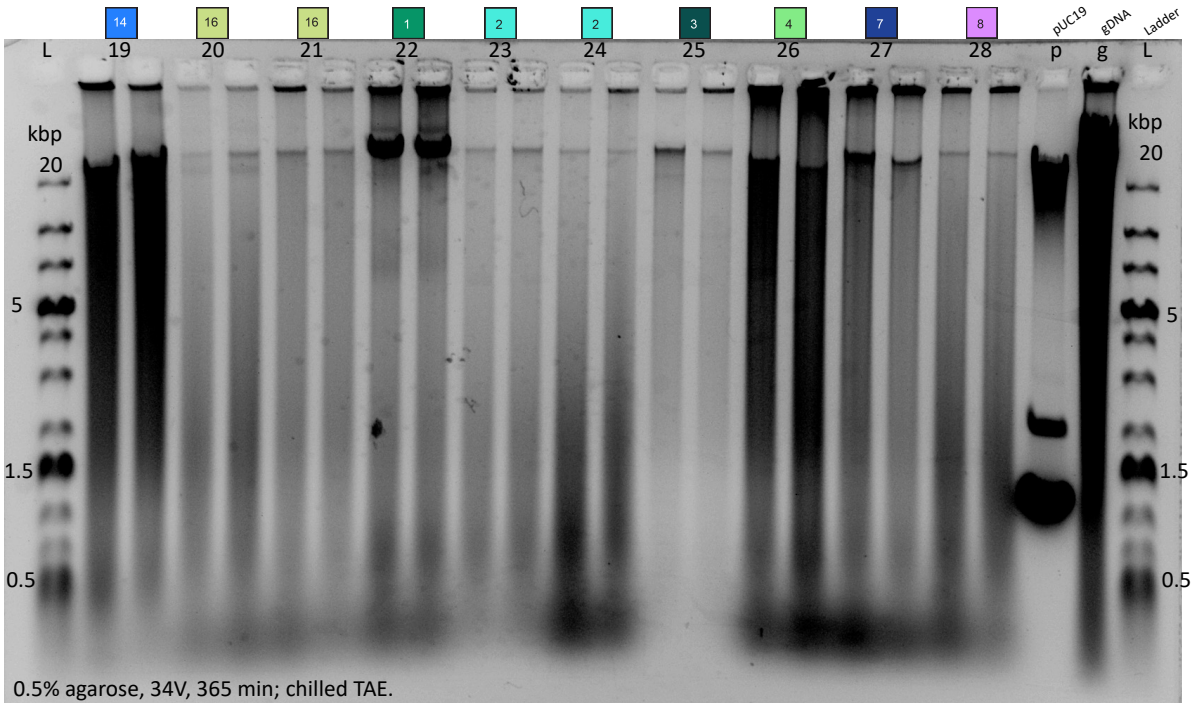


Figure 5.21. Conventional gel electrophoresis of microscale Salinispora plasmids (n=2 extractions per strain) numbered by quadrant location. Gel includes 1kbp ladder (L); a pUC19 2.3kbp plasmid and Salinispora gDNA for comparative controls.

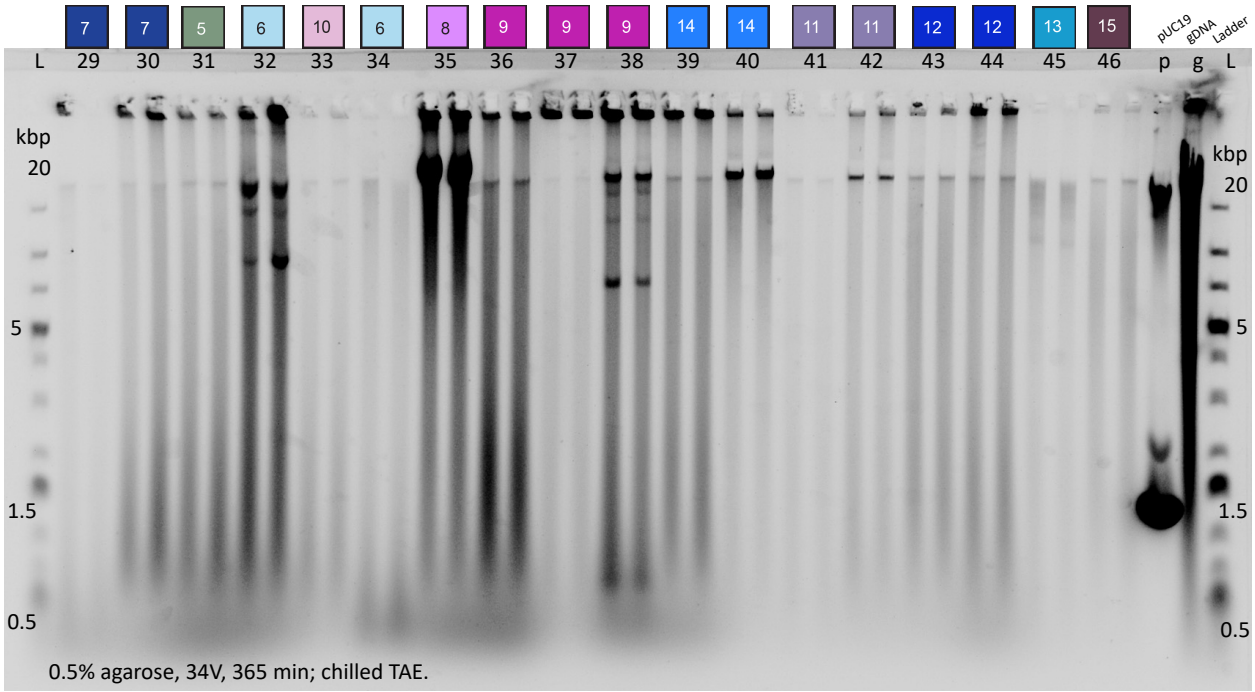


Figure 5.22. Conventional gel electrophoresis of microscale *Salinispora* plasmids (n=2 extractions per strain) numbered by quadrant location. Gel includes 1kbp ladder (L); a pUC19 2.3kbp plasmid and *Salinispora* gDNA for comparative controls.

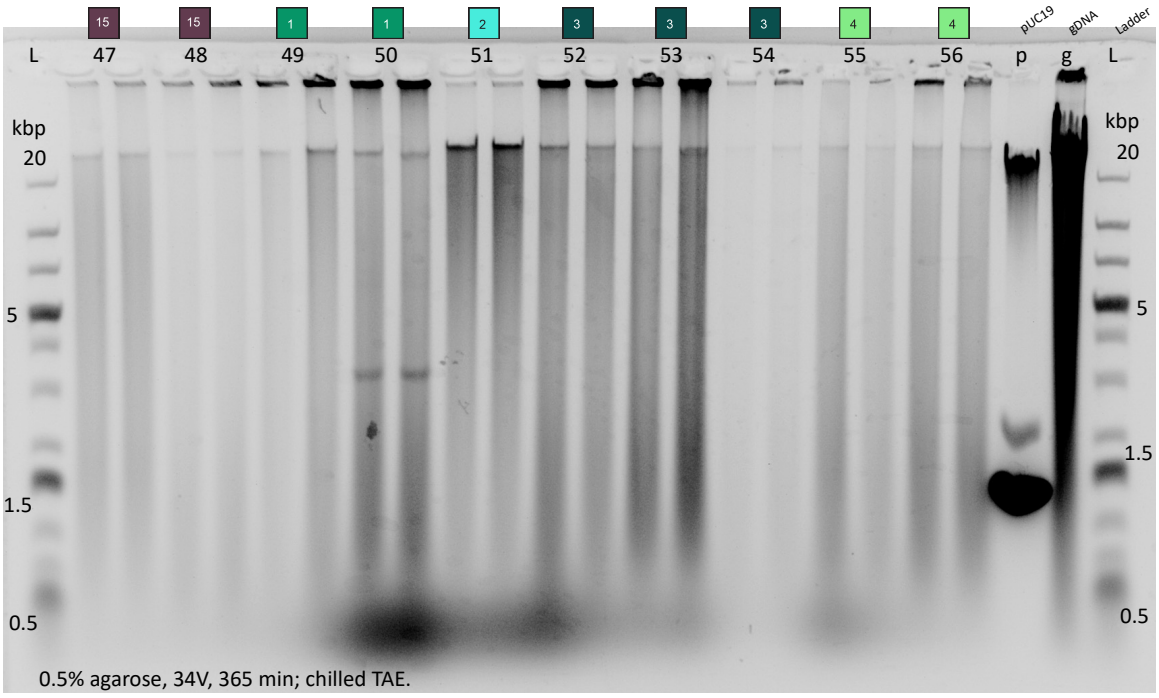


Figure 5.23. Conventional gel electrophoresis of microscale *Salinispora* plasmids (n=2 extractions per strain) numbered by quadrant location. Gel includes 1kbp ladder (L); a pUC19 2.3kbp plasmid and *Salinispora* gDNA for comparative controls.

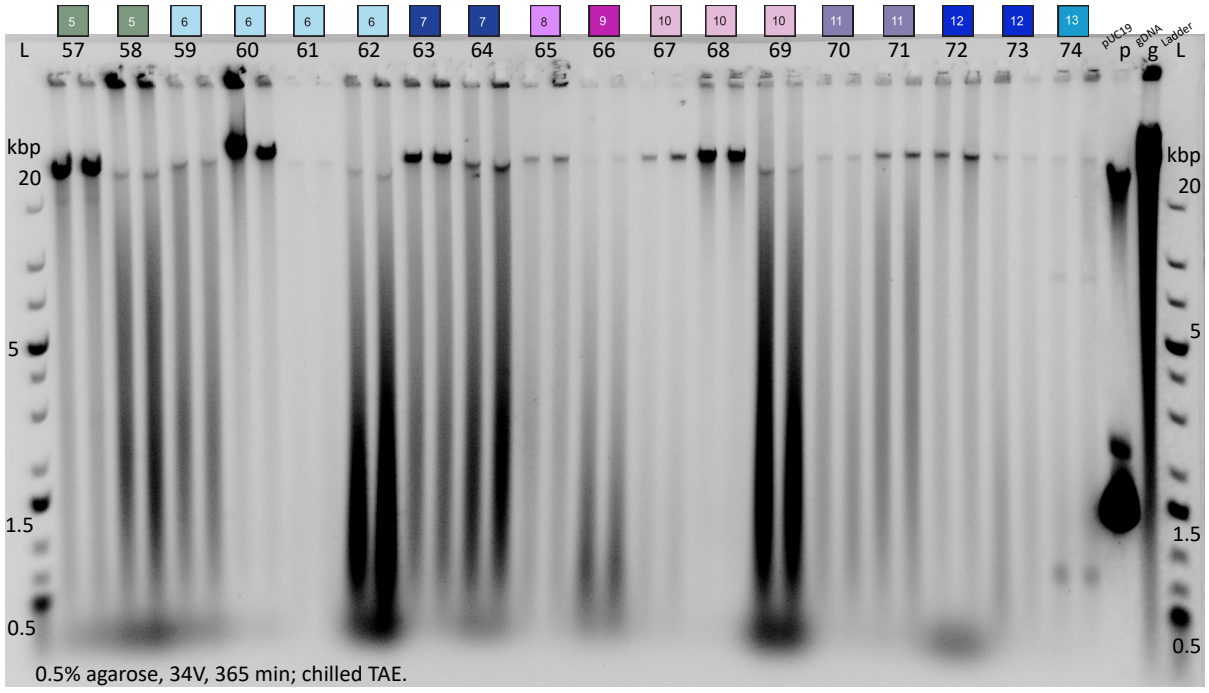


Figure 5.24. Conventional gel electrophoresis of microscale *Salinispora* plasmids (n=2 extractions per strain) numbered by quadrant location. Gel includes 1kbp ladder (L); a pUC19 2.3kbp plasmid and *Salinispora* gDNA for comparative controls.

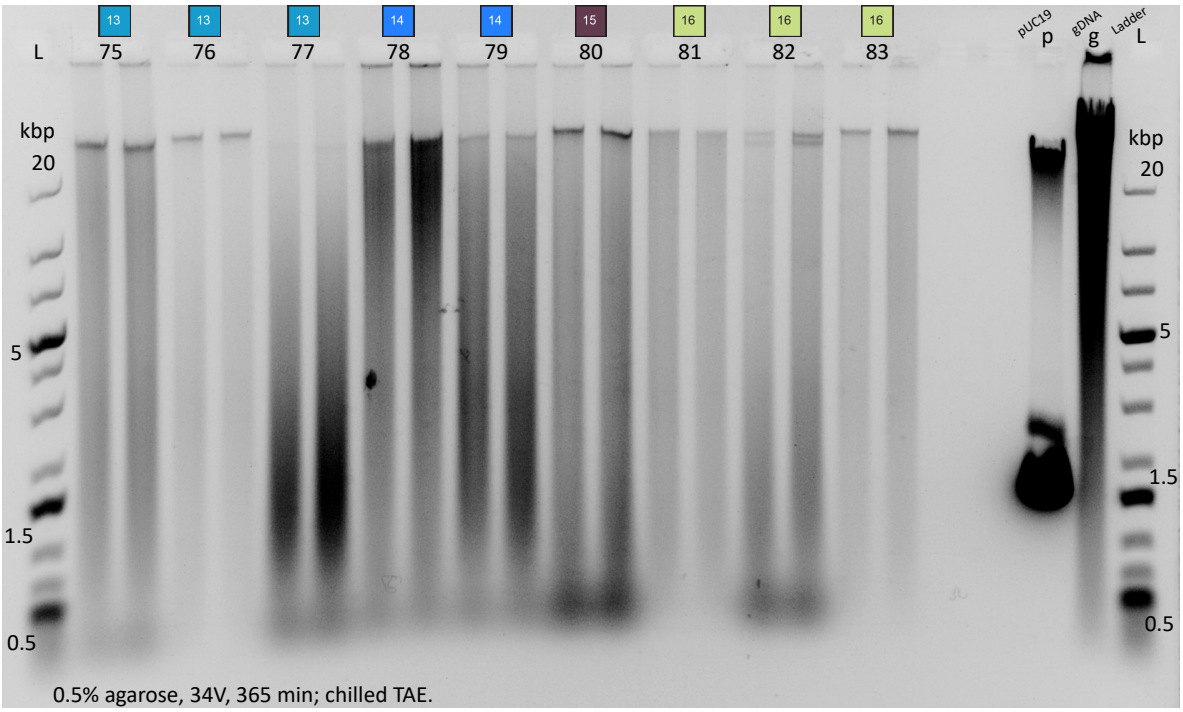
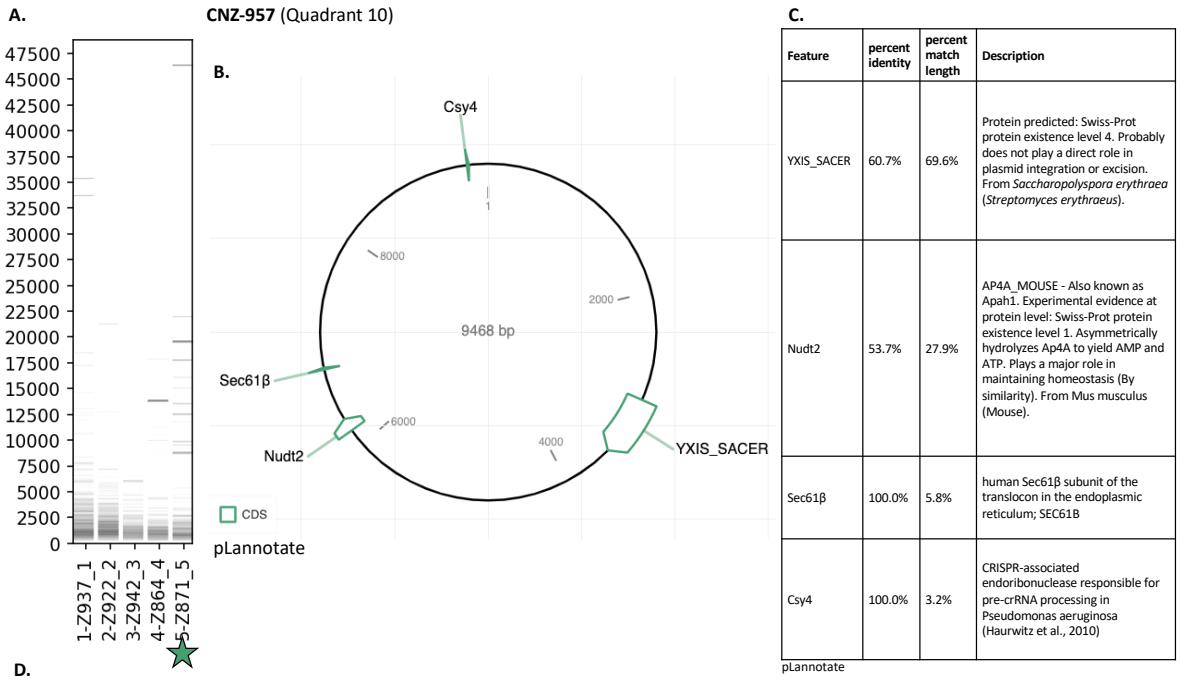


Figure 5.25. Conventional gel electrophoresis of microscale *Salinispora* plasmids (n=2 extractions per strain) numbered by quadrant location. Gel includes 1kbp ladder (L); a pUC19 2.3kbp plasmid and *Salinispora* gDNA for comparative controls.



C.

Feature	percent identity	percent match length	Description
YXIS_SACER	60.7%	69.6%	Protein predicted: Swiss-Prot protein existence level 4. Probably does not play a direct role in plasmid integration or excision. From *Saccharopolyspora erythraea* (*Streptomyces erythraeus*).
Nudt2	53.7%	27.9%	AP4A_MOUSE - Also known as Apah1. Experimental evidence at protein level: Swiss-Prot protein existence level 1. Asymmetrically hydrolyzes Ap4A to yield AMP and ATP. Plays a major role in maintaining homeostasis (By similarity). From Mus musculus (Mouse).
Sec61β	100.0%	5.8%	human Sec61β subunit of the translocon in the endoplasmic reticulum; SEC61B
Csy4	100.0%	3.2%	CRISPR-associated endoribonuclease responsible for pre-crRNA processing in Pseudomonas aeruginosa (Haurwitz et al., 2010)

pLannotate

D.

Query	Hit	% id	align length	mismatches	gap open	start query align	end query align	start subj align	end subj align	evalue	bitscore
5-CNZ-871_9468bp_circ	SaCNZ924_FJ_00021	97.894	2469	48	2	1	2469	9877	7413	0	4268
5-CNZ-871_9468bp_circ	SaCNZ924_FJ_00021	92.222	2070	149	11	4414	6476	14968	12904	0	2920
5-CNZ-871_9468bp_circ	SaCNZ924_FJ_00021	91.724	1885	119	27	7605	9468	11746	9878	0	2582
5-CNZ-871_9468bp_circ	SaCNZ924_FJ_00021	94.727	1536	70	10	2503	4029	6559	5026	0	2377
5-CNZ-871_9468bp_circ	SaCNZ957_FJ_00098	96.805	1878	34	20	5242	7106	1	1865	0	3112
5-CNZ-871_9468bp_circ	SaCNZ957_FJ_00098	100	154	0	0	6697	6850	1609	1456	5.97E-74	285
5-CNZ-871_9468bp_circ	SaCNZ957_FJ_00101	99.09	1539	6	7	7813	9347	1	1535	0	2758
5-CNZ-871_9468bp_circ	SaCNZ957_FJ_00101	100	48	0	0	8272	8319	507	460	5.02E-15	89.8

Figure 5.26. Long-reads plasmid sequencing of microscale *S. arenicola* strain CNZ-957 (gel sample #10), from sub-quadrant 10.

A) Digital gel of the sequencing run;

B) Predicted plasmid sequence map predicted by pLannotate;

C) Description of predicted plasmid features;

D) Top BLASTn matches of the plasmid agains all 99 microscale Salinispora genomes.

5.7.8 Discussion

In this Chapter 5 Appendix, we have developed protocols for extraction and visualization of native *Salinispora* plasmids. These methods were developed and tested on both the macro- and microscale *Salinispora* genome strains, resulting in evidence that *Salinispora* harbor native plasmids of varying sizes from small (~2kbp) to large (> 20kb-> 50kbp from PFGE samples). We applied these plasmid extraction methods to the microscale *Salinispora* strains because this dataset was the first of its kind where we could explore the plasmid "mobilome" a spatial scale of 99 isolates from a $1m^2$ plot. Because the genome diversity of the microscale *Salinispora* does not seem to follow a sub-quadrant patter, we hypothesized that *Salinispora* strains from the different sub-quadrants would have similar plasmids, thus perhaps indicative of plasmid gene flow between strains "distant" from one another in the quadrant. Native plasmids visualized by conventional gel electrophoresis were observed in many different sub-quadrants (**Figure 5.20-5.25, Figure 5.18**), however, sequencing the contents of the plasmids to understand gene flow proved to be difficult **(Figure 5.26**).

Our work in this chapter aimed to characterize the diversity and content of the macroscale and microscale *Salinispora* mobilome, and thus enable future investigations of the ecology of *Salinispora* plasmids. Specifically, the gene content and functions of the plasmids should be analyzed and if interesting, experiments to assess the ability of the plasmids to be transferred via conjugation and transformation methods could be performed. To achieve these aims, one approach would be to introduce a GFP gene in a *Salinispora* plasmid and create a genome-GFP-tagged *Salinispora* strain to visualize if plasmid spreading occurs in growing hyphae, as demonstrated in *Streptomyces* (Thoma and Muth, 2016; Thoma *et al.*, 2016). An alternative method could be to measure of *Salinispora* plasmid transformation efficiency by creating wall-deficient cells which

have both been observed naturally and created for similar DNA-transfer purposes in *Streptomyces* (Ramijan *et al.*, 2018).

5.7.9 Acknowledgements

I want to thank Prof. Eric Allen for letting us borrow his entire Pulsed-Field Gel Electrophoresis (PFGE) system for the years it took to complete this work and for the helpful advice and experimental protocols that he and his lab developed. Thank you to members of the Jensen laboratory for accommodating our very long gel electrophoresis runs and plasmid extraction protocols. I want to acknowledge helpful discussions I had about plasmids in *Salinispora* with Natalie Millán-Aguiñaga, Krystle Chavarria, Eric Allen, Henrique Machado, Rachel Diner, and Tracey J. Mincer.

Chapter 5 Appendix is coauthored with David Vereau-Gorbitz, and Paul R. Jensen. The dissertation author was the primary investigator and author of this chapter.

5.8 References

Arredondo-Alonso, S., Rogers, M.R.C., Braat, J.C., Verschuuren, T.D., Top, J., Corander, J., Willems, R.J.L., and Schürch, A.C. (2018) Mlplasmids: a User-Friendly Tool To Predict Plasmid-and Chromosome- Derived Sequences for Single Species.

Arredondo-Alonso, S., Willems, R.J., Van Schaik, W., and Schürch, A.C. (2017) On the (im)possibility of reconstructing plasmids from whole- genome short-read sequencing data. *Microb Genet*.

Asnicar, F., Thomas, A.M., Beghini, F., Mengoni, C., Manara, S., Manghi, P., Zhu, Q., Bolzan, M., Cumbo, F., May, U., Sanders, J.G., Zolfo, M., Kopylova, E., Pasolli, E., Knight, R., Mirarab, S., Huttenhower, C., and Segata, N. (2020) Precise phylogenetic analysis of microbial isolates and genomes from metagenomes using PhyloPhlAn 3.0. *Nat Commun* **11**: 1–10.

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., Pyshikin, A. V., Sirotkin, A. V., Vyahhi, N., Tesler, G., Alekseyev, M.A., and Pevzner, P.A. (2012) SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *J Comput Biol* **19**: 455–477.

Bauman, K.D., Shende, V. V., Chen, P.Y.-T., Trivella, D.B.B., Gulder, T.A.M., Vellalath, S., Romo, D., and Moore, B.S. (2022) Enzymatic assembly of the salinosporamide γ-lactam-β-lactone anticancer warhead. *Nat Chem Biol*.

Bickhart, D.M., Kolmogorov, M., Tseng, E., Portik, D.M., Korobeynikov, A., Tolstoganov, I., Uritskiy, G., Liachko, I., Sullivan, S.T., Shin, S.B., Zorea, A., Andreu, V.P., Panke-Buisse, K., Medema, M.H., Mizrahi, I., Pevzner, P.A., and Smith, T.P.L. (2022) Generating lineage-resolved, complete metagenome-assembled genomes from complex microbial communities. *Nat Biotechnol*.

Blin, K., Shaw, S., Kloosterman, A.M., Charlop-Powers, Z., van Wezel, G.P., Medema, M.H., and Weber, T. (2021) antiSMASH 6.0: improving cluster detection and comparison capabilities. *Nucleic Acids Res* 1–7.

Blodgett, J.A. V, Oh, D., Cao, S., Currie, C.R., Kolter, R., and Clardy, J. (2010) Common biosynthetic origins for polycyclic tetramate macrolactams from phylogenetically diverse bacteria.

Bordeleau, E., Ghinet, M.G., and Burrus, V. (2012) Diversity of integrating conjugative elements in actinobacteria. *Mob Genet Elements* **2**: 119–124.

Bruns, H., Crüsemann, M., Letzel, A.-C., Alanjary, M., Mcinerney, J.O., Jensen, P.R., Schulz, S., Moore, B.S., and Ziemert, N. (2017) Function-related replacement of bacterial siderophore pathways. *ISME J* **12**: 320–329.

Bucarey, S.A., Penn, K., Paul, L., Fenical, W., and Jensen, P.R. (2012) Genetic Complementation of the Obligate Marine Actinobacterium Salinispora tropica with the Large Mechanosensitive Channel Gene mscL Rescues Cells from Osmotic Downshock. **78**: 4175–4182.

Cao, S., Blodgett, J.A.V., and Clardy, J. (2010) Targeted discovery of polycyclic tetramate macrolactams from an environmental Streptomyces strain. *Org Lett* **12**: 4652–4654.

Carlin, D.E., Demchak, B., Pratt, D., Sage, E., and Ideker, T. (2017) Network propagation in the cytoscape cyberinfrastructure. 1–9.

Carroll, L.M., Wiedmann, M., and Kovac, J. (2020) Proposal of a taxonomic nomenclature for the bacillus cereus group which reconciles genomic definitions of bacterial species with clinical and industrial phenotypes. *MBio* **11**:.

Cavicchioli, R., Ripple, W.J., Timmis, K.N., Azam, F., Bakken, L.R., Baylis, M., Behrenfeld, M.J., Boetius, A., Boyd, P.W., Classen, A.T., Crowther, T.W., Danovaro, R., Foreman, C.M., Huisman, J., Hutchins, D.A., Jansson, J.K., Karl, D.M., Koskella, B., Mark Welch, D.B., Martiny, J.B.H., Moran, M.A., Orphan, V.J., Reay, D.S., Remais, J. V., Rich, V.I., Singh,

B.K., Stein, L.Y., Stewart, F.J., Sullivan, M.B., van Oppen, M.J.H., Weaver, S.C., Webb, E.A., and Webster, N.S. (2019) Scientists' warning to humanity: microorganisms and climate change. *Nat Rev Microbiol* **17**: 569–586.

Chase, A.B., Sweeney, D., Guillén-matus, D.G., and Jensen, P.R. (2021) Vertical Inheritance Facilitates Interspecies Diversification in Biosynthetic Gene Clusters and Specialized Metabolites. *MBio*.

Chen, L.X., Anantharaman, K., Shaiber, A., Murat Eren, A., and Banfield, J.F. (2020) Accurate and complete genomes from metagenomes. *Genome Res* **30**: 315–333.

Corre, C., Song, L., O'Rourke, S., Chater, K.F., and Challis, G.L. (2008) 2-Alkyl-4-hydroxymethylfuran-3-carboxylic acids, antibiotic production inducers discovered by *Streptomyces coelicolor* genome mining. *Proc Natl Acad Sci* **105**: 17510–17515.

Creamer, K.E., Ditmars, F.S., Basting, P.J., Kunka, K.S., Hamdallah, I.N., Bush, S.P., Scott, Z., He, A., Penix, S.R., Gonzales, A.S., Eder, E.K., Camperchioli, D.W., Berndt, A., Clark, M.W., Rouhier, K.A., and Slonczewski, J.L. (2017) Benzoate- and Salicylate-Tolerant Strains of *Escherichia coli* K-12 Lose Antibiotic Resistance during Laboratory Evolution. *Appl Environ Microbiol* **83**: 1–19.

Creamer, K.E., Kudo, Y., Moore, B.S., and Jensen, P.R. (2021) Phylogenetic analysis of the salinipostin γ-butyrolactone gene cluster uncovers new potential for bacterial signalling-molecule diversity. *Microb Genomics* **7**: 1–14.

Darriba, D., Taboada, G.L., Doallo, R., and Posada, D. (2012) JModelTest 2: More models, new heuristics and parallel computing. *Nat Methods* **9**: 772.

Demko, A.M. (2021) Chemical ecology of marine microbial communities: An assessment of bacterial diversity and dynamics in tropical marine sediments.

Demko, A.M., Patin, N. V., and Jensena, P.R. (2021) Microbial diversity in tropical marine sediments assessed using culture-dependent and culture-independent techniques. *Environ Microbiol* 1–50.

Dib, J.R., Wagenknecht, M., Farías, M.E., and Meinhardt, F. (2015) Strategies and approaches in plasmidome studies-uncovering plasmid diversity disregarding of linear elements? *Front Microbiol* **6**: 1–12.

Dong, X., Stothard, P., Forsythe, I.J., and Wishart, D.S. (2004) PlasMapper: A web server for drawing and auto-annotating plasmid maps. *Nucleic Acids Res* **32**: 660–664.

Duncan, K.R., Crüsemann, M., Lechner, A., Sarkar, A., Li, J., Ziemert, N., Wang, M., Bandeira, N., Moore, B.S., Dorrestein, P.C., and Jensen, P.R. (2015) Molecular networking and pattern-based genome mining improves discovery of biosynthetic gene clusters and their products from salinispora species. *Chem Biol* **22**: 460–471.

Edgar, R.C. (2004) MUSCLE: Multiple sequence alignment with high accuracy and high

throughput. *Nucleic Acids Res* **32**: 1792–1797.

Eustáquio, A.S., McGlinchey, R.P., Liu, Y., Hazzard, C., Beer, L.L., Florova, G., Alhamadsheh, M.M., Lechner, A., Kale, A.J., Kobayashi, Y., Reynolds, K. a, and Moore, B.S. (2009) Biosynthesis of the salinosporamide A polyketide synthase substrate chloroethylmalonyl-coenzyme A from *S*-adenosyl-L-methionine. *Proc Natl Acad Sci U S A* **106**: 12295–300.

Evelien, M. and Henk, P.Æ. (2008) Actinomycete integrative and conjugative elements. 127–143.

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

Fischbach, M.A., Walsh, C.T., and Clardy, J. (2008) The evolution of gene collectives: How natural selection drives chemical innovation. *PNAS* **105**: 4601–4608.

Flemming, H.-C. and Wuertz, S. (2019) Bacteria and archaea on Earth and their abundance in biofilms. *Nat Rev Microbiol*.

George, S., Pankhurst, L., Hubbard, A., Votintseva, A., Stoesser, N., Sheppard, A.E., Mathers, A., Norris, R., Navickaite, I., Eaton, C., Iqbal, Z., Crook, D.W., and Phan, H.T.T. (2018) Resolving plasmid structures in Enterobacteriaceae using the MinION nanopore sequencer: assessment of MinION and MinION/Illumina hybrid data assembly approaches. *Microb Genomics* **4**:.

Ghinet, M.G., Bordeleau, E., Beaudin, J., and Brzezinski, R. (2011) Uncovering the Prevalence and Diversity of Integrating Conjugative Elements in Actinobacteria. *PLoS One* **6**:.

Gurevich, A., Saveliev, V., Vyahhi, N., and Tesler, G. (2013) QUAST: Quality assessment tool for genome assemblies. *Bioinformatics* **29**: 1072–1075.

Hamdallah, I., Torok, N., Bischof, K.M., Majdalani, N., Chadalavada, S., Mdluli, N., Creamer, K.E., Clark, M., Holdener, C., Basting, P.J., Gottesman, S., and Slonczewski, J.L. (2018) Experimental Evolution of Escherichia coli K-12 at High pH and RpoS Induction. *Appl Environ Microbiol* AEM.00520-18.

Harden, M.M., He, A., Creamer, K., Clark, M.W., Hamdallah, I., Martinez, K. a., Kresslein, R.L., Bush, S.P., and Slonczewski, J.L. (2015) Acid-Adapted Strains of *Escherichia coli* K-12 Obtained by Experimental Evolution. *Appl Environ Microbiol* **81**: 1932–1941.

Hü, N., Ilhan, J., Wein, T., Kadibalban, A.S., Hammerschmidt, K., Dagan, T., Lang, A., Beatty, J.T., and Rice, P. (2017) An evolutionary perspective on plasmid lifestyle modes. *Curr Opin Microbiol* **38**: 74–80.

Hunt, M., Silva, N. De, Otto, T.D., Parkhill, J., Keane, J.A., and Harris, S.R. (2015) Circlator: Automated circularization of genome assemblies using long sequencing reads. *Genome Biol* **16**: 1–10.

Jain, C., Rodriguez-R, L.M., Phillippy, A.M., Konstantinidis, K.T., and Aluru, S. (2018) High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat Commun* **9**: 1–8.

Jensen, P.R. (2010) Linking species concepts to natural product discovery in the post-genomic era. *J Ind Microbiol Biotechnol* **37**: 219–224.

Jensen, P.R. (2016) Natural Products and the Gene Cluster Revolution. *Trends Microbiol* **24**: 968–977.

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

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

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

Jensen, P.R., Williams, P.G., Oh, D., Zeigler, L., and Fenical, W. (2007) Species-Specific Secondary Metabolite Production in Marine Actinomycetes of the Genus *Salinispora*. *Appl Environ Microbiol* **73**: 1146–1152.

Johnsborg, O., Eldholm, V., and Havarstein, L.S. (2007) Natural genetic transformation: prevalence, mechanisms and function. *Res Microbiol* **158**: 767–778.

Kautsar, S.A., Blin, K., Shaw, S., Navarro-Muñoz, J.C., Terlouw, B.R., van der Hooft, J.J.J., van Santen, J.A., Tracanna, V., Suarez Duran, H.G., Pascal Andreu, V., Selem-Mojica, N., Alanjary, M., Robinson, S.L., Lund, G., Epstein, S.C., Sisto, A.C., Charkoudian, L.K., Collemare, J., Linington, R.G., Weber, T., and Medema, M.H. (2020) MIBiG 2.0: a repository for biosynthetic gene clusters of known function. *Nucleic Acids Res* **48**: D454–D458.

Kav, A.B., Sasson, G., Jami, E., Doron-Faigenboim, A., Benhar, I., and Mizrahi, I. (2012) Insights into the bovine rumen plasmidome. *Proc Natl Acad Sci* **109**: 5452–5457.

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C., Thierer, T., Ashton, B., Meintjes, P., and Drummond, A. (2012) Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **28**: 1647–1649.

Kim, T.K., Garson, M.J., and Fuerst, J.A. (2005) Marine actinomycetes related to the '*Salinospora*' group from the Great Barrier Reef sponge *Pseudoceratina clavata*. *Environ Microbiol* **7**: 509–518.

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

Kinashi, H., Shimaji, M., and Sakai, a (1987) Giant linear plasmids in Streptomyces which code for antibiotic biosynthesis genes. *Nature* **328**: 454–456.

Koren, S., Walenz, B.P., Berlin, K., Miller, J.R., Bergman, N.H., and Phillippy, A.M. (2017) Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res* **27**: 1–15.

Kothari, A., Wu, Y.-W., Chandonia, J.-M., Charrier, M., Rajeev, L., Rocha, A.M., Joyner, D.C., Hazen, T.C., Singer, S.W., and Mukhopadhyay, A. (2019) Large Circular Plasmids from Groundwater Plasmidomes Span Multiple Incompatibility Groups and Are Enriched in Multimetal Resistance Genes. *MBio* **10**: 1–15.

Kudo, Y., Awakawa, T., Du, Y.-L., Jordan, P.A., Creamer, K.E., Jensen, P.R., Linington, R.G., Ryan, K.S., and Moore, B.S. (2020) Expansion of Gamma-Butyrolactone Signaling Molecule Biosynthesis to Phosphotriester Natural Products. *ACS Chem Biol* **15**: 3253–3261.

Lemon, J.K., Khil, P.P., Frank, K.M., and Dekker, J.P. (2017) Rapid nanopore sequencing of plasmids and resistance gene detection in clinical isolates. *J Clin Microbiol* **55**: 3530–3543.

Letzel, A.-C., Li, J., Amos, G.C.A., Millán-Aguiñaga, N., Ginigini, J., Abdelmohsen, U.R., Gaudêncio, S.P., Ziemert, N., Moore, B.S., and Jensen, P.R. (2017) Genomic insights into specialized metabolism in the marine actinomycete *Salinispora*. *Environ Microbiol* **19**: 3660–3673.

Li, R., Xie, M., Dong, N., Lin, D., Yang, X., Wong, M.H.Y., Chan, E.W.-C., and Chen, S. (2018) Efficient generation of complete sequences of MDR-encoding plasmids by rapid assembly of MinION barcoding sequencing data. *Gigascience* **7**: 1–9.

Lloyd, K.G., Steen, A.D., Ladau, J., Yin, J., and Crosby, L. (2018) Phylogenetically Novel Uncultured Microbial Cells Dominate Earth Microbiomes. *mSystems* **3**: 1–12.

Magoč, T. and Salzberg, S.L. (2011) FLASH: Fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* **27**: 2957–2963.

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

McGuffie, M.J. and Barrick, J.E. (2021) PLannotate: Engineered plasmid annotation. *Nucleic Acids Res* **49**: W516–W522.

van der Meij, A., Worsley, S.F., Hutchings, M.I., and van Wezel, G.P. (2017) Chemical ecology of antibiotic production by actinomycetes. *FEMS Microbiol Rev* **41**: 392–416.

Mell, J.C. and Redfield, R.J. (2014) Natural Competence and the Evolution of DNA Uptake Specificity. *J Bacteriol* **196**: 1471–1483.

Millán-Aguiñaga, N., Chavarria, K.L., Ugalde, J.A., Letzel, A.-C., Rouse, G.W., and Jensen, P.R. (2017) Phylogenomic Insight into *Salinispora* (Bacteria, Actinobacteria) Species Designations. *Sci Rep* **7**: 3564.

Mincer, T.J., Fenical, W., and Jensen, P.R. (2005) Culture-dependent and culture-independent diversity within the obligate marine actinomycete genus *Salinispora*. *Appl Environ Microbiol* **71**: 7019–7028.

Mincer, T.J., Jensen, P.R., Kauffman, C.A., and Fenical, W. (2002) Widespread and Persistent Populations of a Major New Marine Actinomycete Taxon in Ocean Sediments. *Appl Environ Microbiol* **68**: 5005–5011.

Navarro-Muñoz, J.C., Selem-Mojica, N., Mullowney, M.W., Kautsar, S.A., Tryon, J.H., Parkinson, E.I., De Los Santos, E.L.C., Yeong, M., Cruz-Morales, P., Abubucker, S., Roeters, A., Lokhorst, W., Fernandez-Guerra, A., Cappelini, L.T.D., Goering, A.W., Thomson, R.J., Metcalf, W.W., Kelleher, N.L., Barona-Gomez, F., and Medema, M.H. (2019) A computational framework to explore large-scale biosynthetic diversity. *Nat Chem Biol* 1–9.

Parks, D.H., Imelfort, M., Skennerton, C.T., Hugenholtz, P., and Tyson, G.W. (2015) CheckM: Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res* **25**: 1043–1055.

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

Penn, K. and Jensen, P.R. (2012) Comparative genomics reveals evidence of marine adaptation in *Salinispora* species. *BMC Genomics* **13**: 86.

Petit, R.A. and Read, T.D. (2020) Bactopia: a Flexible Pipeline for Complete Analysis of Bacterial Genomes. *mSystems* **5**: 1–18.

Probandt, D., Eickhorst, T., Ellrott, A., Amann, R., and Knittel, K. (2018) Microbial life on a sand grain: from bulk sediment to single grains. *ISME J* **12**: 623–633.

Qi, Y., Nepal, K.K., and Blodgett, J.A. V. (2021) A comparative metabologenomic approach reveals mechanistic insights into Streptomyces antibiotic crypticity . *Proc Natl Acad Sci* **118**: e2103515118.

Rambaut, A. (2016) FigTree v1.4.3.

Ramijan, K., Ultee, E., Willemse, J., Zhang, Z., Wondergem, J.A.J., van der Meij, A., Heinrich, D., Briegel, A., van Wezel, G.P., and Claessen, D. (2018) Stress-induced formation of cell wall-deficient cells in filamentous actinomycetes. *Nat Commun* **9**:.

Reuther, J., Gekeler, C., Tiffert, Y., Wohlleben, W., and Muth, G. (2006) Unique conjugation mechanism in mycelial streptomycetes : a DNA-binding ATPase translocates unprocessed

plasmid DNA at the hyphal tip. **61**: 436–446.

Román-Ponce, B., Millán-Aguiñaga, N., Guillen-Matus, D., Chase, A.B., Ginigini, J.G.M., Soapi, K., Feussner, K.D., Jensen, P.R., and Trujillo, M.E. (2020) Six novel species of the obligate marine actinobacterium *Salinispora*, *Salinispora cortesiana* sp. nov., *Salinispora fenicalii* sp. nov., *Salinispora goodfellowii* sp. nov., *Salinispora mooreana* sp. nov., ... *Int J Syst Evol Microbiol* **70**: 4668–4682.

Royer, G., Decousser, J.W., Branger, C., and Dubois, M. (2019) PlaScope : a targeted approach to assess the plasmidome from genome assemblies at the species level. 1–8.

Rozov, R., Kav, A.B., Bogumil, D., Shterzer, N., Halperin, E., Mizrahi, I., and Shamir, R. (2017) Recycler: An algorithm for detecting plasmids from de novo assembly graphs. *Bioinformatics* **33**: 475–482.

RStudio Team (2021) RStudio: Integrated Development Environment for R. *RStudio*.

Salam, N., Jiao, J.Y., Zhang, X.T., and Li, W.J. (2020) Update on the classification of higher ranks in the phylum Actinobacteria. *Int J Syst Evol Microbiol* **70**: 1331–1355.

Seemann, T. (2014) Prokka: Rapid prokaryotic genome annotation. *Bioinformatics* **30**: 2068–2069.

Segata, N., Börnigen, D., Morgan, X.C., and Huttenhower, C. (2013) PhyloPhlAn is a new method for improved phylogenetic and taxonomic placement of microbes. *Nat Commun* **4**:.

Souvorov, A., Agarwala, R., and Lipman, D.J. (2018) SKESA: Strategic k-mer extension for scrupulous assemblies. *Genome Biol* **19**: 1–13.

Stamatakis, A. (2014) RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**: 1312–1313.

Stuttard, C. (1983) Localized hydroxylamine mutagenesis, and cotransduction of threonine and lysine genes, in Streptomyces venezuelae. *J Bacteriol* **155**: 1219–1223.

Tang, X., Li, J., Mill??n-Agui??aga, N., Zhang, J.J., O'Neill, E.C., Ugalde, J.A., Jensen, P.R., Mantovani, S.M., and Moore, B.S. (2015) Identification of Thiotetronic Acid Antibiotic Biosynthetic Pathways by Target-directed Genome Mining. *ACS Chem Biol* **10**: 2841–2849.

Taylor, T.L., Volkening, J.D., DeJesus, E., Simmons, M., Dimitrov, K.M., Suarez, D.L., and Afonso, C.L. (2019) Rapid, multiplexed, whole genome and plasmid sequencing of foodborne pathogens using long-read nanopore technology. **1**: 0–26.

Thoma, L., Dobrowinski, H., Finger, C., Guezguez, J., Linke, D., Sepulveda, E., and Muth, G. (2015) A Multiprotein DNA Translocation Complex Directs Intramycelial Plasmid Spreading during Streptomyces Conjugation. *MBio* **6**:.

Thoma, L. and Muth, G. (2016) Conjugative DNA-transfer in Streptomyces, a mycelial organism.

Plasmid.

Thoma, L., Vollmer, B., and Muth, G. (2016) Fluorescence microscopy of Streptomyces conjugation suggests DNA-transfer at the lateral walls and reveals the spreading of the plasmid in the recipient mycelium. *Environ Microbiol* **18**: 598–608.

Tianero, M.D., Balaich, J.N., and Donia, M.S. (2019) Localized production of defence chemicals by intracellular symbionts of Haliclona sponges. *Nat Microbiol*.

Tuttle, R.N., Demko, A.M., Patin, N. V, Kapono, C.A., Donia, M.S., Dorrestein, P., and Jensen, P.R. (2019) Detection of Natural Products and Their Producers in Ocean Sediments. *Appl Environ Microbiol* **85**: 1–15.

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

Undabarrena, A., Valencia, R., Cumsille, A., Leiva, L.Z.-, Nallar, E.C.-, Gomez, F.B.-, and Cámara, B. (2021) Rhodococcus comparative genomics reveals a phylogenomic- - dependent non- - ribosomal peptide synthetase distribution : insights into biosynthetic gene cluster connection to an orphan metabolite.

Vidgen, M.E., Hooper, J.N.A., and Fuerst, J.A. (2012) Diversity and distribution of the bioactive actinobacterial genus *Salinispora* from sponges along the Great Barrier Reef. *Antonie Van Leeuwenhoek* **101**: 603–618.

Vielva, L., De Toro, M., Lanza, V.F., and De La Cruz, F. (2017) PLACNETw: A web-based tool for plasmid reconstruction from bacterial genomes. *Bioinformatics* **33**: 3796–3798.

Walker, B.J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C.A., Zeng, Q., Wortman, J., Young, S.K., and Earl, A.M. (2014) Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* **9**:.

Wang, M. and Lai, E. (1995) Pulsed field separation of large supercoiled and open-circular DNAs and its application to bacterial artificial chromosome cloning. *Electrophoresis* **16**: 1–7.

Wick, R.R., Schultz, M.B., Zobel, J., and Holt, K.E. (2015) Bandage: Interactive visualization of de novo genome assemblies. *Bioinformatics* **31**: 3350–3352.

Wietz, M., Millán-aguiñaga, N., and Jensen, P.R. (2014) CRISPR-Cas systems in the marine actinomycete Salinispora : linkages with phage defense , microdiversity and biogeography. *BMC Genomics* **15**: 936.

Von Wintersdorff, C.J.H., Penders, J., Van Niekerk, J.M., Mills, N.D., Majumder, S., Van Alphen, L.B., Savelkoul, P.H.M., and Wolffs, P.F.G. (2016) Dissemination of antimicrobial resistance in microbial ecosystems through horizontal gene transfer. *Front Microbiol* **7**: 1–10.

Yu, G. (2020) Using ggtree to Visualize Data on Tree-Like Structures. *Curr Protoc Bioinforma* **69**: 1–18.

Yu, G., Smith, D.K., Zhu, H., Guan, Y., and Lam, T.T.Y. (2017) Ggtree: an R Package for Visualization and Annotation of Phylogenetic Trees With Their Covariates and Other Associated Data. *Methods Ecol Evol* **8**: 28–36.

Zerbino, D.R. and Birney, E. (2008) Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* **18**: 821–829.

Zhang, J.J., Moore, B.S., Tang, X., and Moore, B.S. (2018) Engineering Salinispora tropica for heterologous expression of natural product biosynthetic gene clusters.

Ziemert, N., Lechner, A., Wietz, M., Millán-Aguiñaga, N., Chavarria, K.L., and Jensen, P.R. (2014) Diversity and evolution of secondary metabolism in the marine actinomycete genus *Salinispora*. *Proc Natl Acad Sci U S A* **111**: E1130-9.

Ziemert, N., Podell, S., Penn, K., Badger, J.H., Allen, E., and 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.

CHAPTER 6. Evolutionary radiation of lanthipeptide RiPPs in micro- and macroscale *Salinispora*

6.1 Abstract

The bacterial phylum Actinobacteria includes a wide diversity of ecologically, genetically, and chemically rich species. While classification efforts have typically defined bacterial species based on genetic distance, evolutionary patterns of biosynthetic genes can be another method to differentiate closely related bacterial taxa. Ribosomally synthesized and post-translationally modified peptides (RiPPs) are specialized metabolites that originate from post-translational modification of precursor peptides, resulting in a wide array of compound diversity. A recent analysis of lanthipeptide RiPPs in marine picocyanobacteria revealed extraordinary compound diversity caused by varying core precursor peptide sequences. Do all lanthipeptide RiPP precursor peptide sequences show such an extreme radiation of sequence diversity? The marine-obligate bacterial genus *Salinispora* provides an opportunity to examine the evolutionary history and distribution of lanthipeptide RiPP precursor peptide sequence diversity on two different scales. *Salinispora* species are very closely related yet seem to have species-specific chemical arsenals and significant patterns of specialized metabolite biosynthetic gene cluster (BGC) distributions across species (Letzel *et al.*, 2017). In this study, I assessed the distribution of lanthipeptide precursor peptides across a "macroscale" and a new "microscale" collection of *Salinispora* genomes. I have isolated and whole-genome sequenced a collection of microscale *Salinispora* strains cultured from a 1-meter^2 marine sediment quadrant in Fiji. Both the macro- and microscale *Salinispora* strains have the potential to produce a vast assortment of structurally diverse and yet-to-be-isolated RiPP lanthipeptides from every known class (I-V) of lanthipeptides. I discovered the *Salinispora* genus contains four out of the five known classes of lanthipeptides, including the newest class V type. Most of the *Salinispora* RiPP lanthipeptide precursor peptides are unlike any

precursor peptides from characterized RiPP BGCs and those extracted from 100,000 reference genomes. Further insights including possible evolutionary trajectories leading to the radiation of lanthipeptide RiPP precursor peptide diversity at the macro and microscale are discussed, which will support future efforts to selectively classify and characterize the first lanthipeptide RiPP from *Salinispora*.

6.2 Introduction

The marine obligate actinobacterial genus *Salinispora* is a rich source of biosynthetic potential. A previous comprehensive analysis of all BGCs in 118 macroscale (defined here as *Salinispora* strains isolated from worldwide locations) *Salinispora* genomes identified 176 total BGCs, including polyketide synthase (PKS), non-ribosomal peptide synthetase (NRPS), PKS-NRPS hybrid, ribosomally synthesized and post-translationally modified peptide (RiPP), terpene, siderophore, and a general "other" classes of BGCs (Letzel *et al.*, 2017). A recent analysis increased this estimate to 3,041 complete or fragmented BGCs across the 118 *Salinispora* genomes, belonging to 305 gene cluster families (Chase *et al.*, 2021). However, to date, the number of compounds that have been isolated and linked to a *Salinispora* BGC from the PKS, NRPS, PKS-NRPS, terpene, siderophore, and "other" classes of BGCs is small relative to the number of estimated BGCs (Jensen *et al.*, 2015; Letzel *et al.*, 2017; Chase *et al.*, 2021). Similarly, evolutionary analyses of *Salinispora* BGCs have primarily focused on PKS, NRPS, hybrid clusters, and siderophores (Ziemert *et al.*, 2014; Tang *et al.*, 2015; Bruns *et al.*, 2017; Letzel *et al.*, 2017; J. J. Zhang *et al.*, 2018; Chase *et al.*, 2021; Williams *et al.*, 2022) as genome-mining tools to identify and characterize these classes have been developed due to the predictable modular

organization and large number of characterized BGCs that can help inform genome searches. This includes my own analysis of the salinipostin gene cluster, described in Chapter 4 of this dissertation (Creamer *et al.*, 2021) and the recent description of the pacificamide compound and putative BGC, described in Appendix A of this dissertation (Castro-Falcón *et al.*, 2022). However, in the past few years, there has been an explosion of interest in the RiPP class of specialized metabolites following a concerted effort to characterize RiPP molecules and functions across bacteria (Arnison *et al.*, 2013).

Ribosomally synthesized and post-translationally modified peptides (RiPPs) are different from PKS and NRPS derived compounds because they are the direct products of the cell's ribosomal machinery (Arnison *et al.*, 2013; Ortega and Van Der Donk, 2016; Repka *et al.*, 2017; Hudson and Mitchell, 2018; Montalbán-López *et al.*, 2021). The general scheme of RiPP biosynthesis starts with the transcription and translation of a precursor peptide gene in the RiPP BGC. The precursor peptide is 20-110 residues in length (composed of the 20 traditional proteinogenic amino acids) and subsequently undergoes post-translational modifications (PTMs) by enzymes that recognize the leader part of the precursor peptide and act on the core portion of the peptide (Arnison *et al.*, 2013; Ortega and Van Der Donk, 2016; Repka *et al.*, 2017; Hudson and Mitchell, 2018; Montalbán-López *et al.*, 2021). Finally, the leader half of the peptide is cleaved and the modified core peptide, typically the final active RiPP, is transported to its final destination out of the cell by transporter genes present in the RiPP BGC (Arnison *et al.*, 2013; Ortega and Van Der Donk, 2016; Repka *et al.*, 2017; Hudson and Mitchell, 2018; Montalbán-López *et al.*, 2021).

RiPPs are incredibly diverse with more than >40 subclasses identified based on their salient features including structural motifs and common biosynthetic machinery (Arnison *et al.*, 2013; Ortega and Van Der Donk, 2016; Montalbán-López *et al.*, 2021). These classes of RiPPs all

include potent, bioactive specialized metabolites and therefore identifying all RiPP BGCs in *Salinispora* could lead to the discovery of useful RiPPs and exciting chemical diversity.

While there are many classes of RiPPs, this study focuses on lanthipeptides. Lanthipeptides are lanthionine (Lan) and methyllanthionine (MeLan) containing RiPPs that are found across all three domains of life. They have antibiotic (lantibiotic), antifungal, antiviral, antinociceptive, antiallodynic, signaling molecule, and morphogenetic protein properties (Zhang *et al.*, 2012; van der Donk and Nair, 2014; Repka *et al.*, 2017). The lanthionine structure consists of two alanine residues that are linked through a thioether that connects their beta-carbons; many lanthipeptides also contain methyllanthionines which carry an additional methyl group on one of the beta-carbons (Yu *et al.*, 2013). Biosynthetically, lanthionines and methyllanthionines originate from Ser and Thr residues in the precursor peptide that are first dehydrated to generate dehydroalanine (Dha) and dehydrobutyrine (Dhb) residues, respectively; next, the thiol of a Cys is added across the carbon–carbon double bond of these dehydroamino acids in a Michael-type addition to produce the Lan and MeLan structures, respectively (Yu *et al.*, 2013). There are five classes of lanthipeptides as differentiated by their biosynthetic machinery, including: class I lanthipeptides (reactions require two different enzymes, a dehydratase LanB, and a cyclase LanC); class II lanthipeptides (reactions carried out by a single lanthipeptide synthetase, LanM, containing an N-terminal dehydratase domain and a C-terminal LanC-like cyclase domain); class III (reactions catalyzed by the trifunctional enzyme LanKC with a lyase domain, kinase domain, and putative cyclase domain); class IV (synthesized by the trifunctional enzyme LanL with a lyase domain, kinase domain and LanC-like cyclase domain); and class V (contains a putative lyase/dehydratase, kinase, and cyclases, and feature both lanthionine and AviMeCys motifs) (Zhang *et al.*, 2012; Arnison *et al.*, 2013; Repka *et al.*, 2017; Li *et al.*, 2021). Past and present efforts to identify

lanthipeptides, including radical new LanB behavior of a RiPP peptide serving as a scaffold for a non-ribosomal peptide extension and chemical modification (Ting *et al.*, 2019), lipolanthine biosynthesis combining class III lanthionine synthases with PKS/FAS (Wiebach, Vincent *et al.*, 2018), and three novel members comprising the newest class V type (Kloosterman *et al.*, 2020; Ortiz-López *et al.*, 2020; Xu *et al.*, 2020), have contributed to the understanding of lanthipeptide biosynthetic logic, which then can inform genome-mining approaches. There has been a recent expansion of RiPP genome-mining tools, including updated versions of antiSMASH 4.0, 5.0, and 6.0 (Blin *et al.*, 2017, 2019, 2021), BAGEL 3 and 4 (van Heel *et al.*, 2013, 2018), PRISM 3 and 4 (Skinnider *et al.*, 2017, 2020), RIPPMiner Genome (Agrawal *et al.*, 2017, 2021), and RODEO 1.0 and 2.0 (Tietz *et al.*, 2017; Hudson *et al.*, 2019; Walker *et al.*, 2020), to name a few.

In a recent RiPP genome-mining analysis, it was predicted that lanthipeptides are the second most common RiPP in prokaryotic genomes (~14,000 lanthipeptide RiPP BGCs) and that lanthipeptide BGCs encode the second-most novel chemical collection of RiPP compounds (~1,000 unique lanthipeptides) (Skinnider *et al.*, 2016). However, what is most interesting about this predicted diversity of lanthipeptides is that due to the biosynthetic paradigm of RiPPs, the structural diversity originates from the 20 proteinogenic amino acids that form the core precursor peptide. One particularly striking example of this diversity was observed in the marine *Synechococcus* and *Prochlorococcus* lanthipeptide prochlorosins (Bobeica and van der Donk, 2018). Prochlorosin precursor peptides have undergone an incredible evolutionary radiation as it was discovered that the prochlorosin leader peptide amino acid composition was highly conserved whereas the core peptide showed a high level of diversity with very low amino acid identity (Cubillos-Ruiz *et al.*, 2017). This indicates that in the case of the marine picocyanobacterial prochlorosins, extreme chemical diversification is driven by small changes in the core section of

the precursor peptide whereas the Lan modifying enzymes are less promiscuous in recognizing a conserved leader peptide sequence. Interestingly, a follow-up study to the original report of prochlorosin rapid core structure diversification discovered that the prochlorosin genes were associated with a single-stranded $TnpA_{REP}$ family of transposases—both co-localizing and co-occurring in a phylogenic pattern indicating that the transposases could be responsible for promoting structural diversification of the prochlorosin genes via recombination and other methods (Laurenceau *et al.*, 2020). Inspired by this first-of-its-kind evolutionary analysis of RiPPs, I proposed that our extensive collection of both macroscale and microscale *Salinispora* genomes present an opportunity to identify RiPPs and ask similar evolutionary questions. This builds on a recently published analysis describing the lanthipeptide diversity predicted from *Salinispora* genomes (Kittrell *et al.*, 2020), which I aimed to expand upon by 1) using updated genome-mining tools to see if any new classes of lanthipeptides could be observed in *Salinispora* beyond what was originally reported; 2) expand the analysis of lanthipeptide BGCs and precursor peptides to both the macro- and new microscale collection of *Salinispora* genomes; 3) compare the predicted lanthipeptide RiPP biosynthetic potential to characterized lanthipeptides and 4) investigate the evolutionary dynamics of lanthipeptide precursor peptides in *Salinispora*.

6.3 Methods

6.3.1 Identification of Salinispora RiPP BGCs and precursor peptides

The micro- and macroscale *Salinispora* genomes for this study were isolated and whole genome sequenced in the preceding Chapter 5 of this dissertation. To assess the biosynthetic

potential of the 99 microscale and 118 macroscale *Salinispora* strains, all genomes were analyzed with antiSMASH 6.0 (with all options on, including --taxon bacteria, --cb-general, cb-knownclusters, --cb-subclusters, --cc-mibig, --asf, --rre, --pfam2go, --tigrfam, --smcog-trees, --clusterhmmer, --fullhmmer, --genefinding-tool prodigal) (Blin *et al.*, 2021). Additionally, BiG-SCAPE and CORASON (Navarro-Muñoz *et al.*, 2019) were used to calculate similarity between biosynthetic gene clusters including gene cluster families (GCFs) and gene cluster family clans. BiG-SCAPE was run with the following settings: Pfam 35.0 database; --include_singletons only when the comparison to the MIBiG 2.0 (Kautsar *et al.*, 2020) option (--mibig) was toggled off (and vice versa when the –mibig option was toggled on, singletons were excluded); --cutoffs 0.1 0.3 0.5 0.75 1.0; and --mode auto.

Next, manual inspection of all antiSMASH 6.0 html output files was used to identify predicted RiPP BGCs, and all predicted lanthipeptide BGC precursor peptides were copied from antiSMASH into a FASTA sequence file. If antiSMASH 6.0 predicted the core, leader, Dha, and Dhb residues, the predictions were noted. However, sequences were transformed to contiguous sequences (gaps between leader and core removed) and all Dha were changed back to S and all Dhb were changed back to T before downstream analyses. All extracted precursor peptides were labeled with the *Salinispora* species, strain, isolation location, the antiSMASH BGC cluster number, a unique "lan" number, the protein accession ID, and the predicted type of BGC or Lan class, as predicted by antiSMASH (ie. _putativeclassII or _putativeclassI). For some lanthipeptide RiPP BGCs, antiSMASH did not predict the core peptides, and thus the small genes encoded near the core Lan biosynthetic enzymes were analyzed to see if they could be the precursor peptides. Specifically, I searched for Pfam HMM domains associated with precursor peptides as previously reported (Walker *et al.*, 2020), including: class I Lant_dehydr_N Pfam HMM (PF04738.13),

Lant_dehydr_C Pfam HMM (PF14028.6), and TIGR04363 (LD_lanti_pre: FxLD family lantipeptide; class II DUF4135 Pfam HMM (PF13575.6); class III and IV LANC_Like Pfam + Pkinase Pfam HMM (PF00069.25); and class V based on organization of the genes compared to the three reported class V BGCs. Any putative precursor peptide (small gene near the core Lan biosynthetic genes) that did not include a Cysteine (C) residue was discarded as it could not be a potential lanthipeptide precursor. Short domains that looked like the correct size to be a lanthipeptide precursor peptide but were functionally annotated differently and thus avoided included PP-binding and MbtH proteins. In some instances, the predicted precursor peptide sequences were analyzed with RiPP-Miner (Agrawal *et al.*, 2021) or RODEO 2.0 (Walker *et al.*, 2020) to determine if they were indeed RiPP-like. Custom R scripts were used to count the number of lanthipeptide BGCs from which precursor peptides were identified and the total number of precursor peptides per BGC and *Salinispora* genome.

6.3.2 Comparison of Salinispora precursor peptides with known RiPP compounds

In order to put the *Salinispora* lanthipeptide precursor peptide diversity into context, I extracted out all RiPP precursor peptides from the MIBiG 2.0 database of characterized BGCs (Kautsar *et al.*, 2020) and all putative class I-IV RiPP precursor peptides identified from 100,000 reference genome sequences (Walker *et al.*, 2020). Briefly, a custom script was used to extract all CDS regions from the concatenated MIBiG 2.0 Genbank sequence database, and then targeted searches for “pre-pep” “precursor” “structural gene” “propeptide” “propep” and the *gene*A, along with a list of all RiPP-BGCs to manually search the 353 precursor peptides identified. Precursor peptides from Walker *et al*. 2020 were extracted from the supplemental Microsoft Excel files; and

each precursor was appended with the class type I-IV, the genus species of the encoding strain, and the protein accession ID.

Sequence similarity networks were constructed with EFI-EST (Gerlt *et al.*, 2015; Gerlt, 2017; Zallot *et al.*, 2019, 2021) and visualized with Cytoscape (Carlin *et al.*, 2017). Amino acid alignments were performed in Geneious (Kearse *et al.*, 2012).

6.4 Results

With the help of updated RiPP genome-mining tools, I sought to ascertain the class types of lanthipeptide RiPPs in the genus *Salinispora*, and build an updated assessment of lanthipeptide diversity from the previous report (Kittrell *et al.*, 2020). In the 99 new microscale *Salinispora* genomes, I identified 463 RiPP BGCs, which was the second largest class of BGCs. Our comparative BGC analysis with BiG-SCAPE clustered the RiPP BGCs into 38 gene cluster families (GCFs) with an average of 18 BGCs per BGC family, which were distributed across the 99 genomes in a species-specific pattern for some GCFs (**Figure 6.1**). For example, the FAM_00507 GCF contained 95 BGCs, which were present in almost all of the *Salinispora arenicola* microscale genomes, but not in *S. pacifica* and *S. oceanensis* (**Figure 6.1**). Likewise, *S. pacifica* contained unique RiPP GCFs not found in other strains; and many other RiPP GCFs were distributed across the 99 genomes (**Figure 6.1**). When the 118 macroscale *Salinispora* genome BGCs were combined with the microscale BGCs, I identified a total of 902 RiPP BGCs clustered into 98 RiPP GCFs (average of 14 BGCs per GCF; and the largest GCF contained 145 RiPP BGCs). By plotting the presence and absence of all RiPP GCFs, I observed some species-specific distribution of RiPP GCFs (for example, in *S. tropica* at the top) and many of the new microscale *Salinispora* had similar GCF compositions that were shared with one or two other macroscale

Salinispora strains (**Figure 6.2**). However, because RiPP diversity arises from differences at the precursor peptide level, I focused on individual RiPP BGCs and their precursor peptides instead of at the GCF level.

From targeted RiPP BGC mining, I identified a total of 518 lanthipeptide BGCs. The *Salinispora arenicola* microscale genomes had the greatest number of lanthipeptide BGCs (n=254) followed by the macroscale *S. arenicola* (n=181) and *S. pacifica* (n=16) (**Figure 6.3**). However, these numbers do not reflect the uneven distribution of genomes for each species, thus I calculated the number of lanthipeptide RiPP BGCs per *Salinispora* genome (**Figure 6.4**). While I observed that both micro- and macroscale *S. arenicola* contained ~2.5 lanthipeptide BGCs per genome (and with >100 sequenced genomes in the species, this accounts for the greatest proportion of lanthipeptide BGCs in *Salinispora*), it was surprising to find that the only *S. cortesiana* genome in the genome collection contained four lanthipeptide BGCs (**Figure 6.4**).

Figure 6.1. Presence (red) and absence of RiPP gene cluster families (GCFs) in the 99 microscale *Salinispora*. The Y-axis dendrogram (*Salinispora* genomes) is clustered by absence/presence of RiPP GCF (X-axis, also clustered by presence/absence).

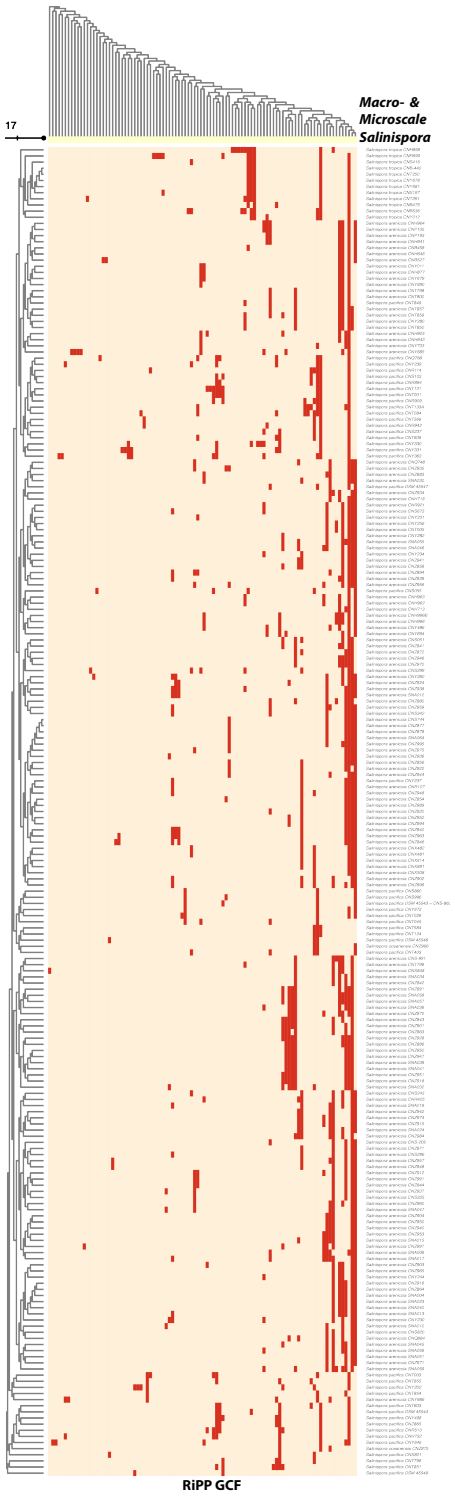


Figure 6.2. Presence (red) and absence of RiPP gene cluster families (GCFs) across all 217 micro- and macroscale *Salinispora*. The Y-axis dendrogram (*Salinispora* genomes) is clustered by absence/presence of RiPP GCF (X-axis, also clustered by presence/absence).

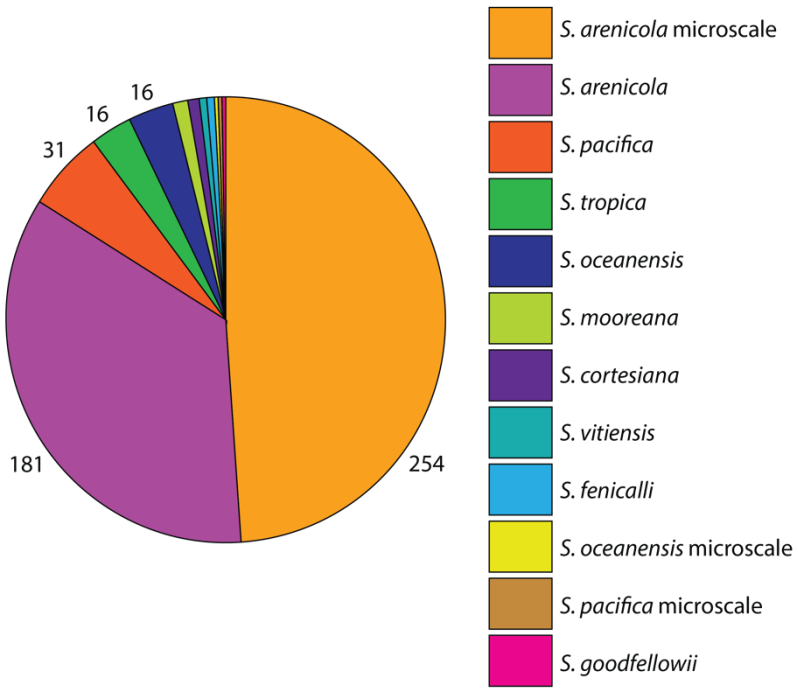


Figure 6.3. Number of lanthipeptide RiPP BGCs in all 217 *Salinispora* genome, colored by *Salinispora* species.

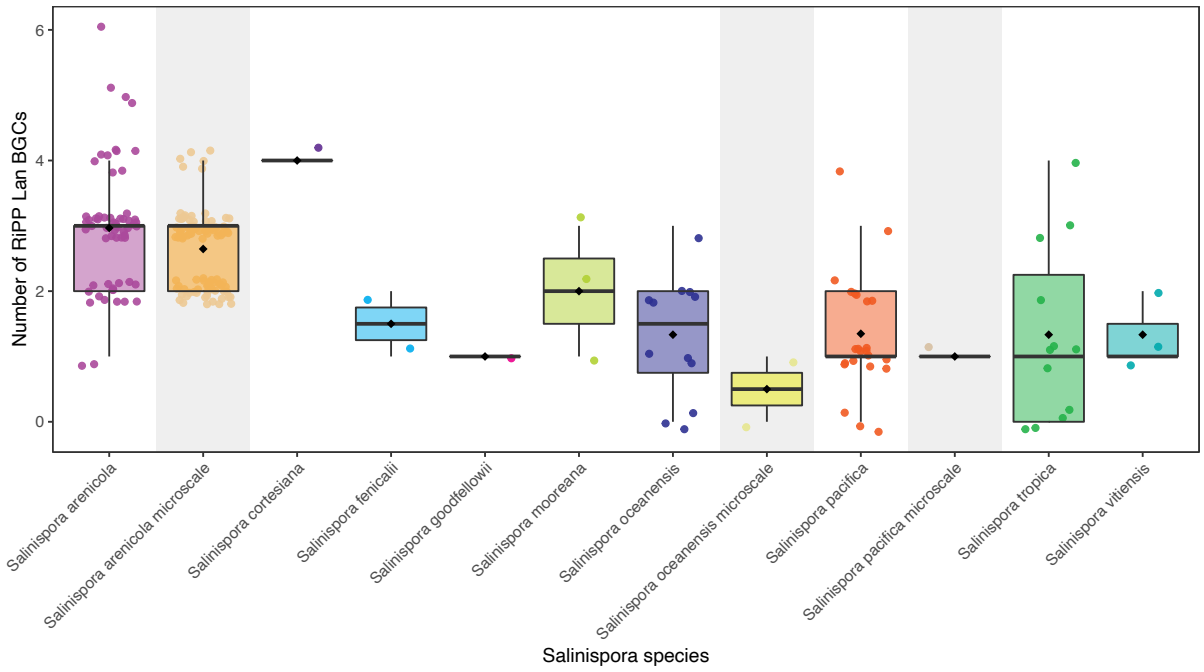


Figure 6.4. Number of lanthipeptide RiPP BGCs per *Salinispora* species genome (from all 217 *Salinispora* genomes). Shaded bars denote new microscale *Salinispora.*

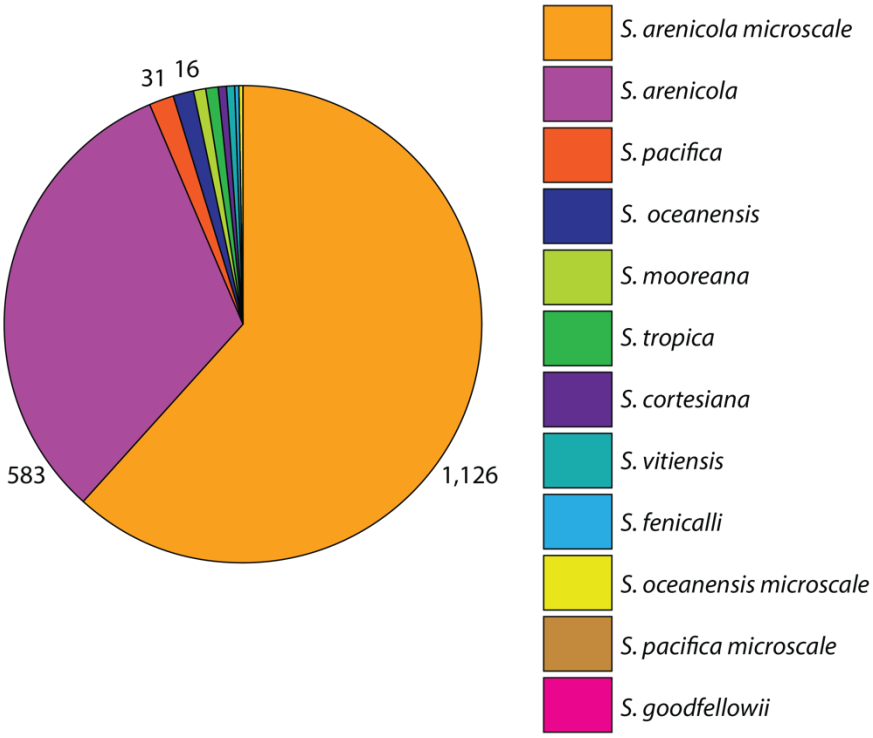


Figure 6.5. Number of lanthipeptide RiPP BGC precursor peptides from all 217 *Salinispora* genomes, colored by *Salinispora* species.

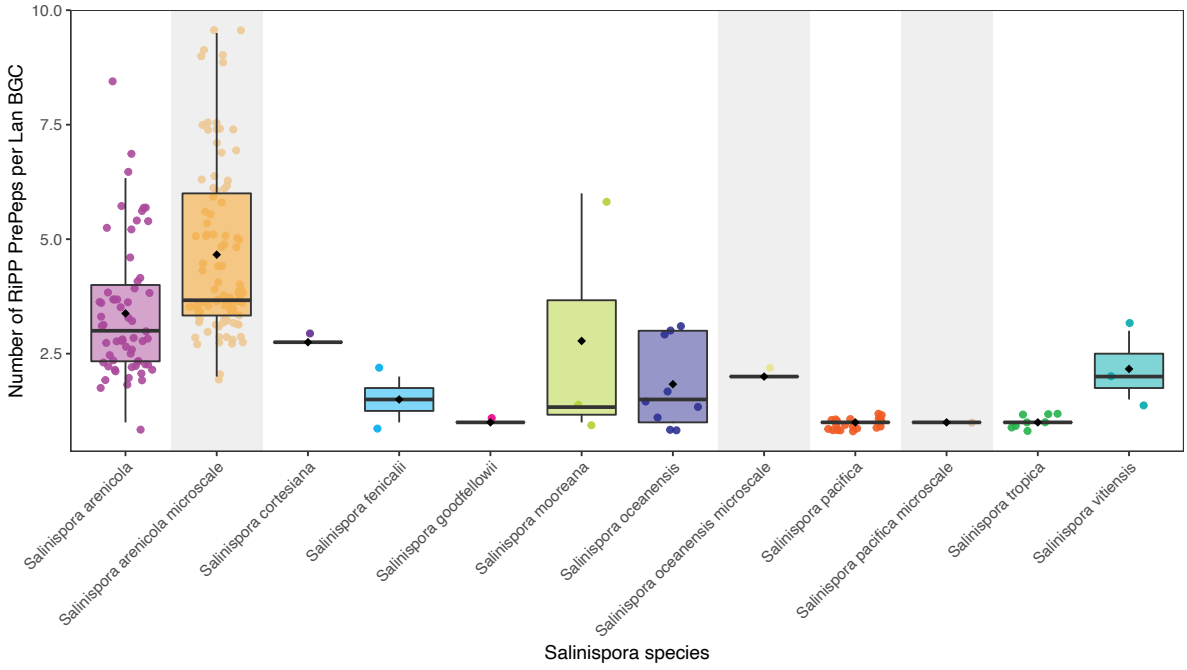


Figure 6.6. Number of precursor peptides per lanthipeptide RiPP BGCs in all 217 *Salinispora* species genomes. Shaded bars denote new microscale *Salinispora*.

Next, I compared the number of precursor peptides identified from the *Salinispora* RiPP lanthipeptide BGCs, which totaled 1,825. Again, the microscale *S. arenicola* contained the largest number of precursor peptides (n=1,126) followed by the macroscale *S. arenicola* (n=583) (**Figure 6.5**). Interestingly, when I calculated the number of precursor peptides per lanthipeptide RiPP BGC in each *Salinispora* species, I discovered that the micro- and macroscale *S. arenicola* each contained ~3-4 precursor peptides per lanthipeptide BGC, which was in stark contrast to the other species with only ~1-2 precursor peptides per lanthipeptide BGC (**Figure 6.6**). Many species like both macro- and microscale *S. pacifica*, *S. tropica*, and *S. goodfellowii* contained only one RiPP precursor peptide per lanthipeptide BGC (**Figure 6.6**). This difference in precursor peptide number could indicate different rates of evolutionary diversification of lanthipeptides between *Salinispora* species as observed in the case of the prochlorosins (Cubillos-ruiz *et al.*, 2017).

To better understand the potential diversity of precursor peptides in all *Salinispora*, I created a sequenced similarity network of all precursor peptides, colored by the *Salinispora* species (**Figure 6.7**). I uncovered 132 distinct clusters (including singletons) of lanthipeptide precursor peptides, and thus potential unique compounds, which was astounding (**Figure 6.7**). There are some species-specific clusters (clusters 13, 44, 34, 38, 45, etc.) which could be unique chemical products in each *Salinispora* species arsenal of compounds (**Figure 6.7**). However, I also observed precursor peptides that were shared across divergent species, i.e., clusters with many different colored nodes (cluster 2, 47, 69, etc.) (**Figure 6.7**). These could be interesting lanthipeptide BGCs to explore further, as they could be BGCs that have been horizontally transferred between species. It also appears as though our expansion of the microscale *Salinispora* genome dataset captured more chemical diversity, especially in the case of cluster 31 where our single *S. pacifica* microscale contained a unique species-specific lanthipeptide precursor peptide (**Figure 6.7**). Overall, the

microscale and macroscale *S. arenicola* shared much chemical diversity, however there are some precursor peptides that are unique to either the macro- or microscale genomes, thus indicating that I did not only rediscover the same lanthipeptide chemical diversity with our new genomes.

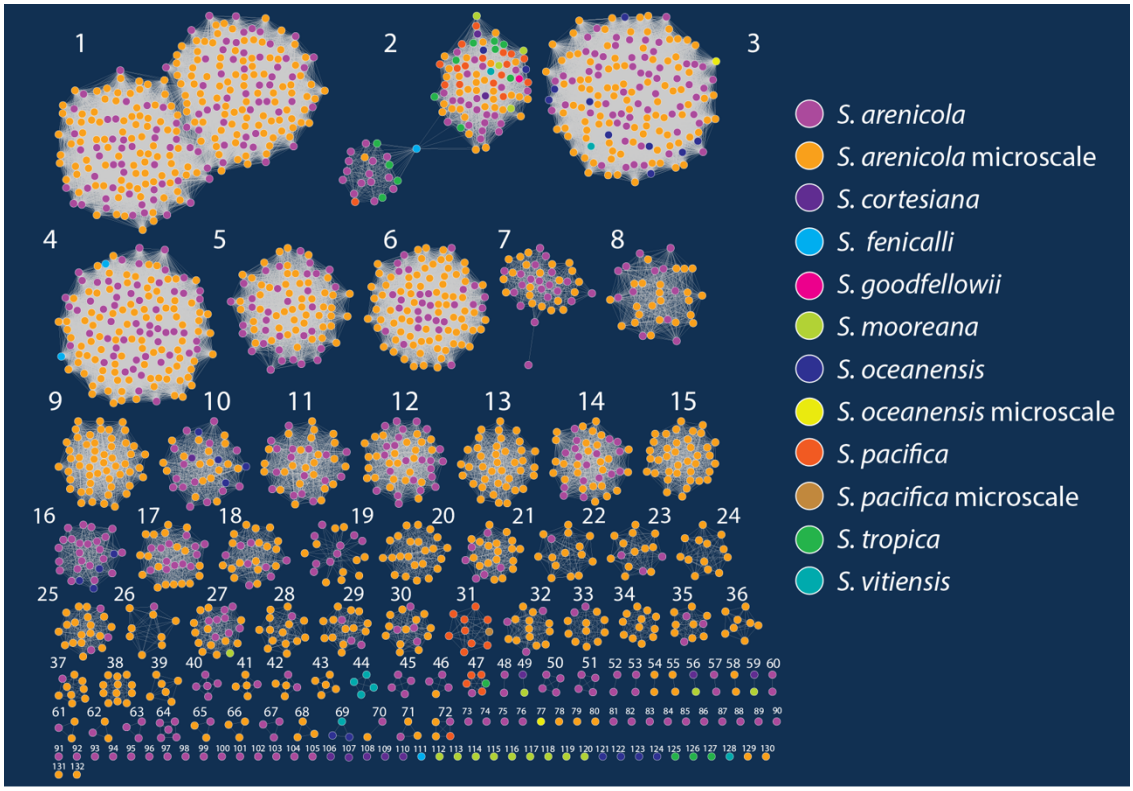


Figure 6.7. Sequence similarity network of lanthipeptide precursor peptides from all 217 *Salinispora* genomes.

Nodes are colored by Salinispora species (1,825 nodes; 68,872 edges; only showing edges >50% sequence identity).

Next, I wanted to understand if these *Salinispora* lanthipeptide precursor peptides shared any similarity based on the location the *Salinispora* was isolated from. While most of the precursor peptides were discovered from *Salinispora* isolated in Fiji (because all microscale *Salinispora* were isolated from Fiji), I did observe multiple location-specific clusters of precursor peptides (**Figure 6.8**). In many cases, this matched with *Salinispora* species that were isolated from the

same location, however cluster 5 and 50 are examples of four precursor peptides from *Salinispora arenicola* macroscale that share similarity even though they were isolated from the Bahamas, Yucatán, and the Sea of Cortez (**Figure 6.8**). Any clusters within the same species yet detected across distant locations could be evidence of the BGC being fixed prior to distribution. It is also possible that some of the location-specific precursor peptides could yield novel chemistry. To truly compare the diversity of lanthipeptide precursors from Fiji, I next colored the sequence similarity network by the location from which the strains were isolated from the $1m^2$ quadrant (**Figure 6.9**).

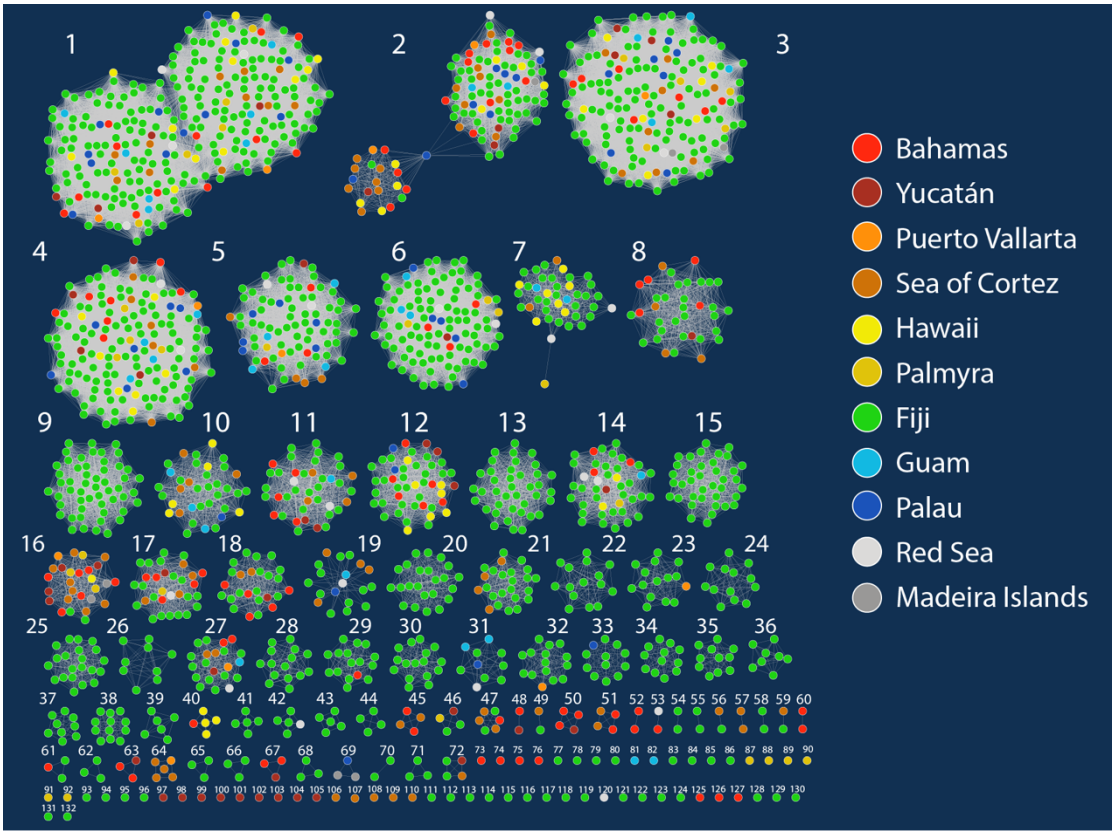


Figure 6.8. Sequence similarity network of lanthipeptide precursor peptides from all 217 Salinispora genomes.

Nodes are colored by *Salinispora* isolation location (1,825 nodes; 68,872 edges; only showing edges >50% sequence identity).

Even among the 16 sub-quadrant locations, it did not appear that lanthipeptide precursor peptide diversity was separated on a microscale (**Figure 6.9**). Cluster 38 could have a slight distribution of all the precursor peptides were from the lower half of the sub-quadrant; similarly in cluster 24 with the left triangular half of the quadrant; however, these patterns are only speculative. This result supports our earlier conclusion (Chapter 5 of this dissertation) that there was little to no microscale genome similarity driven by sub-quadrant isolation location.

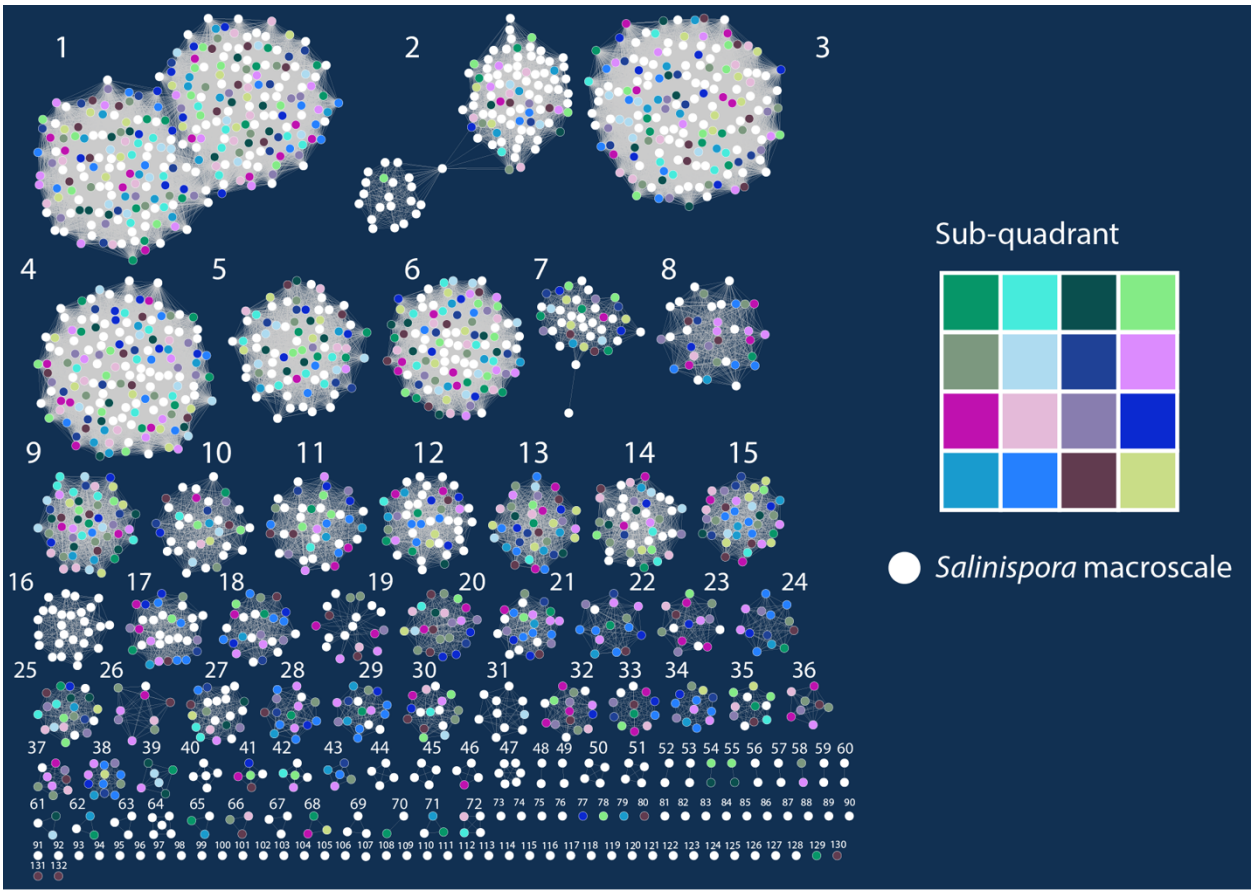


Figure 6.9. Sequence similarity network of lanthipeptide precursor peptides from all 217 *Salinispora* genomes.

Nodes are colored by microscale *Salinispora* sub-quadrant location or macroscale (1,825 nodes; 68,872 edges; only showing edges >50% sequence identity).

While I believe that we have captured the potential *Salinispora* lanthipeptide chemical diversity in the genome sequences, I wanted to understand if any precursor peptides shared similarities with characterized RiPPs. I did this in two ways, utilizing the community power of central repositories and openly accessible data in publications. From the MIBiG 2.0 database of BGCs linked to their products, I mined out 353 precursor peptides from RiPP BGCs. By building a sequence similarity network of the MIBiG precursor peptides and our *Salinispora* precursor peptides, I discovered that only 2 clusters shared similarity to known RiPPs (**Figure 6.10**). The first cluster (number 59) is a doublet from *S. mooreana* and *S. cortesiana* (100% identical), and it shared high sequence similarity to the Gram-positive antibiotics actagardine (GarA, 88% identity), michiganin A (ClvA, 61% identity), and mersacidin (MrsA, 37% identity) (**Figure 6.10**). The predicted core peptides from both *Salinispora* species have the conserved lantibiotic 3-ring structure, indicating that the two *Salinispora* products could have similar cell wall biosynthesis inhibitory properties. The other cluster with similarity to MIBiG was cluster 122, which is predicted to be a class V lanthipeptide, the first report of the newly described class V lanthipeptides in *Salinispora* (**Figure 6.10**). Singleton 122 shares similarity with thioviridamide, neothioviridamide, and thioholgamide precursor peptides. Other MIBiG precursor peptides clustered more closely to one another than *Salinispora* precursor peptides, forming clusters of related compounds as labeled **(Figure 6.10**).

While it was exciting that there appears to be a large potential of novel lanthipeptide chemical diversity, I wanted to ascertain if any other precursor peptides from bacterial genomes shared similarity, and if any additional information about the putative class of lanthipeptide could be assigned. I thus created a sequence similarity network of all *Salinispora* precursor peptides with the collection of lanthipeptide class I-IV precursor peptides from 100,000 reference bacterial

genomes (Walker *et al.*, 2020) **(Figure 6.11**). These precursor peptides were identified using the RODEO 2.0 tool which can accurately predict the class of lanthipeptides (Walker *et al.*, 2020); thus this was a powerful dataset to compare with ours, with the goal to assign classes I-IV to our unknown clusters. From the Walker 2020 dataset I mined out: 2,698 class I; 3,002 class II; 2,304 class III, and 401 class IV lanthipeptide precursor peptides. When clustered together, I observed seven *Salinispora* precursor peptides that clustered with class I-IV nodes **(Figure 6.11**). Cluster 59 clustered with many class II precursor peptides; this is the cluster that also shared similarity to the known class II lantibiotic actagardine, so this clustering made sense **(Figure 6.11**). Cluster 4, a large cluster of many micro- and macroscale *Salinispora*, and two *S. fenicalli* nodes shared similarity to class II precursor peptides, thus giving us confidence that cluster 4 is class II **(Figure 6.11**).

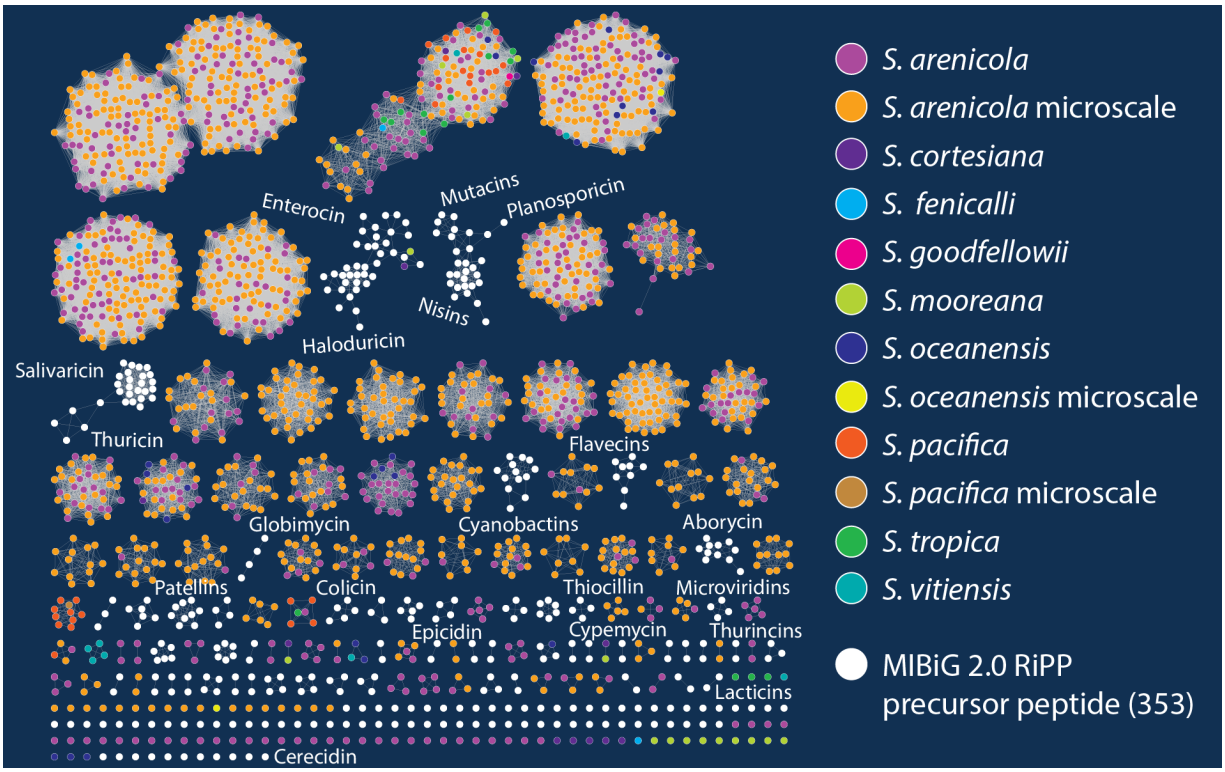


Figure 6.10. Sequence similarity network of lanthipeptide precursor peptides from all 217 Salinispora genomes and 353 precursor peptides from MIBiG 2.0.

Nodes are colored by Salinispora species and MIBiG (2,178 nodes; 69,908 edges; only showing edges with > 28% sequence identity).

Cluster 3 also clustered with a single class II precursor peptide from an *Actinoplanes*, giving us confidence that the large group of cluster 3 precursor peptides from many species of *Salinispora* are class II and could have an interesting evolutionary history with the *Actinoplanes* BGC. **(Figure 6.11)** Cluster 2 clustered with many class I precursor peptides; similarly, cluster 125 (a singleton from *S. tropica*) clustered with three class I precursor peptides **(Figure 6.11)**. The *Salinispora pacifica* species-specific cluster 31 clustered with class III precursor peptides and likewise, the species-specific *S. vitiensis* cluster 44 clustered with other class III precursor peptides **(Figure 6.11)**. This shows that *Salinispora* is a lanthipeptide-privileged genus, with precursor peptides and BGCs from four out of five classes of lanthipeptides. In fact, cluster 126 from *S. tropica* is encodes an additional class V precursor peptide, as identified by antiSMASH 6.0, that

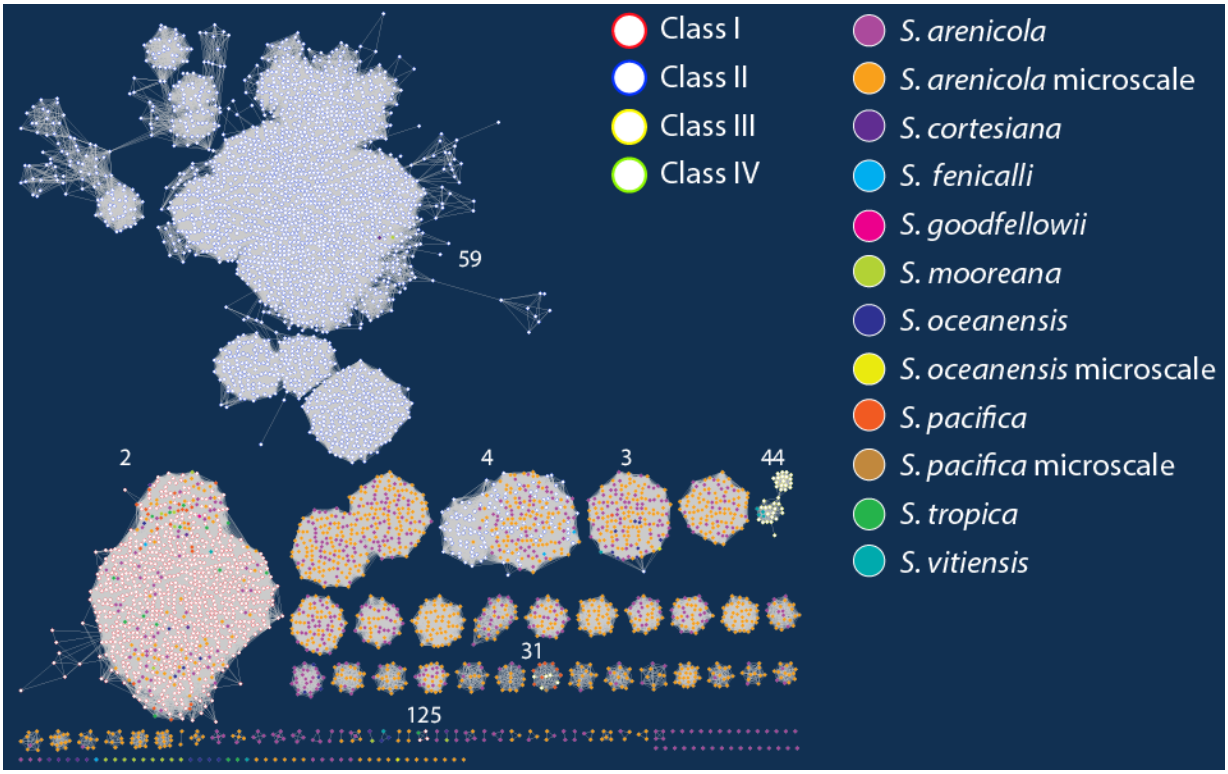


Figure 6.11. Sequence similarity network of lanthipeptide precursor peptides from all 217 Salinispora genomes and 8,405 precursor peptides from Walker et al. 2020.

Nodes are colored by *Salinispora* species and Walker et al. 2020 class I, II, III, or IV (10,230 nodes; 494,667 edges > 22% identity are shown).

did not cluster with other class V precursors like cluster 122. Both class V lanthipeptide BGCs (cluster 122, 126) look different from the three class V lanthipeptides that have been described to date, so this could be a promising lead for the expansion of the class V lanthipeptides.

To investigate the amino acid diversity of the precursor peptides that shared similarity to known lanthipeptides, I created multiple sequence alignments. Alignments of cluster 59 to the actagardine (GarA), michiganin A (ClvA), and mersacidin (MrsA) precursor peptides demonstrated that the leader had conserved residues, likely important for the recognition by the LanM enzyme, and the core peptide is mostly conserved except for one isoleucine amino acid change compared to actagardine and three other amino acid differences between the other two compounds (**Figure 6.12**). The key amino acids that form the 3-ring structure important for the Gram-positive antibiotic activity is conserved however, which means cluster 59 could be targeted with specific bioassay-guided fractionation isolation schemes. In fact, I inspected previously published transcriptomic data available for the *S. mooreana* CNT-150 (Amos *et al.*, 2017) and discovered that both the precursor peptide and key LanM gene, along with neighboring transcriptional regulators and ABC transporters were significantly upregulated at both 96 and 216 hours of growth (**Figure 6.13**).

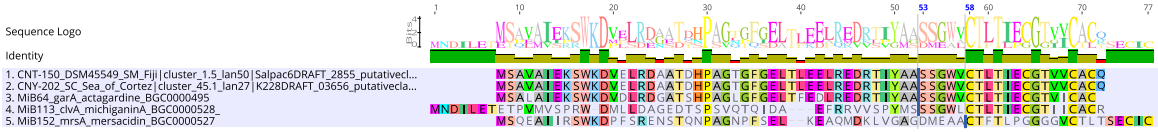


Figure 6.12. Multiple sequence alignment of cluster 59 (*S. mooreana* and *S. cortesiana*) with actagardine (GarA), michiganin A (ClvA) , and mersacidin (MrsA) precursor peptide.

Cursor bar indicates where the leader and core peptide would be split.

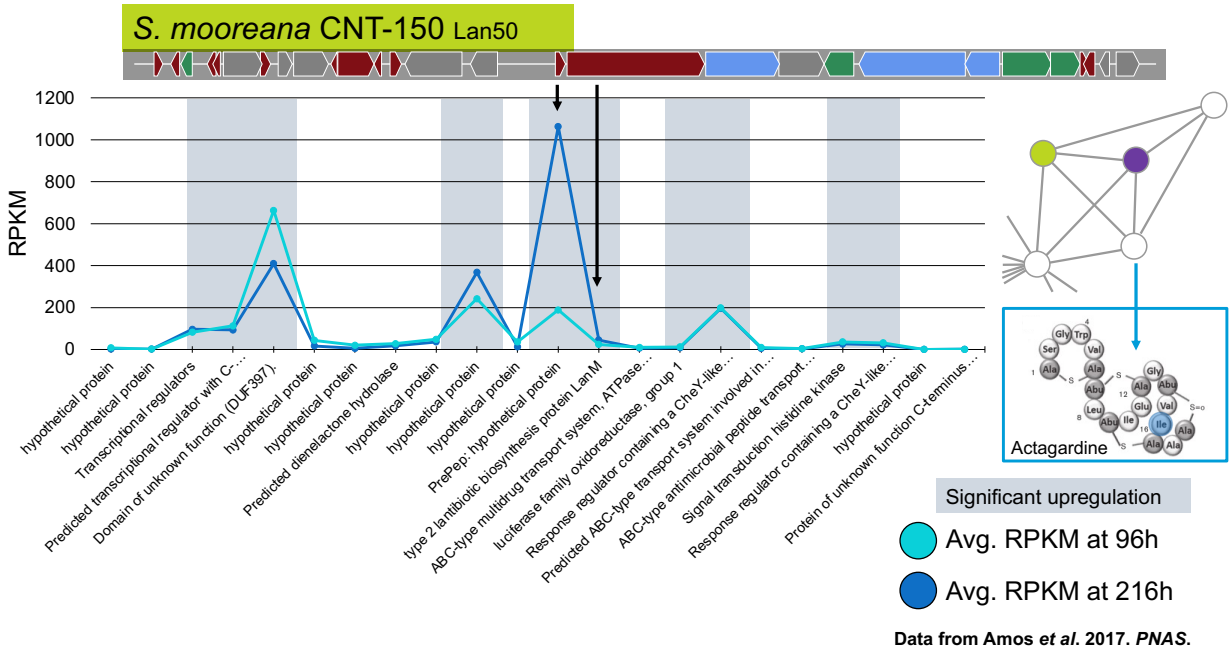


Figure 6.13. Expression (RPKM) of *Salinispora mooreana* CNT-150 grown for 96h and 216h showed key regulator, precursor peptide, LanM, and ABC transporter genes were significantly upregulated (shaded bars).

Significant upregulation was defined as in Amos et al. 2017 "> 27.1 reads per kilobase of transcript per million mapped reads (RPKM)".

I similarly inspected the alignment of cluster 122, the lanthipeptide class V precursor peptide with similarity to thioviridamide, neothrioviridamide, and thioholgamide. The leader part of the peptide was not very conserved, likely due to the evolutionary distance of the biosynthetic enzyme responsible for recognizing the leader RiPP Recognition Element (RRE) site, which for class V lanthipeptides has not been identified as they do not always contain a Lan-like gene. The core part of the peptide shared low similarity; however, the conserved amino acids T, A, and ending HC in the *Salinispora oceanensis* CNT-029 precursor peptide seemed to be somewhat conserved compared to the reported thioviridamine structure (Frattaruolo *et al.*, 2017). In fact, a *Salinispora* cluster similar to the thioviridamine BGC was reported (Frattaruolo *et al.*, 2017), however it was not described. The key class V moiety of AviCys is formed by cyclization between

a serine/cysteine-derived dehydroalanine and a C-terminal cysteine via oxidative decarboxylation and thus the conserved "HC" with the cysteine at the end as seen in the *S. oceanensis* cluster 122 seems to have the important amino acids.

Figure 6.14. Multiple sequence alignment of the class V lanthipeptide precursor peptides from *Salinispora oceanensis* CNT-029, and thioviridamide, neothrioviridamide, and thioholgamide.

The cursor bar indicates where the leader and core are predicted to split.

Finally, I was interested in understanding if particular parts of the precursor peptides in various clusters were undergoing diversification, that is, if there were amino acid differences in the leader or core part of the peptide. I created multiple sequence alignments for all key clusters in our network (data not shown in this dissertation) and observed that for some clusters, there was high conservation, yet in others, there were species-specific differences in the leader part of the peptide and amino acid differences in the core part of the peptide. It was challenging to predict where the leader and core might split from the precursor peptides based on those that have been described to date. I hypothesized that some of these RiPP BGCs have been horizontally transferred between diverse bacteria as I observed that cluster 59 from *S. mooreana* and *S. cortesiana* shared high similarity to *Micromonospora* sp. CNB-394 across the entire precursor peptide; this could be investigated by building phylogenetic trees of the BGC and core encoding enzymes (**Figure 6.15**). Similarly, cluster 44 was only seen in *S. vitiensis* and when all sequences in the cluster were aligned, they showed high leader peptide amino acid conservation, though the core amino acid composition was different in comparison with many *Streptomyces* bacteria (**Figure 6.16**). This could indicate that the lanthipeptide BGC was horizontally exchanged and the key LanKM enzyme

is conserved as the leader RiPP Recognition Element (RRE) element seems to be conserved, but there has been diversification of the ultimate core peptide product (**Figure 6.16**). Finally, I observed that in some classes of lanthipeptide precursor peptides, there was very high amino acid conservation within in the same *Salinispora* species, as seen in cluster 33 which was found in both micro- and macroscale *S. pacifica* strains (**Figure 6.17**). It is fascinating that the microscale *S. pacifica* isolated from the same location years after the macroscale *S. pacifica* strains still had complete conservation of the lanthipeptide RiPP precursor peptide, which could indicate that this is a specific class III lanthipeptide BGC is not undergoing rapid diversification compared to other clusters with more amino acid variation.

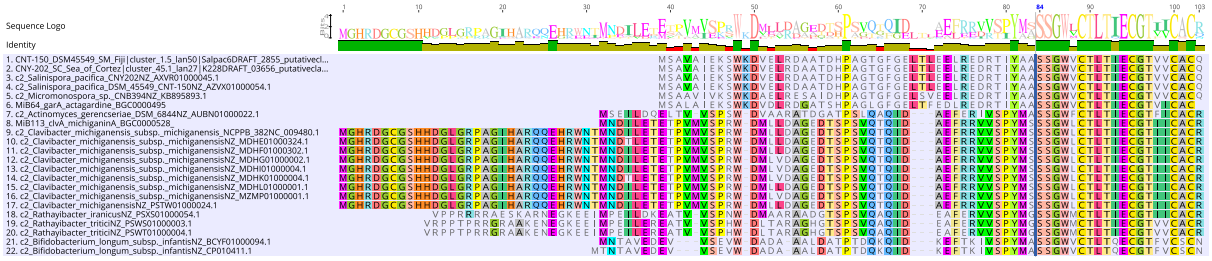


Figure 6.15. Multiple sequence alignment of the cluster 59 class II precursor peptides from *Salinispora mooreana* and *S. cortesiana* which show high similarity to a precursor peptide from *Micromonospora* sp. CNB-394 and other diverse bacteria.

The cursor bar indicates where the leader and core are predicted to split.

Figure 6.16. Multiple sequence alignment of the cluster 44 class III precursor peptides from *S. vitiensis* which show conserved leader sequence but not core compared to many *Streptomyces* bacteria.

The cursor bar indicates where the leader and core are predicted to split.

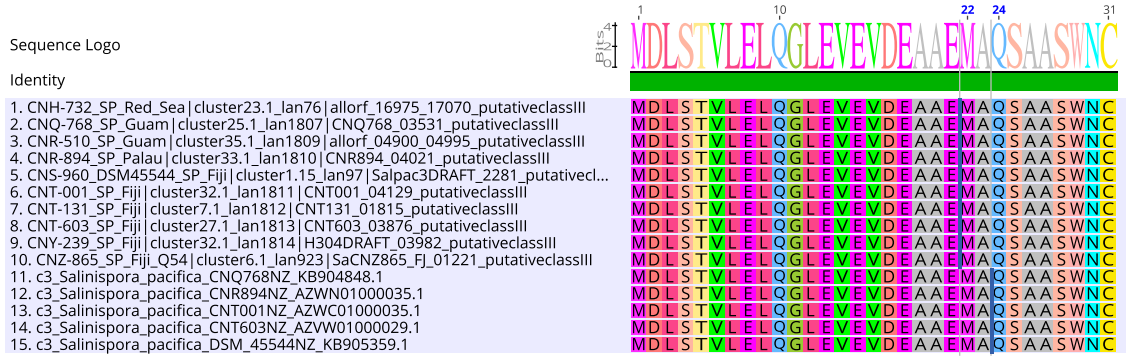


Figure 6.17. Multiple sequence alignment of the cluster 33 class III precursor peptides from *Salinispora pacifica* which show high similarity to another, including a new microscale strain.

The cursor bar indicates where the leader and core peptide are predicted to split.

6.5 Discussion

A previous analysis of the lanthipeptide BGCs in *Salinispora* identified 182 BGCs that comprised 7 groups based on sequence similarities of their core biosynthetic lanthionine enzymes (Kittrell *et al.*, 2020). From these BGCs, class I and II lanthipeptide precursor peptides were predicted and 92% of all *Salinispora* genomes were found to have a predicted lanthipeptide BGC and core peptide (Kittrell *et al.*, 2020). Based on a sequence similarity network of precursor peptides and comparison of the complete BGC organization, it was reported that the *Salinispora* lanthipeptide BGC scaffolds were novel compared to characterized RiPPs. I sought to build on this important foundational work by comparing the lanthipeptide biosynthetic diversity in our new microscale *Salinispora* genomes to the macroscale *Salinispora*. In this chapter, I used genome-mining techniques to identify all RiPP BGCs in all 217 micro- and macroscale *Salinispora* strains and discovered 528 lanthipeptide BGCs and their precursor peptides. Most *Salinispora* genomes contained at least one lanthipeptide BGC, however, *S. arenicola* lanthipeptide BGCs on average had more than one precursor peptide encoded in the BGC. This could indicate that *S. arenicola* utilizes a wider diversity of lanthipeptide compounds for various functions or the BGCs are undergoing diversification. Overall, the *Salinispora* genus has a wide distribution of RiPP gene cluster families (GCFs), which each contain related BGCs. Even for the microscale *S. arenicola* genomes, there were differences in RiPP GCF distribution which could mean these BGCs are being gained, lost, exchanged, or modified commonly even within a species on a close spatial scale.

Instead of comparing full lanthipeptide BGCs to one another, I defined each lanthipeptide precursor peptide sequence as its own operational biosynthetic unit for comparison, akin to the use of ketosynthase (KS) and condensation (C) domains as units to define PKS and NRPS BGC

patterns in *Salinispora* (Ziemert *et al.*, 2014). Additionally, my designation makes logical sense as each precursor peptide is the core scaffold that determines the unique lanthipeptide structure. Surprisingly, I identified a total of 1,825 lanthipeptide precursor peptide sequences. By creating a sequence similarity network (Gerlt *et al.*, 2015; Gerlt, 2017) of the *Salinispora* lanthipeptide precursor peptides, there were 132 clusters and singletons of related precursor peptides, clearly indicating there is both a large diversity of lanthipeptide precursor peptides across the micro- and macroscale *Salinispora*. I expanded this analysis by mining out every biosynthetically characterized RiPP precursor peptide from the MIBiG 2.0 BGC database and a recent analysis of 100,000 reference bacterial genomes which helped to further identify 7 clusters. I describe for the first time that *Salinispora* has class I, II, III, and V lanthipeptide biosynthetic potential—four out of the five lanthipeptide molecule types, which expands upon a previous analysis (Kittrell *et al.*, 2020). Many of the clusters of precursor peptides had relatively high amino acid conservation in both the leader and core peptide, but species-specific patterns among the clusters were observed, and multiple clusters that were rare in *Salinispora* had closest hits to other Actinobacteria including *Actinoplanes, Micromonospora,* and *Streptomyces*. This could be further evidence that some *Salinispora* have acquired RiPP gene clusters from closely related bacteria. Further evolutionary analyses of the core BGC genes will help elucidate these patterns.

This collection of *Salinispora* RiPP BGCs and precursor peptides aggregated were acquired from two different spatial scales and will serve as an important baseline for targeting lanthipeptide RiPP products for elucidation. Specifically, lanthipeptide RiPPs are ideal for heterologous expression because 1) they are smaller in size than PKS and NRPS BGCs, and 2) due to their biosynthesis being reliant on the ribosomal translation of the precursor peptide, lanthipeptide RiPPs do not require genus or phyla specific enzymes and have proven amendable

to expression, modification, and characterization in *E. coli* vectors (Zhao and Van Der Donk, 2016; Y. Zhang *et al.*, 2018; Zhang *et al.*, 2019; Viel *et al.*, 2021). In fact, many recent RiPP studies have successfully captured entire lanthipeptide RiPP BGCs in *E. coli* and subsequently employed genetic modification techniques to modify and study the lanthipeptide RiPP chemical structures (Zhang *et al.*, 2014, 2015; Zhao and Van Der Donk, 2016; Burkhart *et al.*, 2017; Montalbán-López *et al.*, 2017; Chen *et al.*, 2018; Hegemann and van der Donk, 2018; Ren *et al.*, 2018; Si *et al.*, 2018; Y. Zhang *et al.*, 2018).

To test for structural diversification of the lanthipeptide core peptides in *Salinispora*, lanthipeptide RiPP BGCs could be captured in *E. coli* or compared with similar captured clusters. As class I-V lanthipeptides are known to be potent lantibiotics targeting the cell-wall integrity and biosynthetic components, the products of these lanthipeptide BGCs could have exciting biological activities against important pathogens (Li *et al.*, 2021). The chemical characterization of *Salinispora* RiPPs could be facilitated by modern mass spectrometry analysis tools helpful for identifying RiPP peptide ionization fragments linked to RiPP BGCs, as described in (Kersten *et al.*, 2011; Mohimani *et al.*, 2014; Wang *et al.*, 2016; Mohimani, Gurevich, Alexander, *et al.*, 2017; Mohimani, Gurevich, Mikheenko, *et al.*, 2017; Gurevich *et al.*, 2018; Reher *et al.*, 2020).

Finally, to understand the evolutionary diversification of lanthipeptide RiPPs in *Salinispora*, phylogenetic trees of the core lanthipeptide enzymes could be constructed. Coupled with phylogenetic trees of clusters of related *Salinispora* precursor peptides, it will make it possible to establish class or taxa specific patterns driving the evolutionary dynamics in *Salinispora* lanthipeptide RiPP BGCs, especially between the micro- and macroscale *Salinispora* genomes.

6.6 Acknowledgements

I would like to thank Prof. Greg Rouse, Prof. Jon Chekan, and Dr. Sheila Podell for helpful analysis suggestions and discussions.

Chapter 6 is coauthored with Paul R. Jensen. The dissertation author was the primary investigator and author of this chapter.

6.7 References

Agrawal, P., Amir, S., Deepak, Barua, D., and Mohanty, D. (2021) RiPPMiner-Genome: A Web Resource for Automated Prediction of Crosslinked Chemical Structures of RiPPs by Genome Mining. *J Mol Biol* **433**: 166887.

Agrawal, P., Khater, S., Gupta, M., Sain, N., and Mohanty, D. (2017) RiPPMiner: a bioinformatics resource for deciphering chemical structures of RiPPs based on prediction of cleavage and cross-links. *Nucleic Acids Res* **13**: 470–478.

Amos, G.C.A., Awakawa, T., Tuttle, R.N., Letzel, A.-C., Kim, M.C., Kudo, Y., Fenical, W., Moore, B.S., and Jensen, P.R. (2017) Comparative Transcriptomics as a Guide to Natural Product Discovery and Biosynthetic Gene Cluster Functionality. *PNAS* **114**: E11121–E11130.

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

Blin, K., Shaw, S., Kloosterman, A.M., Charlop-Powers, Z., van Wezel, G.P., Medema, M.H., and

Weber, T. (2021) antiSMASH 6.0: improving cluster detection and comparison capabilities. *Nucleic Acids Res* 1–7.

Blin, K., Shaw, S., Steinke, K., Villebro, R., Ziemert, N., Lee, S.Y., Medema, M.H., and Weber, T. (2019) antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline. *Nucleic Acids Res* **47**: W81–W87.

Blin, K., Wolf, T., Chevrette, M.G., Lu, X., Schwalen, C.J., Kautsar, S.A., Suarez Duran, H.G., de los Santos, E.L.C., Kim, H.U., Nave, M., Dickschat, J.S., Mitchell, D.A., Shelest, E., Breitling, R., Takano, E., Lee, S.Y., Weber, T., and Medema, M.H. (2017) antiSMASH 4.0—improvements in chemistry prediction and gene cluster boundary identification. *Nucleic Acids Res* **45**: 1–6.

Bobeica, S.C. and van der Donk, W.A. (2018) The Enzymology of Prochlorosin Biosynthesis, 1st ed. Elsevier Inc.

Bruns, H., Crüsemann, M., Letzel, A.-C., Alanjary, M., Mcinerney, J.O., Jensen, P.R., Schulz, S., Moore, B.S., and Ziemert, N. (2017) Function-related replacement of bacterial siderophore pathways. *ISME J* **12**: 320–329.

Burkhart, B.J., Kakkar, N., Hudson, G.A., Van Der Donk, W.A., and Mitchell, D.A. (2017) Chimeric Leader Peptides for the Generation of Non-Natural Hybrid RiPP Products. *ACS Cent Sci* **3**: 629–638.

Carlin, D.E., Demchak, B., Pratt, D., Sage, E., and Ideker, T. (2017) Network propagation in the cytoscape cyberinfrastructure. 1–9.

Castro-Falcón, G., Creamer, K.E., Chase, A.B., Kim, M.C., Sweeney, D., Glukhov, E., Fenical, W., and Jensen, P.R. (2022) Structure and Candidate Biosynthetic Gene Cluster of a Manumycin-Type Metabolite from Salinispora pacifica. *J Nat Prod* **85**: 980–986.

Chase, A.B., Sweeney, D., Guillén-matus, D.G., and Jensen, P.R. (2021) Vertical Inheritance Facilitates Interspecies Diversification in Biosynthetic Gene Clusters and Specialized Metabolites. *MBio*.

Chen, H., Zhang, Y., Li, Q., Zhao, Y., Chen, Y., and Li, Y. (2018) De Novo Design To Synthesize Lanthipeptides Involving Cascade Cysteine Reactions: SapB Synthesis as an Example. 8–13.

Creamer, K.E., Kudo, Y., Moore, B.S., and Jensen, P.R. (2021) Phylogenetic analysis of the salinipostin γ-butyrolactone gene cluster uncovers new potential for bacterial signalling-molecule diversity. *Microb Genomics* **7**: 1–14.

Cubillos-ruiz, A., Berta-thompson, J.W., Becker, J.W., and Donk, W.A. Van Der (2017) Evolutionary radiation of lanthipeptides in marine cyanobacteria. *Proc Natl Acad Sci U S A* 201700990.

Cubillos-Ruiz, A., Berta-Thompson, J.W., Becker, J.W., Van Der Donk, W.A., and Chisholm, S.W. (2017) Evolutionary radiation of lanthipeptides in marine cyanobacteria. *Proc Natl*

Acad Sci U S A 201700990.

van der Donk, W.A. and Nair, S.K. (2014) Structure and mechanism of lanthipeptide biosynthetic enzymes. *Curr Opin Struct Biol* **29**: 58–66.

Frattaruolo, L., Lacret, R., Cappello, A.R., and Truman, A.W. (2017) A Genomics-Based Approach Identifies a Thioviridamide-Like Compound with Selective Anticancer Activity. *ACS Chem Biol* **12**: 2815–2822.

Gerlt, J.A. (2017) Genomic Enzymology: Web Tools for Leveraging Protein Family Sequence-Function Space and Genome Context to Discover Novel Functions. *Biochemistry* **56**: 4293–4308.

Gerlt, J.A., Bouvier, J.T., Davidson, D.B., Imker, H.J., Sadkhin, B., Slater, D.R., and Whalen, K.L. (2015) Enzyme function initiative-enzyme similarity tool (EFI-EST): A web tool for generating protein sequence similarity networks. *Biochim Biophys Acta - Proteins Proteomics* **1854**: 1019–1037.

Gurevich, A., Mikheenko, A., Shlemov, A., Korobeynikov, A., Mohimani, H., and Pevzner, P.A. (2018) Increased diversity of peptidic natural products revealed by modification-tolerant database search of mass spectra. *Nat Microbiol* **3**: 319–327.

van Heel, A.J., de Jong, A., Montalbán-López, M., Kok, J., and Kuipers, O.P. (2013) BAGEL3: Automated identification of genes encoding bacteriocins and (non-)bactericidal posttranslationally modified peptides. *Nucleic Acids Res* **41**: 448–453.

van Heel, A.J., de Jong, A., Song, C., Viel, J.H., Kok, J., and Kuipers, O.P. (2018) BAGEL4: a user-friendly web server to thoroughly mine RiPPs and bacteriocins. *Nucleic Acids Res* **46**: 1–4.

Hegemann, J.D. and van der Donk, W.A. (2018) Investigation of Substrate Recognition and Biosynthesis in Class IV Lanthipeptide Systems. *J Am Chem Soc* jacs.8b01323.

Hudson, G.A., Burkhart, B.J., DiCaprio, A.J., Schwalen, C.J., Kille, B., Pogorelov, T. V., and Mitchell, D.A. (2019) Bioinformatic Mapping of Radical S -Adenosylmethionine-Dependent Ribosomally Synthesized and Post-Translationally Modified Peptides Identifies New Cα, Cβ, and Cγ-Linked Thioether-Containing Peptides . *J Am Chem Soc* **141**: 8228–8238.

Hudson, G.A. and Mitchell, D.A. (2018) RiPP antibiotics: biosynthesis and engineering potential. *Curr Opin Microbiol* **45**: 61–69.

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

Kautsar, S.A., Blin, K., Shaw, S., Navarro-Muñoz, J.C., Terlouw, B.R., van der Hooft, J.J.J., van Santen, J.A., Tracanna, V., Suarez Duran, H.G., Pascal Andreu, V., Selem-Mojica, N., Alanjary, M., Robinson, S.L., Lund, G., Epstein, S.C., Sisto, A.C., Charkoudian, L.K., Collemare, J., Linington, R.G., Weber, T., and Medema, M.H. (2020) MIBiG 2.0: a repository

for biosynthetic gene clusters of known function. *Nucleic Acids Res* **48**: D454–D458.

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C., Thierer, T., Ashton, B., Meintjes, P., and Drummond, A. (2012) Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **28**: 1647–1649.

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

Kittrell, C.G., Shah, S.C., Halbert, M.E., Scott, D.H., and Limbrick, E.M. (2020) Genomic analysis suggests Salinispora is a rich source of novel lanthipeptides. *Mol Genet Genomics*.

Kloosterman, A.M., Cimermancic, P., Elsayed, S.S., Du, C., Hadjithomas, M., Donia, M.S., Fischbach, M.A., van Wezel, G.P., and Medema, M.H. (2020) Expansion of RiPP biosynthetic space through integration of pan-genomics and machine learning uncovers a novel class of lantibiotics.

Laurenceau, R., Raho, N., Cariani, Z., Bliem, C., Osman, M.A.M., and Chisholm, S.W. (2020) Association of lanthipeptide genes with TnpAREP transposases in marine picocyanobacteria. *bioRxiv* 2020.03.09.984088.

Letzel, A.-C., Li, J., Amos, G.C.A., Millán-Aguiñaga, N., Ginigini, J., Abdelmohsen, U.R., Gaudêncio, S.P., Ziemert, N., Moore, B.S., and Jensen, P.R. (2017) Genomic insights into specialized metabolism in the marine actinomycete *Salinispora*. *Environ Microbiol* **19**: 3660–3673.

Li, C., Alam, K., Zhao, Y., Hao, J., Yang, Q., Zhang, Y., Li, R., and Li, A. (2021) Mining and Biosynthesis of Bioactive Lanthipeptides From Microorganisms. *Front Bioeng Biotechnol* **9**: 1–13.

Mohimani, H., Gurevich, A., Alexander, K.L., Benjamin, C., Leão, T., Glukhov, E., Moss, N.A., Luzzatto-knaan, T., Vargas, F., Nothias, L., Singh, N.K., Sanders, J.G., Benitez, A.S., Thompson, L.R., Hamid, M., Morton, J.T., Shlemov, A., Korobeynikov, A., Friedberg, I., Knight, R., Venkateswaran, K., Gerwick, W., Dorrestein, P.C., and Pevzner, P.A. (2017) MetaRiPPquest : A Peptidogenomics Approach for the Discovery of Ribosomally Synthesized and Post- translationally Modified Peptides. 1–25.

Mohimani, H., Gurevich, A., Mikheenko, A., Garg, N., Nothias, L.F., Ninomiya, A., Takada, K., Dorrestein, P.C., and Pevzner, P.A. (2017) Dereplication of peptidic natural products through database search of mass spectra. *Nat Chem Biol* **13**: 30–37.

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

Montalbán-López, M., van Heel, A.J., and Kuipers, O.P. (2017) Employing the promiscuity of

lantibiotic biosynthetic machineries to produce novel antimicrobials. *FEMS Microbiol Rev* **41**: 5–18.

Montalbán-López, M., Scott, T.A., Ramesh, S., Rahman, I.R., Van Heel, A.J., Viel, J.H., Bandarian, V., Dittmann, E., Genilloud, O., Goto, Y., Grande Burgos, M.J., Hill, C., Kim, S., Koehnke, J., Latham, J.A., Link, A.J., Martínez, B., Nair, S.K., Nicolet, Y., Rebuffat, S., Sahl, H.G., Sareen, D., Schmidt, E.W., Schmitt, L., Severinov, K., Süssmuth, R.D., Truman, A.W., Wang, H., Weng, J.K., Van Wezel, G.P., Zhang, Q., Zhong, J., Piel, J., Mitchell, D.A., Kuipers, O.P., and Van Der Donk, W.A. (2021) New developments in RiPP discovery, enzymology and engineering. *Nat Prod Rep* **38**: 130–239.

Navarro-Muñoz, J.C., Selem-Mojica, N., Mullowney, M.W., Kautsar, S.A., Tryon, J.H., Parkinson, E.I., De Los Santos, E.L.C., Yeong, M., Cruz-Morales, P., Abubucker, S., Roeters, A., Lokhorst, W., Fernandez-Guerra, A., Cappelini, L.T.D., Goering, A.W., Thomson, R.J., Metcalf, W.W., Kelleher, N.L., Barona-Gomez, F., and Medema, M.H. (2019) A computational framework to explore large-scale biosynthetic diversity. *Nat Chem Biol* 1–9.

Ortega, M.A. and Van Der Donk, W.A. (2016) New Insights into the Biosynthetic Logic of Ribosomally Synthesized and Post-translationally Modified Peptide Natural Products. *Cell Chem Biol* **23**: 31–44.

Ortiz-López, F.J., Carretero-Molina, D., Sánchez-Hidalgo, M., Martín, J., González, I., Román-Hurtado, F., de la Cruz, M., García-Fernández, S., Reyes, F., Deisinger, J.P., Müller, A., Schneider, T., and Genilloud, O. (2020) Cacaoidin, First Member of the New Lanthidin RiPP Family. *Angew Chemie - Int Ed* 1–6.

Reher, R., Kim, H.W., Zhang, C., Mao, H.H., Wang, M., Nothias, L.F., Caraballo-Rodriguez, A.M., Glukhov, E., Teke, B., Leao, T., Alexander, K.L., Duggan, B.M., Van Everbroeck, E.L., Dorrestein, P.C., Cottrell, G.W., and Gerwick, W.H. (2020) A Convolutional Neural Network-Based Approach for the Rapid Annotation of Molecularly Diverse Natural Products. *J Am Chem Soc*.

Ren, H., Biswas, S., Ho, S., Van Der Donk, W.A., and Zhao, H. (2018) Rapid Discovery of Glycocins through Pathway Refactoring in Escherichia coli. *ACS Chem Biol*.

Repka, L.M., Chekan, J.R., Nair, S.K., and Van Der Donk, W.A. (2017) Mechanistic Understanding of Lanthipeptide Biosynthetic Enzymes. *Chem Rev* **117**: 5457–5520.

Si, T., Tian, Q., Min, Y., Zhang, L., Sweedler, J. V., van der Donk, W.A., and Zhao, H. (2018) Rapid structure-activity screening of lanthipeptide analogs via in-colony removal of leader peptides in Escherichia coli. *J Am Chem Soc* **140**: jacs.8b05544.

Skinnider, M.A., Johnston, C.W., Edgar, R.E., Dejong, C.A., Merwin, N.J., Rees, P.N., and Magarvey, N.A. (2016) Genomic charting of ribosomally synthesized natural product chemical space facilitates targeted mining. *Proc Natl Acad Sci U S A* 201609014.

Skinnider, M.A., Johnston, C.W., Gunabalasingam, M., Merwin, N.J., Kieliszek, A.M., MacLellan, R.J., Li, H., Ranieri, M.R.M., Webster, A.L.H., Cao, M.P.T., Pfeifle, A., Spencer,

N., To, Q.H., Wallace, D.P., Dejong, C.A., and Magarvey, N.A. (2020) Comprehensive prediction of secondary metabolite structure and biological activity from microbial genome sequences. *Nat Commun* **11**: 1–9.

Skinnider, M.A., Merwin, N.J., Johnston, C.W., and Magarvey, N.A. (2017) PRISM 3: Expanded prediction of natural product chemical structures from microbial genomes. *Nucleic Acids Res* **45**: W49–W54.

Tang, X., Li, J., Mill??n-Agui??aga, N., Zhang, J.J., O'Neill, E.C., Ugalde, J.A., Jensen, P.R., Mantovani, S.M., and Moore, B.S. (2015) Identification of Thiotetronic Acid Antibiotic Biosynthetic Pathways by Target-directed Genome Mining. *ACS Chem Biol* **10**: 2841–2849.

Tietz, J.I., Schwalen, C.J., Patel, P.S., Maxson, T., Blair, P.M., Tai, H.C., Zakai, U.I., and Mitchell, D.A. (2017) A new genome-mining tool redefines the lasso peptide biosynthetic landscape. *Nat Chem Biol* **13**: 470–478.

Ting, C.P., Funk, M.A., Halaby, S.L., Zhang, Z., Gonen, T., and van der Donk, W.A. (2019) Use of a scaffold peptide in the biosynthesis of amino acid–derived natural products. *Science (80-)* **365**: 280–284.

Viel, J.H., Jaarsma, A.H., and Kuipers, O.P. (2021) Heterologous Expression of Mersacidin in Escherichia coli Elucidates the Mode of Leader Processing . *ACS Synth Biol* **10**: 600–608.

Walker, M.C., Eslami, S.M., Hetrick, K.J., Ackenhusen, S.E., Mitchell, D.A., and Van Der Donk, W.A. (2020) Precursor peptide-targeted mining of more than one hundred thousand genomes expands the lanthipeptide natural product family. *BMC Genomics* **21**: 1–17.

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., Crüsemann, M., Boudreau, P.D., Esquenazi, E., Sandoval-Calderón, 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., Boya P, C.A., 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., Rodríguez, 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., Müller, R., Waters, K.M., Shi, W., Liu, X., Zhang, L., Knight, R., Jensen, P.R., Palsson, B.Ø., Pogliano, K., Linington, R.G., Gutiérrez, M., Lopes, N.P., Gerwick, W.H., Moore, B.S., Dorrestein, P.C., and Bandeira, N. (2016) Sharing and community curation of mass spectrometry data

with Global Natural Products Social Molecular Networking. *Nat Biotechnol* **34**: 828–837.

Wiebach, Vincent, Mainz, A., Siegert, M.-A.J., Lesquame, G., Tirat, S., Dreux-Zigha, A., Aszodi, J., Le Beller, D., and Süssmuth, R.D. (2018) The anti-staphylococcal lipolanthines are ribosomally synthesized lipopeptides. *Nat Chem Biol* **14**:.

Williams, D.E., Morgan, K.D., Dalisay, D.S., Matainaho, T., Perrachon, E., Viller, N., Delcroix, M., Gauchot, J., Niikura, H., Patrick, B.O., Ryan, K.S., and Andersen, R.J. (2022) Natural Products Produced in Culture by Biosynthetically Talented Salinispora arenicola Strains Isolated from Northeastern and South Pacific Marine Sediments. *Molecules* **27**:.

Xu, M., Zhang, F., Cheng, Z., Bashiri, G., Wang, J., Hong, J., Wang, Y., Xu, L., Chen, X., Huang, S.X., Lin, S., Deng, Z., and Tao, M. (2020) Functional Genome Mining Reveals a Class V Lanthipeptide Containing a d-Amino Acid Introduced by an F420H2-Dependent Reductase. *Angew Chemie - Int Ed* **59**: 18029–18035.

Yu, Y., Zhang, Q., and Van Der Donk, W.A. (2013) Insights into the evolution of lanthipeptide biosynthesis. *Protein Sci* **22**: 1478–1489.

Zallot, R., Oberg, N., and Gerlt, J.A. (2021) Discovery of new enzymatic functions and metabolic pathways using genomic enzymology web tools. *Curr Opin Biotechnol* **69**: 77–90.

Zallot, R., Oberg, N., and Gerlt, J.A. (2019) The EFI Web Resource for Genomic Enzymology Tools: Leveraging Protein, Genome, and Metagenome Databases to Discover Novel Enzymes and Metabolic Pathways. *Biochemistry*.

Zhang, J.J., Moore, B.S., Tang, X., and Moore, B.S. (2018) Engineering Salinispora tropica for heterologous expression of natural product biosynthetic gene clusters.

Zhang, J.J., Tang, X., and Moore, B.S. (2019) Genetic platforms for heterologous expression of microbial natural products. *Nat Prod Rep*.

Zhang, Q., Doroghazi, J.R., Zhao, X., Walker, M.C., and van der Donka, W.A. (2015) Expanded natural product diversity revealed by analysis of lanthipeptide-like gene clusters in Actinobacteria. *Appl Environ Microbiol* **81**: 4339–4350.

Zhang, Q., Yang, X., Wang, H., and Van Der Donk, W.A. (2014) High divergence of the precursor peptides in combinatorial lanthipeptide biosynthesis. *ACS Chem Biol* **9**: 2686–2694.

Zhang, Q., Yu, Y., Velasquez, J.E., and van der Donk, W.A. (2012) Evolution of lanthipeptide synthetases. *Proc Natl Acad Sci* **109**: 18361–18366.

Zhang, Y., Chen, M., Bruner, S.D., and Ding, Y. (2018) Heterologous Production of Microbial Ribosomally Synthesized and Post-translationally Modified Peptides. *Front Microbiol* **9**: 1801.

Zhao, X. and Van Der Donk, W.A. (2016) Structural Characterization and Bioactivity Analysis of the Two-Component Lantibiotic Flv System from a Ruminant Bacterium. *Cell Chem Biol* **23**:

246–256.

Ziemert, N., Lechner, A., Wietz, M., Millán-Aguiñaga, N., Chavarria, K.L., and Jensen, P.R. (2014) Diversity and evolution of secondary metabolism in the marine actinomycete genus *Salinispora*. *Proc Natl Acad Sci U S A* **111**: E1130-9.

CHAPTER 7. Final Remarks

Advanced genomic sequencing methods have revolutionized the way we can study marine bacteria and the specialized metabolites that they produce. In this dissertation, I explored how BGC diversification over evolutionary time can contribute to nature's chemical diversity. Broadly, these goals fit into three research challenges. Below, I summarize the key findings, propose future work, and discuss broader impacts of each of the research presented in this dissertation.

Challenge 1:

1) We need tools to help find new natural products and patterns of specialized metabolite biosynthetic potential.

In Chapter 2, I developed a new version of the NaPDoS2 webtool which identifies and classifies polyketide KS and non-ribosomal peptide C domains in genomic, metagenomic, and targeted-amplicon sequencing data (Ziemert *et al.*, 2012). In the development of this updated webtool, we made major upgrades to the KS classification scheme and the reference sequence database. Both of these items are significant: 1) the detailed classification scheme is unlike any available genome-mining tool; and 2) with a wider breadth of taxonomic KS domain sequence coverage, it allows extended identification of novel polyketide chemistry in clades yet to have been described. This is built on the foundations of evolutionary phylogenetic theory: by adding more homologous sequences to a database that cover a wider breadth of taxa, you can make wider predictions; and the more information you know about your key reference sequences, the more functional and categorical predictions you can make about hits that clade closely with your known

sequences. Additionally, to support analysis of larger sequencing datasets, our tool went through many iterative tests to ensure it would facilitate users' analyses. I approached these challenges with the focus of improving usability, clarity, and speed; testing sensitivity, selectivity, and accuracy; and demonstrating the applicability and functionalities of the NaPDoS2 webtool for nine different sequence datasets.

It was important to me that we establish confidence in the detection and classification capabilities of NaPDoS2. This had not been done before in any genome mining tool (to the best of our knowledge). I took inspiration from the evolutionary connection between a domain's sequence dictating its structure and thus function. This was very powerful because KS domains have a specific structure that relates to their function, and thus sequence divergence between various KS and closely related enzymes that perform a similar reaction— but not a KS specific reaction— could be used as negative and positive controls. This allowed us to perform ROC analyses and other precision and accuracy benchmarks to assess how well the tool worked for a variety of datasets and suggest improved default settings. I took this idea further and established a similar dataset for the NRPS C-domains, which weren't the focus of NaPDoS2, but should be updated in future versions. Having a negative and positive control sequence dataset for C domains will facilitate similar confidence and accuracy of prediction for the diversity of non-ribosomal peptide sequences.

The strength of our NaPDoS2 webtool comes from users being able to use it for their own dataset analyses. While this webtool is perhaps different from other command-line only workflows that I've used in my own work, it was important to me to consider that a web-based tool is more accessible than tools that require specific knowledge to implement them. We should work to remove barriers that prevent analyses and hypothesis investigation, and one way to address this is

for tools to provide relevant examples of how to best utilize the tool and interpret the results. To support these aims, I led the creation of a very extensive user's guide/webtool documentation which includes detailed descriptions of how the tool works, the logic in the classification system, and nine different NaPDoS2 tutorials that users can work through from gathering, analyzing, and interpreting datasets. I hope this will be helpful to anyone using NaPDoS2 and that it can reach beyond the usefulness only a niche research field, but will continue to no bring biologists, chemists, and bioinformaticians together in their analysis questions.

The best way to measure the application of NaPDoS2 was to apply it to my own data. In this aim, I collaborated with Hans W. Singh who led a large analysis of 240 Gbp of metagenomic data from diverse environments (not included in this dissertation). We then took this a step further and analyzed 617,968 genomes across the tree of life (Chapter 3). This analysis and new dataset of 53,713 KSs went beyond previous biosynthetic potential measurements in the usual suspects of bacteria and fungi, and instead considered archaea, viruses, plasmids, plants, algae, protists, and animals, and thus their putative KS relatedness. Additionally, NaPDoS2 was used to give more information about KS domains within the pacificamide BGC (Appendix A); and NaPDoS2 was used to determine unique biosynthetic potential in my microscale *Salinispora* genome collection. I hope a broader impact of using genome-mining tools like NaPDoS2 to find biosynthetic potential in new taxa will in tandem support actions to preserve, sustainably manage, and non-exploitively explore natural product resources on earth.

I believe that the future of NaPDoS2 includes the incorporation of AI (artificial intelligence) and machine learning techniques (Xu and Jackson, 2019; Cech *et al.*, 2021; Walker and Clardy, 2021). This would allow analyses to move far beyond the reference sequence dataset, and even make connections to types of KS and C domains in very divergent taxa. The utilization

of AI and machine learning will hopefully alleviate the incredible burden of time and work it is to keep a database like NaPDoS2 up to date when there are new types of polyketides and non-ribosomal peptides characterized every day. Additionally, I believe the implementation of AI and machine learning to the tool would facilitate the linkage of bioinformatic biosynthetic potential predictions to actual metabolomic predictions (Walker and Clardy, 2021). For example, new tools released as part of the GNPS (Global Natural Products Social Molecular Networking) (Wang *et al.*, 2016) suite of metabolomic tools now allow users to identify mass spectrometry data belonging to a polyketide or peptide type family (Mohimani *et al.*, 2017; Dührkop *et al.*, 2020). Machine learning could be used to link those predictions of actual chemicals found in an environmental sample to the KS domains also found in the sample and what class/core skeleton structure the KSs are predicted to help build, and thus connect these two queries. This would facilitate targeted isolation efforts in combination with targeted BGC capture and heterologous expression (Zhang *et al.*, 2019) or even metagenome contig re-assembly around the required key KS domains (Sanders *et al.*, 2019; Robinson *et al.*, 2021). In this way, NaPDoS2 would continue to evolve with the fast-paced computational algorithms that will certainly be essential for connecting "big data" – big sequence data and metabolomic data—to "big" results from our hypotheses.

Challenge 2:

2) We do not know the evolutionary patterns and drivers of BGC diversification in bacteria, especially the marine obligate actinomycete *Salinispora*. Do *Salinispora* grow clonally or exist as multiple sub-populations within a microscale-sampled marine sediment quadrant? Do closely related *Salinispora* share the same BGC biosynthetic

potential? Does HGT contribute to BGC diversification in *Salinispora*? Do different types of BGCs undergo different rates of evolution?

To achieve this goal, I genome sequenced 99 new *Salinispora* strains that were isolated from a $1m^2$ quadrant of marine sediment collected nearby a coral reef in Fiji. We called this a "microscale" collection of genomes due to the isolation scheme with specific spatial patterning of 16 sub-quadrants within the $1m^2$ plot; especially in comparison with the "macroscale" *Salinispora* genomes that were isolated from locations worldwide. From our culturing efforts, we discovered that the microscale *Salinispora* were not clonal. In fact, there was significant differences in biosynthetic diversity and even evidence of two or more diverging sub-populations of *Salinispora arenicola* within the quadrant. A close-up examination of one clade of four closely related *S. arenicola* strains from different sub-quadrants showed that they were different because they contained at least two unique BGCs. From the results presented in this dissertation, I propose that *Salinispora* exists as multiple sub-populations on a microscale level, and we can predictively extrapolate this knowledge to other macroscale locations. This means that the diversity within the genus is more complicated than described to date (Jensen and Mafnas, 2006; Jensen, 2010; Vidgen *et al.*, 2012; Ziemert *et al.*, 2014; Millán-Aguiñaga *et al.*, 2017; Román-Ponce *et al.*, 2020; Chase *et al.*, 2021). It seems that some closely related *Salinispora* contain some of the same biosynthetic potential, however this varied for RiPP lanthipeptide BGCs. While many strains share the same precursor peptide scaffolds, there are species-specific differences—divergences in both the BGC and precursor product. Future efforts should assess the diversity between RiPP lanthipeptide precursor peptides in micro- and macroscale *Salinispora* with statistics beyond what I presented in this work—this could a way to assess if the diversity and richness between the two spatial scales

was different or the same. Either result would indicate that what we know about *Salinispora* diversity can be inferred—and even captured—from a much smaller scale than previously recognized. For example, I was able to culture 3 species of *Salinispora* in the microscale dataset and the *S. arenicola* were distributed across the entire species clade, thus representing most of the previous *S. arenicola* species' diversity previously captured. Perhaps this will save carbon emissions of reducing travel to worldwide locations when such diversity can be found in one location. It also raises questions about the hypothesis that bacteria, *Salinispora*, are everywhere but the environment selects. Future work should focus on understanding the recombination barriers between *Salinispora* and if the sub-species diversification is driven by BGCs or other genetic material.

I find the results of genomic and biosynthetic diversity of the microscale *Salinispora* to be fascinating because we do not know how *Salinispora* grows in the environment. Does it's hyphae curl around sand grains, extend through the sediment interstitial space, form a microbial mat, or does it solely exist as spores? Just because I was able to isolate *Salinispora* doesn't mean that there are actively growing *Salinispora* in the environment. In fact, previous studies assessing marine sediment community composition predicted that *Salinispora* was a very rare member (Bull and Goodfellow, 2019; Tuttle *et al.*, 2019). We will be able to further explore this hypothesis with the microscale dataset as in conjunction with the genomes, we also deeply sequenced (> 200 million reads each!) five metagenomes from the sediment quadrant plots in Fiji. One of those plots was the quadrant that the microscale *Salinispora* genomes were isolated from. Thus, when we assemble the metagenomes, we will measure the abundance of *Salinispora* in the community and see if we can assemble *Salinispora* metagenome-assembled genomes (MAGs). It will be curious to see if able to detect *Salinispora* in such deeply sequenced metagenomes—marine sediments are

notoriously rich in microbes and thus 200 million reads still might not be enough (Saary *et al.*, 2017; Thompson *et al.*, 2017; Petras *et al.*, 2019). To our knowledge, these are the most deeply sequenced metagenomes of the marine environment based on what has been published in the literature. If we can assemble *Salinispora* bins and subsequent MAGs, it will be fascinating to see any match the genome strains that we isolated, or if they are different and show additional genetic diversification. Additionally, one aspect I am very interested in exploring is the contribution of horizontally transferred genes in the metagenome community. I want to investigate if we can assess, based on shared phylogenetic signatures, any genetic material from the metagenomic community that is shared with *Salinispora*. This might not be the case, and instead investigating the HGT dynamics of the entire metagenome could be a powerful analysis, as demonstrated in recent work cheese rinds and microalgae (Bonham *et al.*, 2017; Song *et al.*, 2021). I think this would be a powerful way to explore hypotheses of what contributes and acts as a barrier to gene flow in *Salinispora*, and if vertical inheritance really plays as important a ole as recently suggested (Chase *et al.*, 2021). Additionally, we could explore the micro-niche chemistry in relation to *Salinispora*. It has been described that there are various micro-niches with specific oxygen thresholds in the surface layer of the marine sediment (Bertics and Ziebis, 2009), and it is curious that *Salinispora* has been found to grow best in highly aerated monocultures. This makes me wonder what the depth and active growth distribution of *Salinispora* is in marine sediment, how this might contribute to the spatial organization and diversity of *Salinispora* sub-populations.

Goal 3:

3) What could be the mechanism for *Salinispora* BGC diversification, including the exchange and acquisition of BGCs? Could this be mediated by plasmids or a BGC "mobilome"?

The third goal of my dissertation was one that I was most excited to explore with my research. I became fascinated with the evolutionary dynamics observed in specialized metabolite BGCs (Jensen, 2016; Ruzzini and Clardy, 2016), however, there was a big open question of the mechanisms by which diversification and exchange is happening. There have been some fabulous reviews exploring the various evolutionary trajectories of BGCs (Chevrette *et al.*, 2020), specific domains (Nivina *et al.*, 2019, 2021; Grininger, 2020), and gene clusters in general (Wisecaver and Rokas, 2015; Rokas *et al.*, 2020). However, there is a noticeable gap from predicting horizontal gene transfer (HGT) of a BGC and the actual mechanisms facilitating the exchange. Thus, I explored one aspect of genetic material exchange in *Salinispora* —plasmids. I uncovered evidence of native plasmids in *Salinispora*, however future work will be needed to sequence and characterize them. I spent a lot of time optimizing protocols and trying many methods to isolate and sequence plasmids, but it will take new long-read sequencing to fully sequence these plasmids, and a combination of protocols that I developed to go from cell cultures to purified plasmids free of gDNA and RNA. I don't find it surprising that plasmids have not been observed in *Salinispora* before— this project was inspired by a former graduate student in the lab who observed faint plasmid gel bands when performing large-scale gDNA extractions. Over time, plasmids can be lost in laboratory cultures. Plus, the short reads sequencing technology use to sequence the original set of macroscale *Salinispora* genomes easily could have missed plasmids as they likely weren't

assembled or other physical steps during library preparation selected against the odds of seeing a plasmid contig. These challenges of plasmid characterization and sequencing still exits, but I hope the evidence I've shown here inspires others to consider plasmids and extrachromosomal elements to be important in cultured mono-isolates in laboratory settings, especially from strains that historically were believed not to have plasmids.

I think there are many exciting future experiments that could be used to test the rate, mechanisms, and barriers to BGC exchange in *Salinispora*. These future efforts should lean into the approaches developed for studying antibiotic resistance (AMR) spread between bacteria. There has been a lot of work of the dynamics plasmids, phages/viruses, and other gene transfer agents that transfer AMR (MacLean and San Millan, 2015; Frost *et al.*, 2018; Danko *et al.*, 2021). These dynamics coupled with a better understanding the cell physiology could be used understand if and how exchange is happening. For example, in *Streptomyces* it has been shown that there is a growth morphology that lacks definitive cell walls, and that these "wall-less" protoplast cells are able to take up DNA (Okanishi *et al.*, 1974; Hopwood *et al.*, 1977; Ramijan *et al.*, 2018; Claessen and Errington, 2019). In the marine environment, DNA and other nucleic acids have been shown to be stable for a some amount of time via hydrostatic interactions with sand grains and interstitial space (Stewart *et al.*, 1991). What if, *Salinispora* growing on and between the sand grains of the marine environment also at some point during their growth become competent (wall-less cells, hyphal exchange, protoplasts, etc.) and this is a mechanism by which free DNA containing BGCs, perhaps liberated by phage/viruses, can be transferred? Couple this with genomic recombination, gene loss, duplication, replication errors, and other evolutionary processes that all contribute to selection and retainment of certain genetic material. All of these thoughts are hypotheses, but nonetheless, I think these ideas will be the most exciting to explore in future work. Understanding these

mechanisms would also contribute to knowledge of how plasmids, viruses, gene transfer agents, and other undiscovered (or the newly discovered Borgs!) (Al-Shayeb *et al.*, 2021) extrachromosomal elements play in evolutionary adaptations, ecosystem processes, and applications to human health.

As I consider the broader impact of the work presented in this dissertation, I can't help but to feel that none of this research matters when there is a raging two-year global COVID-19 pandemic, a racial reckoning across the US, a growing global climate change catastrophe, unprecedented levels of gun violence, and, most recently, the stripping of the fundamental human right of bodily autotomy for people with uteruses. In spite of this, when I consider how the field of marine microbial natural products fits into the larger picture, we know that the ocean is full of incredible biodiversity—both of life and biosynthetic potential, which could be the source of the next life-saving drug. Bioactive drug molecules are desperately needed to combat antibiotic resistance that is a major threat to human health (Zhang *et al.*, 2011). However, I think the field of marine natural products has the opportunity to be leaders in sustainable and future-oriented practices to continue these discoveries. We must protect biodiversity in locations that have long been targets for "exotic" expeditions to find natural products. Cultural resources and knowledge should be preserved and protected (Stefanoudis *et al.*, 2021). Additionally, the marine natural products field can lead in advocating for utilizing sustainable methods of green chemistry, catalysts, and alternative chemical processes. However, for there to be a market for pharmaceutical and other specialized metabolites, it will take private pharmaceutical companies to once again invest in the development of antibiotics and therapeutics that might not be profitable but will save lives. Large transformations will be needed to revitalize the once-active pipeline of a marine

natural product to an approved drug that has an impact on society (Harvey *et al.*, 2015; de la Torre and Albericio, 2022).

When I started my Ph.D. in 2016, the amount of carbon dioxide levels in the atmosphere was 404 (parts per million) ppm. Today, 6 years later, it is 421 ppm. Global climate change is the largest threat to global biodiversity and human life, exacerbating inequalities, impacts, and environmental injustices between nations and peoples with socioeconomic and resource differences (IPCC, 2014). The impacts of the global climate change catastrophe are already being felt, from increased wildfires, extreme temperatures, rising sea levels, ocean acidification, altered rainfall patterns, and increased extreme weather events, to name a few (IPCC, 2014). Microbes not only will be impacted by global climate change, but they have the potential to be a part of the solution (Cavicchioli *et al.*, 2019; Hutchins *et al.*, 2019). Climate change will impact everyone, and especially those with the least resources to combat it. Marine microbes play a huge role in global climate patterns including climate regulation, oxygen production, biogeochemical cycling, and symbioses; thus climate change threats such as ocean warming acidification, ocean warming, and human overuse of oceanic systems can cause drastic ecosystem functional shifts and impacts to the global food web (Burge *et al.*, 2014; Cavicchioli *et al.*, 2019). On land, microbes will be (and are being) affected by higher carbon dioxide levels in the atmosphere, leading to higher productivity and thus higher carbon emissions (Cavicchioli *et al.*, 2019). Temperature fluctuations and changing climate patterns will affect nutrient cycling and methane production in agricultural systems (Busby *et al.*, 2017). Climate change has also been linked to exacerbated impacts of microbial pathogens including marine diseases, crop blights, increased vector-borne diseases, and antimicrobial resistance spread (Burge *et al.*, 2014). Microbes are both being affected and can be part of the solution, and thus a broader impact I'd like to see in the context of this dissertation work

is how we can apply microbial solutions to global climate change. This should include an improved understanding of microbial interactions that we could harness to combat climate change and specific dynamics that we are already observing. For example, *Wolbachia* bacteria has been used to reduce the transmission of mosquito-borne pathogens (Reveillaud *et al.*, 2019). In agriculture, microbial consortia are being applied for better nutrient cycling; understanding the microbial and sinks of methane production can have large impacts on sustainable efforts of carbona and nitrogen turnover (O'Callaghan, 2016; Busby *et al.*, 2017). We can use microbes to manipulate organic material and even recycle plastics, and other biotechnological advances can be applied to increasing sustainable food output, capturing emissions, and assisting ecosystem services (Cavicchioli *et al.*, 2019). As reported in this dissertation, the diversification of specialized metabolite BGCs over evolutionary time in microbes is contributing to nature's chemical diversity; and thus, I hope we can continue to learn and harness the evolvability and adaptability of microbes for solving our greatest challenges.

7.1 Conclusion References

Al-Shayeb, B., Schoelmerich, M.C., West-Roberts, J., Valentin-Alvarado, L.E., Sachdeva, R., Mullen, S., Crits-Christoph, A., Wilkins, M.J., Williams, K.H., Doudna, J.A., and Banfield, J.F. (2021) Borgs are giant extrachromosomal elements with the potential to augment methane oxidation. *bioRxiv*.

Bertics, V.J. and Ziebis, W. (2009) Biodiversity of benthic microbial communities in bioturbated coastal sediments is controlled by geochemical microniches. *ISME J* **3**: 1269–1285.

Bonham, K.S., Wolfe, B.E., and Dutton, R.J. (2017) Extensive horizontal gene transfer in cheese-associated bacteria. 1–23.

Bull, A.T. and Goodfellow, M. (2019) Dark, rare and inspirational microbial matter in the extremobiosphere: 16 000 m of bioprospecting campaigns. *Microbiology* 1–13.

Burge, C. a, Mark Eakin, C., Friedman, C.S., Froelich, B., Hershberger, P.K., Hofmann, E.E., Petes, L.E., Prager, K.C., Weil, E., Willis, B.L., Ford, S.E., and Harvell, C.D. (2014) Climate change influences on marine infectious diseases: implications for management and society. *Ann Rev Mar Sci* **6**: 249–77.

Busby, P.E., Soman, C., Wagner, M.R., Friesen, M.L., Kremer, J., Bennett, A., Morsy, M., Eisen, J.A., Leach, J.E., and Dangl, J.L. (2017) Research priorities for harnessing plant microbiomes in sustainable agriculture. *PLoS Biol* **15**: 1–14.

Cavicchioli, R., Ripple, W.J., Timmis, K.N., Azam, F., Bakken, L.R., Baylis, M., Behrenfeld, M.J., Boetius, A., Boyd, P.W., Classen, A.T., Crowther, T.W., Danovaro, R., Foreman, C.M., Huisman, J., Hutchins, D.A., Jansson, J.K., Karl, D.M., Koskella, B., Mark Welch, D.B., Martiny, J.B.H., Moran, M.A., Orphan, V.J., Reay, D.S., Remais, J. V., Rich, V.I., Singh, B.K., Stein, L.Y., Stewart, F.J., Sullivan, M.B., van Oppen, M.J.H., Weaver, S.C., Webb, E.A., and Webster, N.S. (2019) Scientists' warning to humanity: microorganisms and climate change. *Nat Rev Microbiol* **17**: 569–586.

Cech, N.B., Medema, M.H., and Clardy, J. (2021) Benefiting from big data in natural products: Importance of preserving foundational skills and prioritizing data quality. *Nat Prod Rep* **38**: 1947–1953.

Chase, A.B., Sweeney, D., Guillén-matus, D.G., and Jensen, P.R. (2021) Vertical Inheritance Facilitates Interspecies Diversification in Biosynthetic Gene Clusters and Specialized Metabolites. *MBio*.

Chevrette, M.G., Gutiérrez-García, K., Selem-Mojica, N., Aguilar-Martínez, C., Yañez-Olvera, A., Ramos-Aboites, H.E., Hoskisson, P.A., and Barona-Gómez, F. (2020) Evolutionary dynamics of natural product biosynthesis in bacteria. *Nat Prod Rep* **37**: 566–599.

Claessen, D. and Errington, J. (2019) Cell Wall Deficiency as a Coping Strategy for Stress. *Trends Microbiol* **xx**: 1–9.

Danko, D., Bezdan, D., Afshin, E.E., Ahsanuddin, S., Bhattacharya, C., Butler, D.J., Chng, K.R., Donnellan, D., Hecht, J., Jackson, K., Kuchin, K., Karasikov, M., Lyons, A., Mak, L., Meleshko, D., Mustafa, H., Mutai, B., Neches, R.Y., Ng, A., Nikolayeva, O., Nikolayeva, T., Png, E., Ryon, K.A., Sanchez, J.L., Shaaban, H., Sierra, M.A., Thomas, D., Young, B., Abudayyeh, O.O., Alicea, J., Bhattacharyya, M., Blekhman, R., Castro-Nallar, E., Cañas, A.M., Chatziefthimiou, A.D., Crawford, R.W., De Filippis, F., Deng, Y., Desnues, C., Dias-Neto, E., Dybwad, M., Elhaik, E., Ercolini, D., Frolova, A., Gankin, D., Gootenberg, J.S., Graf, A.B., Green, D.C., Hajirasouliha, I., Hastings, J.J.A., Hernandez, M., Iraola, G., Jang, S., Kahles, A., Kelly, F.J., Knights, K., Kyrpides, N.C., Łabaj, P.P., Lee, P.K.H., Leung, M.H.Y., Ljungdahl, P.O., Mason-Buck, G., McGrath, K., Meydan, C., Mongodin, E.F., Moraes, M.O., Nagarajan, N., Nieto-Caballero, M., Noushmehr, H., Oliveira, M., Ossowski, S., Osuolale, O.O., Özcan, O., Paez-Espino, D., Rascovan, N., Richard, H., Rätsch, G., Schriml, L.M., Semmler, T., Sezerman, O.U., Shi, L., Shi, T., Siam, R., Song, L.H., Suzuki, H., Court, D.S., Tighe, S.W., Tong, X., Udekwu, K.I., Ugalde, J.A., Valentine, B., Vassilev, D.I., Vayndorf, E.M., Velavan, T.P., Wu, J., Zambrano, M.M., Zhu, J., Zhu, S., Mason, C.E., Abdullah, N., Abraao, M., Adel, A., Afaq, M., Al-Quaddoomi, F.S., Alam, I., Albuquerque, G.E., Alexiev, A., Ali, K., Alvarado-Arnez, L.E., Aly, S., Amachee, J., Amorim, M.G., Ampadu, M., Amran, M.A.-F., An, N., Andrew, W., Andrianjakarivony, H., Angelov, M., Antelo, V., Aquino, C., Aranguren, Á., Araujo, L.F., Vasquez Arevalo, H.F., Arevalo, J., Arnan, C., Alvarado Arnez, L.E., Arredondo, F., Arthur, M., Asenjo, F., Aung, T.S., Auvinet, J., Aventin, N., Ayaz, S., Baburyan, S., Bakere, A.-M., Bakhl, K., Bartelli, T.F., Batdelger, E., Baudon, F., Becher, K., Bello, C., Benchouaia, M., Benisty, H., Benoiston, A.-S., Benson, J., Benítez, D., Bernardes, J., Bertrand, D., Beurmann, S., Bitard-Feildel, T., Bittner, L., Black, C., Blanc, G., Blyther, B., Bode, T., Boeri, J., Boldgiv, B., Bolzli, K., Bordigoni, A., Borrelli, C., Bouchard, S., Bouly, J.-P., Boyd, A., Branco, G.P., Breschi, A., Brindefalk, B., Brion, C., Briones, A., Buczansla, P., Burke, C.M., Burrell, A., Butova, A., Buttar, I., Bynoe, J., Bönigk, S., Bøifot, K.O., Caballero, H., Cai, X.W., Calderon, D., Cantillo, A., Carbajo, M., Carbone, A., Cardenas, A., Carrillo, K., Casalot, L., Castro, S., Castro, Ana V., Castro, Astred, Castro, A.V.B., Cawthorne, S., Cedillo, J., Chaker, S., Chalangal, J., Chan, A., Chasapi, A.I., Chatziefthimiou, S., Chaudhuri, S.R., Chavan, A.K., Chavez, F., Chem, G., Chen, X., Chen, M., Chen, J.-W., Chernomoretz, A., Chettouh, A., Cheung, D., Chicas, D., Chiu, S., Choudhry, H., Chrispin, C., Ciaramella, K., Cifuentes, E., Cohen, J., Coil, D.A., Collin, S., Conger, C., Conte, R., Corsi, F., Cossio, C.N., Costa, A.F., Cuebas, D., D'Alessandro, B., Dahlhausen, K.E., Darling, A.E., Das, P., Davenport, L.B., David, L., Davidson, N.R., Dayama, G., Delmas, S., Deng, C.K., Dequeker, C., Desert, A., Devi, M., Dezem, F.S., Dias, C.N., Donahoe, T.R., Dorado, S., Dorsey, L., Dotsenko, V., Du, S., Dutan, A., Eady, N., Eisen, J.A., Elaskandrany, M., Epping, L., Escalera-Antezana, J.P., Ettinger, C.L., Faiz, I., Fan, L., Farhat, N., Faure, E., Fauzi, F., Feigin, C., Felice, S., Ferreira, L.P., Figueroa, G., Fleiss, A., Flores, D., Velasco Flores, J.L., Fonseca, M.A.S., Foox, J., Forero, J.C., Francis, A., French, K., Fresia, P., Friedman, J., Fuentes, J.J., Galipon, J., Garcia, M., Garcia, L., García, C., Geiger, A., Gerner, S.M., Ghose, S.L., Giang, D.P., Giménez, M., Giovannelli, D., Githae, D., Gkotzis, S., Godoy, L., Goldman, S., Gonnet, G.H., Gonzalez, J., Gonzalez, A., Gonzalez-Poblete, C., Gray, A., Gregory, T., Greselle, C., Guasco, S., Guerra, J., Gurianova, N., Haehr, W., Halary, S., Hartkopf, F., Hastings, J.J.A., Hawkins-

Zafarnia, A., Hazrin-Chong, N.H., Helfrich, E., Hell, E., Henry, T., Hernandez, S., Hernandez, P.L., Hess-Homeier, D., Hittle, L.E., Hoan, N.X., Holik, A., Homma, C., Hoxie, I., Huber, M., Humphries, E., Hyland, S., Hässig, A., Häusler, R., Hüsser, N., Petit, R.A., Iderzorig, B., Igarashi, M., Iqbal, S.B., Ishikawa, S., Ishizuka, S., Islam, S., Islam, R., Ito, K., Ito, S., Ito, T., Ivankovic, T., Iwashiro, T., Jackson, S., Jacobs, J., James, M., Jaubert, M., Jerier, M.-L., Jiminez, E., Jinfessa, A., De Jong, Y., Joo, H.W., Jospin, G., Kajita, T., Ahmad Kassim, A.S., Kato, N., Kaur, A., Kaur, I., de Souza Gomes Kehdy, F., Khadka, V.S., Khan, S., Khavari, M., Ki, M., Kim, G., Kim, H.J., Kim, S., King, R.J., Knights, K., KoLoMonaco, G., Koag, E., Kobko-Litskevitch, N., Korshevniuk, M., Kozhar, M., Krebs, J., Kubota, N., Kuklin, A., Kumar, S.S., Kwong, R., Kwong, L., Lafontaine, I., Lago, J., Lai, T.Y., Laine, E., Laiola, M., Lakhneko, O., Lamba, I., de Lamotte, G., Lannes, R., De Lazzari, E., Leahy, M., Lee, H., Lee, Y., Lee, L., Lemaire, V., Leong, E., Leung, M.H.Y., Lewandowska, D., Li, C., Liang, W., Lin, M., Lisboa, P., Litskevitch, A., Liu, E.M., Liu, T., Livia, M.A., Lo, Y.H., Losim, S., Loubens, M., Lu, J., Lykhenko, O., Lysakova, S., Mahmoud, S., Majid, S.A., Makogon, N., Maldonado, D., Mallari, K., Malta, T.M., Mamun, M., Manoir, D., Marchandon, G., Marciniak, N., Marinovic, S., Marques, B., Mathews, N., Matsuzaki, Y., Matthys, V., May, M., McComb, E., Meagher, A., Melamed, A., Menary, W., Mendez, K.N., Mendez, A., Mendy, I.M., Meng, I., Menon, A., Menor, M., Meoded, R., Merino, N., Meydan, C., Miah, K., Mignotte, M., Miketic, T., Miranda, W., Mitsios, A., Miura, R., Miyake, K., Moccia, M.D., Mohan, N., Mohsin, M., Moitra, K., Moldes, M., Molina, L., Molinet, J., Molomjamts, O.-E., Moniruzzaman, E., Moon, S., de Oliveira Moraes, I., Moreno, M., Mosella, M.S., Moser, J.W., Mozsary, C., Muehlbauer, A.L., Muner, O., Munia, M., Munim, N., Muscat, M., Mustac, T., Muñoz, C., Nadalin, F., Naeem, A., Nagy-Szakal, D., Nakagawa, M., Narce, A., Nasu, M., Navarrete, I.G., Naveed, H., Nazario, B., Nedunuri, N.R., Neff, T., Nesimi, A., Ng, W.C., Ng, S., Nguyen, G., Ngwa, E., Nicolas, A., Nicolas, P., Nika, A., Noorzi, H., Nosrati, A., Noushmehr, H., Nunes, D.N., O'Brien, K., O'Hara, N.B., Oken, G., Olawoyin, R.A., Oliete, J.Q., Olmeda, K., Oluwadare, T., Oluwadare, I.A., Ordioni, N., Orpilla, J., Orrego, J., Ortega, M., Osma, P., Osuolale, I.O., Osuolale, O.M., Ota, M., Oteri, F., Oto, Y., Ounit, R., Ouzounis, C.A., Pakrashi, S., Paras, R., Pardo-Este, C., Park, Y.-J., Pastuszek, P., Patel, S., Pathmanathan, J., Patrignani, A., Perez, M., Peros, A., Persaud, S., Peters, A., Phillips, A., Pineda, L., Pizzi, M.P., Plaku, Alma, Plaku, Alketa, Pompa-Hogan, B., Portilla, M.G., Posada, L., Priestman, M., Prithiviraj, B., Priya, S., Pugdeethosal, P., Pugh, C.E., Pulatov, B., Pupiec, A., Pyrshev, K., Qing, T., Rahiel, S., Rahmatulloev, S., Rajendran, K., Ramcharan, A., Ramirez-Rojas, A., Rana, S., Ratnanandan, P., Read, T.D., Rehrauer, H., Richer, R., Rivera, A., Rivera, M., Robertiello, A., Robinson, C., Rodríguez, P., Rojas, N.A., Roldán, P., Rosario, A., Roth, S., Ruiz, M., Boja Ruiz, S.E., Russell, K., Rybak, M., Sabedot, T.S., Sabina, M., Saito, I., Saito, Y., Malca Salas, G.A., Salazar, C., San, K.M., Sanchez, J., Sanchir, K., Sankar, R., de Souza Santos, P.T., Saravi, Z., Sasaki, K., Sato, Y., Sato, M., Sato, S., Sato, R., Sato, K., Sayara, N., Schaaf, S., Schacher, O., Schinke, A.-L.M., Schlapbach, R., Schori, C., Schriml, J.R., Segato, F., Sepulveda, F., Serpa, M.S., De Sessions, P.F., Severyn, J.C., Shaaban, H., Shakil, M., Shalaby, S., Shari, A., Shim, H., Shirahata, H., Shiwa, Y., Siam, R., Da Silva, O., Silva, J.M., Simon, G., Singh, S.K., Sluzek, K., Smith, R., So, E., Andreu Somavilla, N., Sonohara, Y., Rufino de Sousa, N., Souza, C., Sperry, J., Sprinsky, N., Stark, S.G., La Storia, A., Suganuma, K., Suliman, H., Sullivan, J., Supie, A.A.M., Suzuki, C., Takagi, S., Takahara, F., Takahashi, N., Takahashi, K., Takeda, T., Takenaka, I.K., Tanaka, S., Tang, A., Man Tang, Y., Tarcitano, E., Tassinari, A., Taye, M., Terrero, A., Thambiraja,

E., Thiébaut, A., Thomas, S., Thomas, A.M., Togashi, Y., Togashi, T., Tomaselli, A., Tomita, M., Tomita, I., Tong, X., Toth, O., Toussaint, N.C., Tran, J.M., Truong, C., Tsonev, S.I., Tsuda, K., Tsurumaki, T., Tuz, M., Tymoshenko, Y., Urgiles, C., Usui, M., Vacant, S., Valentine, B., Vann, L.E., Velter, F., Ventorino, V., Vera-Wolf, P., Vicedomini, R., Suarez-Villamil, M.A., Vincent, S., Vivancos-Koopman, R., Wan, A., Wang, C., Warashina, T., Watanabe, A., Weekes, S., Werner, J., Westfall, D., Wieler, L.H., Williams, M., Wolf, S.A., Wong, B., Wong, Y.L., Wong, T., Wright, R., Wunderlin, T., Yamanaka, R., Yang, J., Yano, H., Yeh, G.C., Yemets, O., Yeskova, T., Yoshikawa, S., Zafar, L., Zhang, Y., Zhang, S., Zhang, A., Zheng, Y., and Zubenko, S. (2021) A global metagenomic map of urban microbiomes and antimicrobial resistance. *Cell* 1–18.

Dührkop, K., Nothias, L.F., Fleischauer, M., Reher, R., Ludwig, M., Hoffmann, M.A., Petras, D., Gerwick, W.H., Rousu, J., Dorrestein, P.C., and Böcker, S. (2020) Systematic classification of unknown metabolites using high-resolution fragmentation mass spectra. *Nat Biotechnol*.

Frost, I., Smith, W.P.J., Mitri, S., Millan, A.S., Davit, Y., Osborne, J.M., Francis, J.M.P.-, Maclean, R.C., and Foster, K.R. (2018) Cooperation, competition and antibiotic resistance in bacterial colonies.

Grininger, M. (2020) The role of the iterative modules in polyketide synthase evolution. *Proc Natl Acad Sci U S A* **117**: 8680–8682.

Harvey, A.L., Edrada-Ebel, R., and Quinn, R.J. (2015) The re-emergence of natural products for drug discovery in the genomics era. *Nat Rev Drug Discov* **14**: 111–129.

Hopwood, D.A., Wright, H.M., and Bibb, M.J. (1977) Genetic recombination through protoplast fusion in Streptomyces. *Nature* **268**:.

Hutchins, D.A., Jansson, J.K., Remais, J. V, and Rich, V.I. (2019) Climate change microbiology — problems and perspectives. *Nat Rev Microbiol* **17**: 391–396.

IPCC (2014) Climate Change 2014 Synthesis Report Summary Chapter for Policymakers. *IPCC Rep* 31.

Jensen, P.R. (2010) Linking species concepts to natural product discovery in the post-genomic era. *J Ind Microbiol Biotechnol* **37**: 219–224.

Jensen, P.R. (2016) Natural Products and the Gene Cluster Revolution. *Trends Microbiol* **24**: 968–977.

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

de la Torre, B.G. and Albericio, F. (2022) The Pharmaceutical Industry in 2021. An Analysis of FDA Drug Approvals from the Perspective of Molecules. *Molecules* **27**:.

MacLean, R.C. and San Millan, A. (2015) Microbial Evolution: Towards Resolving the Plasmid Paradox. *Curr Biol* **25**: R764–R767.

Millán-Aguiñaga, N., Chavarria, K.L., Ugalde, J.A., Letzel, A.-C., Rouse, G.W., and Jensen, P.R. (2017) Phylogenomic Insight into *Salinispora* (Bacteria, Actinobacteria) Species Designations. *Sci Rep* **7**: 3564.

Mohimani, H., Gurevich, A., Mikheenko, A., Garg, N., Nothias, L.-F., Ninomiya, A., Takada, K., Dorrestein, P.C., and Pevzner, P.A. (2017) Dereplication of peptidic natural products through database search of mass spectra. *Nat Chem Biol* **13**:.

Nivina, A., Herrera Paredes, S., Fraser, H.B., and Khosla, C. (2021) GRINS: Genetic elements that recode assembly-line polyketide synthases and accelerate their diversification. *Proc Natl Acad Sci* **118**: e2100751118.

Nivina, A., Yuet, K.P., Hsu, J., and Khosla, C. (2019) Evolution and Diversity of Assembly-Line Polyketide Synthases. *Chem Rev* **119**: 12524–12547.

O'Callaghan, M. (2016) Microbial inoculation of seed for improved crop performance: issues and opportunities. *Appl Microbiol Biotechnol* **100**: 5729–5746.

Okanishi, M., Suzuki, K., and Umezawa, H. (1974) Formation and reversion of Streptomycete protoplasts: cultural condition and morphological study. *J Gen Microbiol* **80**: 389–400.

Petras, D., Minich, J.J., Kunselman, E., Wang, M., White, M.E., and Eric, E. (2019) Non-target tandem mass spectrometry enables the prioritization of anthropogenic pollutants in seawater along the northern San Diego coast. *ChemArxiv* 1–25.

Ramijan, K., Ultee, E., Willemse, J., Zhang, Z., Wondergem, J.A.J., van der Meij, A., Heinrich, D., Briegel, A., van Wezel, G.P., and Claessen, D. (2018) Stress-induced formation of cell wall-deficient cells in filamentous actinomycetes. *Nat Commun* **9**:.

Reveillaud, J., Bordenstein, Sarah R., Cruaud, C., Shaiber, A., Esen, Ö.C., Weill, M., Makoundou, P., Lolans, K., Watson, A.R., Rakotoarivony, I., Bordenstein, Seth R., and Eren, A.M. (2019) The Wolbachia mobilome in Culex pipiens includes a putative plasmid. *Nat Commun* **10**:.

Robinson, S.L., Piel, J., and Sunagawa, S. (2021) A roadmap for metagenomic enzyme discovery. *Nat Prod Rep*.

Rokas, A., Mead, M.E., Steenwyk, J.L., Raja, H.A., and Oberlies, N.H. (2020) Biosynthetic gene clusters and the evolution of fungal chemodiversity. *Nat Prod Rep*.

Román-Ponce, B., Millán-Aguiñaga, N., Guillen-Matus, D., Chase, A.B., Ginigini, J.G.M., Soapi, K., Feussner, K.D., Jensen, P.R., and Trujillo, M.E. (2020) Six novel species of the obligate marine actinobacterium *Salinispora*, *Salinispora cortesiana* sp. nov., *Salinispora fenicalii* sp. nov., *Salinispora goodfellowii* sp. nov., *Salinispora mooreana* sp. nov., ... *Int J Syst Evol Microbiol* **70**: 4668–4682.

Ruzzini, A.C. and Clardy, J. (2016) Gene Flow and Molecular Innovation in Bacteria. *Curr Biol* **26**: R859–R864.

Saary, P., Forslund, K., Bork, P., and Hildebrand, F. (2017) RTK: efficient rarefaction analysis of large datasets. *Bioinformatics* **33**: 2594–2595.

Sanders, J.G., Nurk, S., Salido, R.A., Minich, J., Xu, Z.Z., Zhu, Q., Martino, C., Fedarko, M., Arthur, T.D., Chen, F., Boland, B.S., Humphrey, G.C., Brennan, C., Sanders, K., Gaffney, J., Jepsen, K., Khosroheidari, M., Green, C., Liyanage, M., Dang, J.W., Phelan, V. V., Quinn, R.A., Bankevich, A., Chang, J.T., Rana, T.M., Conrad, D.J., Sandborn, W.J., Smarr, L., Dorrestein, P.C., Pevzner, P.A., and Knight, R. (2019) Optimizing sequencing protocols for leaderboard metagenomics by combining long and short reads. *Genome Biol* **20**: 1–14.

Song, W., Wemheuer, B., Steinberg, P.D., Marzinelli, E.M., and Thomas, T. (2021) Contribution of horizontal gene transfer to the functionality of microbial bio fi lm on a macroalgae. *ISME J*.

Stefanoudis, P. V., Licuanan, W.Y., Morrison, T.H., Talma, S., Veitayaki, J., and Woodall, L.C. (2021) Turning the tide of parachute science. *Curr Biol* **31**: R184–R185.

Stewart, G.J., Sinigalliano, C.D., and Garko, K.A. (1991) Binding of exogenous DNA to marine sediments and the effect of DNA/sediment binding on natural transformation of Pseudomonas stutzeri strain ZoBell in sediment columns. *FEMS Microbiol Lett* **85**: 1–8.

Thompson, L.R., Sanders, J.G., Mcdonald, D., Amir, A., Ladau, J., Locey, K.J., Prill, R.J., Tripathi, A., Gibbons, S.M., Ackermann, G., Navas-molina, J.A., Janssen, S., Kopylova, E., Vázquez-baeza, Y., González, A., Morton, J.T., Mirarab, S., Xu, Z.Z., and Jiang, L. (2017) A communal catalogue reveals Earth ' s multiscale microbial diversity.

Tuttle, R.N., Demko, A.M., Patin, N. V, Kapono, C.A., Donia, M.S., Dorrestein, P., and Jensen, P.R. (2019) Detection of Natural Products and Their Producers in Ocean Sediments. *Appl Environ Microbiol* **85**: 1–15.

Vidgen, M.E., Hooper, J.N.A., and Fuerst, J.A. (2012) Diversity and distribution of the bioactive actinobacterial genus *Salinispora* from sponges along the Great Barrier Reef. *Antonie Van Leeuwenhoek* **101**: 603–618.

Walker, A.S. and Clardy, J. (2021) A Machine Learning Bioinformatics Method to Predict Biological Activity from Biosynthetic Gene Clusters. *J Chem Inf Model*.

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., Crüsemann, M., Boudreau, P.D., Esquenazi, E., Sandoval-Calderón, 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., Boya P, C.A., 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., Rodríguez, 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., Müller, R., Waters, K.M., Shi, W., Liu, X., Zhang, L., Knight, R., Jensen, P.R., Palsson, B.Ø., Pogliano, K., Linington, R.G., Gutiérrez, M., Lopes, N.P., Gerwick, W.H., Moore, B.S., Dorrestein, P.C., and Bandeira, N. (2016) Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nat Biotechnol* **34**: 828–837.

Wisecaver, J.H. and Rokas, A. (2015) Fungal metabolic gene clusters-caravans traveling across genomes and environments. *Front Microbiol* **6**: 1–11.

Xu, C. and Jackson, S.A. (2019) Machine learning and complex biological data. *Genome Biol* **20**: 1–4.

Zhang, J.J., Tang, X., and Moore, B.S. (2019) Genetic platforms for heterologous expression of microbial natural products. *Nat Prod Rep*.

Zhang, Q., Lambert, G., Liao, D., Kim, H., Robin, K., Tung, C. -k., Pourmand, N., and Austin, R.H. (2011) Acceleration of Emergence of Bacterial Antibiotic Resistance in Connected Microenvironments. *Science (80-)* **333**: 1764–1767.

Ziemert, N., Lechner, A., Wietz, M., Millán-Aguiñaga, N., Chavarria, K.L., and Jensen, P.R. (2014) Diversity and evolution of secondary metabolism in the marine actinomycete genus *Salinispora*. *Proc Natl Acad Sci U S A* **111**: E1130-9.

Ziemert, N., Podell, S., Penn, K., Badger, J.H., Allen, E., and 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.

APPENDIX A. Structure and Candidate Biosynthetic Gene Cluster of a Manumycin-Type Metabolite from *Salinispora pacifica*

A.1 Introduction to Appendix A

Appendix A describes the discovery and characterization of the manumycin-type metabolite pacificamide from *Salinispora pacifica*. Using methods that I developed in Chapter 4, I assisted the lead author Dr. Gabriel Castro-Falcón in describing the distribution and diversity of the pacificamide (*pac*) BGC. Dr. Castro-Falcón had preliminarily identified the BGC in a *Salinispora pacifica* CNT-855, and his goal was to see if there were other similar BGCs— including BGCs with and without known products— with gene organization that could help him propose the biosynthetic schema for pacificamide. I used all *pac1-33* genes in the *pac* BGC to search the NCBI reference (refseq), nonredundant (nr), metagenomic (env_nr) and patented (pattaa) protein sequence databases using cblaster (Gilchrist and Chooi, 2021; Gilchrist *et al.*, 2021). Manual curation of the thousands of BGC hits resulted in 39 top *pac*-like BGCs, including 1 other *pac* BGC in a different *Salinispora pacifica* strain and other *pac*-like BGCs in the families *Micromonosporaceae*, *Streptomycetaceae*, and *Pseudonocardiaceae*. I additionally searched the *pac* BGC against the MIBiG 2.0 (Kautsar *et al.*, 2020) database of BGCs with known products, and our results showed that the *pac* BGC was unique and rare in its specific gene content and organization. To have further BGCs to compare to, my co-authors genome-sequenced the manumycin-like daryamide and novodaryamide producer *Streptomyces* sp. CNQ-085 (Asolkar *et al.*, 2006; Castro-Falcón *et al.*, 2018). With Dr. Castro-Falcón, I queried the *pac* BGC against the new genome assembly and stitched together the two genome contigs comprising the daryamide BGC (*dar*).

By comparing *pac* to other manumycin-like BGCs including *dar*, Dr. Castro-Falcón proposed a biosynthetic pathway for pacificamide based on the genes present in *pac*. I calculated

a phylogenomic tree of the class *Actinomycetia* (Salam *et al.*, 2020) with one representative genome from all 60 families (334 conserved marker genes from each genome was used to calculate the phylogeny) (Asnicar *et al.*, 2020). We observed that the three families that contain manumycin-type BGCs claded separately in the tree, and thus the manumycin-type and *pac*-like BGCs seem to have been horizontally exchanged across distant taxa. Pacificamide was found to have week antibiotic activity (MIC of 50 μm) against a Gram-positive *Bacillus oceanisediminis*, thus making one wonder about the ecological role of this molecule in *Salinispora pacifica*.

Similar to the approach in Chapter 4 (Creamer *et al.*, 2021), the ability to use biosynthetic knowledge about a BGC to identify similar gene clusters is a powerful approach to understanding the diversity and distribution of important gene clusters. By carefully studying the BGCs in this context, we observed a unique *pac* BGC that produces a manumycin-type pacificamide molecule that is only found in three families of class *Actinomycetia* with variable gene content and organization among them. This is another example of the curious evolutionary patterns that can be observed in BGCs and the chemical diversity resulting from such genetic differences.

Appendix A (Section A.3), in full, is a reprint of the material as it appears in the *Journal of Natural Products Genomics* 85(4), 980-986. Castro-Falcón, G.; Creamer, K.E.; Chase, A.B.; Kim, M.C.; Sweeney, D.; Glukhov, E.; Fenical, W.; and Jensen, P.R., 2022. "Structure and Candidate Biosynthetic Gene Cluster of a Manumycin-Type Metabolite from *Salinispora pacifica*". The dissertation author was the second author of this paper.

A.2 Appendix A Introduction References

Asnicar, F., Thomas, A.M., Beghini, F., Mengoni, C., Manara, S., Manghi, P., et al. (2020) Precise phylogenetic analysis of microbial isolates and genomes from metagenomes using PhyloPhlAn 3.0. *Nat Commun* **11**: 1–10.

Asolkar, R.N., Jensen, P.R., Kauffman, C.A., and Fenical, W. (2006) Daryamides A-C, weakly cytotoxic polyketides from a marine-derived actinomycete of the genus Streptomyces strain CNQ-085. *J Nat Prod* **69**: 1756–1759.

Castro-Falcón, G., Millán Aguiñaga, N., Roullier, C., Jensen, P.R., and Hughes, C.C. (2018) Nitrosopyridine probe to detect polyketide natural products with conjugated alkenes: Discovery of novodaryamide and nocarditriene. *ACS Chem Biol* acschembio.8b00598.

Creamer, K.E., Kudo, Y., Moore, B.S., and Jensen, P.R. (2021) Phylogenetic analysis of the salinipostin γ-butyrolactone gene cluster uncovers new potential for bacterial signalling-molecule diversity. *Microb Genomics* **7**: 1–14.

Gilchrist, C.L.M., Booth, T.J., and Chooi, Y.H. (2021) cblaster: A remote search tool for rapid identification and visualisation of homologous gene clusters. *Bioinforma Adv* 1–19.

Gilchrist, C.L.M. and Chooi, Y.-H. (2021) Clinker & Clustermap.Js: Automatic Generation of Gene Cluster Comparison Figures. *Bioinformatics* 1–3.

Kautsar, S.A., Blin, K., Shaw, S., Navarro-Muñoz, J.C., Terlouw, B.R., van der Hooft, J.J.J., et al. (2020) MIBiG 2.0: a repository for biosynthetic gene clusters of known function. *Nucleic Acids Res* **48**: D454–D458.

Salam, N., Jiao, J.Y., Zhang, X.T., and Li, W.J. (2020) Update on the classification of higher ranks in the phylum Actinobacteria. *Int J Syst Evol Microbiol* **70**: 1331–1355.

A.3 Reprint of: "Structure and Candidate Biosynthetic Gene Cluster of a Manumycin-Type Metabolite from *Salinispora pacifica*."

JOURNAL OF NATURAL PRODUCTS

pubs.acs.org/jnp Article

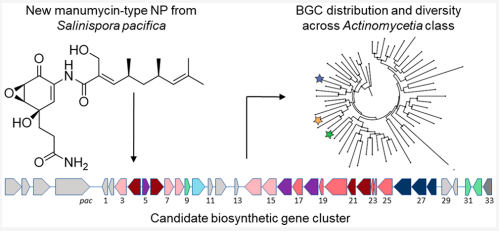
Structure and Candidate Biosynthetic Gene Cluster of a Manumycin-Type Metabolite from *Salinispora pacifica*

Gabriel Castro-Falcón, Kaitlin E. Creamer, Alexander B. Chase, Min Cheol Kim, Douglas Sweeney, Evgenia Glukhov, William Fenical, and Paul R. Jensen*

Cite This: *J. Nat. Prod.* 2022, 85, 980−986 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: A new manumycin-type natural product named pacificamide (**1**) and its candidate biosynthetic gene cluster (*pac*) were discovered from the marine actinobacterium *Salinispora pacifica* CNT-855. The structure of the compound was determined using NMR, electronic circular dichroism, and bioinformatic predictions. The *pac* gene cluster is unique to *S. pacifica* and found in only two of the 119 *Salinispora* genomes analyzed across nine species. Comparative analyses of biosynthetic gene clusters encoding the production of related manumycin-type compounds revealed genetic differences in accordance with the unique pacificamide structure. Further queries of manumycin-type gene clusters from public databases revealed their limited distribution across the phylum Actinobacteria and orphan diversity that suggests additional products remain to be discovered in this compound class. Production of the known metabolite triacsin D is also reported for the first time from the genus *Salinispora*. This study adds two classes of compounds to the natural product collective isolated from the genus *Salinispora*, which has proven to be a useful model for natural product research.



Bacterial natural products have played key roles in both the development of new drugs and the advancement of basic biomedical research.[1,2] The phylum Actinobacteria has been a particularly important source of novel natural products, with most strains originating from soils.[3] More recent explorations of marine microbes have revealed chemically rich actinobacterial lineages that are phylogenetically distinct from their terrestrial counterparts. Included among these is the marine obligate Actinobacterial genus *Salinispora*, which has proven to be a rich source of natural products with cytotoxic, antimalarial, and antibiotic activities.[4−6] The proteasome inhibitor salinosporamide A, which is currently undergoing phase III clinical trials, and the DNA intercalator lomaiviticin A are salient examples of bioactive natural products discovered from *Salinispora* species.[7,8] Our understanding of the natural product biosynthetic potential of the genus *Salinispora* has been further advanced by genome sequencing. Among 119 public genomes representing all nine named species, 305 different gene cluster families devoted to natural product biosynthesis were detected.[9−11] This figure greatly surpasses the number of natural products reported to date from *Salinispora* cultures and suggests that considerable biosynthetic potential remains to be realized.

Herein, we report the new metabolite pacificamide (**1**) from *Salinispora pacifica* CNT-855 and the known metabolite triacsin D from *Salinispora cortesiana* CNY-202. Both compounds, previously unknown from the genus, were initially targeted from a large-scale comparative metabolomics data set that included all nine *Salinispora* species.[11] We focus on the isolation, structure elucidation, and biological activity of the natural product pacificamide, a new member of the manumycin group of metabolites (Figure 1).[12] We identify a candidate pacificamide biosynthetic gene cluster, which we named *pac*, in two *S. pacifica* strains and compare them to characterized manumycin-type BGCs in *Streptomyces* and *Saccharothrix* spp. to correlate gene content with structural differences in the compounds they encode. Finally, we search *pac* homologues in publicly available bacterial genome sequences to reveal the diversity and distribution of manumycin-type BGCs that have yet to be linked to their small-molecule products.

RESULTS AND DISCUSSION

Isolation and Structure Elucidation. HPLC-UV-MS chemical profiling of culture extracts of 30 *Salinispora* spp. led to the detection of two compounds in strains *S. pacifica* CNT-855 and *S. cortesiana* CNY-202 with UV/vis and MS spectra that differed from previously identified *Salinispora*

Received: November 24, 2021
Published: March 9, 2022

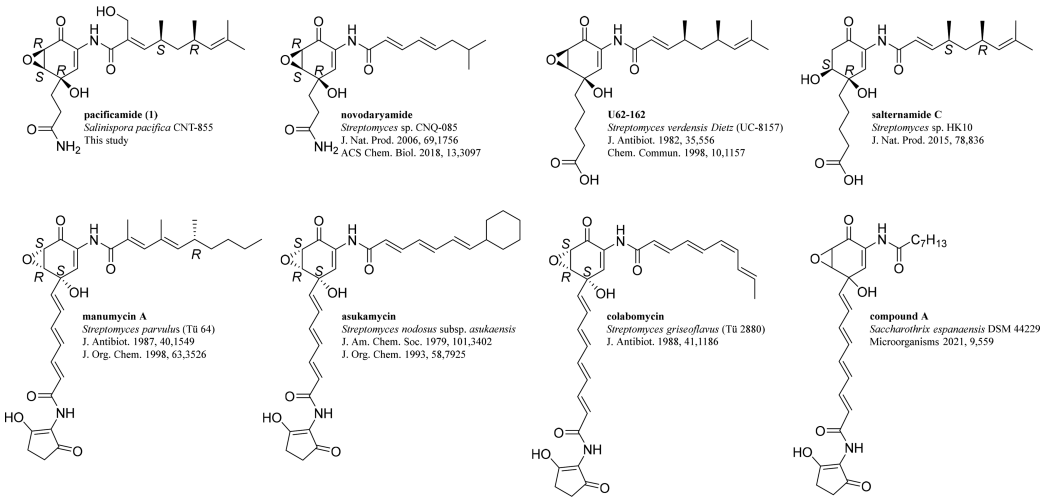


Figure 1. Representative manumycin-type natural products.

natural products. A 6−10 L cultivation of these strains yielded organic extracts, and subsequent targeted purification via C18 reversed-phase chromatography led to the isolation of the known compound triacsin D,[13,14] from CNY-202, and a new compound named pacificamide (**1**), from CNT-855, with a molecular formula assigned as $C_{22}H_{32}N_2O_6$ based on a sodium adduct ion at m/z 443.2149 using HR-ESI-TOF-MS.

Spectroscopic analysis of the ^{1}H NMR data of **1** (in CD_3OD) showed signals for 27 protons, including three olefinic protons at δ_H 7.29 (H-3), 6.49 (H-3′), and 4.88 (H-7′), two oxymethine protons at δ_H 3.66 (H-5) and 3.57 (H-6), an oxymethylene proton at δ_H 4.35 (H_2-10′), three methylene groups at δ_H 2.09 (H_2-7), 2.33 (H_2-8), and 1.34 (H-5′a) and 1.24 (H-5′b), two methine protons at δ_H 2.62 (H-4′) and 2.40 (H-6′), and four methyl groups at δ_H 1.68 (H_3-9′), 1.62 (H_3-13′), 1.01 (H_3-11′), and 0.90 (H_3-12′). The HSQC spectrum showed correlations for 15 carbon signals (Figure 2a), and the HMBC spectrum showed correlations to the remaining seven carbon signals, including three carbonyl carbons at δ_C 190.0 (C-1), 178.0 (C-9), and 169.3 (C-1′), three vinyl carbons at δ_C 130.4 (C-2), 132.1 (C-2′), and 131.0 (C-8′), and an *O*-substituted sp^3 carbon at δ_C 71.4 (C-4) (Figure 2c). The COSY spectrum revealed three spin−spin coupling systems including H-9′ to H-10′, H-3 to H-5, and H-7 to H-8 (Figure 2b). The HMBC spectrum showed correlations that established the connectivity of the three spin−spin coupling systems (Figure 2c). Among observed correlations, H-3 to C-1, C-2, and C-4, H-5 to C-4, and H-6 to C-1 and C-2 suggested the 5,6-epoxy-4-hydroxycyclohex-2-en-1-one of **1**, which is characteristic of many manumycin-type natural products (Figures 1 and 2c).[11] The HMBC correlations of H_2-7 to C-3, C-4, C-5, C-8, and C-9 and H_2-8 to C-9 permitted the propionic amide to be positioned at C-4 of the epoxycyclohexenone. On the basis of the molecular formula and the chemical shift of carbonyl C-1′, we inferred that the aliphatic side chain (C-2′ to C-9′) and the epoxy-cyclohexenone connection was through an amide group. Thus, the planar structure of **1** was established as drawn in Figure 2.

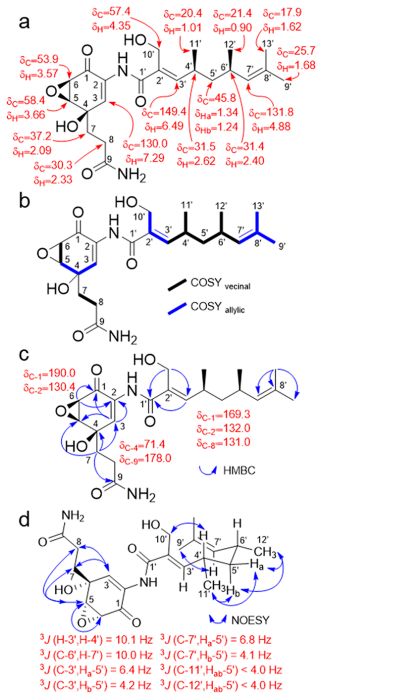
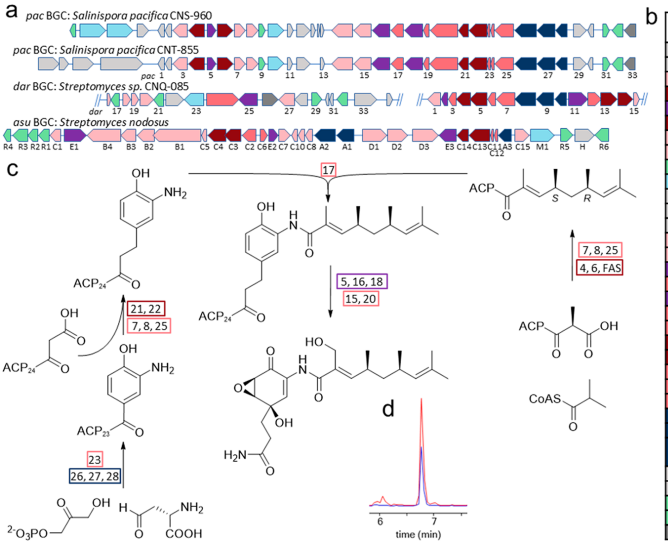


Figure 2. NMR assignments of pacificamide (**1**). (a) ^{1}H and ^{13}C chemicals shifts (in ppm) based on ^{1}H NMR and HSQC data. (b) Spin systems observed by COSY. (c) ^{13}C chemical shifts (in ppm) and key correlations observed by HMBC. (d) Key NOESY correlations and key spin−spin coupling constants (^3J) observed by ^{1}H NMR and HETLOC.

https://doi.org/10.1021/acs.jnatprod.1c01117
J. Nat. Prod. 2022, 85, 980−986



b

#	Predicted function	*asu* homolog	*dar* homolog
1	condensation domain of NRPS		
2	N-acetyltransferase		
3	Zn-dependent alcohol dehydrogenase		
4	beta-ketoacyl-ACP synthase III	*asuC4*	*dar14*
5	flavin reductase	*asuE2*	*dar2*
6	beta-ketoacyl-ACP synthase III	*asuC3*	*dar13*
7	3-hydroxyacyl-CoA dehydrogenase, FadB		
8	enoyl-CoA hydratase/isomerase		
9	transcriptional regulator, MerR		
10	arabinose efflux pump, MFS		*dar23*
11	reverse transcriptase/maturase-like		
12	transposase, IS110		
13	transposase, mutator		
14	acyltransferase		
15	cytochrome P450 hydroxylase		
16	flavin-dependent oxidoreductase, epoxidase	*asuE3*	*dar25*
17	arylamine N-acyltransferase	*asuC2*	*dar12*
18	flavin-dependent oxidoreductase, hydroxylase	*asuE1*	*dar11*
19	thioesterase	*asuC6*	*dar3*
20	asparagine synthetase		*dar24*
21	beta-ketoacyl-ACP synthase II (CLF)	*asuC14*	*dar4*
22	beta-ketoacyl-ACP synthase III (alpha unit)	*asuC13*	*dar5*
23	acyl carrier protein	*asuC12*	*dar6*
24	acyl carrier protein	*asuC11*	
25	enoyl-ACP reductase, FabV		*dar7*
26	AMP-dependent synthetase and ligase	*asuA2*	*dar8*
27	3-amino-4-hydroxybenzoic acid synthase	*asuA1*	*dar9*
28	2-amino-4, 5-dihydroxy-6-oxo-7-(phosphonooxy) heptanoate synthase	*asuA3*	*dar10*
29	transposase		
30	transposase		
31	transcriptional regulator		
32	transcriptional regulator		
33	DsbA family oxidoreductase		*dar26*

Figure 3. Pacificamide BGC and biosynthesis. (a) Candidate pacificamide (*pac*) and daryamide (*dar*, two contigs) BGCs. Asukamycin (*asu*) BGC is shown for reference. Gene numbering for *pac* in CNT-855, *dar*, and *asu* is shown. Genes are color-coded by function [wine: ketosynthases; navy blue: 3,4-AHBA synthesis; purple: oxidoreductases in epoxyquinol synthesis; salmon: biosynthetic (shared); light salmon: biosynthetic (not shared); light green: regulation; light blue: transport; gray: other]. (b) *pac* gene annotations with *asu* and *dar* homologues. (c) Proposed pacificamide biosynthetic pathway (numbers represent *pac* genes). (d) LCMS extracted ion chromatogram (EIC) for pacificamide $[M + Na]^+$ from strains CNT-855 (red) and CNS-960 (blue).

The relative configuration of **1** was assigned by interpretation of ^{1}H, HETLOC and NOESY NMR experiments and DFT-molecular calculations (Figure 2d). The cyclohexenone moiety (C-4 to C-6) including a *cis* epoxide ring was defined as 4*R**, 5*S**, 6*R** due to NOESY correlations of H-3/H-7, H-3/H-8, H-5/H-6, H-5/H-7, and H-5/H-8. Consistent with this, the *syn*-configuration between 4-hydroxy and 5,6-epoxide groups is a general characteristic of all manumycin-type metabolites.[15] On the "upper" side chain of **1**, the C-2′ to C-3′ double bond was defined as *E* due to a NOESY correlation between H-4′ and H-10′. As observed by ^{1}H NMR, a large $^3J_{H,H}$ (10.3 Hz) between vinylic H-3′ and H-4′ suggested these were in *anti*-conformation, which would result in low 1,3-allylic strain. A similar relationship was observed for H-6′ and vinylic H-7′ ($^3J_{H,H}$ = 10.2 Hz). As observed by HETLOC,[16] a large $^3J_{C,H}$ (6.4 Hz) between C-3′ and H-5′a suggested their *anti*-conformation. Similarly, C-7′ and H-5′a ($^3J_{C,H}$ = 6.8 Hz) were also in *anti*-conformation. Moreover, both C-12′ and C-11′ shared small $^3J_{C,H}$ (<4 Hz) with both H-5′a/H-5′b, suggesting their *gauche*-conformation. These results supported a conformation unique to **1** with *syn*-configured 4′S*,6′*R**-dimethyl groups (Figure 2d). The $\Delta\delta^1$H between H-5′a and H-5′b of 0.10 ppm did not support either configuration for the 4,6-dimethyl substituents.[17] Thus, to confirm our configurational assignment, we performed DFT calculations with ^{1}H and ^{13}C NMR chemical shift predictions on *syn*- and *anti*-configured models of the "upper" side chain of **1**.[18,19] Based on the differences between the calculated and experimental chemical shifts for *syn* ($\Delta\delta_{H,ave}$ = 0.051 ppm and $\Delta\delta_{C,ave}$ = 1.5 ppm) and *anti* ($\Delta\delta_{H,ave}$ = 0.091 ppm and $\Delta\delta_{C,ave}$ = 1.5 ppm) models, the *syn* model proved to be a better match to **1**. Furthermore, the structures of the lowest energy optimized conformers for the *syn* model were in agreement with the conformation observed via NMR analysis (Figure 2d).

The absolute configuration of the cyclohexenone moiety was assigned as 4*R*, 5*S*, 6*R* based on the electronic circular dichroism (ECD) spectrum of **1**, which showed a positive Cotton effect at λ_{max} ($\Delta\varepsilon$) 328 nm (+1.05) attributed to the position of the epoxide oxygen with respect to the cyclohexene chromophore in accordance with the "inverse quadrant" rule for epoxyquinols.[20−22] This assignment is consistent with the absolute configuration established for the salternamide natural products (Figure 1).[23] The absolute configuration determination for the "upper" side chain of **1** as 4′*S*,6′*R* is based on genetic evidence that follows below.

Biosynthetic Gene Cluster Analysis. AntiSMASH analysis of the *Salinispora pacifica* CNT-855 genome (NCBI accession AZWS00000000) led to the identification of a candidate BGC for the production of **1** that shared similarity to both the asukamycin BGC (*asu*) from *Streptomyces nodosus* subsp. *asukaensis* and the colabomycin BGC (*col*) from *Streptomyces aureus*.[24−27] A query of 119 *Salinispora* genomes revealed that only one additional strain, *S. pacifica* CNS-960, possessed the candidate pacificamide BGC, which we have named *pac*. A comparison of the *pac* BGC in the two *Salinispora* strains helped to establish BGC boundaries (*pac1*−*33*) and identify two transposase genes in CNS-960 (between *pac12* and *-13*) as the only difference in gene content (Figure 3a). When grown under similar conditions, we observed production of **1** in CNS-960, albeit in reduced quantity relative to strain CNT-855 (Figure 3d).

The *pac* BGC shares several genes with *asu* and *col* that encode essential enzymes in the biosynthesis of their cognate small molecules and served as key references for this

https://doi.org/10.1021/acs.jnatprod.1c01117
J. Nat. Prod. 2022, 85, 980−986

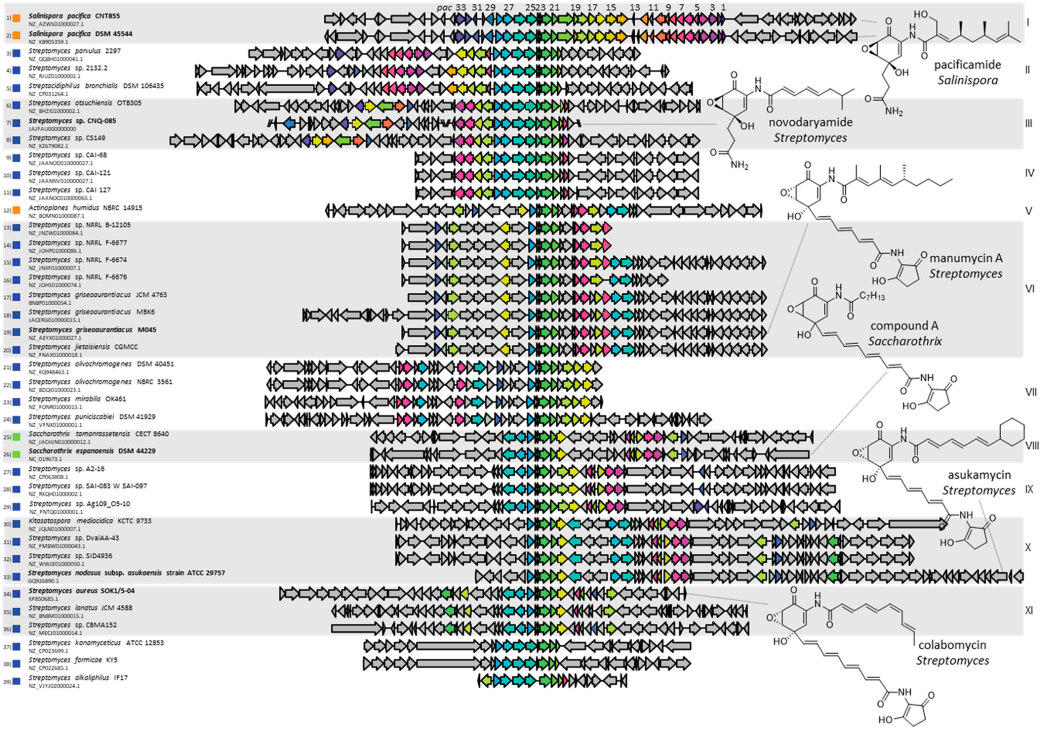


Figure 4. Manumycin-type BGCs identified in bacterial genome sequences. Colored genes represent *pac* gene homologues. Related groups of BGCs (I−XI) are shaded gray if they have been linked to a known metabolite or unshaded if they remain orphan (entries 37−39 are not predicted to encode manumycin-type metabolites). Structures linked to the *pac*, *dar*, *man*, *esp*, *asu*, and *col* BGCs are shown with the source species in bold. Colored squares indicate family level taxa (orange: *Micromonosporaceae*, blue: *Streptomycetaceae*, and green: *Pseudonocardiaceae*).

work.[24−26] *Pac* also shows similarity to the recently identified "compound A" BGC (*esp*) from *Saccharothrix espanaensis* DSM-44229,[28] the manumycin BGC (*man*), which we deduced from mining the manumycin A producer *Streptomyces griseoaurantiacus* M045 genome (RefSeq accession GCF_000204605.1),[29,30] and the daryamide BGC (*dar*), which we identified and named in the course of this study via genome sequencing of the daryamide and novodaryamide producer *Streptomyces* sp. CNQ-085 (NCBI accession no. JAJFAU000000000).[31,32] Key biosynthetic genes shared among these BGCs include three genes (*pac26*−*28*) related to the synthesis and priming of 3-amino-4-hydroxybenzoic acid (3,4-AHBA), two type II PKS genes (*pac21* and *-22*) involved in the production of polyene natural products (as classified by NaPDoS)[33] and associated with the formation of the "lower" side chain that extends from the carbonyl of 3,4-AHBA, two KS III genes (*pac4* and *pac6*) associated with the first condensation step in the biosynthesis of the "upper" acyl chain,[34] one arylamine *N*-acyltransferase gene (*pac17*) involved in the ligation of the "upper" side chain to the amino-hydroxy phenyl moiety, three oxidoreductase genes (*pac5*, *pac16*, and *pac18*) responsible for the formation of the 5,6-epoxy-4-hydroxycyclohex-2-en-1-one moiety, and acyl carrier protein, acyltransferase, and thioesterase encoding genes (Figure 3 and Supporting Information, Figure S11).[24−26] A feature of the biosynthesis that remains unknown is the "upper" side chain extension following priming by the KS III. This is suspected to involve fatty acid synthases encoded in other regions of the genome.[23]

To explain the structural differences between **1** and other manumycin-type compounds, we focused on gene differences among the *pac*, *asu*, *col*, *esp*, *man*, and *dar* BGCs (Figure S11). We found a predicted enoyl-reductase (ER, *fabV* homologue) gene in the *pac* and *dar* BGCs (*pac25* and *dar7*, respectively) that could be responsible for the saturated "lower" side chain in both **1** and the daryamides. This gene is lacking in the *asu*, *col*, *man*, and *esp* BGCs, which instead yield metabolites with a polyunsaturated "lower" side chain. We expect the presence of a similar gene in the BGCs that code for U62-162 and the salternamides, which also display a saturated lower side chain (Figure 1); however genome sequences are not available. Additionally, we found an asparagine synthase gene (a homologue of *nspN* from 4-hydroxy-3-nitrosobenzamide biosynthesis)[35] in the *pac* and *dar* BGCs (*pac20* and *dar24*) that could be responsible for installing the primary amides on the "lower" side chains of **1** and the daryamides. This gene is absent in the *asu*, *col*, *man*, and *esp* BGCs, which instead share three genes encoding the synthesis and ligation of the 2-amino-3-hydroxycyclopente-2-one ring observed in many manumycin-type metabolites.[24−26,36] Finally, we found a cytochrome

https://doi.org/10.1021/acs.jnatprod.1c01117
J. Nat. Prod. 2022, 85, 980−986

P450 gene (*pac15*) that is not observed in the other manumycin-type BGCs and could be responsible for installing the hydroxymethylene of **1**, which is a unique feature within the manumycin natural product family.

It is noteworthy that the *pac* BGC lacks the ketoreductase gene (KR, FabG homologue) and two dehydratase genes (DH, MaoC homologues) shared among all other manumycin-type BGCs (e.g., *asuC7*−*C9* and *dar1,18,19* in Figure 3a). These activities are expected during polyketide extension cycles to achieve the "upper" and "lower" side chains of **1**. Alternatively, we propose that *pac7*, which encodes a dehydrogenase (FadB homologue with documented ketoreduction activity[37]), is responsible for beta-ketoreduction and that *pac8*, which encodes for an enoyl hydratase/isomerase of the crotonase family, is responsible for the predicted dehydration steps, as has been previously observed.[38] During the initial extension leading to the "upper" side chain, we propose that *pac8* is also responsible for a 2-enoyl to 3-enoyl alkene shift resulting in the C-7′/C-8′ alkene in **1**. Another gene of interest is *pac33*, encoding a DsbA family oxidoreductase. While the role of this gene in the *pac* BGC is not known, some manumycins disrupt the mammalian DsbA-DsbB complex through direct covalent modification.[39] Thus, *pac33* may encode a resistant version of this target, as we have previously observed for a *fabB* homologue and the fatty acid inhibitor thiolactomycin.[40]

Next, we asked if bioinformatic analysis of key biosynthetic genes could facilitate prediction of the absolute configuration of **1**. Apart from the suspected involvement in "lower" side chain saturation, the *pac25* enoyl-reductase (FabV homologue) could also be responsible for the saturation on the "upper" side chain and, thus, instrumental in determining the configuration of the 4′ position. The stereospecificity of prokaryotic fatty acid enoyl-reductases reveals that protonation on the enolate intermediate occurs on the 2-*re* face.[41−43] Following this logic, **1** is predicted to have a 4′*S* and 6′*R* absolute configuration, which agrees with the configuration of the salternamides as assigned on the basis of ECD spectroscopy in combination with DFT calculations.[23]

Manumycin-Type BGC Diversity and Distribution. To gain insight into the broader diversity of manumycin-type BGCs, we queried *pac* against the NCBI reference sequence (refseq), nonredundant (nr), metagenomic (env_nr), and patented (pataa) protein databases and the MIBiG 2.0 BGC database using cblaster.[44] This analysis identified 29 manumycin-type BGCs in addition to the seven already mentioned (Figure 4). Hierarchical clustering based on best hit identity values[44] combined with manual comparison between the top cluster matches revealed 11 groups of closely related BGCs, six of which have been linked to specific manumycin-type metabolites (Figure 4). The remaining five BGC groups have unique organizations suggesting additional diversity remains to be discovered in this compound class. When placed in a phylogenomic context, the 36 BGCs are observed in three evolutionary distant families (*Micromonosporaceae*, *Streptomycetaceae*, and *Pseudonocardiaceae*) within the class Actinomycetia, suggesting they have been subject to horizontal gene transfer (Figure S12).

Biological Activities. Pacificamide and triacsin D were tested for antibacterial activity against *Escherichia coli* MG1655 and *Bacillus oceanisediminis* CNY-977 and for cytotoxic activity against the NCI-H460 lung cancer cell line. Pacificamide showed weak activity against *B. oceanisediminis* (MIC of 50 μM), while triacsin D was cytotoxic against the lung cancer cell line (EC_{50} of 5.5 ± 0.9 μM, Figure S15).

In conclusion, we report a new manumycin-type natural product from the marine actinomycete *Salinispora pacifica* along with its candidate biosynthetic gene cluster. The candidate BGC was compared to those reported for related compounds to explore the relationships between genetic and structural diversity. This BGC class is rare among sequenced genomes, and yet-to-be-characterized variants suggest that additional structural diversity remains to be discovered. Furthermore, this study confirms production of a natural product in the triacsin family from *Salinispora* spp., which was previously predicted based on genome mining.[45]

EXPERIMENTAL SECTION

General Experimental Procedures. Optical rotations were recorded on a Jasco P-2000 polarimeter. UV spectra were measured on a Beckman-Coulter DU800 spectrophotometer. ECD spectra were measured on a Jasco J-810 spectropolarimeter. IR spectra were acquired on a JASCO FTIR-4100 spectrometer. 1D and 2D NMR spectroscopic data were obtained on a JEOL 500 MHz or a Bruker 600 MHz NMR spectrometer. NMR chemical shifts were referenced to the residual solvent peaks (δ_H 3.31 and δ_C 49.15 for CD_3OD). High-resolution ESI-TOF mass spectrometric data were acquired on an Agilent 6530 Accurate-Mass Q-TOF mass spectrometer coupled to an Agilent 1260 LC system.

Cultivation of *Salinispora pacifica* CNT-855. A frozen stock of *S. pacifica* CNT-855 was inoculated into 50 mL of medium A1FBC [1% potato starch, 0.4% yeast extract, 0.2% peptone, 0.1% calcium carbonate, 0.01% potassium bromide, 0.04% iron sulfate (pentahydrate), and 2.2% InstantOcean in DI H_2O]. The seed culture was shaken at 200 rpm and 28 °C for 7 days and used to inoculate 1 L of medium A1FBC in a 2.8 L Fernbach flask. This culture was similarly shaken at 200 rpm and 28 °C for 7 days, after which 15 mL was inoculated into each of 10 × 2.8 L Fernbach flasks containing 1 L of medium A1FBC. After 4 days of shaking at 200 rpm and 28 °C, 25 g of sterile XAD-7 adsorbent resin was added to each flask. After three additional days of cultivation, the cultures were filtered through cheesecloth, the cells and resin were extracted with acetone (1 L) for 4 h, the extract was filtered through a cotton plug, and the acetone was removed via rotatory evaporation. The resulting extract was partitioned in a separatory funnel between EtOAc and H_2O (1:1 mixture, 600 mL total) and the organic phase collected, dried over anhydrous sodium sulfate, and concentrated via rotatory evaporation. Using this approach, it was not possible to determine if the compounds were localized to the cells, filtrate, or associated with both.

Purification of Pacificamide. The extract (260 mg) was suspended in EtOAc, mixed with diatomaceous earth, and concentrated via rotatory evaporation to yield a powder that was dried under high-vacuum pump. The powder was loaded into a C18 reversed-phase silica gel column (4 g) that had been equilibrated with H_2O (0.1% TFA). A six-step gradient from 100% H_2O (0.1% TFA) to 100% MeCN (0.1% TFA) was used to create six fractions of differing polarity. Fraction 3 contained pacificamide and was concentrated via rotatory evaporation and lyophilization to give a brown crude (40 mg), which was further separated by HPLC [mobile phase: 38% MeCN in H_2O (0.1% TFA); stationary phase: 5 μm, C8(2), 100 Å, 250 × 10 mm (Phenomenex, Luna) column] to yield pacificamide (t_R 24 min, 4.0 mg).

Pacificamide (1): clear film; $[\alpha]_D^{22}$ +54 (*c* 0.17, MeOH); UV/vis (MeOH) λ_{max} (log ε) 232 (3.37), 280 (3.10) nm; ECD (2.4 mM, MeOH) λ_{max} ($\Delta\varepsilon$) 210 (+1.17), 328 nm (+1.05); IR (ZnSe) ν_{max} 3330, 1662, 1628, 1047, 1024 cm^{-1}; ^1H and 2D NMR, Table S1; HR-ESI-TOF-MS *m*/*z* 443.2149 (calcd for $C_{22}H_{32}N_2O_6Na$, 443.2158).

BGC Bioinformatic Analyses. *Salinispora pacifica* CNT-855 (IMG genome ID 2515154128) and *Salinispora pacifica* CNS-960 (DSM 45544, IMG genome ID 2517287019) were analyzed with

https://doi.org/10.1021/acs.jnatprod.1c01117
J. Nat. Prod. 2022, 85, 980−986

antiSMASH 6.0[27] with detection strictness set to "loose". BGCs similar to *pac* were searched using the *pac1*−33 protein sequences against the NCBI reference (refseq), nonredundant (nr), metagenomic (env_nr), and patented (pataa) protein sequence databases using cblaster[44] with the following parameters: 500 BLASTp hits per query with a maximum 10 000 hits per search and 3 hits required to define a cluster, a maximum 0.01 E-value, minimum 20% identity (30% for refseq), minimum 45% query coverage for a BLASTp hit, and a maximum 20 000 bp distance from the cluster start/end and between intermediate gene hits. The *pac* BGC was queried against the MIBiG 2.0 repository using the same cblaster settings except a maximum 10 000 bp distance from start/end of a cluster to an intermediate gene. Queries resulted in 1518 clusters. The clusters were filtered to remove duplicates, ranked by cblaster cluster score, ordered by the best hit identity values using hierarchical clustering, and manually filtered for genes characteristic of manumycin-type BGCs.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jnatprod.1c01117.

Detailed experimental procedures, UV/vis, ECD and MS spectra, NMR spectroscopic data, DFT models and NMR predictions, BGC annotations and comparisons, phylogenomic analysis, and cytotoxicity results (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Paul R. Jensen − *Center for Marine Biotechnology and Biomedicine, Scripps Institution of Oceanography, University of California, San Diego, La Jolla, California 92093, United States;* orcid.org/0000-0003-2349-1888; Phone: 858-534-7322; Email: pjensen@ucsd.edu

Authors

Gabriel Castro-Falcón − *Center for Marine Biotechnology and Biomedicine, Scripps Institution of Oceanography, University of California, San Diego, La Jolla, California 92093, United States*

Kaitlin E. Creamer − *Center for Marine Biotechnology and Biomedicine, Scripps Institution of Oceanography, University of California, San Diego, La Jolla, California 92093, United States;* orcid.org/0000-0002-0666-2107

Alexander B. Chase − *Center for Marine Biotechnology and Biomedicine, Scripps Institution of Oceanography, University of California, San Diego, La Jolla, California 92093, United States;* orcid.org/0000-0003-1984-6279

Min Cheol Kim − *Center for Marine Biotechnology and Biomedicine, Scripps Institution of Oceanography, University of California, San Diego, La Jolla, California 92093, United States*

Douglas Sweeney − *Center for Marine Biotechnology and Biomedicine, Scripps Institution of Oceanography, University of California, San Diego, La Jolla, California 92093, United States*

Evgenia Glukhov − *Center for Marine Biotechnology and Biomedicine, Scripps Institution of Oceanography, University of California, San Diego, La Jolla, California 92093, United States*

William Fenical − *Center for Marine Biotechnology and Biomedicine, Scripps Institution of Oceanography, University of California, San Diego, La Jolla, California 92093, United States;* orcid.org/0000-0002-8955-1735

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jnatprod.1c01117

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported by the National Institutes of Health (R01GM085770) to P.R.J. including diversity supplement funds to G.C.F. We are grateful to the San Diego IRACDA postdoctoral program for funding to G.C.F. and the National Science Foundation Graduate Research Fellowship Program (under grant no. DGE-1650112) to K.E.C. We thank D. Fishman and J. Granger-Jones from Laser Spectroscopy Laboratories at UC Irvine for ECD measurements, T. Molinski and M. Salib from UCSD Chemistry and Biochemistry Department for access to IR measurements, B. Duggan from the UCSD SSPPS NMR Facility for assistance with NMR experiments, and Y. Su from the UCSD Molecular Mass Spectrometry Facility for HRMS measurements. A high-resolution LC-MS instrument was provided by the National Institutes of Health (S10 OD0106400).

■ REFERENCES

(1) Carlson, E. E. *ACS Chem. Biol.* **2010**, *5* (7), 639−653.
(2) Newman, D. J.; Cragg, G. M. *J. Nat. Prod.* **2020**, *83* (3), 770−803.
(3) Katz, L.; Baltz, R. H. *J. Ind. Microbiol Biotechnol.* **2016**, *43* (2−3), 155−176.
(4) Mincer, T. J.; Jensen, P. R.; Kauffman, C. A.; Fenical, W. *Appl. Environ. Microbiol.* **2002**, *68* (10), 5005−5011.
(5) Maldonado, L. A.; Fenical, W.; Jensen, P. R.; et al. *Int. J. Syst. Evol Microbiol.* **2005**, *55* (5), 1759−1766.
(6) Jensen, P. R.; Moore, B. S.; Fenical, W. *Nat. Prod Rep.* **2015**, *32* (5), 738−751.
(7) Feling, R. H.; Buchanan, G. O.; Mincer, T. J.; Kauffman, C. A.; Jensen, P. R.; Fenical, W. *Angew. Chemie Int. Ed.* **2003**, *42* (3), 355−357.
(8) Kim, L. J.; Xue, M.; Li, X.; et al. *J. Am. Chem. Soc.* **2021**, *143* (17), 6578−6585.
(9) Millán-Aguiñaga, N.; Chavarria, K. L.; Ugalde, J. A.; Letzel, A.-C.; Rouse, G. W.; Jensen, P. R. *Sci. Rep.* **2017**, *7* (1), 3564.
(10) Román-Ponce, B.; Millán-Aguiñaga, N.; Guillen-Matus, D.; et al. *Int. J. Syst. Evol Microbiol.* **2020**, *70* (8), 4668−4682.
(11) Chase, A. B.; Sweeney, D.; Muskat, M. N.; Guillén-Matus, D. G.; Jensen, P. R. *MBio.* **2021**, *12* (6), No. e02361.
(12) Sattler, I.; Thiericke, R.; Zeeck, A. *Nat. Prod Rep.* **1998**, *15* (3), 221−240.
(13) Tanaka, H.; Yoshida, K.; Itoh, Y.; Imanaka, H. *J. Antibiot (Tokyo).* **1982**, *35* (2), 157−163.
(14) Omura, S.; Tomoda, H.; Xu, Q. M.; Takahashi, Y.; Iwai, Y. *J. Antibiot (Tokyo).* **1986**, *39* (9), 1211−1218.
(15) Alcaraz, L.; Macdonald, G.; Ragot, J. P.; Lewis, N.; Taylor, R. J. K. *J. Org. Chem.* **1998**, *63* (11), 3526−3527.
(16) Matsumori, N.; Kaneno, D.; Murata, M.; Nakamura, H.; Tachibana, K. *J. Org. Chem.* **1999**, *64* (3), 866−876.
(17) Schmidt, Y.; Breit, B. *Org. Lett.* **2010**, *12* (10), 2218−2221.
(18) Willoughby, P. H.; Jansma, M. J.; Hoye, T. R. *Nat. Protoc.* **2020**, *15* (7), 2277.
(19) Willoughby, P.; Reisbick, S. *Protoc. Exch.* 2021, DOI: 10.21203/rs.2.1186/v2.
(20) Shen, B.; Whittle, Y. G.; Gould, S. J.; Keszler, D. A. *J. Org. Chem.* **1990**, *55* (14), 4422−4426.
(21) Mohamed, I. E.; Gross, H.; Pontius, A.; et al. *Org. Lett.* **2009**, *11* (21), 5014−5017.

(22) Fu, P.; La, S.; MacMillan, J. B. *J. Nat. Prod.* **2017**, *80* (4), 1096−1101.
(23) Kim, S.-H.; Shin, Y.; Lee, S.-H.; et al. *J. Nat. Prod.* **2015**, *78* (4), 836−843.
(24) Rui, Z.; Petříčková, K.; Škanta, F.; et al. *J. Biol. Chem.* **2010**, *285* (32), 24915−24924.
(25) Rui, Z.; Sandy, M.; Jung, B.; Zhang, W. *Chem. Biol.* **2013**, *20* (7), 879−887.
(26) Petříčková, K.; Pospíšil, S.; Kuzma, M.; et al. *ChemBioChem.* **2014**, *15* (9), 1334−1345.
(27) Blin, K.; Shaw, S.; Kloosterman, A. M.; et al. *Nucleic Acids Res.* **2021**, *49* (W1), W29−W35.
(28) Gorniaková, D.; Petříček, M.; Kahoun, D.; et al. *Microorganisms.* **2021**, *9* (3), 559.
(29) Li, F.; Maskey, R. P.; Qin, S.; et al. *J. Nat. Prod.* **2005**, *68* (3), 349−353.
(30) Li, F.; Jiang, P.; Zheng, H.; et al. *J. Bacteriol.* **2011**, *193* (13), 3417−3418.
(31) Asolkar, R. N.; Jensen, P. R.; Kauffman, C. A.; Fenical, W. *J. Nat. Prod.* **2006**, *69* (12), 1756−1759.
(32) Castro-Falcón, G.; Millán-Aguiñaga, N.; Roullier, C.; Jensen, P. R.; Hughes, C. C. *ACS Chem. Biol.* **2018**, *13* (11), 3097−3106.
(33) Ziemert, N.; Podell, S.; Penn, K.; Badger, J. H.; Allen, E.; Jensen, P. R. *PLoS One.* **2012**, *7* (3), No. e34064.
(34) Nofiani, R.; Philmus, B.; Nindita, Y.; Mahmud, T. *Medchemcomm.* **2019**, *10* (9), 1517−1530.
(35) Noguchi, A.; Kitamura, T.; Onaka, H.; Horinouchi, S.; Ohnishi, Y. *Nat. Chem. Biol.* **2010**, *6* (9), 641−643.
(36) Zhang, W.; Bolla, M. L.; Kahne, D.; Walsh, C. T. *J. Am. Chem. Soc.* **2010**, *132* (18), 6402−6411.
(37) Volodina, E.; Steinbüchel, A. *AMB Express.* **2014**, *4* (1), 69.
(38) Hamed, R. B.; Batchelar, E. T.; Clifton, I. J.; Schofield, C. J. *Cell. Mol. Life Sci.* **2008**, *65* (16), 2507−2527.
(39) Tuladhar, A.; Rein, K. S. *ACS Med. Chem. Lett.* **2018**, *9* (4), 318−322.
(40) Tang, X.; Li, J.; Millán-Aguiñaga, N.; et al. *ACS Chem. Biol.* **2015**, *10* (12), 2841−2849.
(41) Saito, K.; Kawaguchi, A.; Seyama, Y.; Yamakawa, T.; Oduda, S. *Eur. J. Biochem.* **1981**, *116* (3), 581−586.
(42) Hu, K.; Zhao, M.; Zhang, T.; et al. *Biochem. J.* **2013**, *449* (1), 79−89.
(43) Schiebel, J.; Chang, A.; Merget, B.; et al. *Biochemistry.* **2015**, *54* (10), 1943−1955.
(44) Gilchrist, C. L. M.; Booth, T. J.; van Wersch, B; van Grieken, L; Medema, M. H.; Chooi, Y.-H. *Bioinforma Adv.* 2021, vbab016, DOI: 10.1093/bioadv/vbab016.
(45) Twigg, F. F.; Cai, W.; Huang, W.; et al. *ChemBioChem.* **2019**, *20* (9), 1145−1149.

Structure and Candidate Biosynthetic Gene Cluster of a Manumycin-type Metabolite from *Salinispora pacifica*

Gabriel Castro-Falcón, Kaitlin E. Creamer, Alexander B. Chase, Min Cheol Kim, Douglas Sweeney, Evgenia Glukhov, William Fenical and Paul R. Jensen*

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

*Email: pjensen@ucsd.edu, phone: 858-534-7322

Cultivation of *Salinispora cortesiana* CNY-202. *S. cortesiana* CNY-202 was cultured as described for *S. pacifica* CNT-855. However, the culture volume was 6 L and the adsorbent resin was added at day eight and incubated for two additional days before extraction.

Purification of triacsin D. The crude extract (203 mg) from the *S. cortesiana* CNY-202 cultivation was fractionated following the same protocol as the *S. pacifica* CNT-855 extract. Fraction 4 containing triacsin D was separated using HPLC [mobile phase: 58% acetonitrile in water (0.1% TFA); stationary phase: 5 µm, C8(2), 100 Å, 250 x 10 mm (Phenomenex, Luna) column] to give a peak (t_R = 19 min) that was concentrated to give triacsin D (2.0 mg).

Genome sequencing and bioinformatic analyses of Streptomyces sp. CNQ-085. Genomic DNA extraction was performed using the Wizard Genomic DNA Purification Kit (Promega; Madison, WI) with the addition of lysozyme for Gram-positive bacteria. Library preparation and sequencing was performed at the Microbial Genome Sequencing Center (Pittsburg, PA) using an Illumina NextSeq 2500 instrument with 150 bp paired-end reads. Raw reads were quality filtered (trimq=20 minlen=70) with adapters, sequencing artifacts, and phiX removed using the BBMap toolkit.[1] Filtered reads were assembled using the SPAdes genome assembler[2] with a "careful" iterative k-step ranging from k= 31 to 125. The quality of the assembled genome was assessed by creating taxon-annotated-GC coverage plots by mapping back the raw sequence data for coverage of contigs and preliminary taxonomic assignment using MegaBLAST against the NCBI nt database. Gene annotation was performed with PROKKA (Seeman 2014) with the --notrna flag. BGCs were identified with antiSMASH v6.0 to identify the *dar* BGC based on similarity with the *asu* and *col* BGCs.

1. Bushnell, B. BBMap: A Fast, Accurate, Splice-Aware Aligner. United States: N. p., 2014. Web
2. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol. 2012 May;19(5):455-77. doi: 10.1089/cmb.2012.0021.
3. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics. 2014 Jul 15;30(14):2068-9. doi: 10.1093/bioinformatics/btu153.

Phylogenomic relationship of strains with *pac* and manumycin-like BGCs. One representative bacterial genome from each family within the class *Actinomycetia*[4] was downloaded using PhyloPhlAn 3.0[5] "phylophlan_get_reference" and NCBI-genome-download scripts (https://github.com/kblin/ncbi-genome-download) with n=60 genomes. PhyloPhlAn 3.0 was used to identify 334 conserved phylogenetic marker genes across the 60 genomes and calculate a phylogenomic tree (DIAMOND to map conserved marker genes from the 400 "phylophlan" bacterial and archaeal universal gene markers; MAFFT for alignment; trimAl for trimming; and RAxML with 100 rapid bootstraps with PROTCATLG model for final tree calculation). The tree was visualized and colored in FigTree[6] and R (ggtree, RColorBrewer, ggplot2).[7,8,9]

4. Salam N, Jiao JY, Zhang XT, Li WJ. Update on the classification of higher ranks in the phylum Actinobacteria. Int J Syst Evol Microbiol. 2020 Feb;70(2):1331-1355. doi: 10.1099/ijsem.0.003920. Erratum in: Int J Syst Evol Microbiol. 2020 Apr;70(4):2958. PMID: 31808738
5. Asnicar, F., Thomas, A.M., Beghini, F. et al. Precise phylogenetic analysis of microbial isolates and genomes from metagenomes using PhyloPhlAn 3.0. Nat Commun 11, 2500 (2020). https://doi.org/10.1038/s41467-020-16366-7
6. FigTree reference: Rambaut, A. (2016) FigTree v1.4.3. http://tree.bio.ed.ac.uk/software/figtree/
7. Yu, G. (2020) Using ggtree to Visualize Data on Tree-Like Structures. *Curr Protoc Bioinforma* 69: 1–18.
8. Neuwirth, E. and Neuwirth, M.E. (2014) Package 'RColorBrewer.' *Color Palettes*.
9. Wickham, H. (2016) ggplot2: Elegant Graphics for Data Analysis, Springer-Verlag New York.

Molecular calculations. Molecular calculations were performed following reference protocols.[10,11] Conformational searches on syn- and anti-models (**Supplementary Tables S2-S3**) were performed using Spartan Student v8, with conformer rules=normal. DFT calculations on the 20 lowest energy conformers for each diastereomer were performed in Gaussian 09 for geometry optimization and frequency calculation using the M06-2X functional and 6-31+G(d,p) basis set, with the use of a finer integration grid and solvation IEFPCM=methanol. NMR calculations on the optimized structures using GIAO Method with the B3LYP functional and 6-311+G(2d,p) basis set was used to compute NMR shielding tensors. Scaling and referencing factors, derived from linear regression analysis,[10] were applied to correct computed tensor values. The weighted average NMR shielding tensor for each model was attained using the Boltzmann weighting factors and the free energies from frequency calculations. MAEs (i.e. $\Delta\delta_{ave}$) for each model were obtained using the predicted chemical shifts of atoms in positions 2'-13' of models.

10. Willoughby PH, Jansma MJ, Hoye TR. Addendum: A guide to small-molecule structure assignment through computation of (^{1}H and ^{13}C) NMR chemical shifts. *Nat Protoc*. 2020;15(7):2277. doi:10.1038/s41596-020-0293-9
11. Willoughby P, Reisbick S. Generation of Gaussian 09 Input Files for the Computation of 1H and 13C NMR Chemical Shifts of Structures from a Spartan'14 Conformational Search. *Protoc Exch*. Published online 2021. doi:10.21203/rs.2.1186/v2

Biological assays. Antibacterial assays were performed following CSLI reference protocols.[12] Single *Escherichia coli* MG1655 and *Bacillus oceanisediminis* CNY-977 colonies grown on agar plates were inoculated on liquid LB and A1 media (3 mL), respectively, and incubated at 35°C with shaking. After 16 h, cultures were diluted to $5x10^5$ cfu/mL in their respective medium. Pacificamide and triacsin D stock solutions [10μL, 1.25 mM in DMSO] were transferred in triplicate to a 96-well plate serially diluted 5-fold in DMSO (testing concentrations = 50 μM – 0.6 nM and total volume = 8 μL/well). The diluted cultures were transferred into wells (192 μL/well) using a multi-channel pipette and mixed with the test compounds and DMSO controls by pipetting. Initial cell density readouts (OD_{650}) were obtained with a plate reader. Plates were incubated at 35°C with shaking, and after 18 h, cell-densities were re-read using the plate reader. Percent growth was calculated as the difference in absorbance of each well, before and after incubation, divided by the difference between absorbance of inoculated and un-inoculated control wells. Chloramphenicol (positive control) exhibited MICs of 1.0 μM against *Escherichia coli* MG1655 and 40 nM against *Bacillus oceanisediminis* CNY-977.

12. Weinstein, M., 2018. *Methods for Dilution Antimicrobial Susceptibility Tests for Bacteria That Grow Aerobically*. 11th ed. Clinical Laboratory Standards Institute. https://clsi.org/standards/products/microbiology/documents/m07/

Cytotoxicity assays were performed following a standard 48-hour exposure protocol.[13] Batches of NCI-H460 human large cell lung carcinoma cells (purchased from ATCC in December 2019) were frozen in DMSO and subsequently used in no more than 23 passages for each batch. The cells were cultured in RPMI-1640 medium with 10% standard fetal bovine serum and 1% penicillin/streptomycin. The suspended cells (180 μL) at an initial density of 3.33×10^4 cells/mL (pass 9) were seeded into each well of a 96-well microplate and allowed to adhere for 24 h as a monolayer. Pacificamide and triacsin D stock solutions [20μL, 3 mM in DMSO] were transferred in triplicate to a 96-well plate and used to make half-log serial dilutions (total volume of 20 μL/well with RPMI-1640 medium) with testing concentrations spanned 30 μM - 0.01 nM. Before staining with MTT (thiazolyl blue tetrazolium bromide 98%; Sigma-Aldrich), the cells were treated with the indicated test compounds for 48 h. Doxorubicin and 1% DMSO in RPMI 1640 without fetal bovine serum were used as positive and negative controls, respectively. $0D_{630\ and}$ OD_{570} of the stained lysate were measured in SpectraMax M2 microplate reader from Molecular Devices and converted to percent of cell viability with negative control values set as

100% survival. EC_{50} values were obtained from the dose response curves of percent of cell viability values of the triplicate tests versus logarithmic drug concentrations using GraphPad Prism 8.1.2. Doxorubicin (positive control) exhibited EC_{50} of 400 nM in the same experiment.

13. Yu HB, Glukhov E, Li Y, *et al*. Cytotoxic Microcolin Lipopeptides from the Marine Cyanobacterium Moorea producens *J Nat Prod 2019 Sep 27;82(9):2608-2619*

Figure S1. High resolution mass spectrum of pacificamide

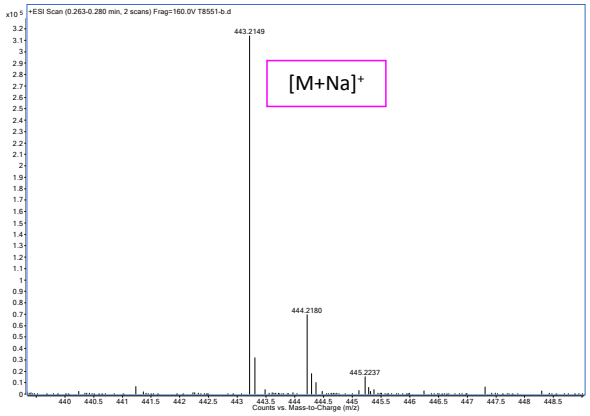


Figure S2. UV/vis spectrum of pacificamide

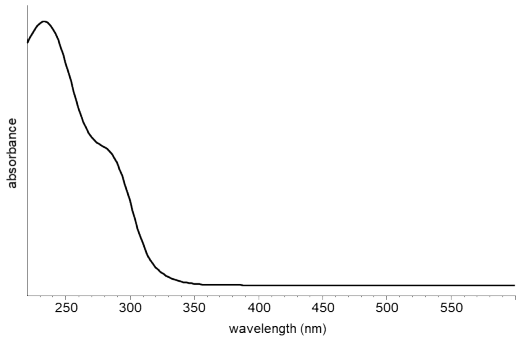


Figure S3. Circular dichroism spectrum of pacificamide

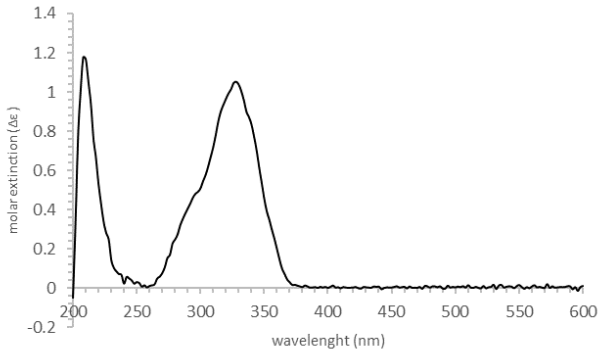
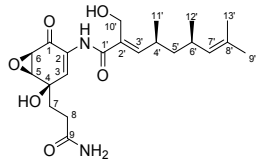


Table S1. NMR Spectroscopic Data (500 MHz, CD_3OD) for pacificamide (**1**)



position	δ_H (J in Hz)[a]	δ_C[b], type	COSY	HMBC	NOESY
1	----	190.0, C	----	----	----
2	----	130.4, C	----	----	----
3	7.29, br dd, (2.7, 0.7)	130.0, CH	H-5	C-1, C-2, C-5	H-7, H-8
4	----	71.4, C	----	----	----
5	3.66, dd, (3.9, 2.8)	58.4, CH	H-3, H-6	C-4[c]	H-6, H-7, H-8
6	3.57, d (3.9)	53.9, CH	H-5	C-1, C-2	H-5
7	2.09, m	37.2, CH_2	H-8	C-3, C-4, C-5, C-8, C-9	H-3, H-5, H-8
8	2.33, m	30.3, CH_2	H-7	C-4 [c], C-7, C-9	H-3, H-5, H-7
9	----	178.0, C	----	----	----
1'	----	169.3, C	----	----	----
2'	----	132.1, C	----	----	----
3'	6.49, d (10.3)	149.4, CH	H-4', H-10'	C-1', C-2', C-4, C-5', C-10', C-11'	H-4', H-5a', H-5b'[c], H-6'[c], H-11
4'	2.62, dq (10.3, 6.6)	31.5, CH	H-3', H-5a', H-5b', H-11'		H-3, H-5a', H-5b' H-10', H-11', H-12'
5'	a 1.34, dt (13.4, 6.7) b 1.24, ddd (13.4, 8.3, 6.8)	45.8, CH_2	H-4', H-6', H-5'b H-4', H-6', H-5'a	C-3', C-4', C-6', C-7', C-11', C-12' C-3', C-4', C-6',C- 7', C-11', C-12'	H-3', H-4', H-6', H-7', H-10', H-11' H-3', H-4', H-6', H-10', H-11'
6'	2.40, m	31.4, CH	H-5a', H-5b', H-7', H-12'		H-5a', H-5b', H-11', H-12', H-13'
7'	4.88 (d, 10.1)[d]	131.8, CH	H-6', H-9', H-13'	C-9', C-13'	H-5b', H-9', H-12'
8'	----	131.0, C	----	----	----
9'	1.68, br d (0.8)	25.7, CH_3	H-7'	C-7', C-8', C-13'	H-7'
10'	4.35, br s	57.4, CH_2	H-3'	C-1', C-2', C-3'	H-4'
11'	1.01, d (6.6)	20.4, CH_3	H-4'		H-3', H-4', H-5'b, H-6'
12'	0.90, d (6.6)	21.4, CH_3	H-6'		H-3', H-4', H-5a', H-5b', H-6', H-7'
13'	1.62, br d (0.8)	17.9, CH_3	H-7'	C-7', C-8', C-9'	H-6'

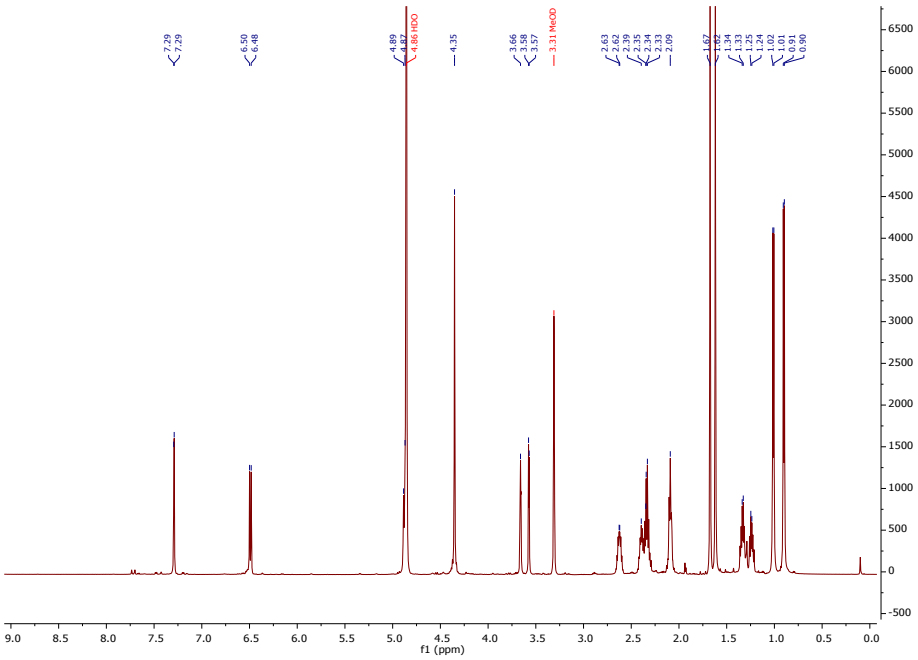
[a] 500 MHz

[b] Assignments were made on the basis of HSQC and HMBC

[c] Denotes weak signal

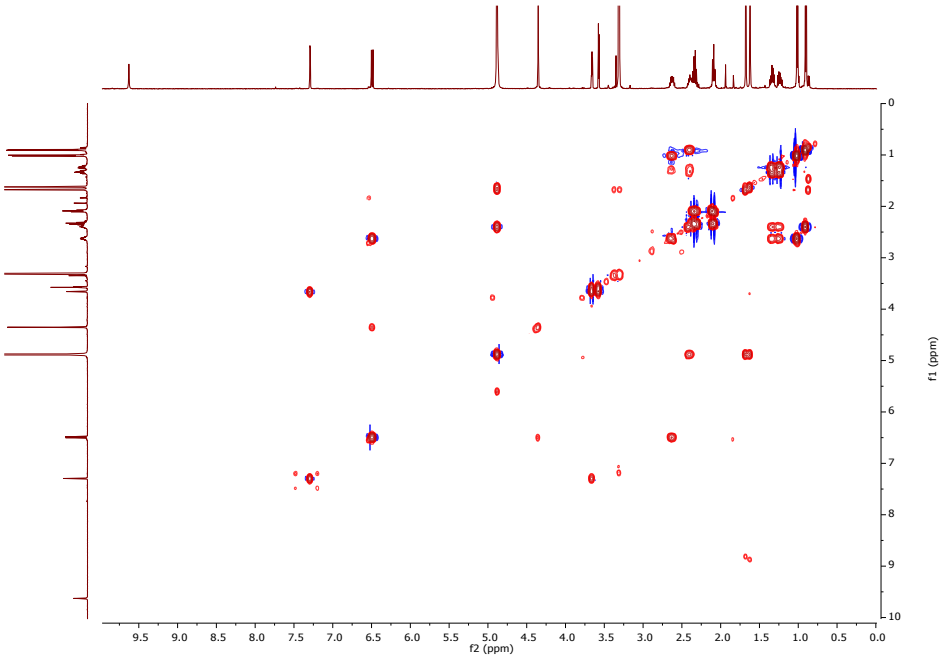
[d] Signal partially obscured

Figure S4. ^{1}H NMR spectrum of pacificamide (500 MHz, CD_3OD)



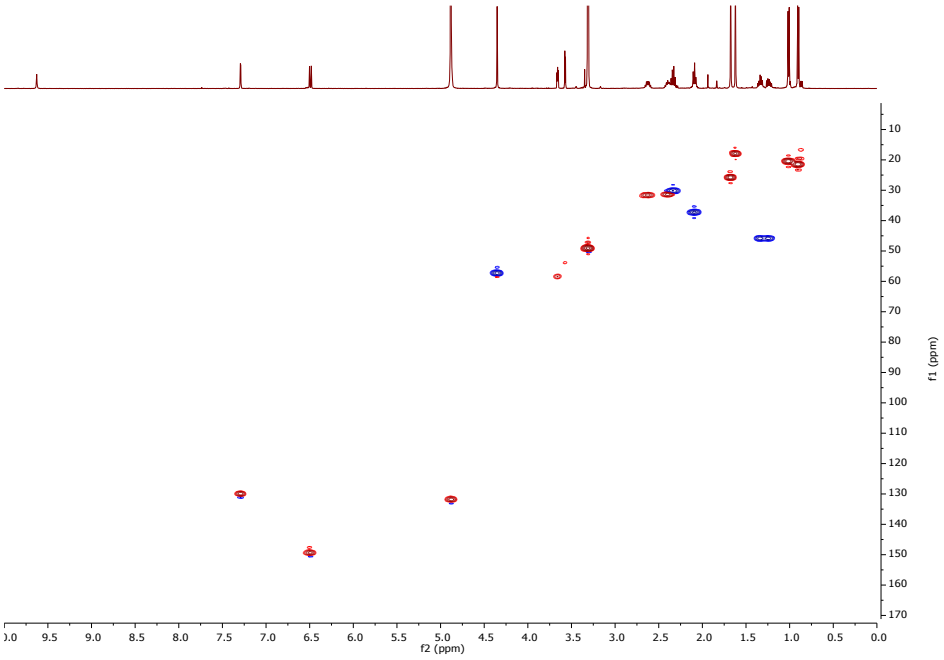
7

Figure S5. COSY spectrum of pacificamide (500 MHz, CD_3OD)



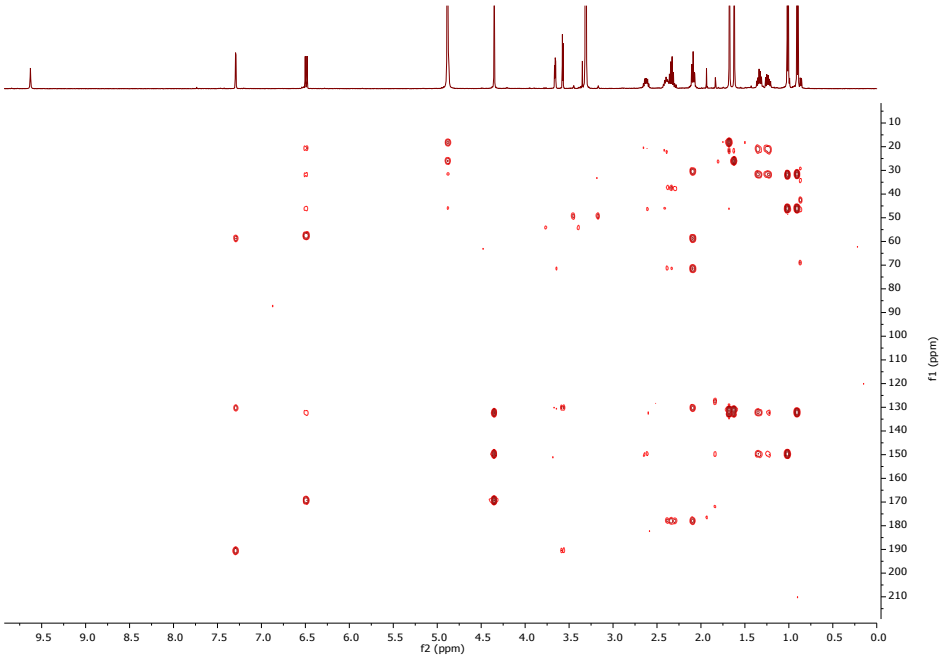
8

Figure S6. HSQC spectrum of pacificamide (500 MHz, CD_3OD)



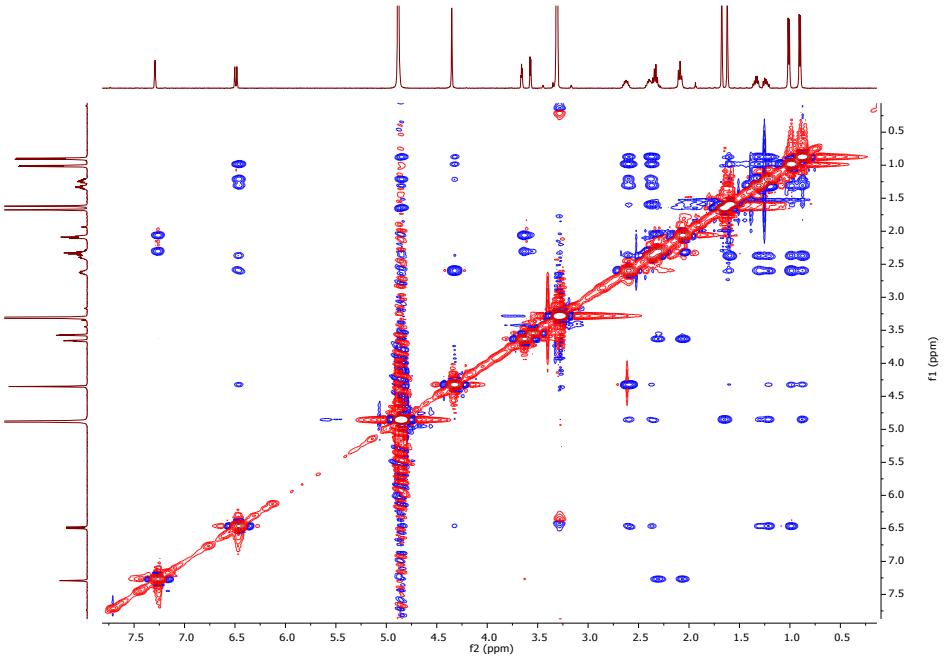
9

Figure S7. HMBC spectrum of pacificamide (500 MHz, CD_3OD)



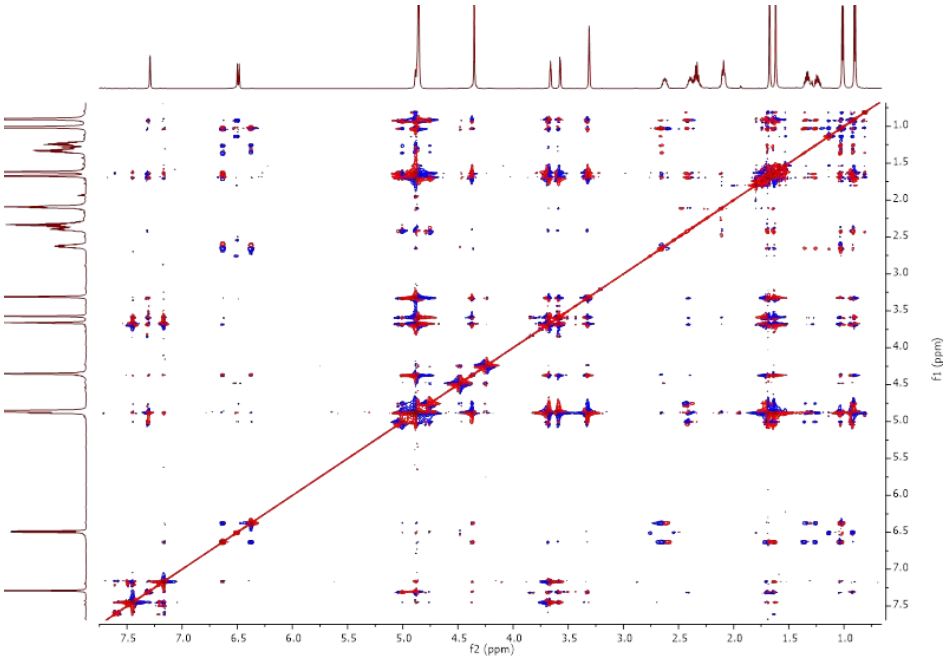
10

Figure S8. NOESY spectrum of pacificamide (500 MHz, CD_3OD)



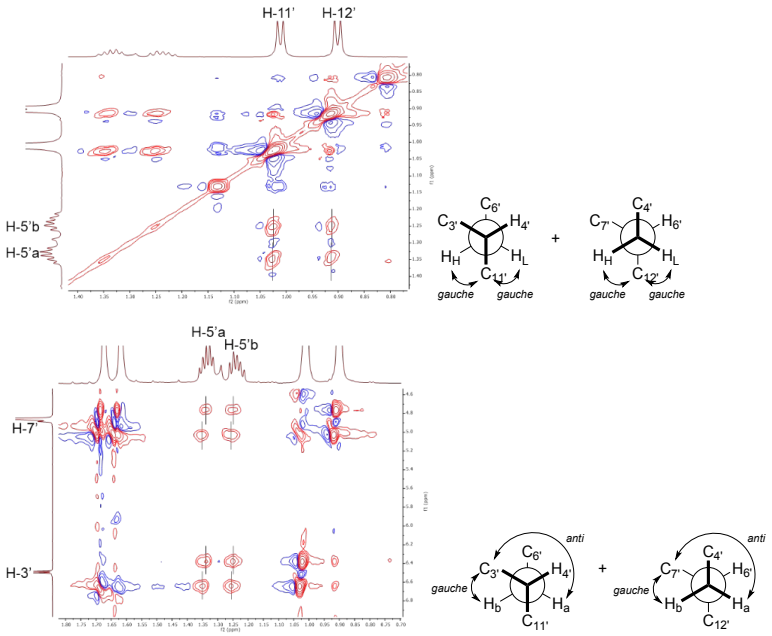
11

Figure S9. HETLOC spectrum of pacificamide in CD_3OD (600 MHz)



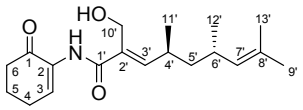
12

Figure S10. Regions of interest in HETLOC spectrum of pacificamide and representative conformations for *syn*-4',6'-dimethyl "upper" side chain.



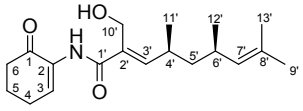
13

Tables S2-S3. Models and comparison between experimental and predicted NMR data



4'*S*, 6'*S* - *anti* model

Position	calc. δ_H	exp. δ_H	$\Delta\ \delta_H$	calc. δ_C	exp. δ_C	$\Delta\ \delta_C$
3'	6.448586	6.49	0.041414	149.0128	149.4	0.387203
4'	2.57391	2.62	0.04609	33.79157	31.5	2.291574
5'	a 1.444829	1.24	0.204829	44.06688	45.8	1.733116
	b 1.391129	1.34	0.051129			
6'	2.2019	2.40	0.1981	33.44366	31.4	2.043661
7'	4.869569	4.88	0.010431	130.1475	131.8	1.652462
8'				133.8438	131.0	2.843832
9'	1.697879	1.68	0.017879	24.4005	25.7	1.299499
10'	4.204854	4.35	0.145146	58.48738	57.4	1.087385
11'	0.987861	1.00	0.012139	19.33467	20.4	1.06533
12'	0.830731	0.90	0.069269	20.27407	21.4	1.125926
13'	1.407941	1.62	0.212059	16.81198	17.9	1.088024
		$\Delta\ \delta_{H\ average}$	0.09168		**$\Delta\ \delta_{C\ average}$**	1.510728



4'*S*, 6'*R* - *syn* model

Position	calc. δ_H	exp. δ_H	$\Delta\ \delta_H$	calc. δ_C	exp. δ_C	$\Delta\ \delta_C$
3'	6.44778	6.49	0.04222	149.3239	149.4	0.07615
4'	2.635473	2.62	0.015473	34.87887	31.5	3.378869
5	a 1.419134	1.34	0.079134	44.50195	45.8	1.298048
	b 1.221673	1.24	0.018327			
6'	2.450613	2.40	0.050613	33.94246	31.4	2.542458
7'	4.865269	4.88	0.014731	132.0822	131.8	0.28222
8'				132.8027	131.0	1.802668
9'	1.592318	1.68	0.087682	24.11121	25.7	1.588785
10'	4.244894	4.35	0.105106	58.43836	57.4	1.038361
11'	0.979211	1.00	0.020789	19.35695	20.4	1.043054
12'	0.863578	0.90	0.036422	19.8282	21.4	1.571797
13'	1.708808	1.62	0.088808	16.11512	17.9	1.784878
		$\Delta\ \delta_{H\ average}$	0.050846		**$\Delta\ \delta_{C\ average}$**	1.491572

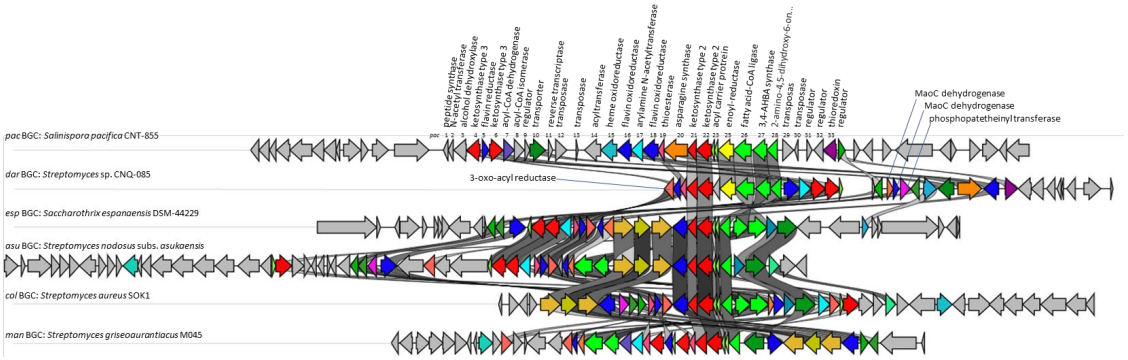
Table S4. Gene annotations for *pac* BGC from *Salinispora pacifica* CNT-855 (IMG genome ID 2515154128)

Gene	Protein ID:	NCBI Conserved Domain (CDD)	NCBI BLAST
pac-1	WP_033661723.1	no identified domains	hypothetical protein
pac-2	WP_027646816.1	Protein N-acetyltransferase	GNAT N-acetyltransferase
pac-3	WP_026324459.1	transposase family ISL3	transposase family ISL3
pac-4	WP_027646815.1	Domain of unknown function (DUF4365)	Domain of unknown function (DUF4365)
pac1	WP_018222396.1	Condensation domain of NRPS	amino acid adenylation domain-containing
pac2	None	No conserved domains	GNAT family N-acetyltransferase
pac3	WP_029128782.1	Zn-dependent alcohol dehydrogenase	Zn-biding dehydrogenase
pac4	WP_029128783.1	beta-ketoacyl-ACP synthase 3	beta-ketoacyl-ACP synthase 3
pac5	WP_033773657.1	FMN-binding domain found in NAD(P)H-flavin oxidoreductases	flavin reductase
pac6	WP_018222392.1	ketoacyl-acyl carrier protein synthase III	ketoacyl-ACP-synthase 3
pac7	WP_050563749.1	3-hydroxyacyl-CoA dehydrogenase FadB	3-hydroxyacyl-CoA dehydrogenase
pac8	WP_018222390.1	Crotonase/Enoyl-CoA hydratase superfamily	enoyl-CoA hydratase/isomerase family protein
pac9	WP_018222389.1	transcription regulator from the MerR superfamily	MerR family transcriptional regulator
pac10	WP_018222388.1	MFS family arabinose efflux permease	MFS transporter
pac11	WP_196233412.1	Reverse transcriptase-like super family	group II intron reverse transcriptase/maturase
pac12	None	IS110 family transposase	IS110 family transposase
pac13	None	Transposase, Mutator family	IS256 family transposase
pac14	WP_050585575.1	Peptidoglycan/LPS O-acetylase	acyltransferase
pac15	WP_080679239.1	cytochrome P450 family 158	cytochrome P450
pac16	None	Flavin-dependent oxidoreductase, luciferase family	LLM class flavin-dependent oxidoreductase
pac17	WP_018222380.1	Arylamine N-acetyltransferase	arylamine N-acetyltransferase
pac18	WP_018222379.1	2-polyprenyl-6-methoxyphenol hydroxylase	FAD-dependent monooxygenase
pac19	WP_018222378.1	Thioesterase, hotdog fold	Thioesterase, hotdog fold
pac20	WP_029128786.1	Asparagine synthase (glutamine hydrolyzing)	Asparagine synthase (glutamine hydrolyzing)
pac21	WP_018222376.1	Beta-ketoacyl synthase 2	Beta-ketoacyl synthase, chain length factor
pac22	WP_018222375.1	Beta-ketoacyl synthase 2	Beta-ketoacyl-ACP synthase
pac23	WP_029128787.1	No conserved domains	acyl carrier protein
pac24	WP_050585576.1	No conserved domains	acyl carrier protein
pac25	WP_155253037.1	Enoyl-ACP-reductase FabV	Enoyl-ACP-reductase FabV
pac26	WP_033773660.1	Acyl-CoA-synthetase	fatty-acid-CoA ligase
pac27	WP_029128788.1	3-dehydro-quinate synthase, class II	3-dehydroquinate synthase II
pac28	WP_026185654.1	fructose-bisphosphate aldolase	fructose-bisphosphate aldolase
pac29	None	ISL3 family transposase	ISL3 family transposase
pac30	None	Transposase, Mutator family	IS256 family transposase
pac31	WP_085981346.1	response regulator, NarL/FixJ family	LuxR transcriptional regulator
pac32	WP_155253038.1	response regulator, NarL/FixJ family	Response regulator transcription factor
pac33	WP_018222363.1	Protein Disulfide Oxidoreductases	DsbA family oxidoreductase

Table S5. Gene annotations for *dar* BGC from *Streptomyces sp.* CNQ-085 (NCBI accession JAJFAU000000000)

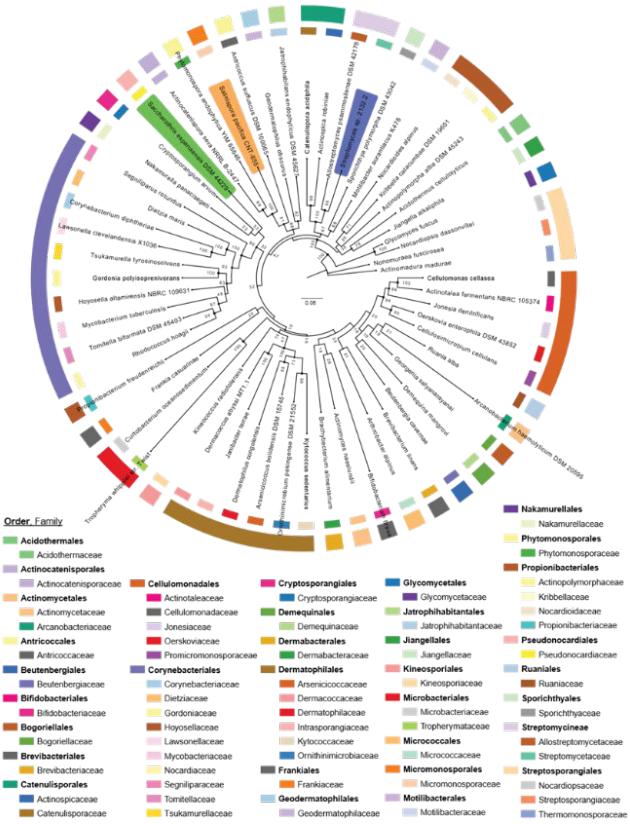
Gene	NCBI Conserved Domain (CDD)	NCBI BLAST
dar1	3-oxoacyl-ACP reductase FabG	3-oxoacyl-ACP reductase FabG
dar2	Flavin reductase	Flavin reductase family protein
dar3	PaaI thioesterase	PaaI family thioesterase
dar4	Beta-ketoacyl-APC synthase, type I and II	Beta-ketoacyl synthase, chain length factor
dar5	Beta-ketoacyl-ACP synthase FadB	Beta-ketoacyl-ACP synthase
dar6	Acyl carrier protein	Acyl carrier protein
dar7	Enoyl-ACP reductase FabV	Enoyl-ACP reductase FabV
dar8	Acyl-CoA synthetase	Acyl-CoA ligase
dar9	3-dehydroquinate synthase II	3-dehydroquinate synthase II
dar10	Fructose-1,6-biphosphate aldolase class I	Fructose-bisphosphate aldolase
dar11	2-polyprenyl-6-methoxyphenol hydroxylase	FAD-dependent monooxygenase
dar12	Arylamine N-acetyltransferase	Arylamine N-acetyltransferase
dar13	Beta-ketoacyl-ACP synthase III, FabH	Beta-ketoacyl-ACP synthase III
dar14	Beta-ketoacyl-ACP synthase III, FabH	Beta-ketoacyl-ACP synthase III
dar15	Acyl carrier protein	Acyl carrier protein
dar16	2-isopropylmalate synthase	2-isopropylmalate synthase
dar17	Transcriptional regulator, AcrR family	transcriptional regulator, TetR family
dar18	Beta-acyl-ACP dehydratase, FabA	MaoC family dehydratase
dar19	Beta-acyl-ACP dehydratase, FabA	MaoC family dehydratase
dar20	Phosphopantetheinyl transferase	Phosphopantetheinyl transferase
dar21	response regulator, NarL/FixJ family	transcriptional regulator, LuxR
dar22	!S630 family transposase	IS630 family transposase
dar23	arabinose efflux permease, MFS family	MFS transporter
dar24	Asparagine synthetase (glutamine-hydrolyzing)	asparagine synthase
dar25	Luciferase-like monooxygenase	LLM class flavin-dependent oxidoreductase
dar26	DsbA-like thioredoxin protein	DsbA family oxidoreductase
dar27	Nucleoside-diphosphate-sugar epimerase	NAD(P)-dependent oxidoreductase
dar28	Hot dog fold protein, FapR regulator	A-factor biosynthesis protein
dar29	transcriptional regulator, AcrR family	TetR/AcrR family transcriptional regulator
dar30	IS110 family transposase	IS110 family transposase
dar31	IS110 family transposase	IS110 family transposase
dar32	transcriptional regulator, AcrR family	Gamma-butyrolactone-binding protein
dar33	IS256 family transposase	IS256 family transposase
dar34	No conserved domains	replication relaxation protein

Figure S11. Comparison of manumycin-type BGCs that have been linked to characterized natural products



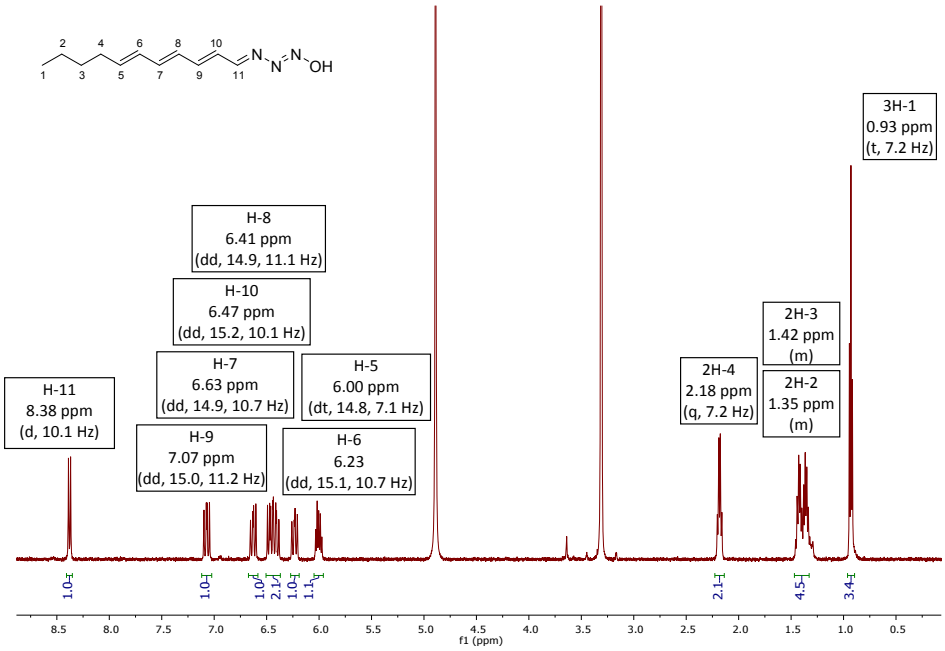
Shared genes are colored by function: ketosynthases (red), oxidoreductases involved in 5,6-epoxy-4-hydroxy-cyclohex-2-en-1-one synthesis (blue) synthesis, genes involved in 3-amino-4-hydroxybenzoic acid (3,4-AHBA; bright green) synthesis, N-aryl aminotransferase (baby blue), enoyl-ACP-reductase (yellow), 3-hydroxybutyryl-CoA dehydrogenase (blue purple), asparagine synthase (orange), FabG 3-oxo acyl reductase (carnation), MaoC family dehydratase (orange and blue), thioredoxin (purple), thioesterase (salmon), phosphopantetheinyl transferase (pink), regulator and transport (dark green) genes installing 2-amino-3-hydroxy-cyclopente-2-one (gold). Note: oxidoreductases in blue and MaoC family dehydratases show homology.

Figure S12. Phylogenomic distribution of manumycin-type BGCs within the class *Actinomycetia* (phylum Actinobacteria)



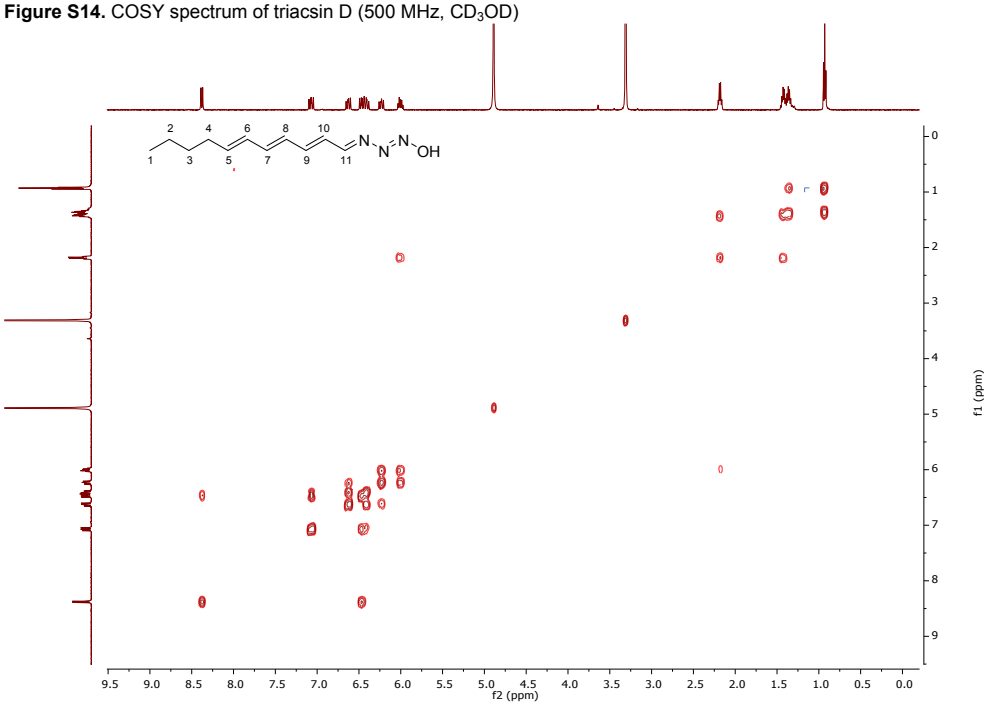
Phylogenomic tree of class *Actinomycetia* with one representative genome from all families (n=60 genomes, 334 conserved marker genes from each genome was used to calculate the phylogeny). The three families that contain manumycin-type BGCs include the highlighted orange (*Micromonosporaceae*), blue (*Streptomycetaceae*), and green (*Pseudonocardiaceae*) taxa. The outer ring color denotes taxonomic order, and the inner ring denotes taxonomic family, as listed.

Figure S13. ^{1}H NMR spectrum triacsin D (500 MHz, CD_3OD)



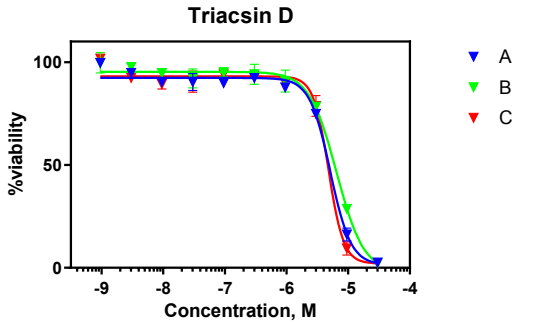
19

Figure S14. COSY spectrum of triacsin D (500 MHz, CD_3OD)



20

Figure S15. Dose response curves for triacsin cytotoxicity (in triplicate) against NCI-H460 lung cancer cell line



	A	B	C
Sigmoidal dose-response (variable slope)			
Best-fit values			
Bottom	1.382	-2.358	2.061
Top	92.33	95.29	93.06
LogEC50	-5.291	-5.189	-5.319
HillSlope	-2.612	-1.959	-3.540
EC50	5.119e-006	6.466e-006	4.793e-006

A.4 Acknowledgements

Appendix A (Section A.3), in full, is a reprint of the material as it appears in the *Journal of Natural Products Genomics* 85(4), 980-986. Castro-Falcón, G.; Creamer, K.E.; Chase, A.B.; Kim, M.C.; Sweeney, D.; Glukhov, E.; Fenical, W.; and Jensen, P.R., 2022. The dissertation author was the second author of this paper.