UC San Diego

UC San Diego Electronic Theses and Dissertations

Title

From nucleotide to gene cluster: Differences driving the specialized metabolism of the marine actinomycete Salinispora

Permalink

https://escholarship.org/uc/item/7zn966db

Author

Guillen Matus, Dulce Guadalupe

Publication Date

2022

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

UNIVERSITY OF CALIFORNIA SAN DIEGO

From nucleotide to gene cluster: Differences driving the specialized metabolism of the marine actinomycete *Salinispora*

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

in

Marine Biology

by

Dulce Guadalupe Guillén Matus

Committee in charge:

Professor Paul R. Jensen, Chair
Professor Lihini Aluwihare
Professor Douglas Bartlett
Professor Michael Burkart
Professor Lena Gerwick

2022

Copyright

Dulce Guadalupe Guillén Matus, 2022

All rights reserved.

The Dissertation of Dulce Guadalupe Guillén Matus is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2022

DEDICATION

A mis padres, Beyda y Manuel, porque sin ustedes nada de esto hubiera sido posible.

EPIGRAPH

"I play with microbes.
There are, of course, many rules to this play...
but when you have acquired knowledge and experience it is very pleasant to break the rules
and to be able to find something nobody has thought of."

-Alexander Fleming

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

First, I would like to thank my advisor, Dr. Paul Jensen for accepting me into his lab and for the support he has given me, for being open to my ideas and for all the feedback, even when you told me *not to go to you for wisdom*, you have been a source of wisdom in many aspects of my development as a scientist, nonetheless. For the good and the bad, thank you.

I would also like to thank all the members of my committee: Lena Gerwick, Lihini Aluwihare, Doug Bartlett, and Michael Burkart, for their advice and support during my time as student, and for all your input in my research. Special thanks to Lena, not everyone would have agreed to meet almost a month in advance while in the other side of the world for a dissertation defense, I greatly appreciate it.

I would also like to thank the SIO Grad Office, Gilbert Bretado, Olivia Padilla, Jessica Gonzalez, Tim De Bold, Joshua Reeves, Shelley Weisel, Dana Jimenez, Maureen McGreevy, and Gregory Jackson for all your help during my time at SIO as an international student. I would like to thank specially to Adrielle Wai, in the business office, for all your help and patience with the many shipments I had to do, even when that was not part of your job description.

I would like to thank all the members of the Jensen laboratory, previous and current, for the laughs, for sharing your knowledge with me, for all the brainstorming sessions, and for making my time in in the lab a positive and fun experience. Hans, Gabriel, Johanna, Alma, Krystle, Alyssa, Julia, Henrique, and Leesa thank you all for your support and your friendship.

Special thanks to my dear friend Kaitlin Creamer, for the long nights at the lab, for taking me to outdoors adventures, and for always been there to listen and to support me. Because I could not have asked for a better friend to share my PhD journey with. It has been a rough journey, but

I would not change a thing because you were there with me. Afterall, as our favorite poster said, all we needed was the right equipment and a slight chemical imbalance in the brain. We did it!

I'm forever grateful to Natalie Millán for always believing in me, for your support and for your friendship, much of what I am as a person and as a scientist is thanks to you.

Lastly, I would like to thank my family, because without you all I would not be half the person I am today. Gracias a mis papás, Manuel, por que me sigues enseñando aunque ya no estés conimigo, por tus sacrificios, por sembrar en mí el amor a seguir aprendiendo, y por ser el mejor papá. Beyda, por todo el amor que me has dado, por alentarme a seguir adelante, por tus consejos y regaños, por siempre apoyar mis locuras, y por ser la mejor mamá. Gracias a mis hermanos, Rosi, Beydi y Nelo, por ser un ejemplo a seguir, por las risas y las peleas, por aguantar todos mis berrinches y por apoyerme incondicionalmente, los amo.

Chapter 4 is coauthored with Gabriel Castro-Falcón and Paul R. Jensen. The dissertation author was the primary investigator and author of this chapter.

Chapter 5 is coauthored with Alexander B. Chase and Paul R. Jensen. The dissertation author was the primary investigator and author of this chapter.

VITA

2006 – 2011	Bachelor of Science in Oceanography Universidad Autonoma de Baja California, Mexico
2012 – 2014	Master of Science in Coastal Oceanography Universidad Autonoma de Baja California, Mexico
2016 – 2022	Doctor of Philosophy in Marine Biology Scripps Institution of Oceanography University of California San Diego. La Jolla, CA

PUBLICATIONS

Schlawis Christian, Tim Harig, Stephanie Ehlers, **Dulce G. Guillén Matus**, Kaitlin E. Creamer, Paul R. Jensen, Stefan Schulz (2020). Extending the Salinilactone Family. ChemBioChem. 21(11):1629-1632.

Román-Ponce Brenda, Natalie Millán-Aguiñaga, **Dulce Guillén Matus,** Alexander B. Chase, Joape G.M. Ginigini, Katy Soapi, Klaus D. Feussner, Paul R. Jensen, Martha E. Trujillo. (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., *Salinispora oceanensis* sp. nov. and *Salinispora vitiensis* sp. nov., and emended description of the genus *Salinispora.* International Journal of Systematic and Evolutionary Microbiology 70 (8) 4668-4682.

Schorn, M.A., *et al.* (2021). A community resource for paired genomic and metabolomic data mining. *Nat Chem Biol* 17, 363–368.

Zamora-Quintero Andrea Y., Monica Torres-Beltrán, **Dulce G. Guillén Matus**, Irasema Oroz-Parra, Natalie Millán-Aguiñaga. (2021) Rare actinobacteria isolated from the hypersaline Ojo de Liebre Lagoon as source of novel bioactive compounds with biotechnological potential. *Submitted to: Microbiology.*

Chase Alexander B., Douglas Sweeney, Mitchell N. Muskat, **Dulce Guillén-Matus,** Paul R. Jensen (2021). Vertical Inheritance Facilitates Interspecies Diversification in Biosynthetic Gene Clusters and Specialized Metabolites. *mBio, 12*.

Capone Dante, Emma J. Chamberlain, Kaitlin E. Creamer, **Dulce Guillén Matus**, Sidney Plummer, Sarah L. Romero, Sabrina L. Ulfer, H. Zoey Yin. (March 2022) "Let's talk about science: why and how we need to do better." *Explorations Now.*

Guillén Matus Dulce G., Gabriel Castro-Falcón, Paul R. Jensen. Assessing the ketomemicin BGC distribution and new analog production in *Salinispora* species. *In preparation.*

Guillén Matus Dulce G., Chase A.B., Paul R. Jensen. New chemotaxonomic markers for bacteria: going beyond cellular wall components. *Literature review, in preparation*

ABSTRACT OF THE DISSERTATION

From nucleotide to gene cluster: Differences driving the specialized metabolism of the marine actinomycete *Salinispora*

by

Dulce Guadalupe Guillén Matus

Doctor of Philosophy in Marine Biology

University of California San Diego, 2022

Professor Paul R. Jensen, Chair

Bacteria are the source of many important medicines, including antibiotics, anticancer, anti-inflammatory, and antiviral. In the last decades, the increased access to bacterial genome sequences revealed a wealth of unexplored biosynthetic potential. This discovery has revolutionized the way we research natural products. One remarkable example of natural product wealth is the marine actinomycete *Salinispora,* a prolific producer of many specialized metabolites with various biological activities and exceptional structures. However, our understanding of the evolutive and ecological forces driving specialized metabolism in bacteria remains limited. Here I studied the genomic differences driving the specialized metabolism of *Salinispora* via an *omics*

approach. I started with a comprehensive assessment of biosynthetic gene cluster distributions and metabolite production patterns in *Salinispora* spp*.,* resulting in 15 unique biosynthetic gene clusters distributed across seven different species. Subsequently, I assessed the expression of these unique biosynthetic gene clusters in standard laboratory conditions for *Salinispora,* finding 80% of the species-specific biosynthetic genes to be silent. I then, focused on ketomemicin and its biosynthetic gene cluster distribution in *Salinispora*. I found two new ketomemicin analogs and discovered the ketomemicin gene cluster is highly conserved in *Salinispora,* but surprisingly scarce in other actinomycetes. Additionally, I investigated the genomic differences behind a rare mutation disrupting the production of the glycosylated carotenoid sioxanthin in six different strains of two *Salinispora* species. I was able to link the albino phenotype to several point mutations in coding and non-coding regions of genes related to early stages of sioxanthin production. The extend of the ecological outcomes of this mutation remains to be determined. My findings provide insights into the genomic differences governing the specialized metabolism in *Salinispora* and suggest its importance as a phenotypic difference in their species. Further, it extends our knowledge on the biosynthetic potential *Salinispora* harbors, arising curiosity on the environmental and ecological dynamics Salinispora has adapted in nature that resulted in the vast natural product repertoire it contains.

Chapter 1

Introduction

1.1 Primary and secondary metabolism

All organisms need to transform a vast variety of organic compounds to live, grow and reproduce. From humans, animals, and plants to microorganisms, they all need to provide themselves with energy and the building blocks to synthesize the biomolecules required for growth and development. Despite the extremely distinct characteristics of all living organisms, the processes for modifying and synthesizing carbohydrates, proteins, fats, and nucleic acids are very similar; these fundamental processes driving the survival and reproduction of life are described as primary metabolism, and the compounds involved as primary metabolites. In contrast, there exist other compounds called secondary, or specialized metabolites also known as natural products, which do not directly contribute to the basal metabolism of the organisms and have a much more limited distribution in nature, albeit still found in almost all levels of life from terrestrial to marine environments (Petersen *et al*., 2020).

Originally, specialized metabolites were designated as secondary, and the reactions involved in their biosynthesis as secondary metabolism. This designation came from the belief that these metabolites were byproducts of primary metabolism; specialized metabolites are usually smaller (<1500 Da) and incorporate building blocks derived from primary metabolism, although their structures resemble nothing like primary metabolites (Harper *et al.,* 2001). The number of building blocks needed in specialized metabolism is surprisingly small for the vast array of compounds that can be built up from them. Some of the most important units employed are

isoprenoids, amino acids, and derivatives of acetyl coenzyme A (Dewick 2011). Further, specialized metabolites can be either synthesized by combining several units of the same type or from a mixture of different types, resulting in extensive structural diversity.

Although the building blocks for specialized metabolites are derived from primary metabolism, their structures and biological functions are different as these metabolites are usually produced under specific conditions to serve specific, largely ecological purposes. Undoubtedly, specialized metabolites play a vital role in the well-being of the producer and increase their likelihood of survival (Petersen *et al*., 2020).

Unlike primary metabolites, specialized metabolites are often unique to a group of organisms and may be considered an individual response of the organisms to their environment (O'Brien & Wright, 2011). Some known functions of specialized metabolites are as chemical defenses against predators (Rohde & Schupp, 2018), attractants towards the same or other species (Blythe *et al.,* 2020), and deterrents or repellents to other organisms. Similarly, they can provide a selective advantage against competing organisms for nutrient acquisition (Garénaux *et al*., 2011), and as protection against high light intensities or temperatures (Lyons *et al*., 2010). Thanks to their vast biological functions and diverse structures, many secondary metabolites have pharmacological properties as well as uses in the food and fragrance industries (Carvalho *et al.*, 2019).

1.2 Biosynthetic gene clusters

Chemists typically classify specialized metabolites according to the metabolic pathways from which they were derived (Harper *et al*., 2001). In some organisms, especially bacteria and fungi, genes involved in the biosynthesis of a specific specialized metabolite generally occur as a gene cluster and these groups of genes are referred to as biosynthetic gene clusters (BGCs) in the

natural products community. There is no rule that all the genes encoding the production of a specialized metabolite should be clustered together, however, this characteristic has proven surprisingly resilient across most microorganisms and in some plants (Nützmann & Osbourn 2014) and animals (Torres *et al*., 2020). The clustering is not essential for gene function, but it seems to be advantageous for regulation (Brakhage, 2013) and, particularly in microorganisms, for the horizontal exchange of genetic information (Diminic *et al*., 2014). Additionally, clustering with the biosynthetic genes we can find other genes encoding for regulation elements, transport, and resistance.

BGCs associated with the biosynthesis of different structural classes of compounds are generally named after the key biosynthetic enzymes such as non-ribosomal peptide synthetases (NRPS) and polyketide synthases (PKS) or after the type of compound produced, e.g., terpenes. NRPS and PKS BGCs are popular targets for natural product discovery as they are known to synthesize diverse antibiotics and other compounds with pharmaceutical activity (Chen *et al*., 2020).

1.2.1 Polyketides and polyketide synthases

Polyketides represent a diverse class of molecules with complex carbon skeletons and functional groups (Gallimore 2009). Polyketides are synthesized either in an assembly-line manner or iteratively, by polyketide synthase enzymes. This is mediated by consecutive Claisen condensation reactions between two ACP bound acetyl groups resulting in a poly-beta-keto backbone also known as polyketide (Dewick 2011).

Polyketide synthases are classified in three different types according to their enzymatic architecture: Type I PKSs are multifunctional enzymes, often organized in catalytic modules. Each module consists of three minimal domains: a ketosynthase (KS) domain, responsible for the

condensation of the two acyl units, an acyltransferase (AT) domain, that determines the two-carbon chain extender that will be used, and an acyl carrier protein (ACP) domain, linked to the polyketide chain and allowing the various domains to be accessed (Ma *et al*., 2019). Large, macrocyclic polyketides containing a cyclic ester are known as macrolides and usually produced by type I PKSs. Type II PKSs are known as minimal PKSs, are composed of a complex of individual monofunctional proteins; and need an AT and ACP domain. Type II PKS generally act iteratively and usually produce aromatic molecules (Hertweck *et al*., 2007). Type III PKSs do not use ACP domains, they act iteratively as well as the type II, but usually only need a single active site in each monomer to catalyze the priming, extension, and cyclization reactions unlike type I and type II PKSs (Rajesh, 2019). Despite the differences in enzymatic architecture, the biochemical aspects of chain construction are effectively the same in all PKS types (Dewick 2011).

Polyketide natural products are of great interest because of their importance for human health and disease treatment. Many of these compounds are well-known for their anti-infective, antitumor, antiparasitic, immunosuppressive, and other pharmaceutical properties (Demian & Sánchez, 2009). Polyketides are widely distributed in nature and are produced in bacteria, fungi (Lane & Moore, 2011), and some animals (Torrest & Schmidt, 2019).

1.2.2 Nonribosomal peptides and NRP synthetases

Nonribosomal peptides are of microbial origin and comprise a large pool of biologically active natural compounds with a broad spectrum of clinical applications, such as antibiotics (Mogi & Kita, 2009), antitumor (Shen *et al*., 2001) and antifungal drugs (Jiang *et al*., 2016), and immunosuppressants (Cohen *et al*., 1984). This medley of biological activities is in part a consequence of how these compounds are synthesized. Nonribosomal peptides result from the consecutive condensation of amino acids by large multimodular enzymes known as nonribosomal

peptide synthetases or NRPSs. This process is not limited to the 20 proteinogenic amino acids, in fact about 500 different monomers have been reported as building blocks for nonribosomal peptides including non-proteinogenic amino acids, fatty acids, and alpha-hydroxy acids (Caboche *et al*., 2010).

NRPSs are composed of distinct modular sections, each of which is responsible for the incorporation of one defined monomer into the final peptide product in an assembly line manner. There are specific modules in a NRPS, which can be further subdivided into catalytic domains. Three domains that are essential for nonribosomal peptide synthesis: the adenylation (A) domain, which is responsible for the activation of the amino acid; the peptidyl carrier protein (PCP) domain, which is responsible for the propagation of the growing peptide chain; and the condensation (C) domain, whioch is responsible for the binding of the amino acids. A thioesterase (TE) domain is responsible for the release of the peptide and is only found in the termination module of a NRPS (Strieker *et al*., 2010).

Peptides derived by nonribosomal processes usually range from 2 to 50 amino acids and thus are much smaller than those produced by the ribosomal. In addition, nonribosomal peptides are often cyclic or polycyclic structures rather than linear (Caboche *et al*., 2008).

1.2.3 Terpenoids and terpene synthases

Terpenoids, also known as isoprenoids, comprise the largest and most structurally diverse group of natural products with over 80 000 known compounds (Rudolf *et al*., 2021). The enzymes responsible for the synthesis of terpenes are known as terpene synthases and are behind the structural diversity of terpenes. The terpene synthase family mediates complex reactions in a cascade manner and tends to have very specific substrate preferences and elegant mechanisms for carbocation formation, stabilization, and cyclization (Baunach *et al*, 2015; Rudolf & Chang.,

2020). The building blocks for terpenoids are isoprene units (five carbon units) joined in a head-to-tail fashion, and they are classified based on the number of isoprene units in the backbone chain starting with monoterpenes (two), sesquiterpenes (three), diterpenes (four), triterpenes (six), and tetraterpenes (eight); with each reaction catalyzed by specific and unique terpene synthases (Dewick 2011). However, terpene synthesis is not initiated with isoprene, but two distinct molecules with a similar structural arrangement; the isopentenyl diphosphate and the dimethylallyl diphosphate are the precursors of terpene synthesis, which subsequently will undergo rearrangements, repetition, and cyclization reactions catalyzed by specific terpene synthases to yield the various terpene classes (Ninkuu *et al*., 2021).

Terpenoids are particularly prevalent in plants, fungi, and certain marine invertebrates, but can be found in all domains of life (Rudolf *et al*., 2021). Many of the aromatic compounds characterizing the odor of flowers, leaves, woods, and roots are terpenoids (Jackson, 2008); some can deter feeding by herbivores and can be used as insecticides in agriculture (Ninkuu *et al*., 2021). Further, terpenoids have many pharmacological properties including antifungal, antimicrobial, antiviral, and antiparasitic activities (Abdallah & Quax, 2017).

1.3 Marine Natural Products

Marine environments are some of the most diverse in terms of species; coral reefs are thought to outnumber even tropical rainforests in species diversity (Haefner 2003) yet the number of species inhabiting the oceans remains undetermined (Simmons *et al*. 2005; Das *et al*. 2006). Marine organisms are under intense competitive pressure for space, light, and nutrients especially in locations with higher diversity and lower nutrient availability like in coral reefs, thus they have developed a range of defense mechanisms to increase survival including behavioral, physical, and chemical strategies such as the production of a wide variety of natural products (Harper *et al*.,

2001). In the marine environment, natural products fulfill various tasks for their producers. They can act as chemical defenses against predators (Rohde & Schupp, 2018), as antimicrobial agents against pathogenic microorganisms (Goecke 2010), and they have an important role in inducing larval settlement of benthic invertebrates (Rohde *et al*., 2015). In fact, marine natural products are important for maintaining and controlling many community functions and population dynamics (Petersen *et al*., 2020).

Besides their ecological impact, many marine natural products have been reported to exhibit a wide range of medical relevant bioactivities, thus serving as promising molecules for the development of new drugs and drug leads (Petersen *et al*., 2020). Some biological activities described for marine natural products are anticancer (Nastrucci *et al*. 2012), antibacterial (Hughes and Fenical 2010), antifungal, antiviral (Mayer *et al*. 2013), and anti-inflammatory (Lauritano *et al*., 2019). They also have activities relevant to the treatment of Alzheimer's disease (Nelson *et al*., 2017).

The structural classes of marine natural products are varied and include PKs, NRPs, terpenes, alkaloids, and peptides (Jiménez 2018) with an unprecedented diversity of structures and functional groups; for instance, more halogenated natural products are discovered from marine than terrestrial environments (Gribble 2015).

Marine organisms have evolved under particular environmental pressures; thus, their natural products bind to targets that can be specific to marine organisms. This has resulted in a range of novel receptor-specific interactions against mammalian receptors and resulted in drugs with novel mechanisms (Molinski 1993). In the same way, studies of receptor-target interactions with marine natural products have led to a greater understanding of marine chemical ecology (Harper *et al*., 2001).

Besides the investigation on marine invertebrates and algae, modern marine biotechnology expanded its interests into the exploration of marine bacteria and fungi; both have shown to be potent producers of bioactive substances with prominent activities against pathogenic bacteria, viruses, and tumor cells (Imhoff *et al*., 2011).

1.4 Marine actinomycetes: promising producers of natural products

The actinomycetes or actinobacteria (phylum *Actinomycetota*) is a prolific bacterial group with a cosmopolitan distribution in marine and terrestrial environments (Jagannathan *et al*., 2021). Marine actinomycete habitats range from shallow to deep waters, polar regions, and deep-sea hydrothermal vents to diverse coral reef ecosystems. The water column, sediments, and the surface of macroorganisms such as algae, sponges, and corals are also favorable ecological niches for marine actinomycetes. The living conditions that marine actinomycetes have evolved to include extremely high pressures (around 1100 atmospheres), low oxygen, and temperatures below 0^{o}C at the deep sea floor. They also include highly acidic conditions and high temperatures (over 100^{o}C) at the hydrothermal vents (Lam 2006). Adaptations to these different habitats may undoubtedly reflect in their genetic and metabolic diversity making marine actinomycetes important sources for bioactive natural products with a variety of distinct structural classes not found in their terrestrial counterparts (Manivasagan *et al*., 2014).

Marine actinomycetes have a vast potential to produce metabolites with pharmaceutical applications. In 2017 alone, more than one hundred novel compounds were described from this group (Carroll *et al*., 2019). In the last few decades many biologically active metabolites from marine actinomycetes have been described and exploited. For instance, the anti-inflammatory compound cyclomarin A from *Streptomyces* sp. (Renner *et al*., 1999), the antibacterial compound chalcomycin A from *Streptomyces* sp. M491 (Wu *et al*., 2007), the cytotoxic agent

salinosporamide A from *Salinispora tropica* (Feling *et al*., 2003), and the inmunosupresor thalassospiramide A from *Thallassospira sp.* (Oh *et al*., 2007) were all reported from marine actinomycetes.

There is still much to learn about the environmental adaptations and evolution of marine actinomycetes and the role these adaptations play in their specialized metabolism. However, there is mounting evidence that marine actinomycetes represent an unparalleled source of novel natural products with potential applications in many industries.

1.5 The genus *Salinispora*

The genus *Salinispora* is a marine obligate actinomycete in the family *Micromonosporaceae* that is comprised by nine species: *Salinispora tropica*, *S. arenicola* (Maldonado *et al*., 2005), *S. pacifica* (Ahmed *et al*., 2013), *S. oceanensis, S. vitiensis, S. mooreana, S. fenicalii, S. goodfellowii,* and *S. cortesiana* (Román-Ponce *et al*., 2020). These species share 99% 16S rRNA gene sequence identity, which creates challenges in resolving their phylogenetic relationships using this particular marker. However, multilocus analysis and whole-genome sequencing have successfully resolved the phylogenetic relatedness among *Salinispora* species (Millán-Aguiñaga *et al*., 2017).

Members of *Salinispora* have been isolated from marine organisms (Jensen *et al*., 2005; Vidgen *et al*, 2012) and most commonly from tropical and subtropical marine sediments (Jensen and Mafnas 2006). They have also been reported from temperate waters (Goo *et al*., 2014). Their colonies form extensively branched substrate hyphae but lack aerial hyphae (Mincer *et al*., 2002). They produce the carotenoid sioxanthin (Richter *et al*., 2015), which gives colonies their characteristic orange color ranging from pale to bright orange. Subsequently, during sporulation, colonies darken into shades ranging dark brown to black (Maldonado *et al*., 2005).

The genus *Salinispora* is a prolific producer of bioactive natural products, devoting around 10% of its genome to this process, and has become a model organism for the study of specialized metabolism (Udwary *et al*., 2007; Jensen *et al*., 2015). Most of its specialized metabolite BGCs are in genomic islands, and it has been suggested this is a relevant ecological advantage (Penn *et al*., 2009). The specialized metabolites reported from *Salinispora* spp. are predominantly new with important biological activities and diverse structures (fig 1.1). One of the best-known metabolites from *Salinispora* is salinosporamide A, a potent proteasome inhibitor under development for the treatment of glioblastoma multiforme in adults (Di *et al*., 2016). Other biologically activite *Salinispora* metabolites include antibiotics, such as arenamycin and rifamycin S, anti-inflammatories such as arenamide A, discovered from *S. arenicola* strain, and cytotoxins including lomaiviticin A from *S. pacifica* and saliniquinones from *S. arenicola* (Jensen *et al*., 2015).

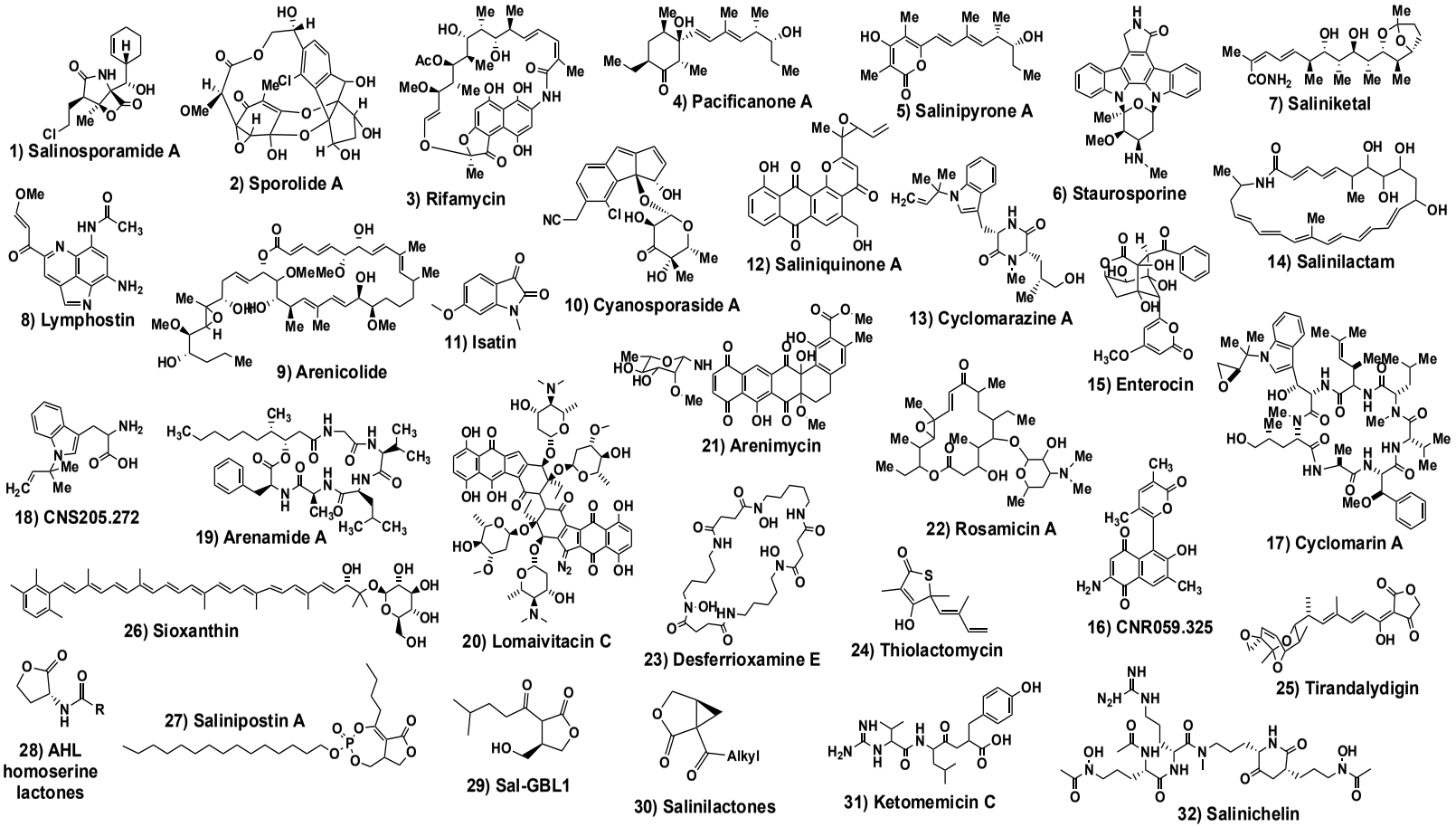

Figure 1.1 Overview of specialized metabolites isolated from *Salinispora*

1.6 Overview of the dissertation

The goal of this dissertation is to assess the correlations between specialized metabolism and speciation in the genus *Salinispora*. In recent years, many studies have recognized phylogenetic patterns in the specialized metabolism of several bacterial groups (Gallagher *et al*., 2013; Adamek *et al*., 2018; Undabarrena *et al*., 2021). The genus *Salinispora* is one of the most prolific producers of specialized metabolites among marine actinomycetes, and one of the most distinct cases of species-specific production of specialized metabolites (Jensen *et al*., 2007). The following chapters focus on different cases of phylogenetic patterns within the specialized metabolism of the most recently described *Salinispora* species by Roman-Ponce and collaborators (2020). From global assessment of biosynthetic gene cluster distributions and metabolite production patterns (presented in chapter 2), I focus on the detection and description of species-specific biosynthetic gene clusters across 118 *Salinispora* genomes. In chapter 3, I assessed the expression of these unique biosynthetic gene clusters in standard laboratory conditions for *Salinispora*. I designed an expression survey of core biosynthetic genes and used *in silico* cutting-edge methodologies to correlate genomics and metabolomics data to associate expressed genes with their putative products.

In chapter 4, I focus on the ketmomemicin gene cluster, its distribution, and metabolite production patterns across *Salinospora* species. To my knowledge, this is the first report of *Salinispora* strains producing ketomemicns under standard growth conditions. I report the discovery of two novel analogs of ketomemicin, and delve into the associated BGC distribution, gene arrangement, gene sequence similarity, and metabolite production within and outside the genus *Salinispora*.

Chapter 5 is a review chapter. It consolidates the phylogenetic patterns found in several bacterial groups and proposes the use of specialized metabolism as a taxonomic tool for bacteria. I go over the current methodologies for bacterial species classification including traditional chemotaxonomic methods and discuss the advantages and disadvantages of implementing a chemotaxonomy for bacteria based on specialized metabolism.

Chapter 6 is a finer look into the differences driving the specialized metabolism of *Salinispora*. I focus in the biosynthetic genes of the terpenoid sioxantin and investigate the origins of a spontaneous albino variation across six different strains and their pigmented counterparts. Using comparative genomics, I was able to link the disruption of sioxantin production to various mutations to a noncoding region of the BGC. I then investigated the ecological roles of sioxantin and hypothesized the repercussions and causes of the lack of coloration for *Salinispora* in the natural environment.

Finally, in the last section of my dissertation I summarize and discuss the significant findings and general conclusions brought from my research and discuss some future research directions.

1.7 References

Abdallah I.I., Quax W.J. A glimpse into the biosynthesis of terpenoids. *KnE Life Sci.* 2017;3:81.

Adamek, M., Alanjary, M., Sales-Ortells, H., Goodfellow, M., Bull, A. T., Winkler, A., Wibberg, D., Kalinowski, J., & Ziemert, N. (2018). Comparative genomics reveals phylogenetic distribution patterns of secondary metabolites in Amycolatopsis species. *BMC Genomics*, *19*(1), 1–15.

Ahmed, L., Jensen, P.R., Freel, K.C., Brown, R., Jones, A.L., Kim, B., & Goodfellow, M. (2013). Salinispora pacifica sp. nov., an actinomycete from marine sediments. *Antonie van Leeuwenhoek, 103*, 1069-1078.

Ahmed, L., Jensen, P.R., Freel, K.C., Brown, R., Jones, A.L., Kim, B., & Goodfellow, M. (2013). Salinispora pacifica sp. nov., an actinomycete from marine sediments. *Antonie van Leeuwenhoek, 103*, 1069-1078.

Asolkar RN, Freel KC, Jensen PR *et al* (2009) Arenamides A–C, cyto- toxic NFκB inhibitors from the marine actinomycete *Salinispora Arenicola*. J Nat Prod 72:396–402

Baunach, M., Franke, J., & Hertweck, C. (2015). Terpenoid biosynthesis off the beaten track: unconventional cyclases and their impact on biomimetic synthesis. *Angewandte Chemie, 54 9*, 2604-26.

Blythe, E.K., Tabanca, N., Demirci, B., & Kendra, P.E. (2020). Chemical Composition of Essential Oil From Tetradenia riparia and Its Attractant Activity for Mediterranean Fruit Fly, Ceratitis capitata. *Natural Product Communications, 15*.

Brakhage, A.A. (2013). Regulation of fungal secondary metabolism. *Nature Reviews Microbiology, 11*, 21-32.

Caboche, S., Leclère, V., Pupin, M., Kucherov, G., & Jacques, P. (2010). Diversity of Monomers in Nonribosomal Peptides: towards the Prediction of Origin and Biological Activity. *Journal of Bacteriology, 192*, 5143 - 5150.

Caboche, S., Pupin, M., Leclère, V., Fontaine, A., Jacques, P., & Kucherov, G. (2008). NORINE: a database of nonribosomal peptides. *Nucleic Acids Research, 36*, D326 - D331.

Carroll, A.R., Copp, B.R., Davis, R.A., Keyzers, R.A., & Prinsep, M.R. (2019). Marine natural products. *Natural product reports, 36 1*, 122-173 .

Chen R, Wong HL, Kindler GS, MacLeod FI, Benaud N, Ferrari BC and Burns BP (2020) Discovery of an Abundance of Biosynthetic Gene Clusters in Shark Bay Microbial Mats. *Front. Microbiol.* 11:1950.

Cohen, D.J., Loertscher, R., Rubin, M.F., Tilney, N.L., Carpenter, C.B., & Strom, T.B. (1984). Cyclosporine: a new immunosuppressive agent for organ transplantation. *Annals of internal medicine, 101 5*, 667-82 .

Das S, Lyla PS, Ajmal Khan S (2006) Marine microbial diversity and ecology: importance and future perspectives. Curr Sci 90:1325–1335

Demain, A.L., & Sánchez, S. (2009). Microbial drug discovery: 80 years of progress. *The Journal of Antibiotics, 62*, 5 - 16.

Dewick, P. (2011). Medicinal Natural Products: A Biosynthetic Approach, 3Rd Edition.

Di, K., Lloyd, G. K., Abraham, V., MacLaren, A., Burrows, F. J., Desjardins, A., Trikha, M., & Bota, D. A. (2016). Marizomib activity as a single agent in malignant gliomas: ability to cross the blood-brain barrier. *Neuro-oncology*, *18*(6), 840–848.

Diminic, J., Starčević, A., Lisfi, M., Baranasic, D., Gacesa, R., Hranueli, D., Long, P.F., Cullum, J., & Zucko, J. (2014). Evolutionary concepts in natural products discovery: what actinomycetes have taught us. *Journal of Industrial Microbiology & Biotechnology, 41*, 211-217.

Feling, R.H., Buchanan, G.O., Mincer, T.J., Kauffman, C.A., Jensen, P.R., & Fenical, W. (2003). Salinosporamide A: a highly cytotoxic proteasome inhibitor from a novel microbial source, a marine bacterium of the new genus salinospora. *Angewandte Chemie, 42 3*, 355-7.

Gallagher, K. A., Rauscher, K., Ioca, L. P., & Jensen, P. R. (2013). Phylogenetic and chemical diversity of a hybrid-isoprenoid- producing streptomycete lineage. *Applied and Environmental Microbiology*, *79*(22), 6894–6902.

Gallimore, A.R. (2009). The biosynthesis of polyketide-derived polycyclic ethers. *Natural product reports, 26 2*, 266-80.

Garénaux, A., Caza, M., & Dozois, C.M. (2011). The Ins and Outs of siderophore mediated iron uptake by extra-intestinal pathogenic Escherichia coli. *Veterinary microbiology, 153 1-2*, 89-98.

Goecke F, Labes A, Wiese J *et al* (2010) Chemical interactions between marine macroalgae and bacteria. Mar Ecol Prog Ser 409:267–300

Goo, K.-S., Tsuda, M., & Ulanova, D. (2014). Salinispora arenicola from temperate marine sediments: new intra-species variations and atypical distribution of secondary metabolic genes. *Antonie van Leeuwenhoek*, *105*(1), 207–219.

Gribble, G.W. (2015). Biological Activity of Recently Discovered Halogenated Marine Natural Products. *Marine Drugs, 13*, 4044 - 4136.

Haefner B (2003) Drugs from the deep: marine natural products as drug candidates. Drug Discov Today 8:536–544

Harper, M.K., Bugni, T.S., Copp, B.R., James, R.D., Lindsay, B.S., Richardson, A.D., Schnabel P.C., Tasdemir, D., VanWagoner, R.M., Verbitski, S.M., Ireland, C.M., (2001). Introduction to the chemical ecology of marine natural products. In: McClintock Jb, Baker BJ editors. Marine chemical ecology. Boca Raton, Fl: CRC Press. P3-70.

Hertweck, C., Luzhetskyy, A., Rebets, Y.V., & Bechthold, A. (2007). Type II polyketide synthases: gaining a deeper insight into enzymatic teamwork. *Natural product reports, 24 1*, 162-90 .

Hughes CC, Fenical W (2010) Antibacterials from the sea. Chem Eur J 16:12512–12525

Imhoff JF, Labes A, Wiese J (2011) Bio-mining the microbial treasures of the ocean: new natural products. Biotechnol Adv 29:468–482

Jackson R. S. (2008) Chemical Constituents of Grapes and Wine. In: Jackson R.S., editor. Food Science and Technology, Wine Science (Third Edition) Academic Press,P 270-331

Jagannathan, S.V., Manemann, E.M., Rowe, S.E., Callender, M.C., & Soto, W. (2021). Marine Actinomycetes, New Sources of Biotechnological Products. *Marine Drugs, 19*.

Jensen, P. R., Gontang, E., Mafnas, C., Mincer, T. J., & Fenical, W. (2005). Culturable marine actinomycete diversity from tropical Pacific Ocean sediments. *Environmental Microbiology*, *7*(7), 1039–1048. https://doi.org/10.1111/j.1462-2920.2005.00785.x

Jensen, P. R., Moore, B. S., & Fenical, W. (2015). The marine actinomycete genus Salinispora: A model organism for secondary metabolite discovery. In *Natural Product Reports* (Vol. 32, Issue 5, pp. 738–751). Royal Society of Chemistry. https://doi.org/10.1039/c4np00167b

Jensen, P. R., Williams, P. G., Oh, D.C., Zeigler, L., & Fenical, W. (2007). Species-specific secondary metabolite production in marine actinomycetes of the genus Salinispora. *Applied and Environmental Microbiology*, *73*(4), 1146–1152.

Jensen, P.R., & Mafnas, C. (2006). Biogeography of the marine actinomycete Salinispora. *Environmental microbiology, 8 11*, 1881-8 .

Jiang, J., Gao, L., Bie, X., Lu, Z., Liu, H., Zhang, C., Lu, F., & Zhao, H. (2016). Identification of novel surfactin derivatives from NRPS modification of Bacillus subtilis and its antifungal activity against Fusarium moniliforme. *BMC Microbiology, 16*.

Jiménez, C. (2018). Marine Natural Products in Medicinal Chemistry. *ACS medicinal chemistry letters, 9 10*, 959-961.

Lam, K.S. (2006). Discovery of novel metabolites from marine actinomycetes. *Current opinion in microbiology, 9 3*, 245-51.

Lane, A.L., & Moore, B.S. (2011). A sea of biosynthesis: marine natural products meet the molecular age. *Natural product reports, 28 2*, 411-28.

Lauritano, C., Ferrante, M.I., & Rogato, A. (2019). Marine Natural Products from Microalgae: An -Omics Overview. *Marine Drugs, 17*.

Lyons, D.A., Scheibling, R.E., & Alstyne, K.L. (2010). Spatial and temporal variation in DMSP content in the invasive seaweed Codium fragile ssp. fragile: effects of temperature, light and grazing. *Marine Ecology Progress Series, 417*, 51-61.

Ma, L., Zhang, Q., Yang, C., Zhu, Y., Zhang, L., Wang, L., Liu, Z., Zhang, G., & Zhang, C. (2019). Assembly Line and Post-PKS Modifications in the Biosynthesis of Marine Polyketide Natural Products. In *Reference Module in Chemistry, Molecular Sciences and Chemical Engineering* (Issue December).

Maldonado, L. a, Fenical, W., Jensen, P. R., Kauffman, C. a, Mincer, T. J., Ward, A. C., Bull, A. T., & Goodfellow, M. (2005). Salinispora arenicola gen. nov., sp. nov. and Salinispora

tropica sp. nov., obligate marine actinomycetes belonging to the family Micromonosporaceae. *International Journal of Systematic and Evolutionary Microbiology*, *55*(Pt 5), 1759–1766.

Manivasagan, P., Kang, K., Sivakumar, K., Li-Chan, E.C., Oh, H., & Kim, S. (2014). Marine actinobacteria: an important source of bioactive natural products. *Environmental toxicology and pharmacology, 38 1*, 172-88 .

Mayer, A.M., Rodríguez, A.D., Taglialatela Scafati, O., & Fusetani, N. (2013). Marine Pharmacology in 2009–2011: Marine Compounds with Antibacterial, Antidiabetic, Antifungal, Anti-Inflammatory, Antiprotozoal, Antituberculosis, and Antiviral Activities; Affecting the Immune and Nervous Systems, and other Miscellaneous Mechanisms of Action. *Marine Drugs, 11*, 2510 - 2573.

Millán-aguiñaga, N., Chavarria, K. L., Ugalde, J. A., Letzel, A., Rouse, W., & Jensen, P. R. (2017). *Phylogenomic Insight into Salinispora (Bacteria , Actinobacteria) Species Designations*. *February*, 1–11.

Mincer, T.J., Jensen, P.R., Kauffman, C.A., & Fenical, W. (2002). Widespread and Persistent Populations of a Major New Marine Actinomycete Taxon in Ocean Sediments. *Applied and Environmental Microbiology, 68*, 5005 - 5011.

Mogi, T., & Kita, K. (2009). Gramicidin S and polymyxins: the revival of cationic cyclic peptide antibiotics. *Cellular and Molecular Life Sciences, 66*, 3821-3826.

Molinski, T.F. (1993). Developments in marine natural products. Receptor-specific bioactive compounds. *Journal of natural products, 56 1*, 1-8 .

Nastrucci C, Cesario A, Russo P (2012) Anticancer drug discovery from the marine environment. Recent Pat Anticancer Drug Discov 7:218–232

Nelson, T.J., Sun, M., Lim, C.S., Sen, A., Khan, T.K., Chirila, F.V., & Alkon, D.L. (2017). Bryostatin Effects on Cognitive Function and PKCε in Alzheimer's Disease Phase IIa and Expanded Access Trials. *Journal of Alzheimer's Disease, 58*, 521 - 535.

Ninkuu, V., Zhang, L., Yan, J., Fu, Z., Yang, T., & Zeng, H. (2021). Biochemistry of Terpenes and Recent Advances in Plant Protection. *International Journal of Molecular Sciences, 22*.

Nützmann, H., & Osbourn, A. (2014). Gene clustering in plant specialized metabolism. *Current opinion in biotechnology, 26*, 91-9 .

O'Brien J, Wright GD (2011) An ecological perspective of microbial secondary metabolism. Curr Opin Biotechnol 22:552–558

Oh, D., Strangman, W.K., Kauffman, C.A., Jensen, P.R., & Fenical, W. (2007). Thalassospiramides A and B, immunosuppressive peptides from the marine bacterium Thalassospira sp. *Organic letters, 9 8*, 1525-8.

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., & Jensen, P. R. (2009). Genomic islands link secondary metabolism to functional adaptation in marine Actinobacteria. *The ISME Journal*, *3*(10), 1193–1203. https://doi.org/10.1038/ismej.2009.58

Petersen, L., Kellermann, M.Y., & Schupp, P.J. (2020). Secondary Metabolites of Marine Microbes: From Natural Products Chemistry to Chemical Ecology. *YOUMARES 9 - The Oceans: Our Research, Our Future*.

Rajesh, T., Tiwari, M.K., Thiagarajan, S., Nair, P., & Jeya, M. (2019). Type III Polyketide Synthases: Current State and Perspectives. *Microorganisms for Sustainability*.

Renner, M.K., Shen, Y., Cheng, X., Jensen, P.R., Frankmoelle, W., Kauffman, C.A., Fenical, W., Lobkovsky, A.E., & Clardy, J. (1999). Cyclomarins A−C, New Antiinflammatory Cyclic Peptides Produced by a Marine Bacterium (Streptomyces sp.). *Journal of the American Chemical Society, 121*, 11273-11276.

Richter, T. K. S., Hughes, C. C., & Moore, B. S. (2015). Sioxanthin, a novel glycosylated carotenoid, reveals an unusual subclustered biosynthetic pathway. *Environmental Microbiology*, *17*(6), 2158–2171. https://doi.org/10.1111/1462-2920.12669

Rohde S, Nietzer S, Schupp PJ (2015) Prevalence and mechanisms of dynamic chemical defenses in tropical sponges. PLoS One 10:e0132236.

Rohde, S., & Schupp, P.J. (2018). Spatial and Temporal Variability in Sponge Chemical Defense. *Chemical Ecology*.

Román-Ponce, B., Millán-Aguiñaga, N., Guillen-Matus, D., Chase, A. B., Ginigini, J., Soapi, K., Feussner, K. D., Jensen, P. R., & 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., *Salinispora oceanensis* sp. nov. and *Salinispora vitiensis* sp. nov., and emended description of the genus *Salinispora*. *International journal of systematic and evolutionary microbiology*, *70*(8), 4668–4682.

Rudolf, J.D., & Chang, C. (2020). Terpene synthases in disguise: enzymology, structure, and opportunities of non-canonical terpene synthases. *Natural product reports*.

Rudolf, J.D., Alsup, T.A., Xu, B., & Li, Z. (2021). Bacterial terpenome. *Natural product reports*.,38, 905-980

Shen, B., Du, L., Sánchez, C., Edwards, D.J., Chen, M., & Murrell, J.M. (2001). The biosynthetic gene cluster for the anticancer drug bleomycin from Streptomyces verticillus ATCC15003 as a model for hybrid peptide–polyketide natural product biosynthesis. *Journal of Industrial Microbiology and Biotechnology, 27*, 378-385.

Simmons TL, Andrianasolo E, McPhail K *et al* (2005) Marine natural products as anticancer drugs. Mol Cancer Ther 4:333–342

Stachowicz, J.J. (2001). Chemical Ecology of Mobile Benthic Invertebrates: Predators and Prey, Allies and Competitors. In: McClintock Jb, Baker BJ editors. Marine chemical ecology. Boca Raton, Fl: CRC Press. P157-194

Strieker, M., Tanović, A., & Marahiel, M.A. (2010). Nonribosomal peptide synthetases: structures and dynamics. *Current opinion in structural biology, 20 2*, 234-40 .

Torres, J.P., Lin, Z., Winter, J.M., Krug, P.J., & Schmidt, E.W. (2020). Animal biosynthesis of complex polyketides in a photosynthetic partnership. *Nature Communications, 11*.

Udwary, D. W., Zeigler, L., Asolkar, R. N., Singan, V., Lapidus, A., Fenical, W., Jensen, P. R., & Moore, B. S. (2007). Genome sequencing reveals complex secondary metabolome in the marine actinomycete Salinispora tropica. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(25), 10376–10381.

Undabarrena, A., Valencia, R., Cumsille, A., Zamora-Leiva, L., Castro-Nallar, E., Barona-Gómez, F., & 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. *Microbial Genomics, 7*.

Vidgen ME, Hooper JN, Fuerst JA (2012) Diversity and distribution of the bioactive actinobacterial genus Salinispora from sponges along the Great Barrier Reef. Antonie Van Leeuwenhoek 101:603–618.

Wu, S.J., Fotso, S., Li, F.C., Qin, S., & Laatsch, H. (2007). Amorphane sesquiterpenes from a marine Streptomyces sp. *Journal of natural products, 70 2*, 304-6.

Chapter 2

Comparative genomics and metabolomics reveal species-specific patterns in specialized metabolism of the marine actinomycete *Salinispora*

2.1 Abstract

The genus *Salinispora* is a prolific producer of natural products with many biological activities and has become a model organism for the study of specialized metabolism in closely related species. The species *S. arenicola* and *S. tropica* are known to produce secondary metabolites in species-specific patterns, however similar patterns have not been observed for the remainder of the *Salinispora* species. Here, I performed comparative genomics and metabolomics to assess the distribution of specialized metabolites and biosynthetic gene clusters in seven species of *Salinispora.* The comparative genomics revealed a total of 15 species-specific biosynthetic gene clusters, with 30% of the variation in the metabolomes explained by species. Although the ecological and evolutionary significance of these findings are yet to be fully understood, the results from this investigation help our understanding the correlation of specialized metabolism with species phylogeny in the marine actinomycete *Salinispora.*

2.2 Introduction

Currently, the easy access to genome sequencing has motivated the development of several techniques to explore the newly generated data. Genome sequencing has played a key role revitalizing natural products research (Ceniceros *et al*., 2017). The realization that bacteria harbor

many uncharacterized biosynthetic gene clusters (BGCs) has revolutionized natural product research pipelines. Similarly, increased access to genome sequencing data increased the opportunities to ask broader questions about specialized metabolism (Jensen 2016).

Genome mining attempts to identify genes associated with specialized metabolism from genome sequences, particularly those that might result in unprecedented products (Diminic *et al*., 2014). Therefore, it has become a standard tool in natural products research and is currently an important technology complementing the classical natural product discovery pipeline (Chevrette *et al*., 2022). In addition, comparative genomics allows a comprehensive exploration of BGCs based on high throughput mining, providing much-needed evidence to target certain BGCs and augmenting the power of genome-guided bioprospecting for natural products (Undabarrena *et al*., 2021). There exist many bioinformatic tools such as NP.searcher (Li *et al*., 2009) and PRISM3.0 (Skinnider *et al*., 2017) including some that can perform in a high throughput manner like antiSMASH (Medema *et al*., 2011) and NaPDoS (Ziemert *et al*., 2012).

Metabolomics is the comprehensive study of small molecules within a biological system and provides a direct measure of detectable specialized metabolites within an organism of interest (Covington *et al*., 2017). Currently, mass spectrometry is one of the most used analytical platforms for metabolomics because of its sensitivity, high throughput, and ability to measure molecules in complex biological samples (Liu & Locasale, 2017). However, metabolomic data collection is just the first step. Ultimately, advances in the exploration of a biological system come from the integration of instrumentation, chemistry, statistics, and computer science (Patti *et al*., 2012). Most comparative metabolomics studies make use of statistical techniques to look for correlations between spectral signals and the biological states of the samples. Some of the statistical tools commonly used include principal component analysis (PCA), partial least squares (PLS),

orthogonal projection onto latent structures (O-PLS), and discriminant analysis (PCA-DA, PLS-DA, O-PLS-DA). (Robinette *et al*., 2012). While metabolomic analyses have historically been performed independently of any knowledge of the associated BGCs, it has become increasingly common to integrate these datasets to generate bioinformatic links between natural products and their associated BGCs (Machado *et al*., 2017).

A model organism to address questions related to specialized metabolism and BGC evolution is the marine actinomycete *Salinispora*. This genus is composed of nine closely related species, in which comparative genomics have allowed us to assess BGCs distribution and diversity and make inferences about the evolutionary processes driving these observations (Letzel *et al*., 2017). The specialized metabolism of *Salinispora* spp. is incredibly diverse with a wide variety of metabolites of different biosynthetic origins including novel structures with important biological activities (Jensen *et al*., 2015). Genome mining has revealed that around 80% of the BGCs *Salinispora* harbors remain uncharacterized. Surprisingly, few of these BGCs are shared at the genus level and up to 54% are observed in only one or two strains (Letzel *et al*., 2017). Further, species-specific compound production has been reported for two of the nine species, *S. tropica* and *S. arenicola*, and has long been hypothesized that all species might follow a similar pattern (Jensen *et al*., 2007; Ziemert *et al.,* 2014).

In recent years, the recognition of lineage-specific patterns in bacterial specialized metabolism has increased, primarily thanks to comparative genomics and metabolomic analysis assessing large groups of closely related strains (Adamek *et al*., 2020). Assessing the correlations between BGC distributions, compound production, and species phylogeny in *Salinispora* gives insight into the relationships between speciation and specialized metabolism and opportunities to further explore if the observations apply to other closely related bacterial taxa. If specialized

metabolites provide important ecological functions in bacteria, it is possible that vertical inheritance plays a significant role in specialized metabolism gene distribution. Under these premises, this chapter tries to assess the correlation between specialized metabolism and species phylogeny in *Salinispora* spp. using comparative genomics and metabolomics. The BGCs of 118 *Salinispora* genome strains were mined and coupled with a metabolomic study of 130 crude extracts to explore the BGC distribution and metabolite production and try to test the hypothesis of species-specificity in *Salinispora* specialized metabolism.

2.3 Methods

2.3.1 Biosynthetic gene cluster distribution

To assess BGC distributions across 118 *Salinispora* genomes, I took advantage of the datasets from Letzel and collaborators (2017), where they assessed *Salinispora* BGC diversity and distribution at the species level to gain insights into BGC evolution. One dataset consisted in the AntiSMASH 2.0 (Blin *et al*., 2013) output for all 118 *Salinispora* genomes. I primarily used this data for BGC identification and comparison across every *Salinispora* strain. For RiPPs, terpenes and "other" biosynthetic types (classified based on AntiSMASH 2.0 output), I compared the product prediction, gene arrangement, and architecture of each gene cluster along with functional gene annotation using the conserved domain database from NCBI (Marchler-Bauer *et al*., 2017). For PKS BGCs, I also performed a detailed analysis of each cluster using product prediction and gene architecture from AntiSMASH 2.0 outputs. To corroboration the PKS type predicted by antiSMASH, I used the NaPDoS pipeline and database (Ziemert *et al*., 2012). In the case of NRPS BGCs, I followed the same analysis pipeline from the antiSMASH 2.0 data output as I did with the other BGCs types and further parsed each NRPS gene using the NRPS/PKS prediction website

(Bachmann & Ravel, 2009). Once I had compared each BGC across all nine Salinispora species, I was able to identify unique BGCs for each of the novel six species and *S. pacifica*. Having the candidate species-specific BGCs, I reanalyzed a subset of 14 genome sequences with antiSMASH 5.0 (Blin *et al.,* 2019) and manually sorted them according to Blast similarity using Multigene blast (Medema *et al*., 2013) to compare cluster architecture and conservation across all *Salinispora* genome strains while paying attention to gene and protein functional predictions. The designation of a species-specific BGC was based on the presence of the same cluster in all strains of one species and its absence in the other 8 species. An exception was made for PKS25, which is shared across *S. oceanensis* and two strains of *S. mooreana.*

2.3.2 Cultivation and organic extraction

To assess metabolite distribution in *Salinispora* spp*.,* I selected at least 3 different strains for each species or all available strains in those species with fewer than 3 genomes (Table 2.1). Each strain was grown in 100 mL modified A1 media in 250 mL Erlenmeyer flasks in triplicate for 7; or 12 days, to capture the exponential and stationary growth phase. Cultures were continuous shaking at 200 rpm followed by liquid-liquid extraction with ethyl acetate at a 1:1 v/v proportion. The organic phase was collected and subsequently dried in vacuo via rotary evaporation. Three media controls were extracted in a similar manner. The modified A1 media formulation was as follows: starch 5 g, yeast 4 g, peptone 2 g, and marine salts (instant ocean ®) 22 g per 1 liter of distilled water.

Table 2.1 Cultured Salinispora strains for metabolomic analysis.

Species	Strain
S. mooreana	CNS-237
	CNY-646
	CNT-150
S. cortesiana	CNY-202
S. fenicalii	CNT-569
	CNR-942
S. vitensis	CNS-801
	CNS-055
	CNT-148
S. goodfellowii	CNY-666
S. pacifica	CNR-114
	CNT-855
	CNY-331
	CNY-498
	CNT-851
S. oceanensis	CNT-029
	CNY-673
	CNY-703
	CNT-403
S. arenicola	CNH-643
S. tropica	CNB-440

2.3.3 Mass spectrometry data acquisition and analysis

A total of 120 crude extracts were diluted in HPLC grade methanol at a concentration of 1 mg/mL. The mass spectrometry acquisition was performed in a liquid chromatographer (HPLC Agilent brand) in tandem with a quadrupole-time of flight mass spectrometer in positive mode. The ionization technique was via electrospray with nitrogen as the ionization gas; the collision energy was set at 30 v. The total elution time was 18 min with the first four minutes of each run diverted to waste for a total of 14 minutes of data acquisition. The mass spectrometry data was then transformed into mzXML files to be further analyzed and filtered using the free software Mzmine 2 (Pluskal *et al*., 2010). I set the noise levels for peak detection at $1x10^5$ and $1x10^4$ of intensity for MS1 and MS2 level, respectively and used the centroid algorithm. For chromatogram

building and deconvoluting, the base-line algorithm was selected, followed by isotope detection and selection of the most abundant isotope as the parent ion. Subsequently, an intensity peak list was built using the joint alignment option to combine similar peaks across all samples. This resulting peak table was used for multivariate statistical analysis.

2.3.4 Statistical analysis

To assess the metabolome variation within and across species, I used two different pipelines for multivariate analysis. First, to explore the metabolite distributions across species and variation within species, I ran a PCA analysis using an Euclidean distance matrix using the R-based web tool Metaboanalyst, and subsequently generated a dendrogram using the same Euclidean distance (Xia & Wishart, 2016; Pang *et al*., 2021). The intensity peak table generated with mzMine was normalized to sample median, scaled using Pareto scaling, and transformed using the square root transformation, prior to statistical analysis. Second, to assess the metabolomic variance among species, I performed a PCoA using the Shannon diversity index for alpha diversity analysis and a Canberra distance matrix for beta diversity; and a PERMANOVA test with 999 permutations, using the metabolomics plugin of the python-based platform Qiime2 (Bolyen *et al*., 2016).

2.3.5 Molecular networking and GNPS annotation

The molecular network was built using the classic methodology (Wang *et al*., 2016) and the following workflow parameters: A molecular network was created using the online workflow (https://ccms-ucsd.github.io/GNPSDocumentation/) on the GNPS website (http://gnps.ucsd.edu). The data was filtered by removing all MS/MS fragment ions within +/- 17 Da of the precursor *m/z*. MS/MS spectra were window filtered by choosing the top 6 fragment ions in the +/- 50Da window throughout the spectrum. The precursor ion mass tolerance was set to 0.1 Da and a MS/MS

fragment ion tolerance of 0.01 Da. A network was then created where edges were filtered to have a cosine score above 0.7 and more than 5 matched peaks. Further, edges between two nodes were only kept in the network if each of the nodes appeared in each other's respective top 10 most similar nodes. Finally, the maximum size of a molecular family was set to 100, and the lowest scoring edges were removed from molecular families until the molecular family size was below this threshold. The spectra in the network were then searched against the GNPS spectral libraries. The library spectra were filtered in the same manner as the input data. All matches kept between network spectra and library spectra were required to have a cosine score above 0.7 and at least 4 matched peaks.

2.4 Results and discussion

2.4.1 Comparative metabolomics

A total of 120 crude extracts were analyzed by tandem mass spectrometry and the resulting data were used to construct a molecular network with the GNPS web interface (Wang *et al.,* 2016; Chong *et al*., 2019). Simultaneously, a feature table from the mass spectrometry data was generated that included intensity, retention time, and mass-to-charge ratio (*m/z*) for statistical analysis.

The resulting GNPS network contained a total of 3081 nodes with 30% having no connections with other nodes (i.e., singeltons). There were library matches to rifamycins, arenamycin, desferrioxamine E, and saliniketal (fig 2.1A). At least one unique molecular family included parent ions from six out of the nine *Salinispora* species: *S. arenicola, S. cortesiana*, *S. fenicalii*, *S. oceanensis*, *S. vitiensis*, *S. mooreana*, and *S. goodfellowii* (fig. 2.1B). For a molecular family to be annotated as species-specific, the mass spectral data had to comply to the following

criteria: 1) Parent ions (nodes) within clusters (molecular families) should appear in at least 2 of the 3 replicates from at least one set of extractions (either 7 or 12 days). 2) Nodes in the cluster should be produced by most strains per species. 3) The parent ion intensity should be at least 3 times higher than the set noise level in the chromatogram or have a clean extracted ion chromatogram. These criteria allowed me to narrow down the species-specific nodes observed in the GNPS network. None of the nodes belonging to the species-specific molecular families could be annotated using the GNPS library search. In proportion to the whole network, the species-specific molecular families were a rare occurrence (16 out of 1336), most nodes were shared among multiple species. Additionally, the Principal component analysis (PCA); doesn't show a clear separation among species (fig. 2.2). In some cases the variation within species was greater than among species, particularly for *S. mooreana, S. fenicalii, S. oceanensis,* and *S. cortesiana*. This variability can be more easily observed in a dendrogram built from the Euclidean distance matrix used for the PCA (fig. 2.3) where the samples and replicates from some species are not clustering together. This might be related to a few, but very intense features shared across samples many samples, for example, features related to media components; since Euclidean similarity matrix may increase the influence of large components (Qi & Voit, 2017). Further, the variability explained by the PCA is 4.4% for component 1 and 3.5% for component 2, with about 15% of the variance explained across the five components and up to 20% when using eight components meaning there is no real separation among species in the metabolomics data.

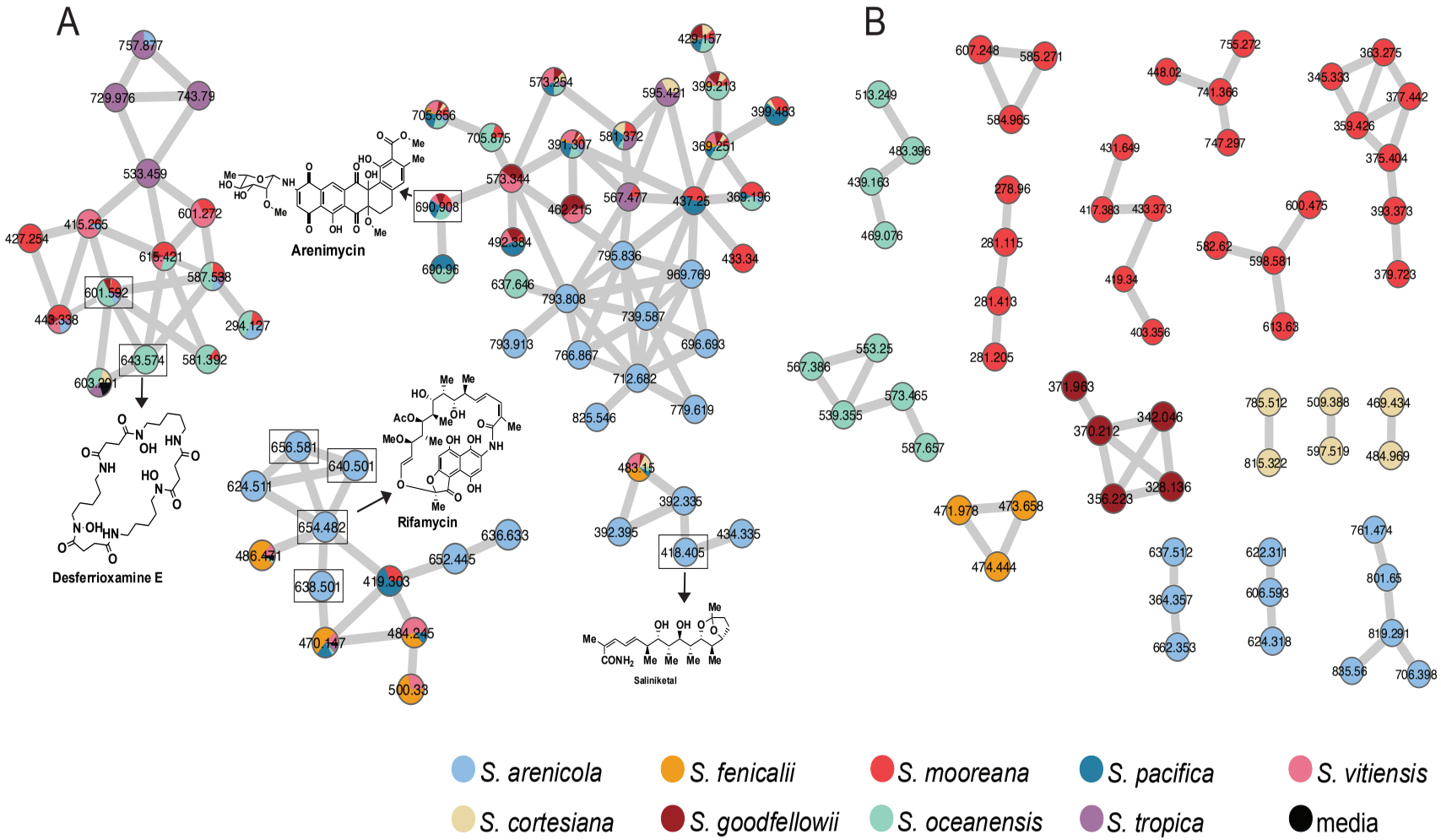

Figure 2.1 Filtered molecular networks showing representative molecular families. Nodes with a cosine score >0.7 are connected. Node label indicates parent ion *m/z*, and color represents species. **A.** Library matches with known *Salinispora* compounds. The matching nodes are within rectangles and the structures are shown. **B.** Species-specific molecular families.

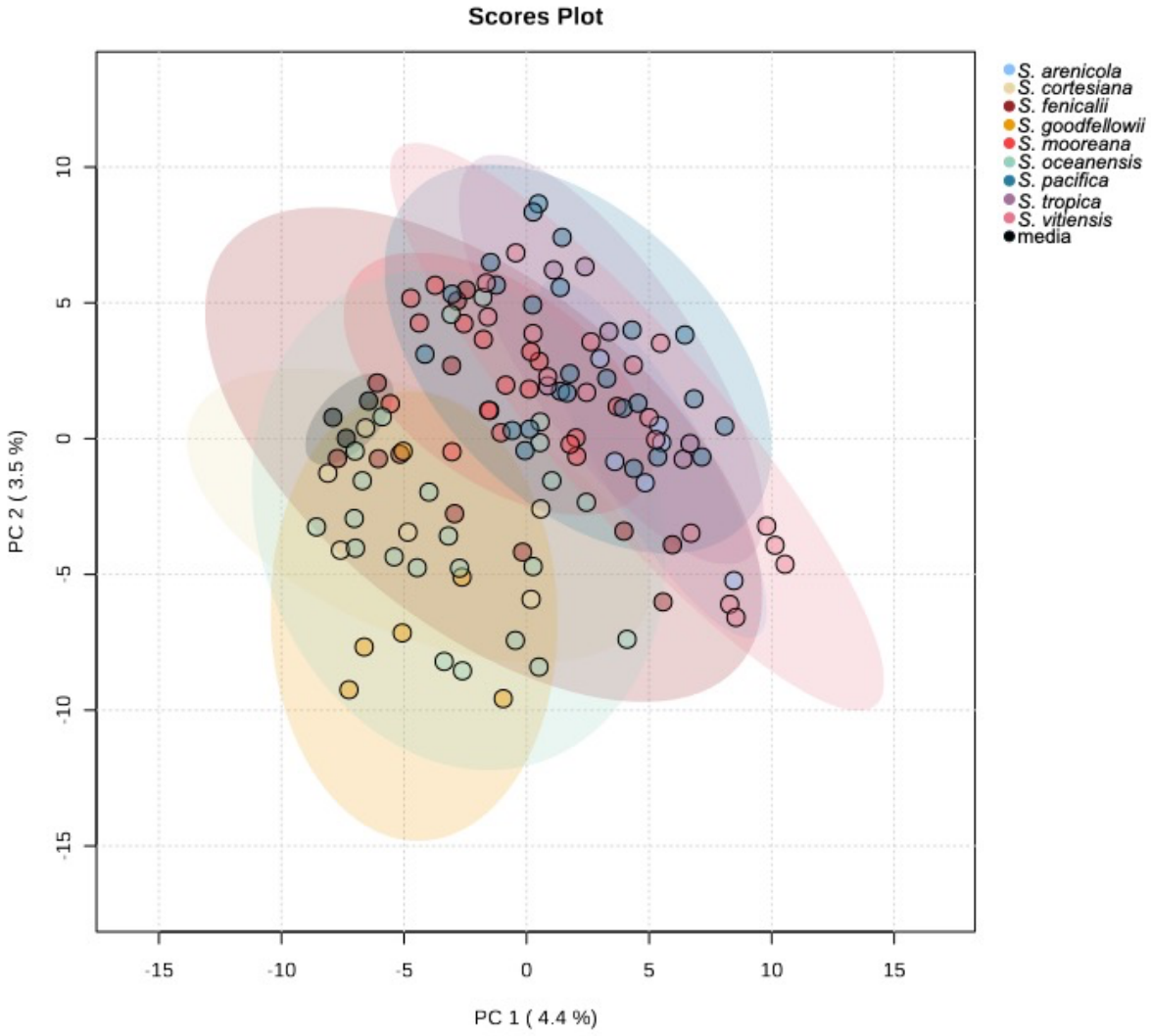

Figure 2.2 Principal component analysis (PCA) showing the metabolome variation among *Salinispora* spp. Each node represents one replicate or medium control (n=120).

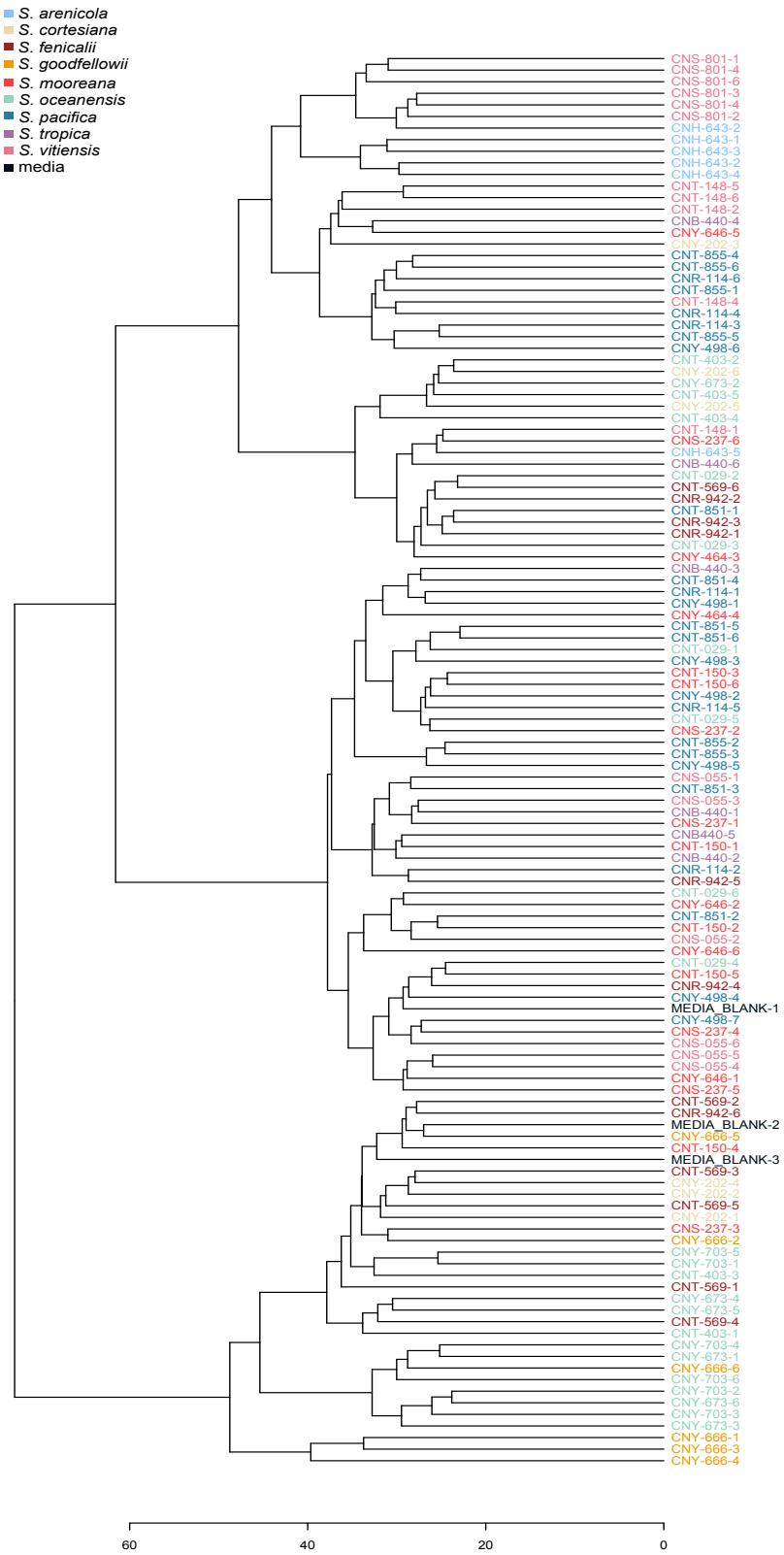

Figure 2.3 Dendrogram using Euclidean distance matrix from *Salinispora* spp. metabolomics (n=120), samples are color coded to show species.

To test for metabolomic variation among species, I performed a PERMANOVA based on the spectral features and their magnitude. This resulted in all species being different from each other except for *S. fenicalii* and *S. cotersiana* (p=0.1), meaning a greater variance among species than within species. However, there was no significant difference in variance between species and media (p=0.98) (table 2.2). Therefore, I decided to remove the features shared between all species with the media and re-analyze the data by PCA (fig. 2.4). After removing the peaks associated with the media control, the variation explained by PCA went up to 29.9%, (PERMANOVA p=0.001) with component 1 explaining around 8.2% and component 2 explaining 3.5%. Thus, although the variation among species is greater than within species, the metabolomes of *Salinispora* species in this data set are very similar to each other even after removing features from the culture media. In a different metabolomics experiment using a similar set of *Salinispora* species, the variance among *Salinispora* species of 44.9% (Chase *et al* 2021). While species explained more of the variance in this study, more *S. arenicola* and *S. tropica* strains were used, which might account for the different outcomes. However, in both my data set and in Chase *et al.*, (2021) species designation explained less than 50% of the variation in the *Salinispora* metabolome, indicating that vertical inheritance might play a significant role in *Salinispora* specialized metabolism, although not as predominantly as I had previously hypothesized. I then questioned whether location influenced the variation in *Salinispora* metabolomes, however, there was no significant difference among samples based on location (PERMANOVA p=0.089). Although, when selecting the representative strains, I prioritized phylogenetic diversity, which resulted in 30% of the strains being isolated from Fiji. Therefore, although there is no significant effect of location in this particular metabolomic study, I cannot confidently conclude that location is not an important factor in the specialized metabolism of *Salinispora*.

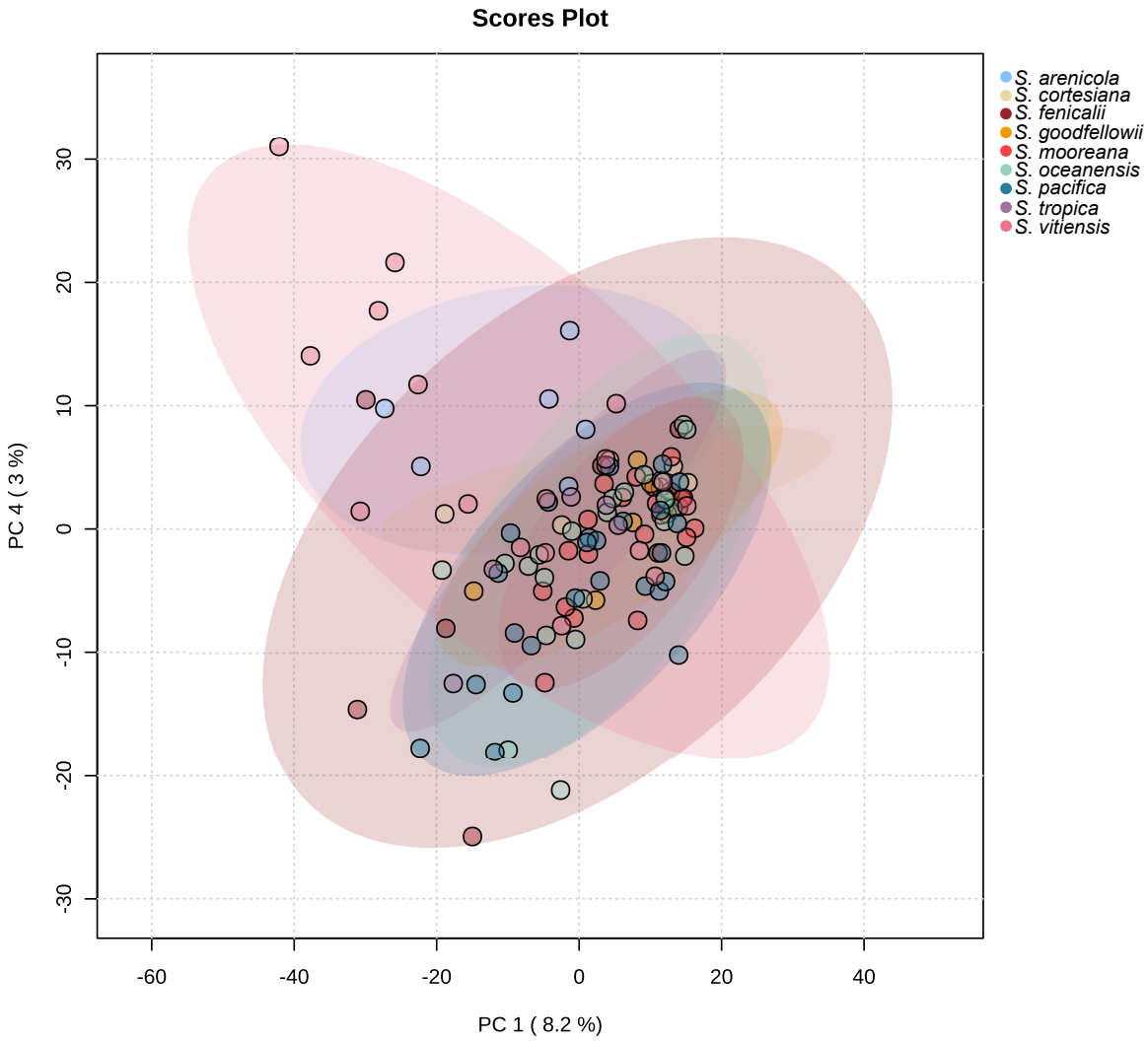

Figure 2.4 Principal component analysis (PCA) showing the metabolome variation in *Salinispora* spp., PC1 and PC4 as shown as they better represent the overlap among samples. Each node represents one sample replicate (n=117).

Table 2.2 Pairwise PERMANOVA test p-value 0.001, indicating the sample size, permutations, and pseudo-F values for each pairwise comparison between species (group 1 and group 2). Samples are grouped by species.

Group 1	**Group 2**	**Sample size**	**Permutations**	**pseudo-F**	**p-value**	**q-value**
arenicola	*cortesiana*	12	999	1.362287	0.002	0.003
	fenicalii	18	999	1.5509020	0.001	0.0017307
	goodfellowii	11	999	1.4126326	0.002	0.003
	media	9	999	0.6050609	0.987	1
	mooreana	24	999	1.7928215	0.001	0.0017307
	oceanensis	28	999	1.9488578	0.001	0.0017307
	pacifica	31	999	2.0608897	0.001	0.0017307
	tropica	12	999	1.3223380	0.001	0.0017307
	vitiensis	23	999	1.7892001	0.001	0.0017307
cortesiana	*fenicalii*	18	999	1.1072510	0.094	0.1175
	goodfellowii	11	999	1.2739460	0.003	0.0042187
	media	9	999	0.4493742	1	1
	mooreana	24	999	1.4092696	0.001	0.0017307
	oceanensis	28	999	1.4070005	0.002	0.003
	pacifica	31	999	1.6130300	0.001	0.0017307
	tropica	12	999	1.1420713	0.003	0.0042187
	vitiensis	23	999	1.5175596	0.001	0.0017307
fenicalii	*goodfellowii*	17	999	1.5679139	0.001	0.0017307
	media	15	999	0.4598712	0.948	1
	mooreana	30	999	1.2275955	0.001	0.0017307
	oceanensis	34	999	1.2251590	0.002	0.003
	pacifica	37	999	1.3485564	0.001	0.0017307
	tropica	18	999	1.2312121	0.006	0.0079411
	vitiensis	29	999	1.3775509	0.001	0.0017307
goodfellowii	media	8	999	0.5284241	1	1
	mooreana	23	999	1.8246405	0.001	0.0017307
	oceanensis	27	999	1.8595835	0.001	0.0017307
	pacifica	30	999	2.1127621	0.001	0.0017307
	tropica	11	999	1.3704553	0.004	0.0054545
	vitiensis	22	999	1.9214826	0.001	0.0017307
media	*mooreana*	21	999	0.4618714	0.983	1
	oceanensis	25	999	0.4742110	0.952	1
	pacifica	28	999	0.7521969	0.812	0.9875675
	tropica	9	999	0.7772469	0.989	1

Table 2.2 Pairwise PERMANOVA test p-value 0.001, indicating the sample size, permutations, and pseudo-F values for each pairwise comparison between species (group 1 and group 2). Samples are grouped by species, continued.

Group 1	**Group 2**	**Sample size**	**Permutations**	**pseudo-F**	**p-value**	**q-value**
media	*vitiensis*	20	999	0.7469974	0.966	1
mooreana	*oceanensis*	40	999	1.4667930	0.001	0.0017307
	pacifica	43	999	1.2612223	0.001	0.0017307
	tropica	24	999	1.2294497	0.017	0.0218571
	vitiensis	35	999	1.3919845	0.001	0.0017307
oceanensis	*pacifica*	47	999	1.5915327	0.001	0.0017307
	tropica	28	999	1.4506056	0.001	0.0017307
	vitiensis	39	999	1.6253720	0.001	0.0017307
pacifica	*tropica*	31	999	1.4218115	0.001	0.0017307
	vitiensis	42	999	1.2244125	0.001	0.0017307
tropica	*vitiensis*	23	999	1.3043088	0.001	0.0017307

2.4.2 Biosynthetic gene cluster diversity

To address biosynthetic gene cluster diversity and its correlation with speciation in *Salinispora* spp*.,* I performed a genome mining analysis of the antiSMASH output of the 118 *Salinispora* genome strains. I focused primarily on the distribution of BGCs among the novel *Salinispora* species and *S. pacifica.* All strains from one species must contain the putative species-specific BGCs, but not be present in other genome strains from different species; an exception was made for PKS25 and NRPS 27 (figure 2.5) which are in all strains of *S. oceanensis* and in two strains of *S. mooreana* and one strain of *S. goodfellowii* genome. I will explain the reasoning behind these exceptions in the next section of this chapter. Species-specific clusters were defined based on biosynthetic type, product prediction, gene cluster architecture, identity, and conservation across strains within one species. For PKS and NRPS BGCs, I incorporated the sequence similarity of KS and C domains to my analysis.

The methodology I followed for the BGC comparisons preceded the publication of the BiG-SCAPE/CORASON pipeline (Navarro-Muñoz *et al*., 2019), which is an important tool that facilitated the way we study biosynthetic diversity and genome mining nowadays. The BiG-SCAPE analysis utilizes antiSMASH output to classify the gene clusters into gene cluster families, which are clusters that are predicted to produce metabolites within the same structural family. However, the full biosynthetic diversity might be obscured by classifying similar BGCs as families, and subtle differences in gene sequence or distribution that can result in significant structural changes in the specialized metabolites. Determining gene clusters boundaries and what constitutes a GCF remains a major challenge. To assess the diversity of BGCs in the new *Salinispora* species, I needed finer scale differentiation that I could only achieve by comparing each gene cluster individually.

My results revealed 14 species-specific biosynthetic gene clusters across the six new species. I decided to include one additional GCF from the genome mining analysis performed by Chase *et al.*, (2021) as this GCF is predicted as a ladderane (named LADG, table 2.3 and fig. 2.5) for a total of 15 species-specific BGCs. These 15 biosynthetic gene clusters are predominantly NRPSs, PKSs, and hybrids, but also include one PKS type I cluster located close to genes related to ribosomal encoded and posttranslational modified peptides (RiPPs) and others predicted to be involved in the synthesis of a butyrolactone-like metabolite (fig. 2.5) The species with the most species-specific BGCs were *S. goodfellowii* and *S. fenicalii* with four species-specific BGCs each. However, these are also the species with fewest number of genome sequences and therefore, the true species-specific distribution of these gene clusters will depend on increasing the sample sizes and having access to more genome strains from *S. goodfellowii* and *S. fenicalii*.

The only characterized biosynthetic gene cluster among the 15 species-specific BGCs is the salinichelin BGC, which encodes a siderophore of NRPS origin that was isolated and described from *S. pacifica* strain CNY-331 (Bruns *et al.,* 2018). During the description of this compound, the species-specific distribution of the cluster was not reported. In fact, Bruns and collaborators hypothesized a functional swap between the desferrioxamine BGC, a commonly produced siderophore in *Salinispora*, and the salinichelin BGC because all the genome strains containing salinichelin lack the desferrioxamine BGC. This finding, together with knowing this is a species-specific BGC for *S. pacifica* emphasizes an important link between specialized metabolism and species evolution in *Salinispora.*

The correlation between species designations and specialized metabolism in *Salinispora* spp. has been studied recently using a different approach. Chase *et al*., (2021) looked at the species patterns in nine well described *Salinispora* metabolites and their related GCFs, including Rifamycin, staurosporine, salinosporamide, lomaiviticin, among others. They concluded that even in cases where the biosynthetic gene cluster may have originally been acquired through horizontal gene transfer (such as rifamycin) the genetic diversity within all nine GCFs was structured and subsequently maintained by processes of vertical descent over evolutionary timescales. These findings, coupled with my findings of species-specific BGCs, point out how important specialized metabolism is for *Salinispora*.

In the next section I will be discussing some of the most interesting cases of species-specific BGCs among *Salinispora* species.

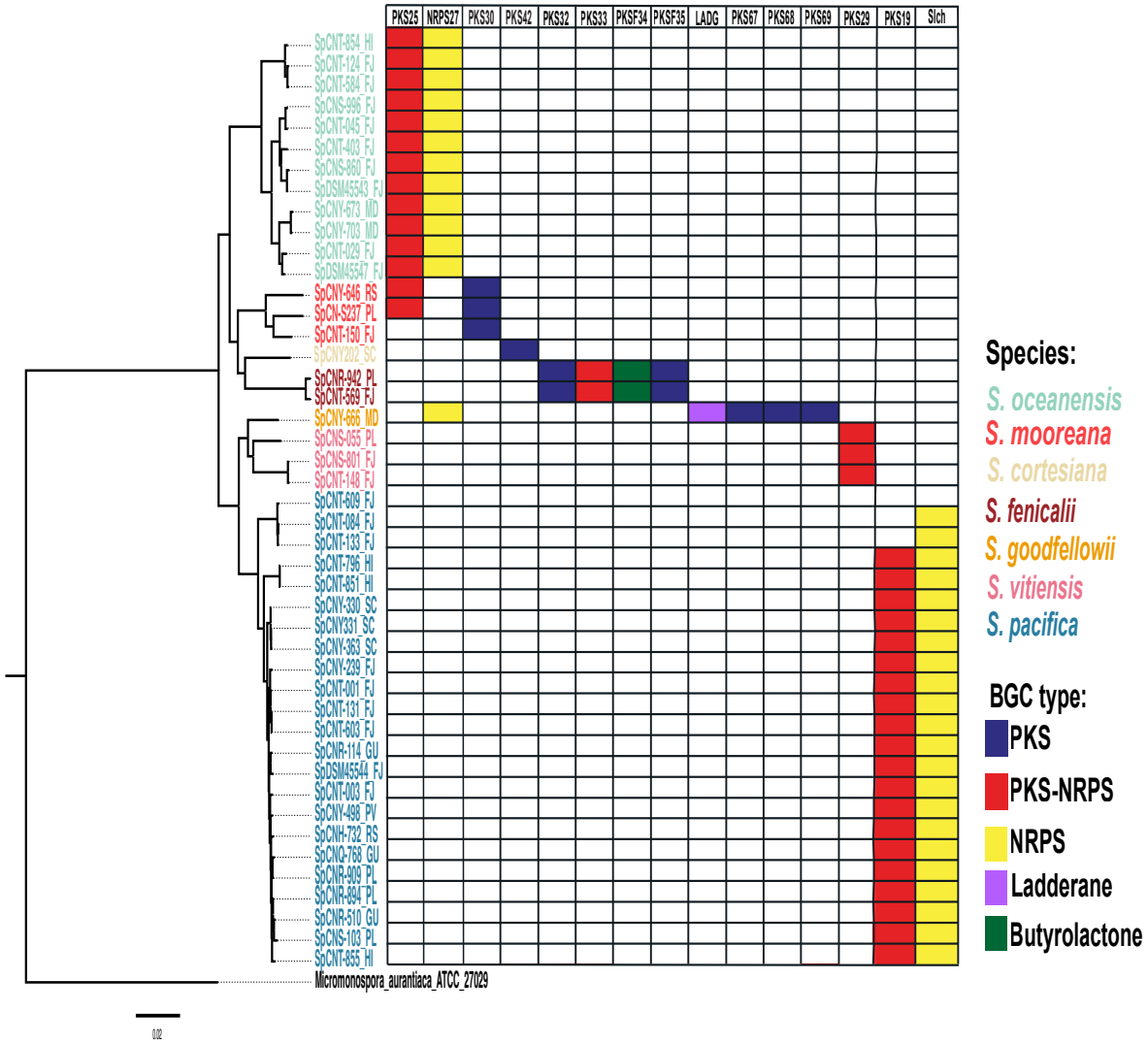

Figure 2.5 Distribution of the species-specific BGCs according to *Salinispora* species phylogeny. Strains are colored by species. The presence/absence table is colored based on BGC type using nomenclature reported in Letzel *et al*., 2016.

Table 2.3 *Salinispora* species-specific biosynthetic gene clusters. Compound predictions are based on antiSMASH v5.0. The BGC name and genomic position is based on the Letzel *et al*.,2016.

Species	Total of strains	BGC name	Description	Compound prediction	Genomic position	Conserved
S. mooreana	3	PKS30	Type I PKS/other	diisonitrile antibiotic SF2768 11%	not mapped	yes
S. cortesiana	1	PKS42	Type I PKS/ other	Tetronasin	GI 4/5	yes
S. fenicalli	2	PKS32	NRPS/Enediyne PKS	N/A	GI 15	yes
		PKS33	Thiopeptide/LAP/terpene/TIPKS/NRPS		GI 13/14	yes
		PKS34	Type II PKS/ butyrolactone	Prejadomycin (53%)	GI 4	yes
		PKS35	trans AT PKS	oocydin A 13%	GI 10	yes
S. vitiensis	3	PKS29	Ladderane, lipolanthine, arylpolyene, lanthipeptide	Kedarcidin 55%	GI 4	yes
S. goodfellowii	1	PKS67	Type I PKS	N/A	GI 4/5	yes
		PKS68	Type I PKS	N/A	GI 15	yes
		PKS69	Type I PKS	N/A	GI 15	yes
		M185_GCF4445	ladderane, NRPS genes	A54145 (nrps part) 5%	not mapped	yes
S. pacifica	23	NRPS16	NRPS	Salinichelin 100%	GI 16	yes
		PKS19	Trans-A NRPS/Type I PKS	Tubulysin 17%	GI 10	no, shared with CNS-055
S. oceanensis	12	NRPS27	NRPS	Maduropeptin 27%	GI 12 (9) GI 4/5 (1)	no, shared with CNY- 666
		PKS25	Type I PKS/ NRPS	Skyllamycin 40%	GI 10/11	no, shared with CNS-237, CNY-464

Table 2.4 NaPDoS results of all ketosynthases in the species-specific BGCs. The Query Id includes the strain name, BGC and KS sequence, the number indicates the putative PKS module.

Query id	Database match id	% identity	align length	e-value	pathway product	domain class
CNT_150_PKS30_KS1	chlorizidine_KS04_FASII	47	345	6.60E-71	chlorizidine	FAS type II
CNT_150_PKS30_KS2	nostophycin_KS03_cisAT	51	425	1.70E-124	nostophycin	cis-AT
CNT_569_PKSF32_KS	maduropeptin_KS01_iPKSenediyne	74	457	1.30E-202	maduropeptin	iterative cis-AT enediyne
CNT_569_PKS33_KS	tubulysin_KS01_cisHybridKS	50	429	2.70E-114	tubulysin	cis-AT hybrid KS
CNT_569_PKSF34_KS1	oviedomycin_KS02_angucyclineIKSb	74	397	2.40E-173	oviedomycin	type II KSa aromatic angucycline I
CNT_569_PKSF34_KS2	urdamycin_KS01_angucyclineIKSa	82	410	2.80E-193	urdamycin	type II KSa aromatic angucycline I
CNT_569_PKSF35_KS1	oxazolomycin_KS03_transAT	59	434	1.60E-143	oxazolomycin	trans-AT
CNT_569_PKSF35_KS2	oxazolomycin_KS05_transAT	47	440	1.20E-98	oxazolomycin	trans-AT
CNT_569_PKSF35_KS3	sorangicin_KS11_transAT	58	435	2.00E-138	sorangicin	trans-AT
CNT_569_PKSF35_KS4	myxovirescin_KS10_transAT	51	437	2.10E-122	myxovirescin	trans-AT
CNT_569_PKSF35_KS5	ripostatin_KS09_transAT	55	424	2.10E-127	ripostatin	trans-AT
CNT_569_PKSF35_KS6	sorangicin_KS20_transAT	53	442	8.70E-137	sorangicin	trans-AT
CNT_569_PKSF35_KS7	SIA7248_KS01_transAT	57	420	6.10E-127	SIA7248	trans-AT

Table 2.4 NaPDoS results of all ketosynthases in the species-specific BGCs. The Query Id includes the strain name, BGC and KS sequence, the number indicates the putative PKS module, continued.

Query id	Database match id	% identity	align length	e-value	pathway product	domain class
CNT_569_PKSF35_KS8	leinamycin_KS07_transAT	54	425	1.00E-126	leinamycin	trans-AT
CNT_569_PKSF35_KS9	chivosazol_KS16_transAT	61	416	3.90E-150	chivosazol	trans-AT
CNT_148_PKS29_KS1	colabomycin_KS02_polyeneKSa	69	378	9.50E-146	colabomycin	type II KSa polyene
CNT_148_PK29_KS2	colabomycin_KS01_polyeneKSb	61	186	8.40E-59	colabomycin	type II KSb polyene
CNT_148_PKS29_KS3	skyllamycin_KS02_polyeneKSa	43	376	7.90E-70	skyllamycin	type II KSa polyene
CNT_148_PKS29_KS4	skyllamycin_KS03_polyeneKSb	40	177	1.00E-27	skyllamycin	type II KSb polyene
CNR_114_PKS19_KS	tubulysin_KS02_cisHybridKS	53	427	7.20E-117	tubulysin	cis-AT hybrid KS

2.4.3 A shared NRPS/PKS biosynthetic gene cluster in two *Salinispora* species: *S. oceanensis* and *S. mooreana*

S. oceanensis has 12 genome strains which contain two species-specific BGCs. One is a NRPS type (NRPS27), with 27% similarity of maduropeptin, and a type I PKS/NRPS hybrid (PKS25) with the NRPS/PKS core genes having 40% similarity to skyllamycin from *Streptomyces sp.* Acta 2897 (fig. 2.6A). Using the NRPS/PKS monomer predictor website (Bachmann *et al.,* 2009), I was able to mine how many NRPS and PKS modules were in each gene and the amino acids and acetate monomers that PKS25 would incorporate (fig. 2.6B). There are three NRPS modules including the loading module, incorporating phenylalanine in the loading module, valine or isoleucine in module 2, and an undetermined amino acid in the third module. The third gene contains one PKS module, which is predicted to incorporate malonyl-CoA. The KS clades with hybrid KS sequences in NaPDoS, giving me confidence that this is indeed a hybrid BGC and not just NRPS and PKS genes from different BGCs that are co-located. There are other biosynthetic genes in the gene cluster that might act as tailoring enzymes and further modify the structure of the product, for example two methyltransferases, and carboxylases.

Upstream the PKS25, there are two type II PKS genes and an individual acyl carrier protein (ACP) (fig 2.6A). These are conserved across all 12 *S. oceanensis* strains and the two *S. mooreana* strains that have PKS25 (fig. 2.7). Type II PKSs are known to have individual genes encoding for the ACP and the α and β ketosynthase subunits, which is the arrangement found in PKS25. To my knowledge, there are no reports of specialized metabolites biosynthesized by hybrid type I and type II PKS. However, more work will be required to determine if this represents one large or two separate BGCs.

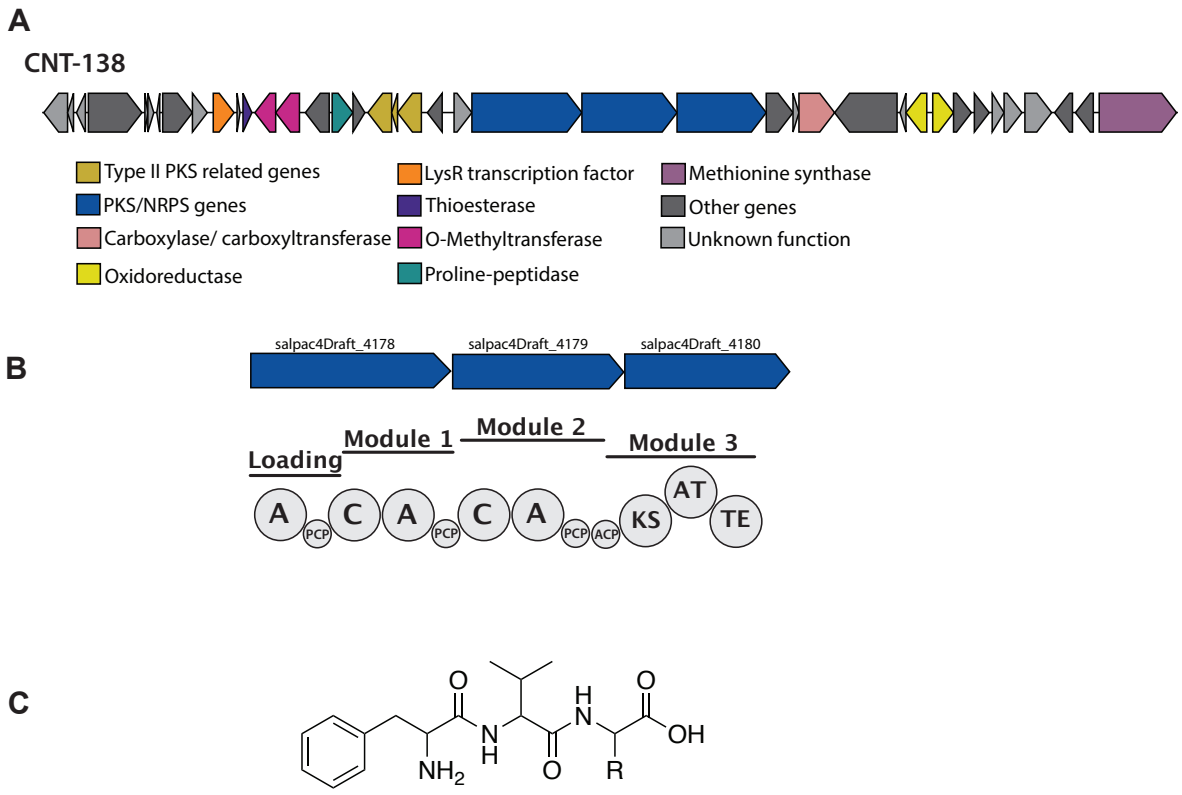

Figure 2.6 Overview of PKS25. **A**. Description of biosynthetic genes in PKS25, color coded by gene annotation. **B**. NRPS/PKS biosynthetic genes and the module architecture. **C.** core product prediction based only NRPS/PKS biosynthesis, using Valine as the amino acid incorporated in module 1; the R indicates the undetermined amino acid from module 2.

PKS25 is located close to the staurosporine BGC and thus the borders can't be easily predicted resulting in antiSMASH annotating the whole region as one gene cluster or neighboring cluster environment. However, after further inspection of other species with the staurosporine gene cluster, I was able to distinguish the staurosporine BGC from PKS25 and define putative boundaries within the *S. oceanensis* and *S. mooreana* genomes (fig. 2.7). I observed subtle but important differences between the versions of PKS25 in the two species; the most important of which is the lack of a predicted transcription regulator of the *lysR* family in *S. mooreana*, which could be very important for the expression of the gene cluster. The *lysR* family of transcription factors regulate diverse set of genes, including those involved in virulence, metabolism, antibiotic

resistance, quorum sensing, and motility (Maddocks & Oyston, 2008). The gene missing from the *S. mooreana* BGC is predicted to encode for a cofactor-binding protein. Without the production of this cofactor, perhaps PKS25 wouldn't be active in *S. mooreana.* It is also possible that this cofactor affects the transcription of other genes since co-localization of *lysR* transcription regulators is not necessary for their functionality (Maddocks & Oyston, 2008). Ideally, transcription experiments could be designed to examine the functionality of PKS25 in *S. mooreana*.

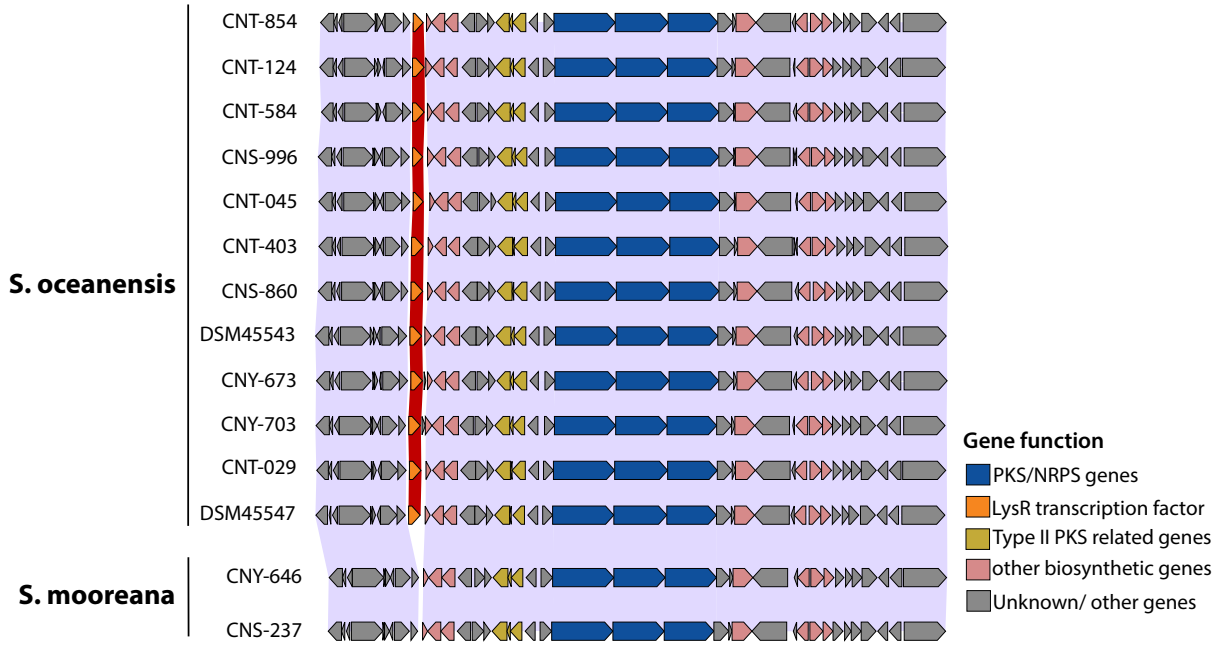

Figure 2.7 Alignment of the BGC PKS25 across S. oceanensis and S. mooreana. The violet shading shows the homology of conserved genes across strains. The red shading highlights the distribution of the *lysR* transcription cofactor.

2.4.4 The Trans-AT PKS biosynthetic gene cluster in *S. fenicalii*

There are four unique biosynthetic gene clusters in *S. fenicalii*. Although the species only contains two genome strains, it is interesting that these four BGC are not shared with any other *Salinispora* genome strains. These four biosynthetic gene clusters contain type I and type II PKS

genes and are co-located with other biosynthetic types like NRPS and butyrolactone like genes (table 2.3).

The PKS35 gene cluster is predicted as a trans-AT PKS by antiSMASH. To date, there are no trans-AT PKS encoded natural products described from *Salinispora*. In general, trans AT-like biosynthetic enzymes have been reported to be scarce in actinomycetes (Paulus *et al*., 2018); thus, I found this BGC to be particularly interesting for genome mining. The cluster contains three PKS genes with nine different trans-AT KS domains and one smaller gene predicted as the acetyltransferase domain (fig.2.8). The NaPDoS pipeline results confirmed these nine KS sequences as trans-AT. However seven out of the nine KS sequences formed an individual phylogenetic clade, which could be hinting on novelty, either biosynthetically or because of the few trans-AT KS domains in the database from actinomycetes, for the remainder two once cladded with a KS sequence from the ripostatin biosynthetic pathway, originally isolated from myxobacteria (Irschik *et al*., 1995), and the second with the tartolon biosynthesis, isolated from several sources including a marine derived *Streptomyces sp.* (Pérez *et al*., 2009), and from a symbiotic gammaproteobacteria isolated from shipworm gills (Elshahawi *et al*., 2013).

Aside from the KS domains, the nine modules contained domains that are usually found to be correlated with trans-AT polyketide biosynthesis, such as a trans-AT docking domain. The loading module contains a FbkH domain, which has been described as a less canonical domain for trans-AT PKS. FbkH-like proteins are C3-acyltransferases reported to be part of the loading domain in trans-AT PKSs and to load glycolytic intermediates onto ACP domains (Helfrich and Piel, 2015; Helfrich *et al*., 2019).

Unfortunately, the rules that are used for predicting cis-AT polyketide structures cannot be applied to trans-AT polyketides. However, bioinformatic studies have revealed a close correlation

between KS domain phylogenies and the biosynthetic intermediate it recognizes, and this relationship has been used to predict monomer incorporation by trans-AT PKSs. One of the most comprehensive tools for trans-AT monomer prediction is transATor (Helfrich *et al*., 2019), which I used to investigate the monomers for PKS35, which are listed in table 2.5. The predicted core polyketide structure has a mass of around 340 Da. However, there are two methyltransferase domains inside the PKS genes in module 5 and module 7 and a cytochrome P450 in the BGC, which if functional would increase the mass of the predicted polyketide. However, I was not able to find any peaks in the mass spectrometry data that represent candidate products for this trans-AT PKS gene cluster.

It is important to highlight the genomic environment where PKS35 is located. In CNT-569, the cluster is at the end of a contig, thus further sequencing should be done to better characterize this BGC. Nonetheless, it is interesting that the cluster is flanked by CRISPR elements in both strains of *S. fenicalii*. In CNT-569, there are CRISPR repeats upstream of the cluster while in CNR-942, there are several genes annotated as CRISPR associated proteins related to Cas7 and Cas3 from the type I cascade system (Morisaka *et al.,* 2019) along with CRISPR repeats. CRISPR and *cas* locus are often located next to each other in bacterial genomes. Although few studies have addressed co-localization, it has been reported that only about 11% of ~6000 bacterial and archaeal genomes have isolated type I CRISPR-Cas systems, meaning without CRISPRs or *cas* elements collocated closely within the (Zhang and Ye, 2017). Although, in some cases the isolated CRISPR regions may be not functional, or function with remote *cas* locus, within the same genome which could be the case for CNT-569 (Zhang and Ye, 2017). Some studies have reported other functions such as regulatory roles for Type I CRISPR-Cas systems beyond bacterial defenses against viral nucleic acids or plasmid invasion, (Mohanraju *et al.,* 2022). For instance, in *Myxococcus xanthus*,

it has been observed that during nutrient limitation the CRISPR-Cas system can modulate expression of endogenous genes to regulate group behavior, cell aggregate formation, and spore development (Viswanathan *et al*., 2007). It would be interesting to further investigate which of these scenarios is occurring in the *S. fenicalii* genome strains and if the CRISPR-Cas genes flanking PKS35 have a role in BGC functionality.

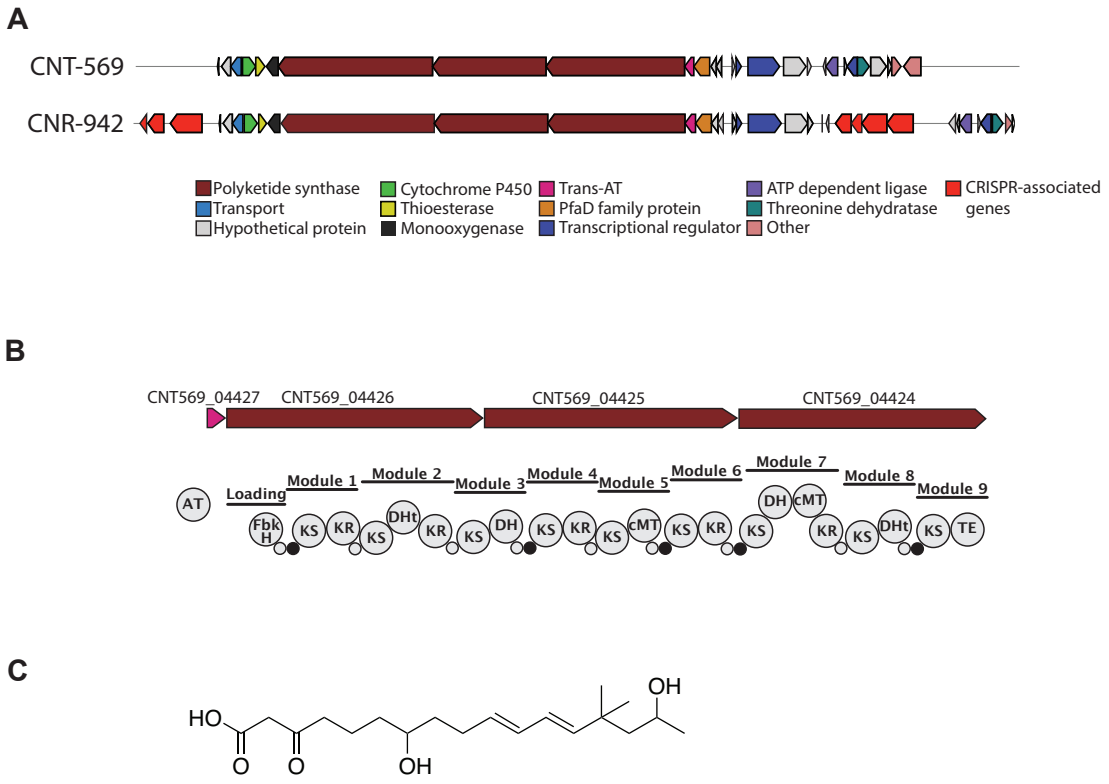

Figure 2.8 The PKS35 BGC and its predicted product. **A**. Biosynthetic gene annotation in two the two strains of *S. fenicalii*. **B**. Trans-AT PKS genes and module architecture and **C.** Core product prediction based the transAtor software.

Table 2.5 Monomer prediction for PKS35. Results resulting from transATor software.

Gene	Ketosynthase	Prediction monomer	E-value
CNT569_04426	Ks1	lactate starter	9.90E-168
	KS2	various specificities (α-Me)/completely reduced/β D-OH	2.40E-160
	KS3	β-OMe /βMe double bond/β-ketoexomethylene	3.60E-148
CNT569_04425	KS4	β-keto/double bonds/amino acids	2.00E-156
	KS5	various specificities/α-Me/B OH	9.80E-145
	KS6	various specificities/completely reduced/ pyran or furan rings	1.10E-174
CNT569_04424	KS7	β D-OH/β-keto/ double bonds	7.90E-193
	KS8	non elongating/double bonds/β D-OH	2.90E-137
	KS9	various specificities/completely reduced/β- D-OH	8.90E-180

2.4.5 The Type I PKS and Trans-Adenylation NRPS BGC in *S. pacifica*

As mentioned before, *S. pacifica* genomes have two different species-specific BGCs: the salnichelin gene cluster and a type I PKS-NRPS hybrid named PKS19. The latter cluster is 17% similar to the tubulysin BGC, an NRPS/PKS type I hybrid from the myxobacterium *Cystobacter sp*. SBCb004. However, they only share sequence homology between the PKS genes. Interestingly, two of the three PKS19 NRPS genes don't have an adenylation domain, which is characteristic of NRPs biosynthetic pathways. Instead, they possess a small gene encoding for adenylation and *mbtH*-like domains upstream of the NRPS/PKS genes (fig. 2.9). MbtH-like proteins are found in NRPS BGCs. They are homologous with MLPs and have been reported essential for the solubility and activity of the NRPSs (Felnagle *et al*., 2010; Boll *et al*., 2011). NRPSs usually act co-linearly and the order of amino acids in the peptide corresponds to the order of the individual modules (Süssmuth & Mainz 2017). However, there are reports of adenylation domains working in a trans manner, for instance in the biosynthesis of WS9326A from *Streptomyces calvus,* where one of the

NRPS genes encodes an A-less module. These BGCs have a single gene encoding for an MbtH-dependent adenylation domain (Bernhardt & Berman, 2020), which is similar to what I observed in PKS19. This led me to hypothesize that PKS19 could be an MbtH dependent, trans-A NRPS BGC. Bernhardt and Berman (2020) determined that MbtH-like proteins promote soluble expression of the trans adenylation domains and are crucial for the biosynthesis of WS9326A and its analogs. While there are not many examples, trans-A NRPS BGCs have been proposed in the biosynthesis of andrimid (Magarvey *et al*., 2008).

Mining this type of BGC might not be as straight forward as with cys-A NRPS genes, nonetheless, the NRPS/PKS monomer predictor website (Bachmann and Ravel, 2009) predicted valine as the monomer amino acid for the trans-adenylation domain. Besides the two trans-A modules, there is a final cys-A module, for which I was not able to predict the amino acid monomer. The PKS gene is located between the NRPS genes and contains the minimal ACP, KS, and AT domains. Interestingly, there is no predicted thioesterase domain, or any other cleaving enzyme for NRPS or PKS biosynthesis. Perhaps one of the uncharacterized genes located in the cluster could act as a cleaving enzyme, but there is not enough information in databases to make an informed hypothesis. Ideally, further enzymatic analysis would be needed to characterize the cluster and its functionality.

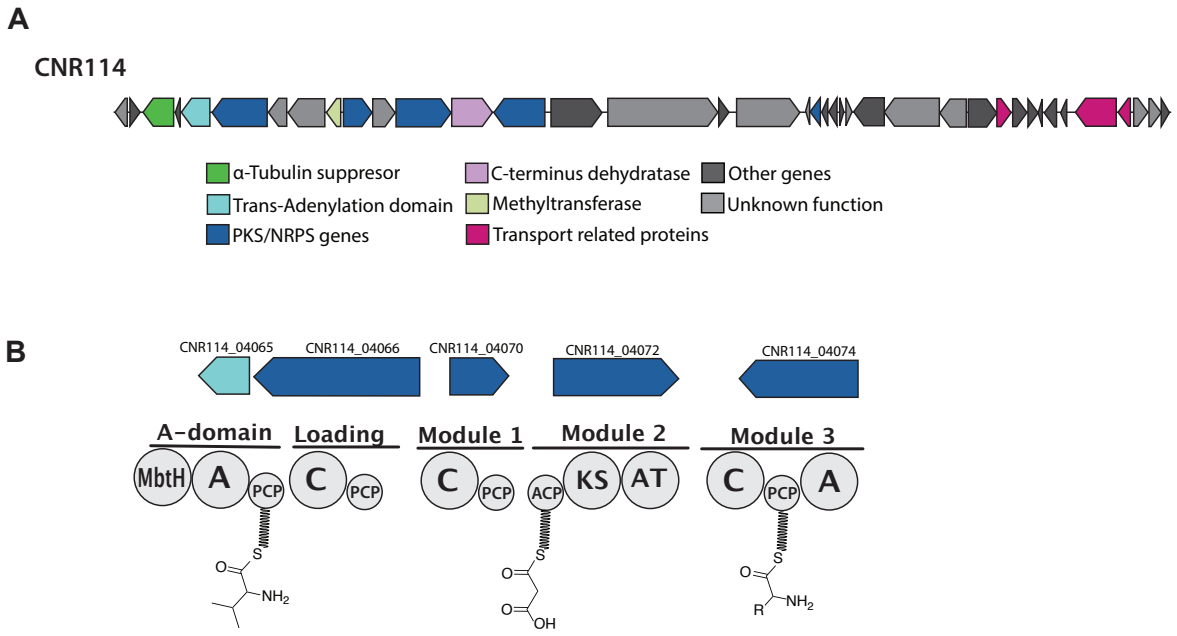

Figure 2.9 *S. pacifica* species-specific BGC. **A**. Trans-Adenylation domain gene cluster observed in CNR-114. Genes colored by functional annotation. **B**. The PKS/NRPS biosynthetic modules. Light blue gene indicates the trans Adenylation NRPS gene. Valine is predicted as a monomer for the first two modules. Malonyl-CoA is predicted as monomer of module 2. R represents the unknown monomer from module 3.

2.5 Conclusion

From the genome mining analysis performed in this study, I have discovered BGCs representing poorly studied biosynthetic pathways, such as a trans-adenylation NRPSs and a trans-AT PKS, which are both very rare in actinomycetes. I also detected other uncharacterized biosynthetic gene clusters with promising novelty, such as clusters with genes encoding for butyrolactone-like products or RiPPs.

I have confirmed that the six new *Salinispora* species each contain at least one species-specific biosynthetic gene cluster and identified a total of 15 species-specific BGCs in total within these six species and *S. pacifica.* However, the metabolomic distribution does not follow the same species pattern, with about 30% of the variation in the data explained by species (PERMANOVA p=0.001). The specialized metabolism of the genus *Salinispora* has been long hypothesized to

follow a species-specific pattern. The findings reported here support the species-specific pattern hypothesis in the specialized metabolism of *Salinispora* and help our understanding of the correlation of specialized metabolism with species phylogeny in the marine actinomycete *Salinispora*.

2.6 Acknowledgements

I would like to thank members of the Jensen Laboratory including Kaitlin Creamer, Henrique Ramalho Machado and Krystle Chavarría, for their assistance in using all the bioinformatic tools used for the comparative genomics done in this chapter. Further, I would like to thank Leesa Klau for introducing me to metaboAnalyst. Special thanks to Itzel Lizama Chamú my mentee from the summer undergraduate research program, for her help with chemical extracts and the comparative genomics of pks19.

This research was supported by the NIH grant to R01GM085770 PRJ and BSM and the CONACYT and UC MEXUS doctoral fellowship.

2.7 References

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

Bernhardt, M., Berman, S.F., Zechel, D.L., & Bechthold, A. (2020). Role of Two Exceptional trans Adenylation Domains and MbtH-like Proteins in the Biosynthesis of the Nonribosomal Peptide WS9324A from Streptomyces calvus ATCC 13382. *Chembiochem, 21*, 2659 - 2666.

Boll, B., Taubitz, T., & Heide, L. (2011). Role of MbtH-like Proteins in the Adenylation of Tyrosine during Aminocoumarin and Vancomycin Biosynthesis*. *The Journal of Biological Chemistry, 286*, 36281 - 36290.

Bolyen, E., Rideout, J.R., Dillon, M.R., Bokulich, N.A., Abnet, C.C., Al-Ghalith, G.A., Alexander, H., Alm, E.J., Arumugam, M., Asnicar, F., Bai, Y., Bisanz, J.E., Bittinger, K., Brejnrod,

A.D., Brislawn, C.J., Brown, C.T., Callahan, B.J., Caraballo-Rodríguez, A.M., Chase, J.H., Cope, E.K., da Silva, R.R., Diener, C., Dorrestein, P.C., Douglas, G.M., Durall, D.M., Duvallet, C., Edwardson, C.F., Ernst, M., Estaki, M., Fouquier, J.T., Gauglitz, J.M., Gibbons, S.M., Gibson, D.L., Gonzalez, A., Gorlick, K., Guo, J., Hillmann, B.M., Holmes, S.P., Holste, H., Huttenhower, C., Huttley, G.A., Janssen, S., Jarmusch, A.K., Jiang, L., Kaehler, B.D., Kang, K.B., Keefe, C.R., Keim, P.S., Kelley, S.T., Knights, D., Koester, I., Kosciólek, T., Kreps, J., Langille, M.G., Lee, J.S., Ley, R.E., Liu, Y., Loftfield, E., Lozupone, C.A., Maher, M., Marotz, C.A., Martin, B.D., McDonald, D., McIver, L.J., Melnik, A.V., Metcalf, J.L., Morgan, S.C., Morton, J.T., Naimey, A.T., Navas-Molina, J.A., Nothias, L., Orchanian, S.B., Pearson, T.R., Peoples, S.L., Petráš, D., Preuss, M.L., Pruesse, E., Rasmussen, L.B., Rivers, A.R., Robeson, M.S., Rosenthal, P., Segata, N., Shaffer, M., Shiffer, A., Sinha, R., Song, S.J., Spear, J.R., Swafford, A.D., Thompson, L.R., Torres, P.J., Trinh, P., Tripathi, A., Turnbaugh, P.J., Ul-Hasan, S., van der Hooft, J.J., Vargas, F., Vázquez-Baeza, Y., Vogtmann, E., von Hippel, M., Walters, W.A., Wan, Y., Wang, M., Warren, J., Weber, K.C., Williamson, C.H., Willis, A.D., Xu, Z.Z., Zaneveld, J.R., Zhang, Y., Zhu, Q., Knight, R., & Caporaso, J.G. (2019). Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nature Biotechnology*, 1-5.

Bruns, H., Crüsemann, M., Letzel, A., Alanjary, M., McInerney, J.O., Jensen, P.R., Schulz, S., Moore, B.S., & Ziemert, N. (2018). Function-related replacement of bacterial siderophore pathways. *The ISME Journal, 12*, 320 - 329.

Ceniceros, A., Dijkhuizen, L., Petrusma, M., & Medema, M.H. (2017). Genome-based exploration of the specialized metabolic capacities of the genus Rhodococcus. *BMC Genomics, 18*.

Chase, A.B., Sweeney, D., Muskat, M.N., Guillén-Matus, D., & Jensen, P.R. (2021). Vertical Inheritance Facilitates Interspecies Diversification in Biosynthetic Gene Clusters and Specialized Metabolites. *mBio, 12*.

Chevrette, M.G., Sélem-Mojica, N., Aguilar, C., Labby, K.J., Bustos-Díaz, E.D., Handelsman, J., & Barona-Gómez, F. (2022). Evolutionary Genome Mining for the Discovery and Engineering of Natural Product Biosynthesis. *Methods in molecular biology, 2489*, 129-155

Covington, B.C., McLean, J.A., & Bachmann, B.O. (2017). Comparative mass spectrometry-based metabolomics strategies for the investigation of microbial secondary metabolites. *Natural product reports, 34 1*, 6-24 .

Diminic, J., Starčević, A., Lisfi, M., Baranasic, D., Gacesa, R., Hranueli, D., Long, P.F., Cullum, J., & Zucko, J. (2014). Evolutionary concepts in natural products discovery: what actinomycetes have taught us. *Journal of Industrial Microbiology & Biotechnology, 41*, 211-217.

Elshahawi, S.I., Trindade-Silva, A.E., Hanora, A.M., Han, A.W., Flores, M.S., Vizzoni, V.F., Schrago, C.G., Soares, C.A., Concepcion, G.P., Distel, D.L., Schmidt, E.W., & Haygood, M.G. (2013). Boronated tartrolon antibiotic produced by symbiotic cellulose-degrading

bacteria in shipworm gills. *Proceedings of the National Academy of Sciences, 110*, E295 - E304.

Felnagle, E.A., Barkei, J.J., Park, H., Podevels, A.M., McMahon, M.D., Drott, D.W., & Thomas, M.G. (2010). MbtH-like proteins as integral components of bacterial nonribosomal peptide synthetases. *Biochemistry, 49 41*, 8815-7.

Helfrich, E.J., & Piel, J. (2016). Biosynthesis of polyketides by trans-AT polyketide synthases. *Natural product reports, 33 2*, 231-316.

Irschik, H., Augustiniak, H., Gerth, K., Höfle, G., & Reichenbach, H.P. (1995). The ripostatins, novel inhibitors of eubacterial RNA polymerase isolated from myxobacteria. *The Journal of antibiotics, 48 8*, 787-92.

Letzel, A., Li, J., Amos, G.C., Millán-Aguiñaga, N., Ginigini, J., Abdelmohsen, U.R., Gaudêncio, S.P., Ziemert, N., Moore, B.S., & Jensen, P.R. (2017). Genomic insights into specialized metabolism in the marine actinomycete Salinispora. *Environmental Microbiology, 19*, 3660–3673.

Li, M. H., Ung, P. M., Zajkowski, J., Garneau-Tsodikova, S., & Sherman, D. H. (2009). Automated genome mining for natural products. *BMC bioinformatics*, *10*, 185.

Liu, X., & Locasale, J.W. (2017). Metabolomics: A Primer. *Trends in biochemical sciences, 42 4*, 274-284.

Maddocks, S.E., & Oyston, P.C. (2008). Structure and function of the LysR-type transcriptional regulator (LTTR) family proteins. *Microbiology, 154 Pt 12*, 3609-23 .

Magarvey, N.A., Fortin, P.D., Thomas, P.M., Kelleher, N.L., & Walsh, C.T. (2008). Gatekeeping versus promiscuity in the early stages of the andrimid biosynthetic assembly line. *ACS chemical biology, 3 9*, 542-54 .

Marchler-Bauer, A., Yu, B., Han, L., He, J., Lanczycki, C.J., Lu, S., Chitsaz, F., Derbyshire, M.K., Geer, R.C., Gonzales, N.R., Gwadz, M., Hurwitz, D.I., Lu, F., Marchler, G.H., Song, J.S., Thanki, N., Wang, Z., Yamashita, R.A., Zhang, D., Zheng, C., Geer, L.Y., & Bryant, S.H. (2017). CDD/SPARCLE: functional classification of proteins via subfamily domain architectures. *Nucleic Acids Research, 45*, D200 - D203.

Medema, M. H., Blin, K., Cimermancic, P., de Jager, V., Zakrzewski, P., Fischbach, M. A., Weber, T., Takano, E., & Breitling, R. (2011). antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. *Nucleic acids research*, *39*(Web Server issue), W339–W346.

Medema, M.H., Takano, E., & Breitling, R. (2013). Detecting Sequence Homology at the Gene Cluster Level with MultiGeneBlast. *Molecular Biology and Evolution, 30*, 1218 - 1223.

Mohanraju P., Saha C., van Baarlen P., Louwen R., Stalls R. H. J, van der Oost J. (2002). Alternative functions of CRISPR–Cas systems in the evolutionary arms race. *Nat Rev Microbiol* **20,** 351–364

Morisaka, H., Yoshimi, K., Okuzaki, Y., Gee, P., Kunihiro, Y., Sonpho, E., Xu, H., Sasakawa, N., Naito, Y., Nakada, S., Yamamoto, T., Sano, S., Hotta, A., Takeda, J., & Mashimo, T. (2019). CRISPR-Cas3 induces broad and unidirectional genome editing in human cells. *Nature Communications, 10*.

Navarro-Muñoz, J.C., Sélem-Mojica, N., Mullowney, M.W., Kautsar, S., Tryon, J.H., Parkinson, E.I., de los Santos, E.L., Yeong, M., Cruz-Morales, P., Abubucker, S., Roeters, A., Lokhorst, W., Fernàndez-Guerra, A., Cappelini, L.T., Goering, A.W., Thomson, R.J., Metcalf, W.W., Kelleher, N.L., Barona-Gómez, F., & Medema, M.H. (2019). A computational framework to explore large-scale biosynthetic diversity. *Nature chemical biology, 16*, 60 - 68.

Pang, Z., Chong, J., Zhou, G., de Lima Morais, D.A., Chang, L., Barrette, M., Gauthier, C., Jacques, P., Li, S., & Xia, J. (2021). MetaboAnalyst 5.0: narrowing the gap between raw spectra and functional insights. *Nucleic Acids Research, 49*, W388 - W396**.**

Patti, G.J., Yanes, Ó., & Siuzdak, G. (2012). Innovation: Metabolomics: the apogee of the omics trilogy. *Nature Reviews Molecular Cell Biology, 13*, 263-269.

Paulus, C., Rebets, Y.V., Zapp, J., Rückert, C., Kalinowski, J., & Luzhetskyy, A. (2018). New Alpiniamides From Streptomyces sp. IB2014/011-12 Assembled by an Unusual Hybrid Non-ribosomal Peptide Synthetase Trans-AT Polyketide Synthase Enzyme. *Frontiers in Microbiology, 9*.

Pérez, M., Crespo, C., Schleissner, C., Rodríguez, P., Zúñiga, P.E., & Reyes, F. (2009). Tartrolon D, a cytotoxic macrodiolide from the marine-derived actinomycete Streptomyces sp. MDG-04-17-069. *Journal of natural products, 72 12*, 2192-4 .

Pluskal, T., Castillo, S., Villar-Briones, A., & Orešič, M. (2010). MZmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinformatics, 11*, 395 - 395.

Qi, Z., & Voit, E.O. (2017). Strategies for Comparing Metabolic Profiles: Implications for the Inference of Biochemical Mechanisms from Metabolomics Data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics, 14*, 1434-1445.

Robinette, S.L., Brüschweiler, R., Schroeder, F.C., & Edison, A.S. (2012). NMR in Metabolomics and Natural Products Research: Two Sides of the Same Coin. *Accounts of Chemical Research, 45*, 288 - 297.

Skinnider, M. A., Merwin, N. J., Johnston, C. W., & Magarvey, N. A. (2017). PRISM 3: expanded prediction of natural product chemical structures from microbial genomes. *Nucleic acids research*, *45*(W1), W49–W54. https://doi.org/10.1093/nar/gkx320

Süssmuth, R.D., & Mainz, A. (2017). Nonribosomal Peptide Synthesis-Principles and Prospects. *Angewandte Chemie, 56 14*, 3770-3821.

Undabarrena, A., Valencia, R., Cumsille, A., Zamora-Leiva, L., Castro-Nallar, E., Barona-Gómez, F., & 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. *Microbial Genomics, 7*.

Viswanathan, P., Murphy, K.A., Julien, B., Garza, A.G., & Kroos, L. (2007). Regulation of dev, an Operon That Includes Genes Essential for Myxococcus xanthus Development and CRISPR-Associated Genes and Repeats. *Journal of Bacteriology, 189*, 3738 - 3750.

Wang, M., Carver, J., Phelan, V.V., Sanchez, L.M., Garg, N., Peng, Y., Nguyen, D.D., Watrous, J.D., Kapono, C.A., Luzzatto-Knaan, T., Porto, C., Bouslimani, A., Melnik, A.V., Meehan, M.J., Liu, W., Crüsemann, M., Boudreau, P.D., Esquenazi, E., Sandoval-Calderón, M., Kersten, R.D., Pace, L.A., Quinn, R.A., Duncan, K.R., Hsu, C., Floros, D.J., Gavilán, R.G., Kleigrewe, K., Northen, T.R., Dutton, R.J., Parrot, D., Carlson, E.E., Aigle, B., Michelsen, C.F., Jelsbak, L., Sohlenkamp, C., Pevzner, P.A., Edlund, A., McLean, J.S., Piel, J., Murphy, B.T., Gerwick, L., Liaw, C., Yang, Y., Humpf, H., Maansson, M., Keyzers, R.A., Sims, A.C., Johnson, A.R., Sidebottom, A.M., Sedio, B.E., Klitgaard, A.K., Larson, C.B., P, C.A., Torres-Mendoza, D., Gonzalez, D.J., Silva, D.B., Marques, L.M., Demarque, D.P., Pociūtė, E., O'Neill, E.C., Briand, E., Helfrich, E.J., 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.S., Silva, R.R., Mascuch, S.J., Tomasi, S., Jenkins, S., Macherla, V.R., Hoffman, T., Agarwal, V., Williams, P.G., Dai, J., Neupane, R.P., Gurr, J.R., Rodriguez, A.M., Lamsa, A., Zhang, C., Dorrestein, K., Duggan, B.M., Almaliti, J., Allard, P., Phapale, P., Nothias, L., Alexandrov, T., Litaudon, M., Wolfender, J., Kyle, J.E., Metz, T.O., Peryea, T., Nguyen, D., Vanleer, D., Shinn, P., Jadhav, A., Müller, R., Waters, K.M., Shi, W., Liu, X., Zhang, L., Knight, R., Jensen, P.R., Palsson, B.O., Pogliano, K., Linington, R.G., Gutiérrez, M., Lopes, N.P., Gerwick, W.H., Moore, B.S., Dorrestein, P.C., & Bandeira, N. (2016). Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nature Biotechnology, 34*, 828-837.

Xia, J., & Wishart, D.S. (2016). Using MetaboAnalyst 3.0 for Comprehensive Metabolomics Data Analysis. *Current Protocols in Bioinformatics, 55*, 14.10.1 - 14.10.91.

Zhang, Q., & Ye, Y. (2017). Not all predicted CRISPR–Cas systems are equal: isolated cas genes and classes of CRISPR like elements. *BMC Bioinformatics, 18*.

Ziemert, N., Podell, S., Penn, K., Badger, J.H., Allen, E.E., & Jensen, P.R. (2012). The Natural Product Domain Seeker NaPDoS: A Phylogeny Based Bioinformatic Tool to Classify Secondary Metabolite Gene Diversity. *PLoS ONE, 7*.

Chapter 3

Expression of species-specific biosynthetic gene clusters in the marine actinomycete *Salinispora*

3.1 Abstract

Actinomycetes are prolific producers of natural products. The increasing access to genomic data has uncovered a wealth of new natural product biosynthetic gene clusters in actinomycete genomes. However, to what extent these biosynthetic gene clusters are expressed or what signals are required to activate them, remains poorly understood. The marine actinomycete *Salinispora* produces a myriad of natural products. Recent findings suggest a species-specific distribution in *Salinispora*'s biosynthetic gene clusters, albeit most of them uncharacterized. Here I performed comparative genomics paired with reverse transcriptase-PCR in new species from the *Salinispora* genus to inform about the functionality of species-specific biosynthetic gene clusters. Transcription experiments revealed activity in only two species-specific biosynthetic gene cluster under the culture conditions used for these experiments. Additionally, I confirmed species-specific expression patterns in a *S. oceanensis* biosynthetic gene cluster. Comparative metabolomics revealed species-specific patterns in *S. pacifica* and *S. oceanensis* metabolite production and several targets to further investigate as products of the species-specific biosynthetic gene clusters.

3.2 Introduction

The rise of rapid and inexpensive DNA sequencing technology has revealed a wealth of new natural product biosynthetic gene clusters in actinomycete genomes—sometimes tenfold higher than the number of natural products previously discovered in each organism (Rutledge &

Challis, 2015). With the rise of genome sequencing and data availability, many omic techniques have been developed to try and mine this newly discovered potential. While the discovery of novel natural products has increased with these newly developed techniques (Pye *et al*., 2017), our ability to identify BGCs from sequence data continues to far exceed our ability to identify the metabolites they encode (Machado *et al,* 2017). Further there remain many questions regarding specialized metabolite production.

The biosynthetic potential uncovered in genome sequences raised the question of why bacteria produce far fewer specialized metabolites under laboratory conditions than their genomes suggest. One of the most compelling reasons for this scarce production is the lack of either direct or indirect environmental signals controlling the expression of BGCs (at the levels of transcription or translation); resulting in poorly or no BGC expression under laboratory conditions (van Wezel & McDowall 2011). However, it remains unknown to what extent BGCs are expressed or what signals are required to activate them. This becomes even more challenging when the products of the BGCs are unknown. Global expression analyses have rarely been applied to distinguish between silent and expressed BGCs, and in the cases where this has been addressed, microarray or cDNA methods have been used. More commonly, expression analyses have been used to search for the products of BGCs in genetic manipulation experiments in the heterologous host (Machado *et al.,* 2017).

Control of transcription plays a critical role in the multistep process that regulates gene expression. Gene transcript levels inside a cell change in response to a wide variety of signals that occur during cell development, differentiation, and normal physiological function as well as in response to disease (O'Driscoll *et al.,* 2004). Reverse transcription polymerase chain reaction (rt-PCR) is used to amplify RNA targets. The RNA template is converted into complementary DNA

(cDNA) by reverse transcriptase. The cDNA then serves as template for PCR; rt-PCR has numerous applications in research, and it plays a major role in cloning and sequencing (Jalali *et al.*, 2017).

While metabolomic analyses have historically been performed independently of any knowledge of genomic information, integrating metabolomic and genomic data has increased in order to generate bioinformatic links between natural products and their associated BGCs. Comparing BGC distributions with metabolomic profiles has been termed "pattern-based genome mining" (Duncan *et al*., 2015) and used to compare large numbers of closely related bacteria (Machado *et al*., 2017). To inform the expression of recently identified species-specific BGCs, I performed a paired omics (metabolomics, genomics and transcriptomics).

3.3 Methods

3.3.1 Cultivation and sample preparation

All strains were grown shaking at 200 rpm at 28^{o}C in triplicate in 50 ml of A1 media for 9 days. At 8 days, activated and sterile amberlite XAD-7 resin was added to each culture in a proportion of 20 g of resin per liter of medium, and was left for passive adsorption during 24 h. The A1 media formulation is as follows: soluble starch 10 g, yeast extract 4 g, peptone 2 g per liter of 0.2 μm filtered of seawater and supplemented with 5 ml/L of KBr (20g/L), $Fe_2(SO_4)_3$ (8 g/L) after autoclaving. After 9 days, aliquots for total RNA and chemical extractions were collected. For total RNA, three 1.5 ml aliquots were centrifuged at 10,000 rpm for 3 minutes, the supernatant was discarded and 200 μl of RNAlater ® was added to the cell pellet and stored at -80^{o}C until extracted. For chemical extractions, the culture with resin was centrifuged at 10,000 rpm for 5

minutes, the supernatant was discarded, and the pellets with cells and resin stored at 80ºC until extracted.

3.3.2 RNA extractions and cDNA synthesis and reverse transcriptase-PCR

The cell pellets were thawed, and a total RNA extraction was performed using the Qiagen Rneasy mini kit ® protocol and reagents. A second DNase I digestion at 2.5% was needed to fully eliminate all the DNA contamination. For expression experiments, I first performed a complementary DNA synthesis (cDNA) using the SuperScript ® III first strand synthesis system following the Oligo$(dT)_{20}$ protocol as suggested for high GC content samples, with an incubation at 65ºC for a total of 85 min. To ensure no DNA contamination, a 16S rRNA PCR was performed using the RNA samples before the cDNA synthesis. A separate PCR was performed and following the cDNA synthesis to confirm the success of the reaction. Subsequently, a classic PCR was performed using Phusion Green Polymerase ®. The primer sequences used for the PCR reactions are listed in table 3.1. The total PCR reaction was 20 µl of with 2-4 µl of cDNA template. Genomic DNA was used in every reaction as a positive control, and one negative control with all the reagents used during the total RNA extraction, cDNA synthesis, and PCR reaction. Subsequently, an electrophoresis gel of 1% agarose in TAE buffer was run for 30-45 min at 40 volts. When the amplification was successful, samples were enzymatically cleaned using the exoSAP-it™ (exonuclease and shrimp alkaline) thermo fisher ® and sent for Sanger sequencing.

Table 3.1 Primer sequences for the species-specific biosynthetic genes tested in rt-PCR

Species	BGC	BGC Type	Gene	Product length (bps)	primer name	sequence 5'-3'
S. oceanensis	PKS25	TIPKS/NRPS	NRPS	908	3019F	CTGATCTGGGAGTACGCCAC
					30926R	AACATGCTGGTTACGCCGTA
			PKS	1405	7130F	CGTCTCCTTCGACATCTCGG
					8534R	GGTCGGACCGTACATGTTGT
	NRPS27	NRPS	NRPS	599	2648F	TCCAGATCAAGTTCCACGGC
					3246R	TCATCCAATCGTGGTCGTGG
S. pacifica	PKS19	trans-A NRPS/TIPKS	PKS	906	1524F	ACCCGTGATGTTCCTGTTCC
					2429 R	GTGACCAGGGCTTTCTCCTC
			NRPS	1083	6684 F	GCATCATCGGGATCCACCAT
					7766 R	GAAGACCTCCGCTATCCAGC
S. mooreana	PKS30	TIPKS	PKS	1585	PKS30T1F	GTCGAATACGCCCTCGCCCG
					PKS30T1R	CCGATACGCGTTCGCCCCTC
S. vitiensis	PKS29	TIIPKS/LanC	TIIPKS	655	29894F	TTGGACTACATGCTGGCTGG
					30548R	CCCTCCCGAAGATCACGATG
			LanC-like	807	45135F	TTGCCGTTCAAGTTTCTGCG
					45941R	CTCGTGAATCCGCTCGATGA
S. fenicalii	PKS32	Iterative TIPKS	PKS	1107	4123F	GCCTTCGGTGGAGTTGATCA
					5229R	TTCAAGGGTGGCGGATACAC
	PKS33	TIPKS/NRPS	NRPS	951	2518F	GTGGTACTGGAGCTGCACAT
					3468R	GAACCAGGACAGCTCGTTGA
			PKS	505	661F	GGAGGAGAAGGTGACACAGC
					1165R	CATCTGGAGACCTGTCGTCG

Table 3.1 Primer sequences for the species-specific biosynthetic genes tested in rt-PCR, continued.

Species	BGC	BGC Type	Gene	Product length (bps)	primer name	sequence 5'-3'
S. fenicalii	PKS34	TIIPKS/Butyrolactone	A-factor like gene	556	55F	GATAAGACACGCGGGTCGAT
					719R	AAGAACACCGGATGCTTGGT
			TIIPKS	720	442F	GGTCATATGGAACGCGTTGC
					1161R	ATCACGTTCTTCGACCCGTC
	PKS35	transAT-PKS	PKS	1479	669F	CCTCTACATCGAGTCGGTGC
					2147R	GTCATGACATCGAGTTGCGC
			Trans-AT	1086	9094F	GCCACCGAAAAGATCATCGC
					10179R	GTAACGGTCGTTGGTCACCT

3.3.3 Chemical extraction and data acquisition

The cells with resin pellets were thawed, if frozen, or immediately extracted with 60 ml acetone and dried at vacuo, subsequently extracted again with 2 x 20 ml of ethyl acetate to remove salts and other water-soluble impurities, dried in vacuo via rotary evaporation and stored at -20°C until analyzed.

Extracts were dissolved at 0.5 mg/ml in HPLC grade methanol and injected onto a Phenomenex Kinetex C18 reversed-phase high-performance LC (HPLC) column (150 mm x 4.6 mm, pore size: 5 µm, and particle size: 100 Å). Samples were analyzed using an Agilent 6530 Accurate-Mass Q-TOF spectrometer coupled to an Agilent 1260 LC system under the following LC conditions with 0.1% trifluoroacetic acid (TFA): 1–2 min (10% acetonitrile [MeCN] in H2O), 2–12 min (10–100% MeCN), 12–15 min (100% MeCN). The divert valve was set to waste for the first 2 min. Q-TOF MS settings during the LC gradient were as follows: positive ion mode mass range 300–2,500 *m/z*, MS scan rate 1/s, MS/MS scan rate 5/s, fixed collision energy 20 eV; source gas temperature 300 °C, gas flow 11 l/min, nebulizer 45 psig. The MS was auto-tuned using Agilent tuning solution in positive mode before each measurement. LC (DAD) data were transformed to mzXML files and mzmine software was used for data analysis and visualization.

3.3.4 Data analysis

A molecular network was created using the online workflow (https://ccms-ucsd.github.io/GNPSDocumentation/) on the GNPS website (http://gnps.ucsd.edu). The data was filtered by removing all MS/MS fragment ions within +/- 17 Da of the precursor *m/z*. MS/MS spectra were window filtered by choosing only the top 6 fragment ions in the +/- 50Da window throughout the spectrum. The precursor ion mass tolerance was set to 0.5 Da and a MS/MS fragment ion tolerance of 0.05 Da. A network was then created where edges were filtered to have

a cosine score above 0.6 and more than 4 matched peaks. Further, edges between two nodes were kept in the network if and only if each of the nodes appeared in each other's respective top 10 most similar nodes. Finally, the maximum size of a molecular family was set to 100, and the lowest scoring edges were removed from molecular families until the molecular family size was below this threshold. The spectra in the network were then searched against GNPS' spectral libraries. The library spectra were filtered in the same manner as the input data. All matches kept between network spectra and library spectra were required to have a cosine score above 0.7 and at least 4 matched peaks.

Further to enhance chemical structural information within the molecular network, information from in silico structure annotations from GNPS Library Search were incorporated into the network using the GNPS MolNetEnhancer workflow (Ernst *et al.,* 2019) (https://ccms-ucsd.github.io/GNPSDocumentation/molnetenhancer/) on the GNPS website (http://gnps.ucsd.edu). Chemical class annotations were performed using the ClassyFire chemical ontology.

To correlate BGCs to metabolomic data, I used NPLInker, a python-based pattern matching software (Hjörleifsson Eldjárn *et al.,* 2021). Computational prioritization of links between BGCs and candidate products were made using the rosetta score. The identification of BGCs was carried out using antiSMASH 6.0 (Blin *et al.,* 2021). The detected BGCs were grouped into Gene Cluster Families (GCF) using BiG-SCAPE 1.0 (Navarro-Muñoz *et al*. 2019), with the underlying assumption that BGCs that belong to the same GCF produce similar metabolites. BiG-SCAPE was run using Longest Common Subcluster alignment mode, and cluster analysis carried out at the default cutoff of 0.3.

3.4 Results and discussion

In Chapter 2, I identified 15 species-specific BGCs in the newly described *Salinispora* species (Román-Ponce *et al.,* 2020). From these 15 BGCS, I investigated the expression of 15 biosynthetic genes in nine BGCs from five different *Salinispora* species. I excluded *S. goodfellowii* and *S. cortesiana* BGCs because they both have only one genome strain. Out of the 15 genes, I detected expression in three genes from two different BGCs (table 3.2): a trans-A NRPS/TIPKS hybrid (PKS19) from *S. pacifica* and a TIPKS/NRPS hybrid from *S. oceanensis* (PKS25). This was a qualitative experiment, so although total RNA starting concentration was adjusted to 2.5 µg for cDNA synthesis, I did not quantify the yields of cDNA before running the final PCR nor the efficiency of the cDNA synthesis reactions. Thus, even when observing bands of different intensity in gels, this may not be correlated with differences in expression among samples. Each positive PCR product was purified enzymatically with exonucleases and shrimp alkaline phosphatase and sent for Sanger sequencing to further corroborate I was amplifying the correct gene.

The small proportion of species-specific genes that were expressed was somewhat disappointing but perhaps not surprising. There is a common discrepancy between the biosynthetic potential of bacterial strains and the actual number of specialized metabolites they produce. When bacteria are grown in pure cultures, many of the activating signals are presumably absent and the expression of specialized metabolite BGCs is thus downregulated (Rutledge & Challis, 2015). I selected A1 medium since is the standard medium used for the cultivation of *Salinispora* strains, and many of the natural products isolated from *Salinispora* were discovered using this medium. Similarly, the extraction method I selected was one of the most common used in our lab for chemical extractions and the use of resin allows the capture of a wide variety of molecules and can increase yields when it is left in the culture for several hours. In an attempt to promote metabolite

production, I decided to use natural seawater (0.2 μm filtered) under the assumption that it contains nutrients and signaling molecules that would favor BGC activation.

There is evidence supporting expression of cryptic BGCs. For example, microarrays have also been used to assess BGC expression. In the case of *S. coelicolor,* an experiment assessing expression with microarrays revealed that eight of 22 BGCs were expressed at higher levels following the transition from exponential to stationary phase (Karoonuthaisiri *et al.,* 2005). Microarrays also revealed that 12 of 18 BGCs were expressed in wild-type *Myxococcus xanthus*, thus suggesting that many orphan BGCs may be transcriptionally active (Bode *et al.,* 2009). However, the environmental conditions or signals that turn on these BGCs remains mostly unknown, and these types of questions are usually answered on a case by case basis when there is motivation to find the products of a promising BGC.

Table 3.2 Expression results for species-specific gene clusters tested via rt-PCR. Expression is indicated by -/+ symbols. – no expression, + expression. *CNS-960: 2/3 of the replicates were positive.

Species	**BGC**	**BGC type**	**Gene**	**Strain**	**Expression +/-**
S. oceanensis	PKS25	TIPKS/NRPS	NRPS	CNT-124	-
				CNT-403	-
				CNY-703	-
				CNS-996	-
			PKS	CNT-124	+
				CNT-403	+
				CNY-703	+
				CNS-996	+
	NRPS27	NRPS	NRPS	CNT-124	-
				CNT-403	-
				CNY-703	-
				CNS-996	-
S. pacifica	PKS19	trans-A NRPS/TIPKS	NRPS	CNS-960	+
				CNR-909	+
				CNY-331	+
				CNY498	-
			TIPKS	CNS-960	+ *
				CNR-909	+
				CNY-331	+
				CNY498	+
S. fenicalii	PKS32	Iterative TIPKS	PKS	CNT-569	-
				CNR-942	-
	PKS33	TIPKS/NRPS	PKS	CNT-569	-
				CNR-942	-
			NRPS	CNT-569	-
				CNR-942	-
	PKS34	TIIPKS/butyrolactone	A-factor-like	CNT-569	-
				CNR-942	-
			TIIPKS	CNT-569	-
				CNR-942	-
S. fenicalii	PKS35	transAT-PKS	transAT	CNT-569	-
				CNR-942	-
			TIPKS	CNT-569	-
				CNR-942	-

Table 3.2 Expression results for species-specific gene clusters tested via rt-PCR. Expression is indicated by -/+ symbols. – no expression, + expression, continued. *CNS-960: 2/3 of the replicates were positive.

Species	**BGC**	**BGC type**	**Gene**	**Strain**	**Expression +/-**
S. mooreana	PKS30	TIPKS	PKS	CNY-646	-
				CNS-237	-
				CNT-150	-
S. vitiensis	PKS29	TIIPKS/LanC	PKS	CNS-055	-
				CNS-801	-
				CNT-148	-
			LanC	CNS-055	-
				CNS-801	-
				CNT-148	-

In *Salinispora* spp., there exist only one previous assessment of gene expression in relation to specialized metabolism. A comparative transcriptomic analysis was done in four *Salinispora* strains from three different species, including *S. tropica*, *S. arenicola*, and the *S. mooreana* strain CNT-150. According to this assessment, the expression baseline between silence and expressed genes was at 27.1 reads per kilobase of transcript per million mapped reads (RPKM) (Amos *et al.,* 2017). From this previous transcriptomic data, I was able to find the transcription levels of the species-specific BGC PKS30 in *S. mooreana*, which ranged from 50 to 37,000 RPKM. It was unfortunate that in my rtPCR experiments I did not detected expression in PKS30. Of course, RNA sequencing and transcription analysis might give a finer resolution than rt-PCR. The TIPKS gene I tested had an expression level in Amos *et al.,* (2017) of 71 RPKM, which is fairly close to the base line. Compared to other genes within the same cluster, like a cytochrome P450 with levels of 1390 RPKM, or other BGCs within the same strain that had expression levels around 10,000, such as the enterocin BGC, it's not surprising that I did not detect expression of PKS30.

In regard to *S. pacifica*, the species-specific BGC is a trans-adenylation NRPS/PKS hybrid (PKS19). Trans-A NRPS biosynthesis was recently discovered and there are only a couple of examples of compounds confirmed (Bernhardt & Berman, 2020) or proposed (Magarvey *et al.,* 2008) to be biosynthesized in this manner. Thus, it is very promising that for both the TIPKS and one of the NRPS genes I detected expression in at least three of the four strains tested. In the case of CNY-498, the expression bands were "less intense" than in the other strains and only one out of the three replicates showed a clear band for the NRPS gene (figure 3.1 A & B). Unfortunately, there was not enough product to obtain sequence data from this reaction. Nonetheless, as mentioned before, band intensity cannot be correlated with levels of expression since I did not standardize the cDNA template concentrations. Therefore, I decided to extract all four strains for comparative metabolomics to try to identify putative products of PKS19. In section 3.4.2, I will discuss the results from these analyses.

The BGC PKS25 is a TIPKS/NRPS hybrid BGC, and the species-specific BGC for *S. oceanensis*. I assessed expression in one NRPS gene and the TIPKS hybrid gene, detecting expression of the TIPKS gene in all four strains tested (figure 3.2A). Differential expression of genes within the same BGC has been reported before for *Salinispora* strains, in the case of PKS4, a BGC encoding for a type II PKS predicted to produce the black spore pigment characteristic of *Salinispora* species. In Amos *et al.,* 2017, they observed full expression of the cluster in the two strains for CNT-150 *S. mooreana* (formerly *S. pacifica* in the cited study) and *S. tropica,* however, in the two strains of *S. arenicola* only half of the BGC was expressed, and when comparing the gene cluster environment they found the S. *arenicola* version missing *luxR* homologs genes present in *the S. tropica* and *S. mooreana* versions, they concluded that half of the BGC is might be regulated by a different mechanism, however they did not mention any observations of differential

sporulation among the three species.. PKS25 is highly conserved across all *S. oceanensis* strains, and all contain the regulatory protein *LysR* upstream the cluster. Perhaps a similar situation is occurring in PKS25, and the NRPS gene is regulated by different regulatory elements. It remains to be tested whether this is the case, this would be even more informative if we were able to identify the product of PKS25 and compare differential expression versus differential metabolite production.

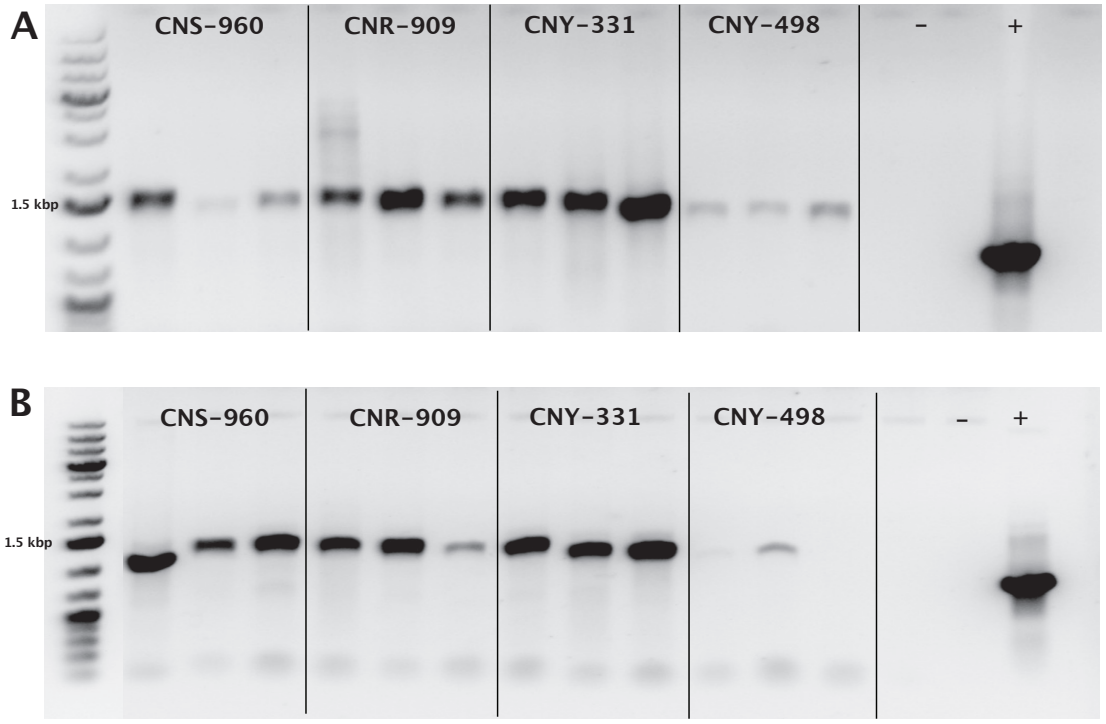

Figure 3.1 rt-PCR results for the four *S. pacifica* strains that contain the PKS19 BGC. **A** TIPKS gene. **B** NRPS gene. The 1500 bp sized band is indicated.

3.4.1 Differential expression of PKS25 between *S. oceanensis* and *S. mooreana*

In chapter 2, when characterizing the species-specific BGC from *S. oceanensis*, I realized that PKS25 was shared with two *S. mooreana* strains (CNS-237 and CNY-646) and that the BGCs in *S. mooreana* are missing a *lysR* transcription factor gene (Maddocks & Oyston, 2008). I hypothesized that due to the missing transcription factor, the *S. mooreana* version of PKS25 might

be inactive. Therefore, I compared PKS25 gene expression between *S. oceanensis* and *S. mooreana*. I used CNT-150 as a negative control since this genome doesn't contain PKS25 (figure 3.2). The rt-PCR showed expression in the gene from *S. oceanensis* strains, but not in the *S. mooreana* strains, supporting the hypothesis of differential expression in different species (fig 3.2). This is not the only case where differential expression is observed to follow a species pattern in *Salinispora*. One example is STPKS1, an orphan BGC that was expressed in *S. mooreana* strain CNT- 150 but was silent in *S. tropica* strain CNB-440. The expressed version of this BGC includes an *araC* activator upstream from the core PKS genes, while in the silent BGC, this activator has been replaced with a family IS285 transposase. Surprisingly, this silent version is largely maintained within a well-supported clade in the *S. tropica* species phylogeny (Amos *et al.,* 2017).

The loss of a key regulatory element suggests that PKS25 in *S. mooreana* might be permanently silenced, or alternatively, that the silent BGC is controlled by different regulatory elements.

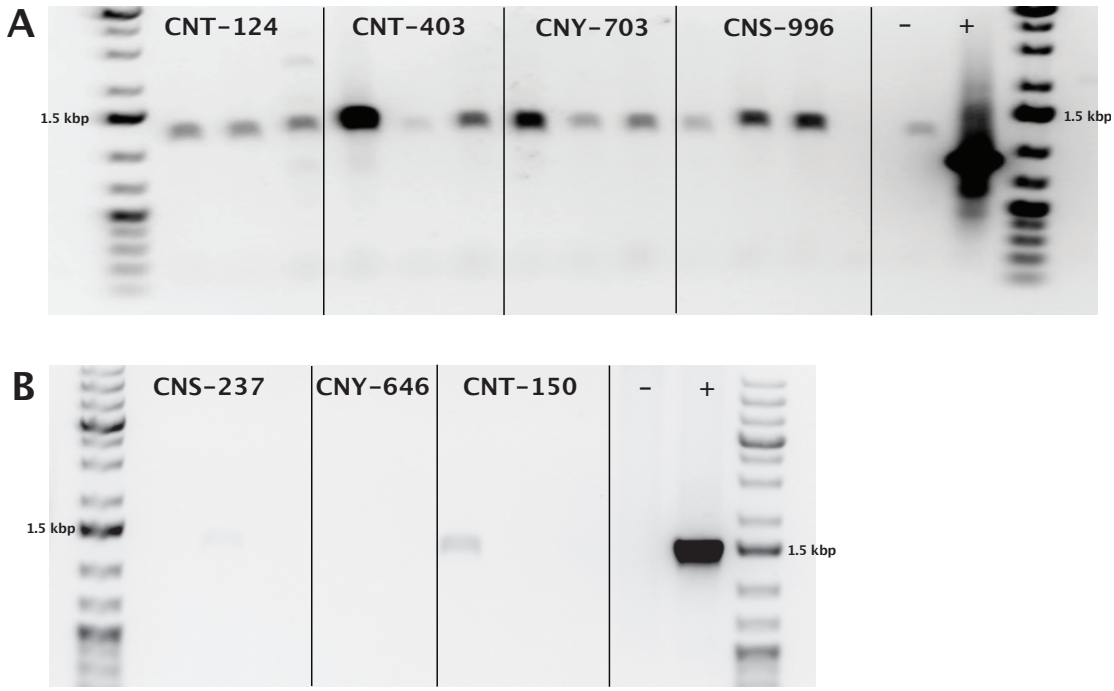

Figure 3.2 rt-PCR results for the TIPKS gene in PKS25 from *S. oceanensis* (**A**) and *S. mooreana* (**B**). The 1500 bp sized band is indicated.

3.4.2 Comparative metabolomics

To find putative products from PKS25 and PKS19, the only two BGCs expressed in my rt-PCR experiments, I did chemical extractions of *S. pacifica* and *S. oceanensis* strains. The mass spec data from all 24 extracts (four strains per species and three biological replicates) was analyzed with GNPS, MolnetEnhancer, and NPLinker. I filtered out all the nodes shared between both species and focused on those that were species-specific. I selected only the ones shared in at least three out of the four strains and in at least two out of the three biological replicates. The resulting molecular network is shown in figure 3.3. A total of 19 molecular families unique to *S. pacifica* and 14 were unique to *S. oceanensis*. From MolnetEnhancer, an *in silico* predictor of functional groups, many different compound classes such as alkaloids, nucleosides, lipids, benzenoids, and polyketides, among others were detected. Both PKS25 and PKS19 are predicted to biosynthesize

NRP/PK hybrids. MolNetEnahncer identified two molecular families related to polyketides, one produced in *S. pacifica* with masses of 405 *m/z*, 419 *m/z*, and 433 *m/z*, and one produced in *S. oceanensis* with mases of 233 *m/z* and 235 *m/z* (both marked with blue squares in figure 3.3). However no amino acid or peptide-related nodes were identified by MolnetEnhancer.

In addition, I analyzed for correlations between mass spectrometry features and biosynthetic gene cluster distribution using NPLinker (Hjörleifsson Eldjárn *et al.,* 2021). This python-based tool integrates mass spectrometry data and BGC data and links them by testing different scoring functions among hypothetical links. In this case, I used the rosetta scoring, which has been successfully used to link natural products and BGCs from polar actinomycetes (Soldatou *et al.,* 2021). In total, three different masses were linked as putative products of the species-specific clusters, two for *S. pacifica* and one for *S. oceanensis* (figure 3.3, green squares). From these pattern-based analysis, only one putative molecular family in *S. pacifica* has a node linked to PKS19 and polyketide-related annotations. In the case of *S. oceanensis*, there was one node with a mass of 300 *m/z* that linked to PKS25. Therefore, I decided to pursue the *S. pacifica*-specific molecular family for structural elucidation, in particular the peaks of 419 *m/z* and 433 *m/z*. I decided to cultivate the strain with the highest peak intensities in the MS chromatogram (CNY-498) despite showing the fainter bands in the electrophoresis gels. Following a similar chemical extraction, I extracted 18 litters of liquid culture and purified the 419 *m/z* and 433 *m/z* compounds via HPLC, and after NMR analysis in collaboration with post-doctoral scholar Gabriel Castro discovered that the compounds are two novel ketomemicin C analogs. The isolation details and further results from the discovery of new ketomemicin analogs will be discussed in chapter 4.

Unfortunately, I was not able to identify the products of the species-specific BGCs I was working with. Although I pursued the masses that seemed the most promising to discover the

product of PKS19, perhaps the metabolites were not abundant enough to be detected or extracted with the methodology I followed, or I simply focused on the wrong targets. This highlights the limitations of pattern-based genome mining. Also, there are other targets in the molecular network that I was not able to explore, perhaps one of those masses is indeed the product of the cluster. This experiment was just a snapshot of what was happening within the specialized metabolism of these strains at the moment I extracted the RNA. Perhaps these genes are expressed under different conditions or at different growth stages. therefore, I cannot rule out completely the functionality of the species-specific BGCs. It was just not possible to link a product to those that were expressed.

Product prediction by genome mining is challenging, although in some cases it is more straight forward than in others. With multinodular NRPS and PKSs, whose building blocks can be predicted in a straighter forward manner, there are many tools for in silico predictors. Further manual inferences can be made from the amino acid sequences of the genes and from the conserved enzymatic active sites. Still, hybrid biosynthetic machinery and understudied biosynthetic classes like trans adenylation domains in NRPS add complexity to product prediction. In addition, specialized metabolites often undergo tailoring reactions resulting in more complex carbon skeletons, making it harder to confidently predict a final product just based on sequence mining. Nonetheless, thanks to the constant study of biosynthetic pathways, our ability to predict structures from BGCs continues to increase.

Coupling Mass spectrometry and genome mining data facilitates the selection of promising chemical features in complex crude extracts or fractions. By coupling comparative genomic and transcriptomic data, it becomes possible to predict which versions of a BGC are silent and perhaps gain more information that will help in predicting the structure of the product. As mentioned before, it would be very interesting to look at differential BGC expression using different culture

media, co-culturing, or elicitors, at various growth stages. Focusing on species-specific gene clusters could help uncover the ecological signals that activate silent gene clusters. Differentiating between expressed and silent BGCs can be particularly useful in selecting strains for product discovery or heterologous expression. For example, I've demonstrated that PKS19 is active in the wild type strains. Since this gene cluster encodes rare biosynthetic machinery, it would be very interesting trying to find the product via heterologous expression. Finding the product of this gene cluster would not only help expand *Salinspora*'s natural product repertoire but also, and more interestingly, it would expand the knowledge the field has about trans-adenylation domains in NRPS biosynthesis.

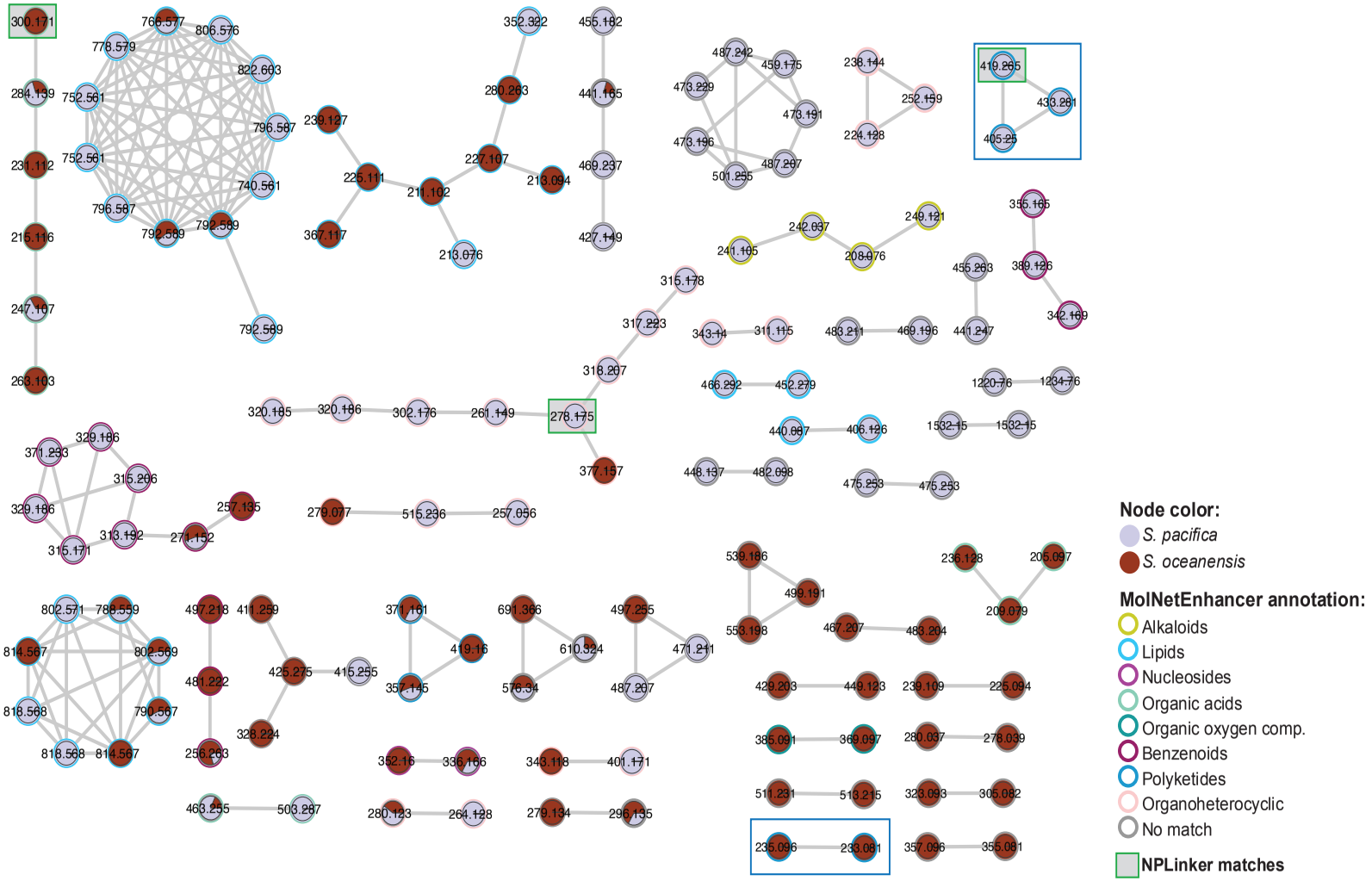

Figure 3.3 Molecular network of only species-specific nodes, from S. oceanensis and S. pacifica. The nodes are color coded by species, and the outer ring represents the MolNetEnhancer annotation. Green squares represent the nodes correlated to the species-specific BGCs PKS25 and PKS19, by NPLinker. The blue square highlights the two molecular families related to polyketides.

3.4.3 Some reflections on rt-PCR product size and troubleshooting.

The rt-PCR cDNA samples always ran slower than the gDNA control in the electrophoresis gel. Manufacturer reports state that usually when a PCR product from cDNA is larger it might be contaminated with gDNA. However, all RNA samples were digested with DNase I, and I performed 16S rDNA PCR reactions with the cleaned RNA to confirm that no gDNA remained. Another hypothesis to explain the larger bands is that I loaded less PCR product from the gDNA controls compared to the cDNA samples. This was done to avoid overexposure of the gDNA controls (as observed in figure 3.2A) since gDNA controls contained higher concentrations than the cDNA and the ladder. Similarly, I added the DNA stain directly to the samples in the well, which might have caused the difference in expected band sizes. However, I tried adding the same volumes for ladder, gDNA, control and rt-PCR product and the difference in size would persist.

Another reason for the larger products from the cDNA samples could be interactions between primers and the remanent of dNTPs and the reverse transcriptase from the cDNA synthesis. If the efficiency of cDNA synthesis was low, an excess of reagents could transfer over to the PCR step, which would result in an excess of dNTPs and favor activity of remanent reverse transcriptase. Interactions between primers and reverse transcriptase in rt-PCR experiments have been documented (Chumakov, 1994), resulting in primer dimerization, which I often observed in my samples. Reverse transcriptase readily uses DNA as a template and is known to use very short oligonucleotide primers to initiate DNA strand synthesis. Thus, it is possible that it can add nucleotides to the 3' end of primers in the PCR mixture (Chumakov, 1994). Similarly, it may be possible that complexes between primer dimers and cDNA template occurred resulting in a larger band size. A proposed solution to deactivate reverse transcriptase is to heat freshly synthesized cDNA samples at 94^{o}C (Chumakov, 1994). However, I observed mixed results when trying this

method before running the PCR reaction, with primer dimers eliminated in some cases but not the difference in product sizes. Lastly, in the cDNA synthesis, there is a final incubation with Rnase H to destroy the RNA template and cleave specific RNA sequences from short cDNA segments (Gilboa *et al.,* 1979). If the final incubation with Rnase H did not efficiently cleave the cDNA from the RNA template, this could have resulted on the different sized complexes forming in the subsequent PCR reaction. However, this type of uncleaved RNA and cDNA templates result in inhibition of PCR (Schultz & Champoux 2008).

Nonetheless after Sanger sequencing, I confirmed that the amplified products corresponded to the target gene and that the sequence were the same as the gDNA controls. Therefore, if any complexes were formed during the PCR, they would not be picked up during sequencing, or perhaps the cause for slower migration would be related most likely to something happening during the electrophoresis step and perhaps not during the PCR.

3.5 Conclusions

Around 80% of the species-specific biosynthetic gene clusters identified in *Salinispora* spp. are silent when growing in A1 media (ref), which is the standard media used to grow *Salinispora*. It remains to be determined if these biosynthetic gene clusters are functional and under what conditions they are activated.

There exists differential expression between *S. oceanensis* and *S. moorena* in genes from the PKS25 gene cluster. The *S. mooreana* strains are missing a transcription regulator from the *lysR* family suggesting that this version of PKS25 is silent or regulated by different transcriptional elements than in *S. oceanensis*. Unfortunately, I was not able to identify the product of PKS25, although comparative genomics revealed 14 target molecular families that could be further investigating with this goal in mind.

The species-specific biosynthetic gene cluster of *S. pacifica* (PKS19) is an unusual trans-A NRPS biosynthetic gene cluster and apparently active when grown in laboratory conditions. This makes PKS19 an interesting target for heterologous expression. Finding the product of this BGC would expand the information available about natural products biosynthesized via rare trans-adenylation NRPSs.

As a next step, it would be interesting to compare expression under different media and using elicitors. Furthermore, global transcriptomics would give more insight into which biosynthetic genes within a cluster are expressed and at what levels under various conditions.

3.6 Acknowledgements

I would like to thank members of the Jensen Laboratory including Kaitlin Creamer and Johanna Gutleben for their insights into rt-PCR methodology and their help with electrophoresis troubleshooting.

This research was supported by the NIH grant to R01GM085770 PRJ and BSM and the CONACYT and UC MEXUS doctoral fellowship.

3.7 References

Amos, G.C., Awakawa, T., Tuttle, R.N., Letzel, A., Kim, M.C., Kudo, Y., Fenical, W., S. Moore, B., & Jensen, P.R. (2017). Comparative transcriptomics as a guide to natural product discovery and biosynthetic gene cluster functionality. *Proceedings of the National Academy of Sciences, 114*, E11121 - E11130.

Bernhardt, M., Berman, S.F., Zechel, D.L., & Bechthold, A. (2020). Role of Two Exceptional trans Adenylation Domains and MbtH-like Proteins in the Biosynthesis of the Nonribosomal Peptide WS9324A from Streptomyces calvus ATCC 13382. *Chembiochem, 21*, 2659 - 2666.

Blin, K., Shaw, S., Kloosterman, A.M., Charlop-Powers, Z., van Wezel, G.P., Medema, M.H., & Weber, T. (2021). antiSMASH 6.0: improving cluster detection and comparison capabilities. *Nucleic Acids Research, 49*, W29 - W35.

Bode, H., Ring, M.W., Schwär, G., Altmeyer, M., Kegler, C., Jose, I.R., Singer, M., & Müller, R. (2009). Identification of Additional Players in the Alternative Biosynthesis Pathway to Isovaleryl-CoA in the Myxobacterium Myxococcus xanthus. *ChemBioChem, 10*.

Chumakov, K. (1994). Reverse transcriptase can inhibit PCR and stimulate primer-dimer formation. *PCR methods and applications, 4 1*, 62-4.

Duncan, K.R., Crüsemann, M., Lechner, A., Sarkar, A., Li, J., Ziemert, N., Wang, M., Bandeira, N., Moore, B.S., Dorrestein, P.C., & Jensen, P.R. (2015). Molecular networking and pattern-based genome mining improves discovery of biosynthetic gene clusters and their products from Salinispora species. *Chemistry & biology, 22 4*, 460-471.

Ernst, M., Kang, K.B., Caraballo-Rodríguez, A.M., Nothias, L., Wandy, J., Chen, C., Wang, M., Rogers, S., Medema, M.H., Dorrestein, P.C., & van der Hooft, J.J. (2019). MolNetEnhancer: Enhanced Molecular Networks by Integrating Metabolome Mining and Annotation Tools. *Metabolites, 9*.

Feunang, Yannick Djoumbou, *et al*. "ClassyFire: Automated chemical classification with a comprehensive, computable taxonomy." J. Cheminform. 8 (2016): 61. PMID: 27867422

Gilboa, E., Mitra, S.W., Goff, S.P., & Baltimore, D. (1979). A detailed model of reverse transcription and tests of crucial aspects. *Cell, 18*, 93-100.

Hjörleifsson Eldjárn, G., Ramsay, A., van der Hooft, J.J., Duncan, K.R., Soldatou, S., Rousu, J., Daly, R., Wandy, J., & Rogers, S. (2021). Ranking microbial metabolomic and genomic links in the NPLinker framework using complementary scoring functions. *PLoS Computational Biology, 17*.

Jalali, M., Zaborowska, J., & Jalali, M. (2017). The Polymerase Chain Reaction: PCR, qPCR, and RT-PCR in: Jalali M., Saldanha F. Y. L., Jalali M. (editors) Basic Science Methods for Clinical Researchers. Elsevier.

Karoonuthaisiri, N., Weaver, D., Huang, J., Cohen, S.N., & Kao, C.M. (2005). Regional organization of gene expression in Streptomyces coelicolor. *Gene, 353 1*, 53-66.

Machado, H., Tuttle, R.N., & Jensen, P.R. (2017). Omics-based natural product discovery and the lexicon of genome mining. *Current opinion in microbiology, 39*, 136-142.

Maddocks, S.E., & Oyston, P.C. (2008). Structure and function of the LysR-type transcriptional regulator (LTTR) family proteins. *Microbiology, 154 Pt 12*, 3609-23.

Magarvey, N.A., Fortin, P.D., Thomas, P.M., Kelleher, N.L., & Walsh, C.T. (2008). Gatekeeping versus promiscuity in the early stages of the andrimid biosynthetic assembly line. *ACS chemical biology, 3 9*, 542-54.

Navarro-Muñoz, J.C., Sélem-Mojica, N., Mullowney, M.W., Kautsar, S., Tryon, J.H., Parkinson, E.I., de los Santos, E.L., Yeong, M., Cruz-Morales, P., Abubucker, S., Roeters, A., Lokhorst, W., Fernàndez-Guerra, A., Cappelini, L.T., Goering, A.W., Thomson, R.J.,

Metcalf, W.W., Kelleher, N.L., Barona-Gómez, F., & Medema, M.H. (2019). A computational framework to explore large-scale biosynthetic diversity. *Nature chemical biology, 16*, 60 – 68.

O'Driscoll, L., Daly, C., Saleh, M., & Clynes, M. (2004). The use of reverse transcriptase-polymerase chain reaction (RT-PCR) to investigate specific gene expression in multidrug-resistant cells. *Cytotechnology, 12*, 289-314.

Pye, C.R., Bertin, M.J., Lokey, R.S., Gerwick, W.H., & Linington, R.G. (2017). Retrospective analysis of natural products provides insights for future discovery trends. *Proceedings of the National Academy of Sciences, 114*, 5601 - 5606.

Román-Ponce, B., Millán-Aguiñaga, N., Guillen-Matus, D., Chase, A. B., Ginigini, J., Soapi, K., Feussner, K. D., Jensen, P. R., & 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., *Salinispora oceanensis* sp. nov. and *Salinispora vitiensis* sp. nov., and emended description of the genus *Salinispora*. *International journal of systematic and evolutionary microbiology*, *70*(8), 4668–4682.

Rutledge, P.J., & Challis, G.L. (2015). Discovery of microbial natural products by activation of silent biosynthetic gene clusters. *Nature Reviews Microbiology, 13*, 509-523.

Schultz, S.J., & Champoux, J.J. (2008). RNase H activity: structure, specificity, and function in reverse transcription. *Virus research, 134 1-2*, 86-103.

Soldatou, S., Eldjarn, G.H., Ramsay, A., van der Hooft, J.J., Hughes, A.H., Rogers, S., & Duncan, K.R. (2021). Comparative Metabologenomics Analysis of Polar Actinomycetes. *Marine Drugs, 19*.

van Wezel, G.P., & McDowall, K.J. (2011). The regulation of the secondary metabolism of Streptomyces: new links and experimental advances. *Natural product reports, 28 7*, 1311-33.

Wang, M., Carver, J., Phelan, V.V., Sanchez, L.M., Garg, N., Peng, Y., Nguyen, D.D., Watrous, J.D., Kapono, C.A., Luzzatto-Knaan, T., Porto, C., Bouslimani, A., Melnik, A.V., Meehan, M.J., Liu, W., Crüsemann, M., Boudreau, P.D., Esquenazi, E., Sandoval-Calderón, M., Kersten, R.D., Pace, L.A., Quinn, R.A., Duncan, K.R., Hsu, C., Floros, D.J., Gavilán, R.G., Kleigrewe, K., Northen, T.R., Dutton, R.J., Parrot, D., Carlson, E.E., Aigle, B., Michelsen, C.F., Jelsbak, L., Sohlenkamp, C., Pevzner, P.A., Edlund, A., McLean, J.S., Piel, J., Murphy, B.T., Gerwick, L., Liaw, C., Yang, Y., Humpf, H., Maansson, M., Keyzers, R.A., Sims, A.C., Johnson, A.R., Sidebottom, A.M., Sedio, B.E., Klitgaard, A.K., Larson, C.B., P, C.A., Torres-Mendoza, D., Gonzalez, D.J., Silva, D.B., Marques, L.M., Demarque, D.P., Pociūtė, E., O'Neill, E.C., Briand, E., Helfrich, E.J., 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.S., Silva, R.R., Mascuch, S.J., Tomasi, S., Jenkins, S., Macherla, V.R., Hoffman, T., Agarwal, V., Williams, P.G., Dai, J., Neupane, R.P., Gurr,

J.R., Rodriguez, A.M., Lamsa, A., Zhang, C., Dorrestein, K., Duggan, B.M., Almaliti, J., Allard, P., Phapale, P., Nothias, L., Alexandrov, T., Litaudon, M., Wolfender, J., Kyle, J.E., Metz, T.O., Peryea, T., Nguyen, D., Vanleer, D., Shinn, P., Jadhav, A., Müller, R., Waters, K.M., Shi, W., Liu, X., Zhang, L., Knight, R., Jensen, P.R., Palsson, B.O., Pogliano, K., Linington, R.G., Gutiérrez, M., Lopes, N.P., Gerwick, W.H., Moore, B.S., Dorrestein, P.C., & Bandeira, N. (2016). Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nature Biotechnology, 34*, 828-837.

Wang, Mingxun, *et al*. "Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking." Nature Biotechnology 34.8 (2016): 828-837. PMID: 27504778, https://www.nature.com/articles/nbt.3597

Chapter 4

Discovery, structural elucidation, and biosynthetic gene cluster distribution of ketomemicin analogs in *Salinispora*

4.1 Abstract

Pseudopeptides are rare occurring natural products with many diverse biological activities, including the treatment for anticancer, for hypertension, thrombosis, diabetes, as well as viral and bacterial infections. Ketomemicins are carboxymethelene containing pseudopeptides, and potential protease inhibitors previously identified through heterologous expression. Here, we isolated two new ketomemicin analogs (**1** - **2**) from the culture extracts of *S. pacifica* CNY-498. Their structures were characterized using NMR and mass spectrometry. We further investigated ketomemicin production in *Salinispora* via molecular networking and public *Salinispora* datasets in the GNPS platform, finding new ketomemicin analogs characterized by MS and broadening the distribution of ketomemicin production in *Salinispora* species. Subsequently, we compared the distribution and phylogenetic patterns of the ketomemicin biosynthetic gene cluster (*ktm*) within and outside *Salinispora*. The cluster is highly conserved within the genus, with some differences in *S. arenicola* compared to the other species, and surprisingly sparse in other bacterial taxa. This study confirms cultures of *Salinispora* spp. as sources of ketomemicins and demonstrates the ubiquity of ketomemicin production in *Salinispora.*

4.2 Introduction

Peptides are indispensable in all levels of life and fulfill many different biological functions. Furthermore, peptides may have different biological activities, as antibacterial or cytotoxic agents, and nerve toxins (Patch *et al*., 2004). In contrast, pseudopeptides, which is a term used to describe peptides with modified backbones, are a rare occurrence in nature (Ogasawara 2019). In drug design, replacing the peptide backbone with nonhydrolyzable isosteres, such as esters, alkenes, and carbomethylene is a strategy utilized to develop pseudopeptides that can act as receptor ligands or enzyme inhibitors (Kawata *et al.,* 2017), such as proteases inhibitors. Proteases, also known as peptidases, are ubiquitous in all living organisms and are crucial in many biological and physiological processes. The clinical applications of protease inhibitors are diverse, including for the treatment of cancer, hypertension, thrombosis, diabetes, as well as viral and bacterial infections (Kaysser 2019).

Peptides are indispensable in all levels of life and fulfill many different biological functions. Furthermore, peptides may have different biological activities as antibacterial or cytotoxic agents and nerve toxins (Patch et al., 2004). In contrast, pseudopeptides, which is a term used to describe peptides with modified backbones, are a rare occurrence in nature (Ogasawara 2019). In drug design, replacing the peptide backbone with nonhydrolyzable isosteres, such as esters, alkenes, and carbomethylenes is a strategy utilized to develop pseudopeptides that can act as receptor ligands or enzyme inhibitors (Kawata *et al.,* 2017) such as proteases inhibitors. Proteases, also known as peptidases, are ubiquitous in all living organisms and are crucial in many biological and physiological processes. The clinical applications of protease inhibitors are diverse and include the treatment of cancer, hypertension, thrombosis, diabetes, as well as viral and bacterial infections (Kaysser 2019).

Ketomemicin A, ketomemicin B1-6, and ketomemicin C are tri-peptide-like natural products previously obtained from heterologous expression of homologous biosynthetic gene clusters from *Micromonospora* sp., *Streptomyces mobaraensis*, and *Salinispora tropica* CNB-440, respectively (Ogasawara *et al.,* 2016). Arphamenines A-B are structurally similar di-peptide-like metabolites previously isolated from the Gram-negative bacterium *Chromobacterium violaceum* that inhibit mammalian endopeptidase B (Umezawa *et al.,* 1983). Both ketomemicins and arphamenines share a carboxymethylene C-CO bond at the C-terminal region of the molecules that replaces the peptidic N-CO bond of a standard peptide. This substitution is suggested to be essential for the protease inhibitory activity of arphamenines. Conceivably, arphamenines bind to the catalytic site of endopeptidase B by mimicking the enzyme's peptidic substrate and, due to the absence of a cleavable peptide bond, remain bound to the enzyme, thereby inhibiting it from normal function. Although ketomemicins and arphamenines are the only naturally occurring pseudopeptides containing the carbonylmethylene moiety, synthetic peptides with a similar structure have been developed as protease inhibitors (Zalman *et al.*, 2000; Kawata *et al*., 2017). To our knowledge, there are no previous reports of the protease inhibitory activities of ketomemicins.

Ketomemicin biosynthesis is performed by a cluster of six genes (fig. 4.2), including an aldolase (*ktmA*) that catalyzes an aldol reaction simultaneously with the decarboxylation of malonyl-CoA resulting in benzylanmalyl-CoA. A PLP-dependent amino acid C-acyltransferase (*ktmB*) then catalyzes a Claisen-type condensation between phenylalanine and benzylfumaryl-CoA as in the biosynthesis of ketomemicin B. This is followed by a dehydratase (*ktmC*) and a (pseudo)peptide ligase (*ktmD*), which is responsible for the amide bond formation between the amidino amino acid and the pseudodipeptide in the final step of ketomemicin biosynthesis. And

lastly, an amidinotransferase (*ktmE*), that catalyzes an amidinotransfer reaction and a dehydrogenase (*ktmF*) that reduces the hydroxyl group resulting from the condensation of *ktmB* (kawata *et al.,* 2017; Ogasawara 2019).

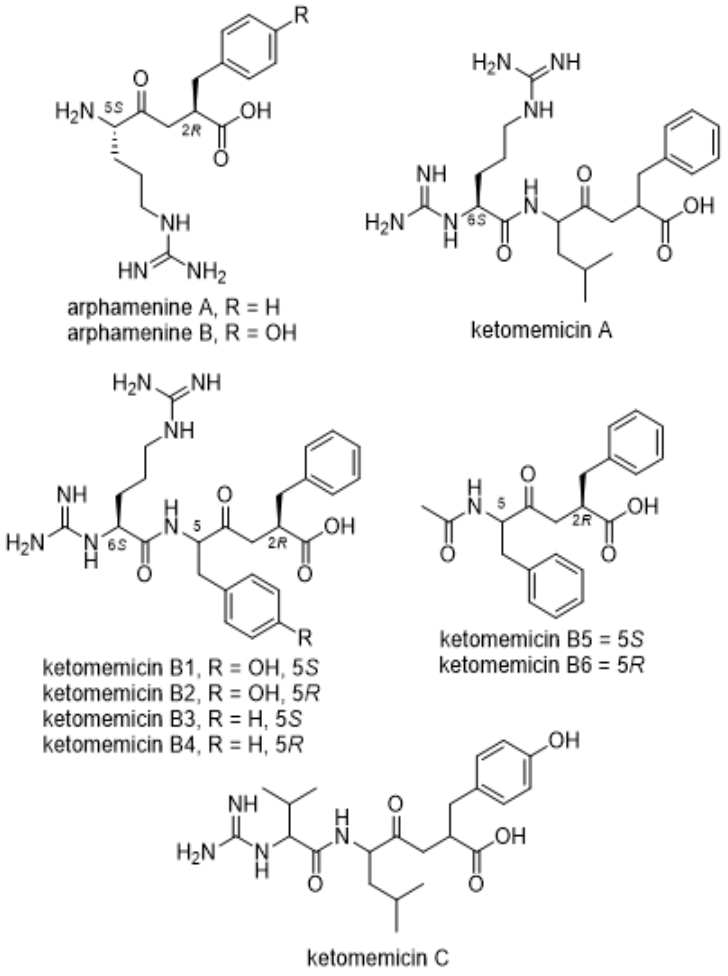

Figure 4.1 Known ketomemicins and arphamenines. ketomemicin A, from a BGC of *Micromonospora sp*., ketomemicin B's from *S. mobaramensis*, and ketomemicin C from *S. tropica.*

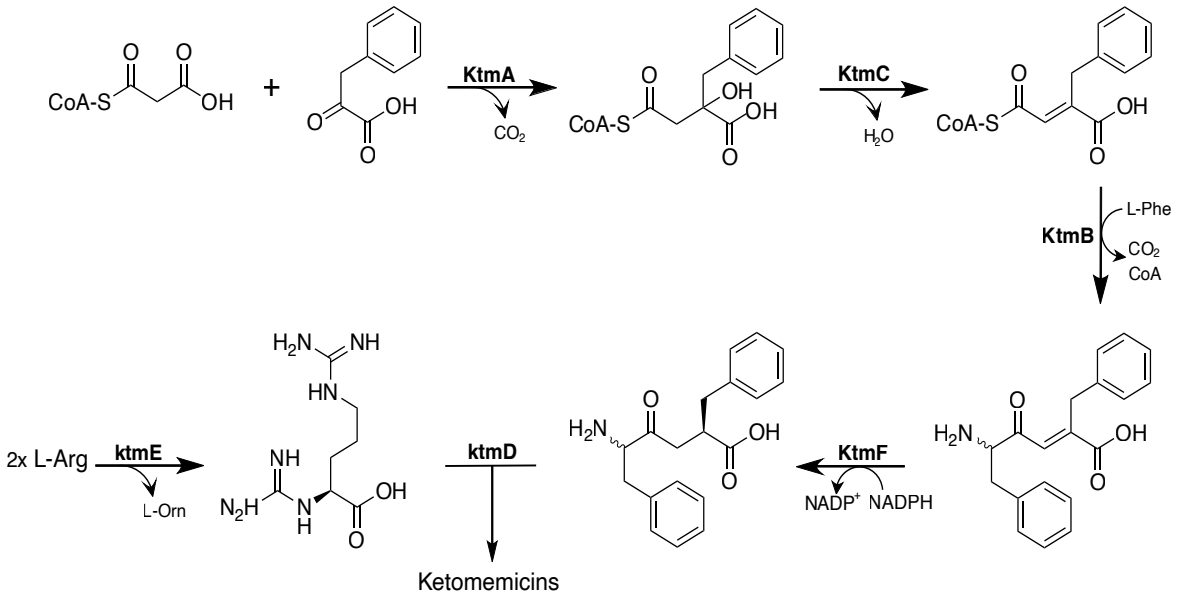

Figure 4.2 Biosynthesis of ketomemicins adapted from Kawata *et al*., 2017 and Ogasawara 2019

In this work, in collaboration with postdoctoral scholar Gabriel Castro, we isolated previously unidentified natural products from the culture extracts of *Salinispora pacifica* CNY-498. Posdoctoral scholar Gabriel Castro assigned their structures, including absolute stereochemistry, revealing these as new ketomemicin derivatives. Additionally, I studied the production of ketomemicin analogs across extracts of *Salinispora* spp. and related this information with the diversity and distribution of the ketomemicin biosynthetic gene cluster (*ktm*) in *Salinispora* genomes.

4.3 Methods

4.3.1 Cultivation and chemical extraction

A frozen stock of *Salinispora pacifica* CNY-498 was inoculated into 50 mL of medium A1 [1% potato starch, 0.4% yeast extract and 0.2% peptone in 22% InstantOcean®]. The seed culture was shaken at 200 rpm and 28℃ for seven days then used to inoculate 1 L of medium A1 in a 2.8 L Fernbach flask. This culture was similarly shaken at 200 rpm and 28℃ for eight days after which 20 mL were inoculated into each of 20 x 2.8 L Fernbach flasks containing 1 L of medium A1FB [A1 supplemented with 0.01% potassium bromide and 0.03% iron (III) sulfate ($5{\cdot}H_2O$)]. After eight days of shaking at 200 rpm and 28℃, 25 g of sterile XAD-7 adsorbent resin was added to each flask. After two additional days of cultivation, the cultures were filtered through cheesecloth, the oversize particles composed of cells and resin were soaked in acetone (3 L) for 2 h with agitation. The acetone extract was filtered through a cotton plug and concentrated via rotatory evaporation. The resulting solution was partitioned in a separatory funnel between Ethyl Acetate and H_2O (1:1 mixture, 1 L total, 2X). The organic phase was collected, dried over anhydrous sodium sulfate, and concentrated via rotatory evaporation to yield a red crude extract (500 mg).

4.3.2 Isolation of ketomemicins

The organic extract was fractionated over C18 column flash chromatography (5g) using a six-step elution gradient from 100% H_2O (0.1% formic acid) to 100% MeCN (0.1% formic acid) to yield six fractions of differing polarity. Fraction 4 (18.8 mg, 60% MeCN) was concentrated, resuspended, and separated over HPLC [mobile phase: 70% MeCN in H_2O (0.1% formic acid) at 3 $ml{\cdot}min^{-1}$; stationary phase: 5 µm, C18(2), 100 Å, 250 x 10 mm (Phenomenex, Luna) column] to

yield subfractions A (2-4 min, 3.8 mg) and B (4-10 min, 11.9 mg). Subfraction A was separated over HPLC [mobile phase: 30% MeCN in H_2O (0.1% formic acid) at 3 ml·min^{-1}; stationary phase: 5 µm, C18(2), 100 Å, 250 x 10 mm (Phenomenex, Luna) column] to yield ketomemicin C-418 (**1**, tR = 14 min, 0.8 mg), ketomemicin C-332A (**2**, tR = 20 min, 0.7 mg) and ketomemicin C-432B (**3**, tR = 22 min, 0.5 mg).

4.3.3 Assessment of ketomemicin production in *Salinispora* and other data sets

Three *Salinispora* metabolomic data sets from Crüsemman *et al.*, 2017 (MSV000078836) Duncan *et al.*, 2015 (MSV000079284), and Chase *et al.*, 2021 (MSV000085890) were selected to run a molecular network with CNY-498 fraction 3, using the Global Natural Products Social Molecular Networking (GNPS) (Wang *et al.,* 2016). The classical molecular networking workflow was as follows: The MS/MS spectra were window filtered by choosing only the top 6 fragment ions in the +/- 50Da window throughout the spectrum. The precursor ion mass tolerance was set to 2.0 Da and a MS/MS fragment ion tolerance of 0.5 Da. The network was then created where edges were filtered to have a cosine score above 0.7 and more than 5 matched peaks. Cytoscape v 3.9.1 was used for network visualization and further manipulation. To further investigate ketmomemicin production outside *Salinispora,* we used MASST, an interface that allows the search a single MS/MS spectrum against public GNPS spectral libraries and all public MS/MS datasets (Wang *et al.,* 2020). The parameters for the MASST search are as follow: The data was filtered by removing all MS/MS fragment ions within +/- 17 Da of the precursor *m/z*. MS/MS spectra were window filtered by choosing only the top 6 fragment ions in the +/- 50 Da window throughout the spectrum. The precursor ion mass tolerance was set to 0.01 Da and a MS/MS fragment ion tolerance of 0.1 Da. The library spectra were filtered in the same manner as the input

data. All matches kept between input spectra and library spectra were required to have a score above 0.7 and at least 4 matched peaks.

4.3.4 Assessment of BGC diversity in *Salinispora* and other taxa

To investigate the ketomemicin BGC in other bacterial taxa I used the Cblaster remote search workflow (Gilchrist *et al.,* 2021). I did an extensive search of the collocated protein sequences for all six genes in the *ktm* BGC against the non-redundant protein sequences and the reference sequences of NCBI databases with a maximum of 5 000 maximum target sequences, a maximum E value of 0.01, minimum identity and minimum coverage of 50% and a maximum distance between collocated genes of 10 kbps. To investigate gene neighboring and cluster architecture we integrated genes located 15 kbp upstream and downstream of the ketomemicin biosynthetic gene cluster. For alignment and visualization of the gene cluster we used Clinker, a python-based workflow for BGC comparison and figure generator (Gilchrist & Chooi 2021).

4.3.5 Enzyme sequence similarity networks

To build the sequence similarity networks (SSNs) for the genes *ktmA*, *ktmB* and *ktmD*, we extracted the amino acid sequences of each *Salinispora* genome strain containing the ketomemicin BGC and ran the EFI-Enzyme similarity tool (Zallot *et al.,* 2019) for sequence comparison using the recommended parameters as follows: minimum coverage of 35% and alignment score of E value $1x10^{-5}$. We used the full network file, where each node represents one single protein sequence, and the connection threshold was set at 97% similarity. Cytoscape V3.9.1 (Shannon *et al.*, 2003) was used for further data analysis, visualization, and manipulation.

4.4 Results and discussion

Ketomemicin C-418 (**1**) was isolated as a thin white film. By HRMS analysis we determined the molecular formula as $C_{22}H_{34}N_4O_4$ (observed 419.2650 *m/z*, calculated 419.2653). The ^{1}H NMR spectrum showed three aromatic signals (7.15, 7.22 and 7.24 ppm), two signals for alpha-protons (3.88 and 4.42 ppm), five de-shielded (2.30/2.88, 2.99 and 2.64/3.05 ppm), and eight shielded (0.91 – 2.18 ppm) aliphatic protons. The HSQC and HMBC spectra revealed the chemical shifts for the 22 carbons in the molecule. The COSY spectrum revealed four proton spin systems. And the HMBC revealed their inter-connectivity. The relative configurations of the novel ketomemicin analogs were determined using molecular modeling and DFT calculations. We performed DFT ^{1}H and ^{13}C NMR chemical shifts calculations for the eight diastereomers of each ketomemicin analog. Postdoctoral Gabriel Castro assigned the relative configurations as 2S*, 11S*, 18S*-**1** since these diastereomers were most similar to the respective experimental NMR ^{1}H and ^{13}C assignments. The absolute configuration of the novel ketomemicins will be further determined. Along with **1**, another two ketomemicins were isolated, ketomemicin C-432A (**2**) ppm and C-432B (an isomer of **2**) with a molecular formula of $C_{23}H_{37}N_4O_4$ (observed 433.351 *m/z*, calculated 433.2809).

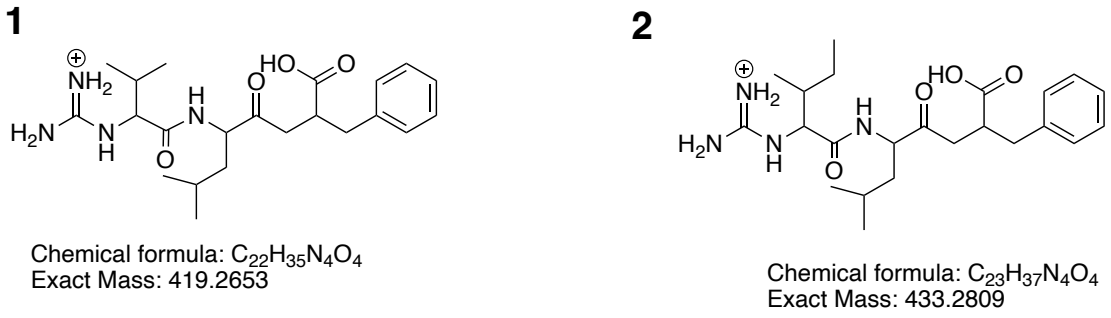

Figure 4.3 Structures of novel ketomemicins isolated from S. pacifica CNY-498. Showing the ion form of ketomemicin C-418 (**1**) and ketomemicin C-432 A (**2**).

Using GNPS molecular networking and the GNPS Massive database, we identified putative ketomemicin analogs with similar mass fragmentation patterns from *Salinispora* spp. previously studied by Crüsemann *et al.,* (2017), resulting in a total of 27 strains across seven species showing production of ketomemicins. From this, 13 strains in four species produced two more putative ketomemicin analogs with masses 449 *m/z* and 499 *m/z* (figure 4.4). The structure of these analogs was assigned by comparing their MS/MS spectra with **1**-**2** (figure 4.3 and 4.4).

Interestingly, we observed that the hydroxylated ketomemicins were mainly produced by *S. tropica* while the un-hydroxylated ketomemicins were mainly produced by *S. pacifica*, *S. fenicalii* and *S. oceanensis* (fig. 4.5). In *S. arenicola* and *S. cortesiana,* we observed low production of non-hydroxylated ketomemicins. These patterns of differential occurrences of ketomemicins could be related to extraction or cultivation methods as the samples from Crüsemann *et al*., (2017) were cultivated on solid media and extracted with multiple solvents including methanol, butanol, and ethyl acetate. The use of three extraction methods corresponded with the greatest occurrence of ketomemicins. In contrast, we did not observe ketomemicins in the Duncan *et al.,* 2015 dataset generated from ethyl acetate extracts of liquid cultures. Agar may be a more stable matrix to

accumulate the excreted specialized metabolites and facilitate their detection thus resulting in a greater variety of ketomemicins (an many other compounds) in the samples. Similarly, the polymeric adsorbing resin XAD-7 used to extract the ketomemicins in this work may facilitate the detection of ketomemicins from *Salinispora* liquid cultures. In contrast, there might be other physiological processes occurring within the cell that might be affecting the production of ketomemicins, and the variation in ketomemicin production we observed are indeed because differential production either due to gene regulation, gene activity, or substrate specificity among the different species.

We next queried 118 *Salinispora* genome strains to assess the distribution of the *ktm* BGC across all nine species. We identified *ktm* in 80 of the 119 *Salinispora* genomes across 7 of 9 species (*S. oceanensis, S. pacifica, S. cortesiana, S. fenicalii, S. tropica, S. arenicola,* and *S. mooreana*) (fig.4.6). Interestingly, one strain of *S. goodfellowii* had an incomplete *ktm* BGC with 4 of 6 biosynthetic genes (*ktmA, ktmB, ktmD and ktmF*). It is interesting to speculate whether "incomplete BGCs" such as these are functional or in a process of losing or acquiring function. Additionally, this analysis revealed intra- and inter-species patterns associated with the genomic environments flanking the six *ktm* genes. We observed conserved genes coding for transcriptional regulators from the *lysR* and *tetR* families and genes involved in specialized metabolism such as monooxygenases, phosphate oxidases, and N-acetyl transferases in almost all species. However, the genomic environment is completely different in *S. arenicola*. Instead of the transcriptional regulator and tailoring enzyme genes, we observed genes related to primary metabolism or uncharacterized functions (fig. 4.6). Perhaps these differences in *S. arenicola* account for the lower production of ketomemicins in this species.

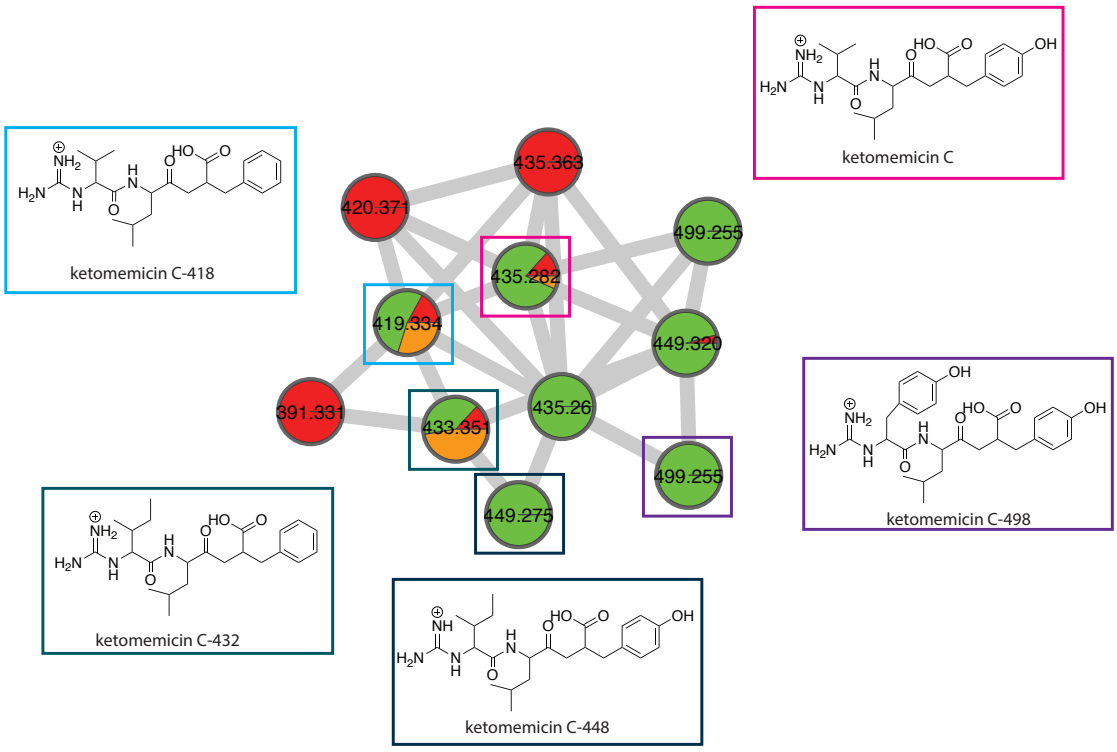

Figure 4.4 Ketomemicins molecular networking from different *Salinispora* datasets. The red nodes represent masses from the CNY-498 semi purified extract; the green nodes represent masses from the extracts in Crüsemann *et al.,* (2017) data set; and the orange color represents masses from the extracts in Chase *et al.,* (2021).

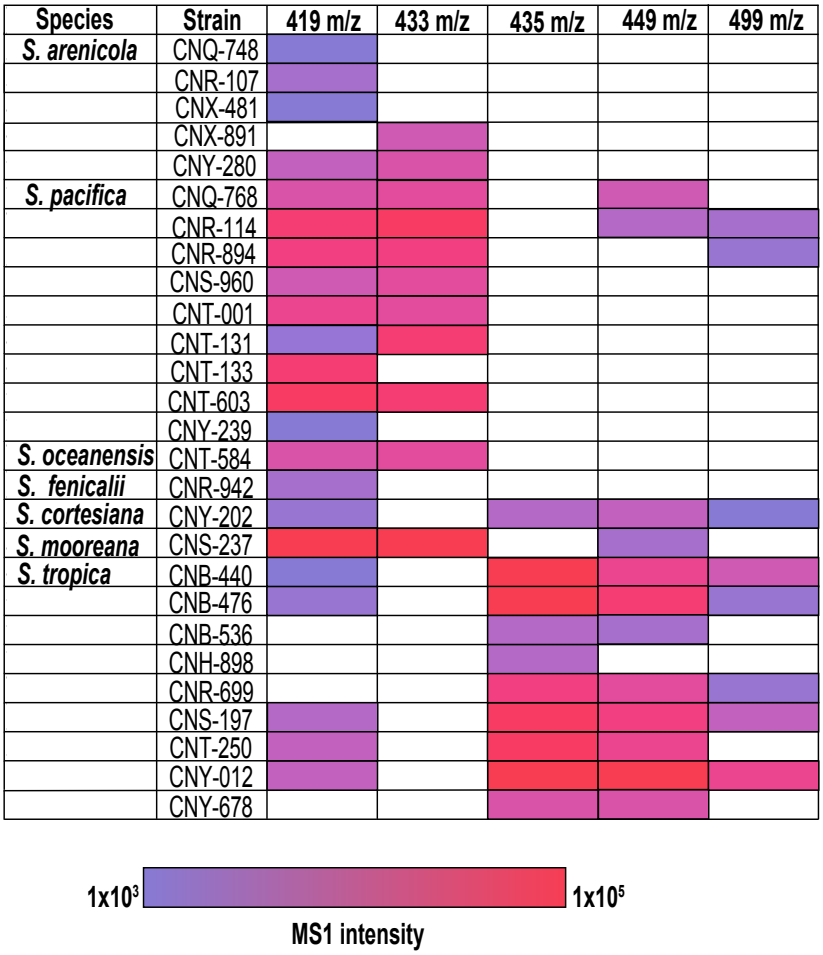

Figure 4.5 Ketomemicin analog production in Salinispora species, the colored panels indicate the intensity in the mass spectrometry chromatogram.

Further, in bacteria, N-acetyl transferases transfer an acetyl group from acetyl-CoA to a large arrange of substrates (Favrot *et al.,* 2016), it would be interesting to investigate if the presence or absence of any of the other biosynthetic genes might indirectly affect production of ketomemicins in *Salinispora.*

Based on previous characterization of the enzymes involved in ketomemicin biosynthesis, we hypothesized that variation in the *ktm* genes would affect production of the different

ketomemicin analogs. Specifically, the aldolase (*ktmA)*, the C-acyltransferase (*ktmB)* and the pseudo-peptide ligase (*ktmD*). Under the premise that the structural differences observed the ketomemicin C analogs, would be the result of variations in the active sites of these enzymes by favoring the incorporation of different substrates. We investigated this hypothesis by comparing the amino acid sequences of the genes coding for these three enzymes across the 80 *Salinispora* strains with *ktm* using sequence similarity networks (showing connections for >97% sequence similarity, fig. 4.7). All three genes showed a similar clustering, following the species phylogeny. For example, in all gene comparisons, *S. tropica* genes and in two cases (ktmA and ktmD) *S. arenicola genes* clustered separately from the other species. *KtmD* showed the most sequence diversity (fig. 4.7C), while *ktmB* was the most conserved (fig 4.7B). Surprisingly, all *ktmB* sequences clustered together in all but the *S. tropica* genes and two *S. pacifica* genes; this result suggested a higher sequence diversity of *ktmB* among two strains (CNT-851 and CNT-796) and the rest of the *S. pacifica,* than among the rest of *Salinispora* species, we must point out that neither strain showed production of ketomemicin. In the biosynthesis of ketomemicin A, ketomemicin C, an all the ketomemicin C analogs produced in *Salinispora, ktmB* catalyzes a Claisen-type condensation between benzylfumaryl-CoA and leucin, thus it is coherent to observe the highest conservation in the protein sequences of *ktmB* genes among *Salinispora* species.

Particularly for the differential production of hydroxylated ketomemicins, *ktmA* might have an influence on which substrate is incorporated. Thus, I expected to observe clustering patterns similar to the production patterns observed (fig.4.5). For instance, *S. tropica* clustering closely with genes from *S. mooreana* and *S. cortesiana* since those species have the most similar production patterns. However, this was not the case. *S. tropica* genes clustered separately, and *S. arenicola* clustered with the reminder species (fig. 4.7A), this suggest that if a difference in the

protein sequence of *ktmB* is causing the differential production of hydroxylated ketomemicins among species, the difference might be in a few amino acid positions, and the method we are utilizing for comparison is too robust to be able to discern those differences. Unfortunately, the active site for *ktmA* gene product has not been fully characterized. Further investigation of the binding sites inside the substrate pocket would enlighten if a change of amino acids in a given position would favor the binding of the hydroxylated substrate over the non-hydroxylated.

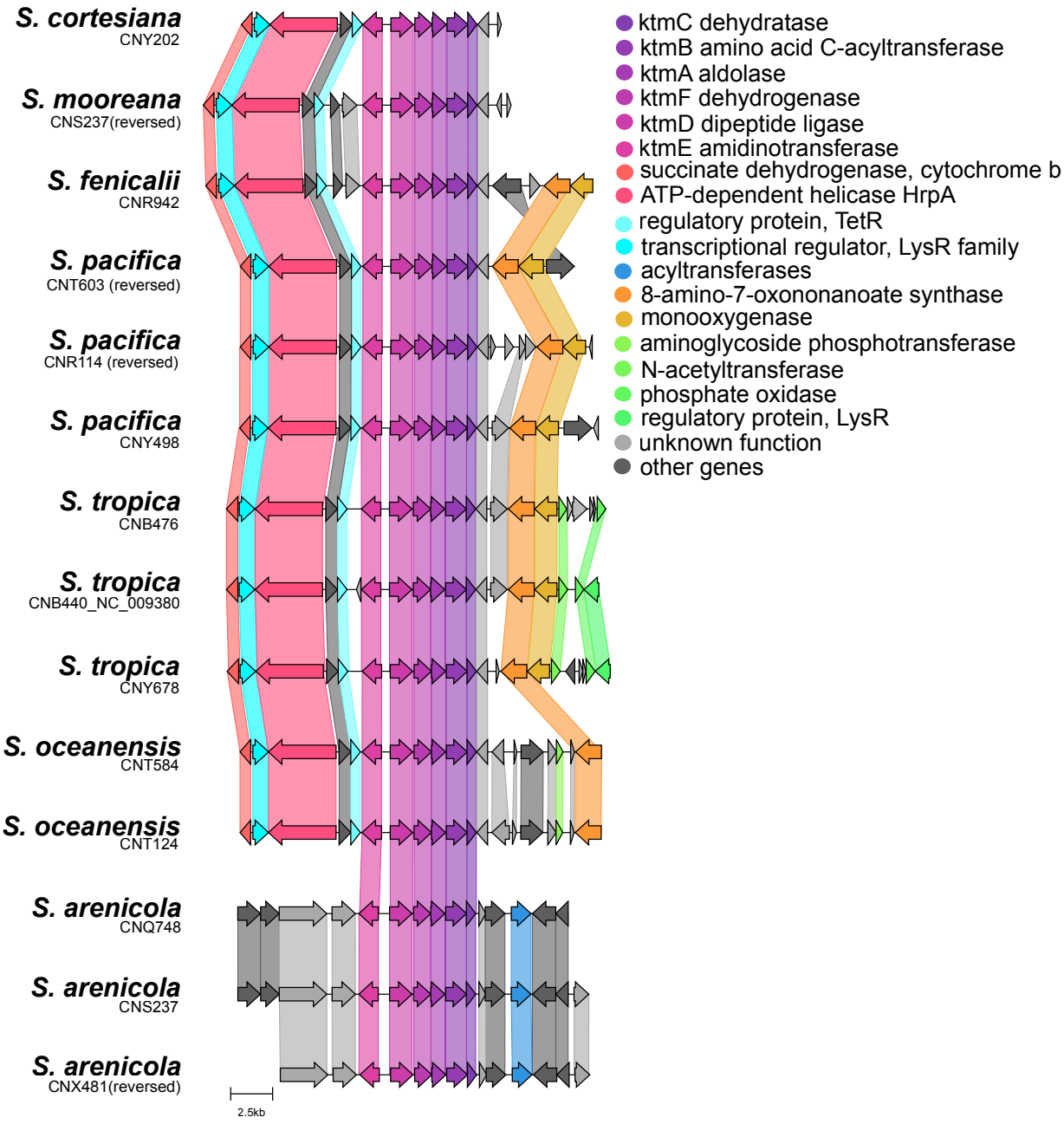

Figure 4.6 Ketomemicin BGC diversity and gene neighborhoods in a subset of *Salinispora* strains. The alignment shows representatives of the most diverse ktm BGC neigbhorhoods among the 80 strans where the cluster was found. The six genes involved in ketomemicin biosynthesis are colored in different shades of pink. The shading connecting genes between species represents sequence homology.

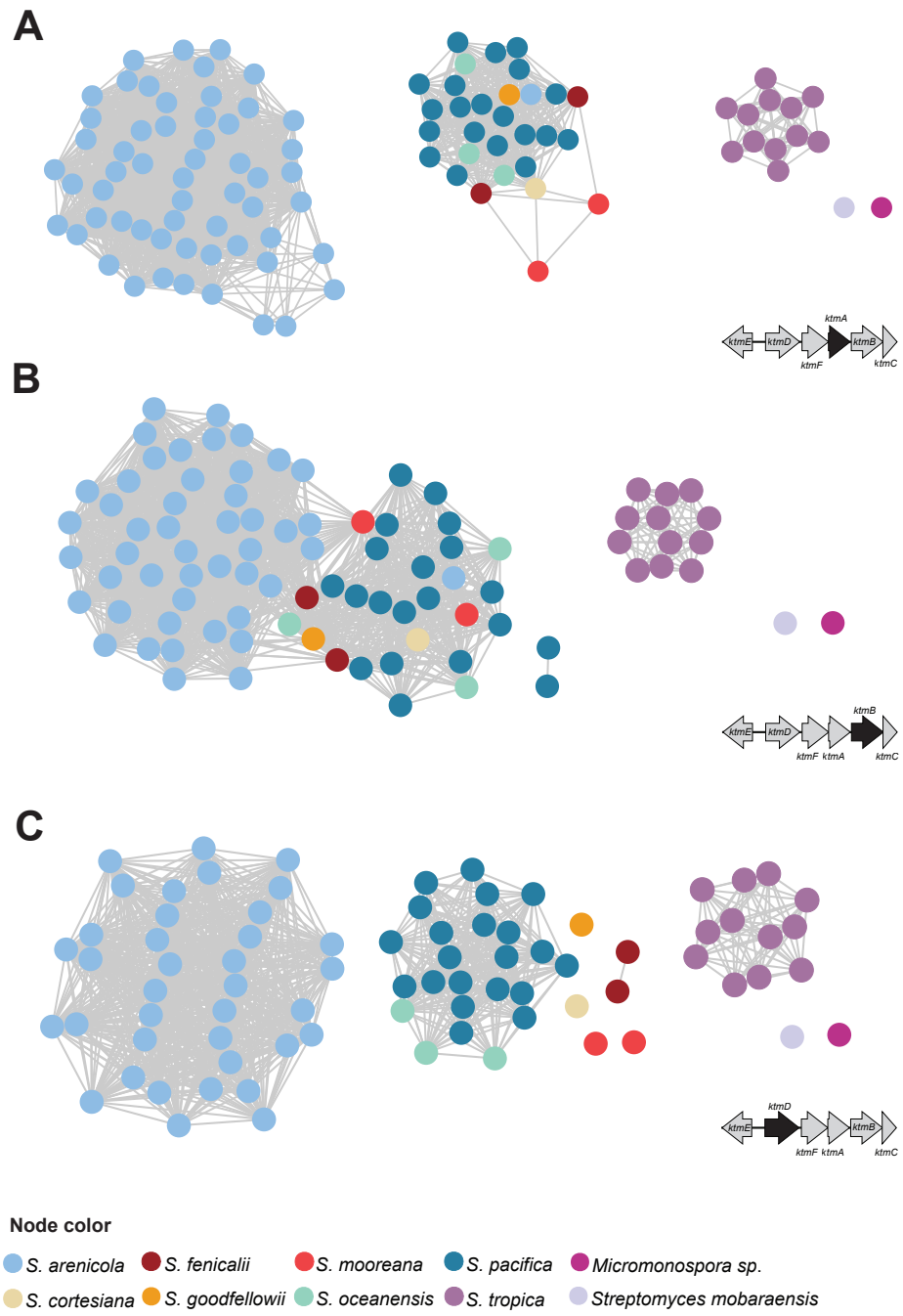

Figure 4.7 Sequence similarity networks constructed from amino acide sequences of: **A.** ktmA (aldolase), **B**. ktmB (C-acyltransefrase), and **C.** ktmD (pseudopeptide ligase). Edges between nodes represent a sequence similarity of <97%. Node colors correspond to species. The position of each gene in the *ktm* BGC is depicted in the lower right corner of each network.

Despite *ktm* being fairly ubiquitous among *Salinispora* spp., its distribution outside the genus is scarce (fig.4.8). From a query on NCBI using Cblaster, we found in total 28 strains containing ketomemicin-like BGCs, 14 in *Streptomyces* spp*.,* nine in *Micromonospora* spp*.,* four in *Kitasatospora* spp., and one in *Actinophytocola xanthii*. All of these bacteria belong to the phylum *Actinomycetota* (formerly Actinobacteria). We also detected BGCs containing *ktmA, ktmB*, *ktmC*, and *ktmF* gene homologs in other taxa but lacking the pseudopeptide ligase *ktmD* gene, which catalyzes the final peptide ligation step in ketomemicin biosynthesis.

To further investigate ketomemicin production outside *Salinispora*, we queried GNPS datasets using the MASST tool (Wang *et al.* 2020) with the mass fragmentation data of **1** and **2**. We found spectra similar to ketomemicin in *Amycolatopsis sp.* (MSV000087905) and *Streptomyces sp.* (MSV000078839) samples. Other representative strains from these two genera contain the *ktm* BGC within their genome, thus making it plausible that additional strains can produce this compound class. In other species of *Amycolatopsis sp*, I found only a partial *ktm* BGC (figure 4.8), and the genomic information from the strains in the GNPS databases are not publicly available to further investigate the presence of *ktm* BGC. However, it would be interesting to further investigate if *Amycolatopsis* proudces similar compounds to ketmomemicin. Nonetheless, these analyses shed light on the distribution of ketomemicin by looking into currently available genomic and metabolomic databases.

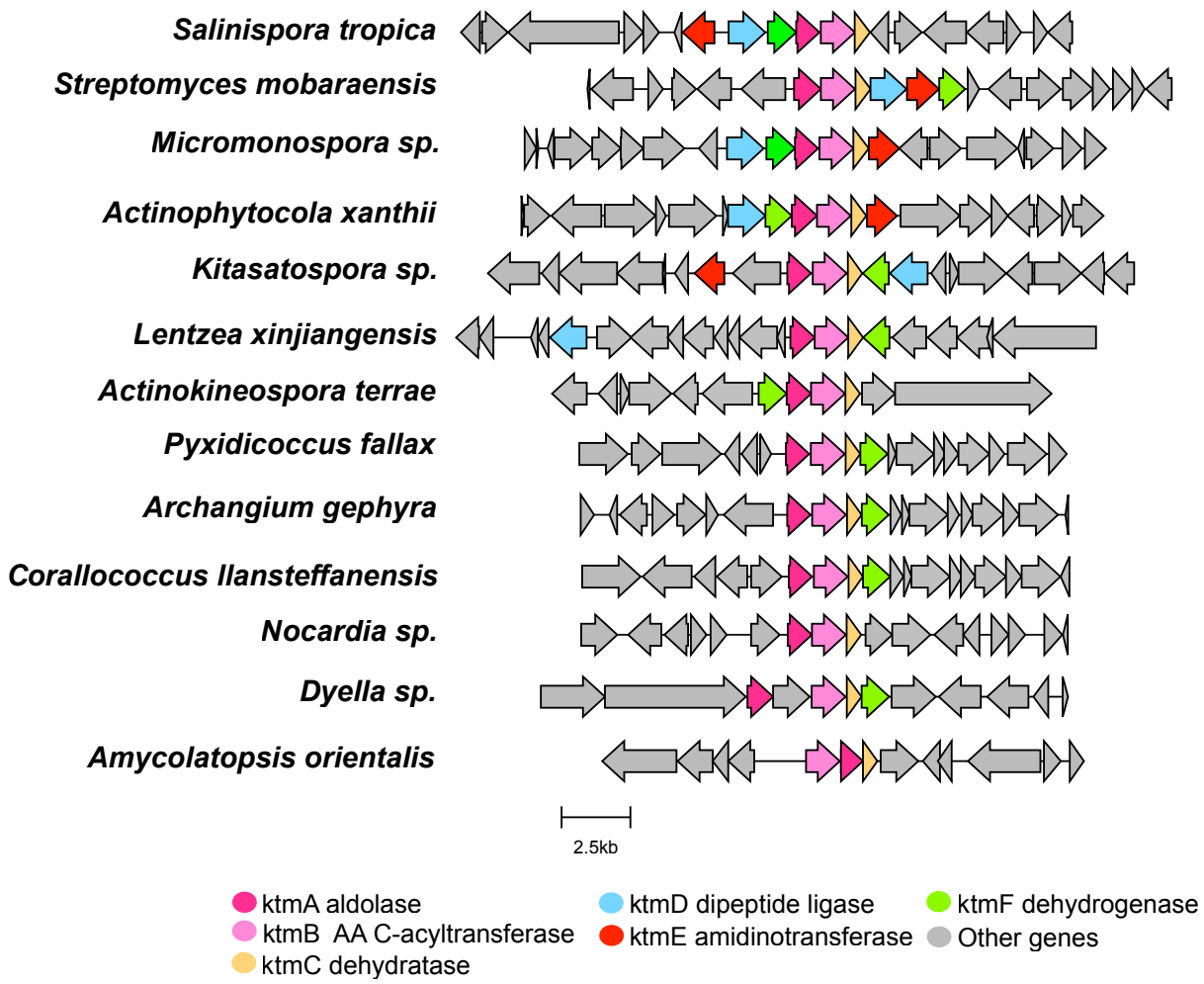

Figure 4.8 Distribution of the ketomemicin BGC in bacterial taxa outside the Salinispora genus.

4.5 Conclusions

We isolated three new ketomemicin analogs, ketomemicin C-418, C-432A and C-432B from the culture extracts of *S. pacifica* CNY-498, confirming *Salinispora* spp. as source of ketomemicins and making this the first report of wild type strains producing ketomemicins. Further, we were able to broaden the distribution of ketomemicin production in *Salinispora* species using public datasets in the GNPS platform, finding additional ketomemicin analogs characterized

by mass spectrometry. Our use of the GNPS database demonstrates the advantage of having freely available and robust metabolomic data sets available to the natural product community.

The ketomemicin biosynthetic gene cluster is highly conserved in *Salinispora* spp. yet rarely observed in other bacterial taxa. We found some cladding patterns of neighboring genes in S. *arenicola* compared to the other species, lacking genes related to transcription regulation that may account for the low production of ketmomemcin C's in *S. arenicola,* although more testing is needed.

Methylene carbonyl containing pseudopeptides are very rare in nature. Besides ketomemicin, arphamenine is the only another pseudo peptide containing this type of bond. This compound was reported to be an inhibitor of aminopeptidase type B, and thus we are investigating the protease inhibition activity of ketomemicin C-418 and believe it highly plausible for ketomemicin to be active against proteases as well.

4.6 Acknowledgements

I would like to thank Alexander Bogdanov for his revisions of spectral data. Special thanks to my co-author Gabriel Castro for contributing with the structural elucidation of ketomemicin analogs and for his assistance with cultivation and chemical extractions. I would also like to thank Reiko Cullum and Bill Fenical for use of their equipment for the large-scale fermentations and chemical extractions.

This research was supported by the NIH grant to R01GM085770 PRJ and BSM and the CONACYT and UC MEXUS doctoral fellowship.

Chapter 4 is coauthored with Gabriel Castro-Falcón and Paul R. Jensen. The dissertation author was the primary investigator and author of this chapter.

4.7 References

Blin, K., Shaw, S., Kloosterman, A.M., Charlop-Powers, Z., van Wezel, G.P., Medema, M.H., & Weber, T. (2021). antiSMASH 6.0: improving cluster detection and comparison capabilities. *Nucleic Acids Research, 49*, W29 - W35.

Chase, A.B., Sweeney, D., Muskat, M.N., Guillén-Matus, D., & Jensen, P.R. (2021). Vertical Inheritance Facilitates Interspecies Diversification in Biosynthetic Gene Clusters and Specialized Metabolites. *mBio, 12*.

Crüsemann, M., O'Neill, E.C., Larson, C.B., Melnik, A.V., Floros, D.J., da Silva, R.R., Jensen, P.R., Dorrestein, P.C., & Moore, B.S. (2017). Prioritizing Natural Product Diversity in a Collection of 146 Bacterial Strains Based on Growth and Extraction Protocols. *Journal of natural products, 80 3*, 588-597.

Duncan, K.R., Crüsemann, M., Lechner, A., Sarkar, A., Li, J., Ziemert, N., Wang, M., Bandeira, N., Moore, B.S., Dorrestein, P.C., & Jensen, P.R. (2015). Molecular networking and pattern-based genome mining improves discovery of biosynthetic gene clusters and their products from Salinispora species. *Chemistry & biology, 22 4*, 460-471.

Favrot, L., Blanchard, J.S., & Vergnolle, O. (2016). Bacterial GCN5-Related N-Acetyltransferases: From Resistance to Regulation. *Biochemistry, 55 7*, 989-1002.

Gilchrist, C.L., & Chooi, Y. (2021). Clinker & Clustermap.js: Automatic Generation of Gene Cluster Comparison Figures. *Bioinformatics* 37(16): 2473–2475

Gilchrist, C.L., Booth, T.J., Wersch, B.V., Grieken, L.V., Medema, M.H., & Chooi, Y. (2021). cblaster: a remote search tool for rapid identification and visualization of homologous gene clusters. *Bioinformatics Advances* 1(1).

Kawata, J., Naoe, T., Ogasawara, Y., Dairi, T. (2017). Biosynthesis of the Carbonylmethylene Structure Found in the Ketomemicin Class of Pseudotripeptides. *Angewandte Chemie - International Edition*, *56*(8), 2026–2029.

Kaysser, L. (2019). Built to bind: biosynthetic strategies for the formation of small-molecule protease inhibitors. *Natural product reports*.

Ogasawara, Y. (2019). New enzymes for peptide biosynthesis in microorganisms. *Bioscience, Biotechnology, and Biochemistry, 83*, 589 - 597.

Ogasawara, Y., Kawata, J., Noike, M., Satoh, Y., Furihata, K., Dairi, T. (2016). Exploring Peptide Ligase Orthologs in Actinobacteria - Discovery of Pseudopeptide Natural Products, Ketomemicins. *ACS Chemical Biology*, *11*(6), 1686–1692.

Patch, J.A., Kirshenbaum, K.S., Seurynck, S.L., Zuckermann, R.N., & Barron, A.E. (2004). Versatile Oligo(N-Substituted) Glycines: The Many Roles of Peptoids in Drug Discovery.

In: Nielsen, P.E. (editor) Pseudo-peptides in drug discovery. Wiley-VCH. 1st edition. 1-31p.

Shannon, P.T., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., & Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research, 13 11*, 2498-504.

Umezawa, H., Aoyagi, T., Ohuchi, S., Okuyama, A., Suda, H., Takita, T., Hamada, M.S., & Takeuchi, T. (1983). Arphamenines A and B, new inhibitors of aminopeptidase B, produced by bacteria. *The Journal of antibiotics, 36 11*, 1572-5.

Wang, M., Jarmusch, A.K., Vargas, F., Aksenov, A.A., Gauglitz, J.M., Weldon, K.C., Petráš, D., da Silva, R.R., Quinn, R.A., Melnik, A.V., van der Hooft, J.J., Caraballo-Rodríguez, A.M., Nothias, L., Aceves, C.M., Panitchpakdi, M., Brown, E., Di Ottavio, F., Sikora, N.C., Elijah, E.O., Labarta-Bajo, L., Gentry, E.C., Shalapour, S., Kyle, K.E., Puckett, S.P., Watrous, J.D., Carpenter, C.S., Bouslimani, A., Ernst, M., Swafford, A.D., Zúñiga, E.I., Balunas, M.J., Klassen, J.L., Loomba, R.S., Knight, R., Bandeira, N., & Dorrestein, P.C. (2020). Mass spectrometry searches using MASST. *Nature Biotechnology, 38*, 23-26.

Wang, Mingxun, *et al*. "Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking." Nature Biotechnology 34.8 (2016): 828-837. PMID: 27504778.

Zallot, R., Oberg, N., & Gerlt, J.A. (2019). The EFI Web Resource for Genomic Enzymology Web Tools: Leveraging Protein, Genome, and Metagenome Databases to Discover Novel Enzymes and Metabolic Pathways. *Biochemistry*. 58 (41): 4169–4182

Zalman, L.S., Brothers, M.A., Dragovich, P.S., Zhou, R., Prins, T.J., Worland, S.T., & Patick, A.K. (2000). Inhibition of Human Rhinovirus-Induced Cytokine Production by AG7088, a Human Rhinovirus 3C Protease Inhibitor. *Antimicrobial Agents and Chemotherapy, 44*, 1236 - 1241.

Chapter 5

A chemotaxonomy beyond cell wall components: specialized metabolites as chemotaxonomic markers for bacteria.

5.1 Abstract

Chemotaxonomy is a method of classification based on the presence or absence of specific biomolecules. Bacterial chemotaxonomy has largely focused on fatty acids or cellular components while specialized metabolites have been employed for the classification of certain chemically rich plants, animals, and fungi. While many groups of bacteria are rich sources of specialized metabolites, and their production in taxon-specific patterns has been reported, these molecules have been largely overlooked as taxonomic markers. Here, we review evidence linking specialized metabolism to bacterial classification and discuss how these metabolites may provide opportunities to resolve ecologically distinct units of diversity that are not easily distinguished using more conserved approaches. We suggest that some specialized metabolites represent taxonomic markers in bacteria with their distributions reflecting ecological adaptations over long evolutionary timeframes.

5.2 Introduction

Chemotaxonomy originated with the use of chemical-based characteristics such as taste, color, smell, and medicinal value to aid in the identification of plants (Wink *et al*., 2010). The use of specialized metabolites, also called secondary metabolites or natural products, became more prevalent in plant classification as more of these compounds were identified (Whalley & Edwards, 1995). Today, specialized metabolites are the primary chemotaxonomic markers used to

differentiate plants (Umoh, 2020), proving effective to resolve ambiguities among closed related organisms (Singh & Geetanjali, 2019; Umoh, 2020). Specialized metabolites are also recognized as useful taxonomic markers in fungi (Frisvad 2007) and some marine invertebrates such as sponges and soft corals (Erpenbeck & Soest, 2006; Figueiredo *et al*., 2016) (Figure 1) based on observations that organisms at the same taxonomic level often produce similar specialized metabolites or metabolite families (Singh & Geetanjali, 2018).

Chemotaxonomy has long been a required component of bacterial classification. However, unlike applications with other organisms, the focus has been on primary metabolites such as whole cell amino acids and sugars, membrane lipids, and respiratory quinones (Buchanan, 1955; Lechevalier *et al*, 1977; Schleifer 2009). Nonetheless, it is widely recognized that many bacterial taxa produce specialized metabolites (Bode & Müeller,2006; Andryukov *et al*., 2019). These compounds have a wide range of chemical structures and ecological functions including chemical defense, signaling, and nutrient acquisition; among others (Abdel-Aziz *et al*., 2017; Mosunova *et al*., 2021). Bacterial specialized metabolites have not been widely employed as taxonomic markers despite studies showing correlations between phylogeny and their production (Jensen *et al*., 2007; Doroghazi *et al*., 2014; Ziemert *et al*., 2014; Adamek *et al*., 2018; Undabarrena *et al*., 2021).

While sequence-based approaches such as ANI have gained prominence as rapid methods to assess species delineations, they provide little insight into the ecological differences among closely related strains. Conversely, functional traits such as specialized metabolites have largely been overlooked in terms of bacterial systematics yet may provide important insights into ecological divergence. In this review, we explore evidence that bacterial specialized metabolites are produced in taxon-specific patterns and thus may represent taxonomic markers in certain groups of bacteria. This is based on the hypothesis that specialized metabolites represent functional

traits of sufficient ecological significance to be fixed within phylogenetically coherent lineages. We discuss the challenges of this approach and opportunities to employ specialized metabolites and their cognate biosynthetic genes to help resolve ecologically meaningful units of bacterial diversity.

5.3 An overview of bacterial classification

Bacterial classification began in the late 19th century using phenotypic traits such as morphology and pathogenicity. Subsequently, descriptions of bacterial physiology and ecology were introduced (Schleifer 2009). Modern DNA techniques were incorporated in the late 20th century providing a more precise way to classify bacteria. These include guanine-cytosine content, DNA:DNA hybridization, genetic fingerprints (such as random amplified polymorphic DNAs), 16S rRNA gene sequencing, and multilocus sequence analysis (Mordarska *et al*., 1972; Roselló-Móra, 2012). Whole genome sequencing has advanced the study of bacterial diversity and evolution and have also provided new approaches to delineate species including genome-to-genome distance, average amino acid identity (AAI), average nucleotide identity (ANI), and core genome analysis among others (Thompson *et al*., 2015). However, even commonly applied whole genome metrics such as ANI do not offer a consensus definition of bacterial species. For instance, for the pathogen *Bacillus cereus sensu lato* an ANI threshold of 92.5% was determined superior to 95%, favoring a lower cut-off for a more accurate diagnosis based on the argument that an ANI of 95% would still cluster pathogenic and non-pathogenic species as one (Zervas *et al*., 2020).

Despite these advances, bacterial taxonomy remains a challenge and the approaches used to distinguish among species continue to evolve. For example, 70% DNA-DNA relatedness and 97% 16S rRNA gene sequence identity were once commonly applied as cut-offs for species delineations but are now widely recognized as too broad. (Schleifer 2009; Thompson *et al*., 2015).

Furthermore, species concepts for bacteria remain controversial (Cohan 2002; Gevers *et al*., 2005). While there is currently no consensus species concept, microbiologists have empirically circumscribed bacteria to species embracing genealogical, genomic (or genetic) and phenotypic premises (Roselló-Móra, 2012). However, many favor genome-based phylogenies for taxonomic classification (Parks *et al*., 2018). Similarly, gene expression might also play an important role, and expression experiments can help differentiate ecological distinct groups, adding another level to taxonomic considerations (Konstantinidis & Tiedje 2005; Liang *et al*., 2014).

Today, bacterial taxonomy requires a polyphasic approach that includes the traditional phenotypic descriptions along with genomic and phylogenetic analyses (Tindall *et al*., 2010). Various elements of the cell continue to be used as chemotaxonomic markers for bacteria including cell membrane fatty acids, polar lipids, respiratory lipoquinones, pigments, the amino acid composition of the peptidoglycan, the presence and size of mycolic acids, and constituents of the cytoplasm like polyamine (Tindall *et al*., 2010). However, these approaches often fail to capture ecological differences among closely related strains.

5.4 Chemotaxonomy of Eukaryotes

The use of specialized metabolites in chemotaxonomy is best developed for plants, with most members of a given monophyletic clade sharing at least one chemical characteristic (Wink 2003; Wink *et al*., 2010 & Umoh 2020). Plants are particularly enriched in the production of alkaloids, which have been widely used as chemotaxonomic markers in the families *Malvacea, Ranunculaceae, Magnoliaceae, Polygnonaceae, Solananceae*, and *Asphodelaceae* (Singh & Geetanjali, 2018; Viljoen *et al*., 1998). Examples include tropane alkaloids such as nicotine in the family *Solanaceae*, (Griffin & Lin, 2000), the alkaloid quinine in the family *Rubiaceae,* and the prenylated polyketides cannabinoids in the family *Cannabaceae* (Singh *et al*., 2019). Flavonoids

are also produced in taxon-specific patterns, with 6-O-coumaroylaleosin representing a diagnostic metabolite for all six *Aloe* species in the family *Asphodelaceae* (Singh & Geetanjali, 2018; Singh *et al*., 2019).

Among fungi, specialized metabolites have proven helpful to distinguish among *Penicillium*, *Aspergillus*, and *Fusarium* species, which are otherwise regarded as difficult to classify (Zain, 2010; Polizzi *et al*., 2012; Helaly *et al*., 2018). For instance, the mycotoxin fusaric acid has been proposed as a chemotaxonomic marker for *Fusarium* spp. in contaminated grains (Bacon *et al*., 1996). Chemotaxonomy has been used for over two decades in the order *Xylariales* (Helaly *et al*., 2018), with the dihydroisocoumarin 5-methylmellein widely distributed throughout the family *Xylariaceae* (figure 5.1) (Whalley & Edwards, 1995). Other examples are the terphenyl quinones atromentin and auranticin observed in the class *Basidiomycetes* (Benedict, 1970).

In the case of soft corals, fatty acids are the most common chemotaxonomic markers (Imbs *et al*., 2007). However, specialized metabolites have been useful for chemotaxonomy at the genus genus level in *Cladiella*, which produces the diterpenoid cladieunicellin, and at the species level in *Lemnalia flava*, which produces the sesquiterpene flavalin (Wei *et al*., 2013 & Wu *et al*., 2018), and in *Antillogorgia elisabethae*, which produces the glycosylated diterpene pseudopterosin (Look *et al*., Blunt *et al*., 2007; Imhoff *et al*., 2011) (figure 5.1).

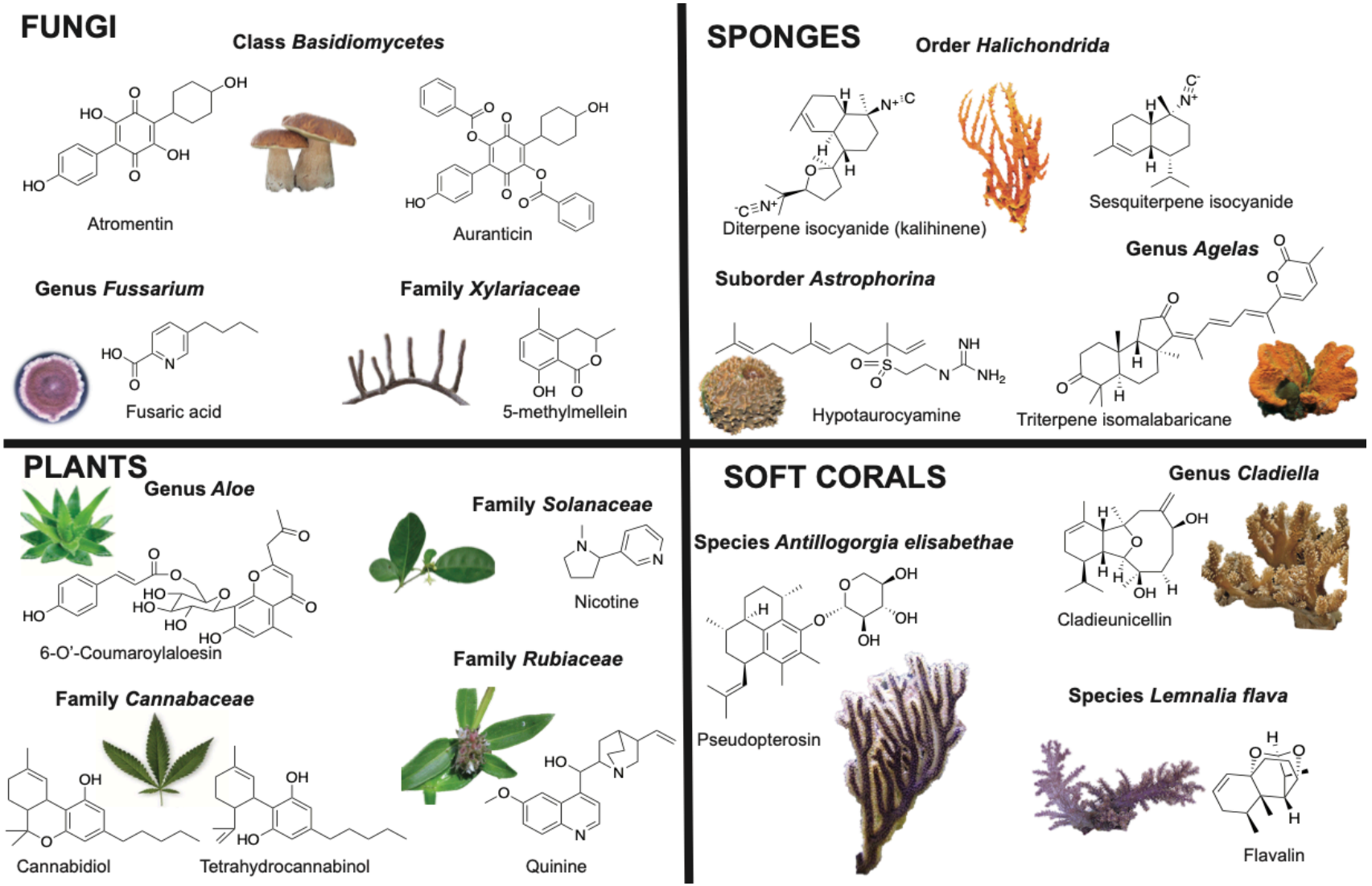

Figure 5.1 Examples of specialized metabolites used as chemotaxonomic markers for sponges, fungi, plants, and soft corals.

Marine sponges have long presented taxonomic challenges, with sterols employed to distinguish among taxa as early as the mid-nineteenth century (Bergquist & Wells, 1983). As the numbers of natural products discovered from sponges increased, so did the diversity of compounds used as taxonomic markers. Despite many sponge species harboring diverse microbial communities that can complicate their specialized metabolite profiles, they are enriched in the production of terpenes and alkaloids that are rarely found in microbes (Habener *et al*., 2016). The distribution of terpene metabolites among sponges has proven useful for their differentiation at the order and family levels and as a complement to morphology-based classification (Bergquist & Wells, 1983). Some examples of terpene metabolites used in sponge chemotaxonomy are sequiterpene and diterpene isocyanides in the order *Halichondrida*, hypotaurocyamine in the suborder *Astrophorina,* and the triterpene isomalabaricane for the genus *Agelas* (figure 5.1) (Erpenbeck & Soest, 2006 & Rohde & Schupp, 2018).

5.5 The basis for specialized metabolism in bacterial chemotaxonomy

While certain primary metabolites and cell wall components are used in bacterial chemotaxonomy, they are generally not very useful for discriminating at the species level nor are they ecologically informative (Thompson *et al*., 2015). Further, there are not public databases or repositories for bacterial chemotaxonomic markers and those available are more focused on identification rather than classification and lack options for adding updated reference material (Roselló-Móra, 2012). Among private databases, the Sherlock microbial identification system (MIDI Inc.), has one of the most comprehensive databases for identification based on FAMEs, and is suitable for the identification of around 15% of the classified prokaryotic diversity (Tindell, 2010, Thompson, 2015).

Conversely, specialized metabolites represent functional traits used by bacteria to interact with each other and the environment (Fischbach *et al*., 2008; Chevrette & Handelsman, 2021). Given that closely related bacteria are known to produce different specialized metabolites, they may offer finer-level taxonomic resolution than traditional taxonomic markers, which likely evolve at a slower pace. These compounds may provide insight into recently diverged populations and evidence for ecological differences among closely related strains. Given that our understanding of the diversity and evolution of specialized metabolism relies heavily on genomic information, a chemotaxonomy based on specialized metabolism can go in parallel with the genome-based phylogenies and other genomic approaches including descriptions of BGC distributions, which will be especially useful when chemical information is not available.

Metabolomic studies in bacteria can be done in a high throughput manner which can facilitate rapid species identification with well-established chemotaxonomic markers. Furthermore, there is an ongoing effort to establishing high throughput methods to assess bacterial specialized metabolism (Tang *et al*., 2016; Zampieri *et al*., 2017; Tobias *et al*., 2019). It's getting easier to share and store metabolomic information, which gives another advantage to specialized metabolites over the classic chemotaxonomic markers. This will facilitate the rapid identification of bacterial species if chemotaxonomic markers can be assigned to a given taxon

5.6 Phylogenetic patterns in bacterial specialized metabolism

Several studies have shown correlations between phylogeny and specialized metabolite biosynthetic gene distributions. The concept of vertical inheritance of specialized metabolite genes is gaining more importance and we are gaining insight into how the evolution of these genes can also be a major driver for bacterial speciation.

Examples of this are found within the phylum *Actinomycetota* (formerly Actinobacteria and commonly called actinomycetes), which is a major source of specialized metabolites including some widely used today in clinical practice (Park *et al*., 2019). As with other groups of bacteria, actinomycete taxonomy has proven challenging with many species renamed or reclassified and incongruencies observed between phenotypic and genotypic groups (Embley and Stackebrandt 1994; Stackebrandt and Schumann 2006). The search for specialized metabolites from actinomycetes has revealed patterns that suggest they could facilitate taxonomic assignments. For instance, Doroghazi *et al*. (2014) found significant correlation between phylogeny and species-level PKS and NRPS gene distributions in 860 actinobacterial genomes. Similarly, recent comparative genomic analyses have revealed biosynthetic gene cluster (BGC) distributions that followed closely with species phylogeny in the genus *Amycolatopsis* (Adamek *et al*., 2018), while clade specific BGC families and phylogenomic distributions supporting vertical inheritance of a non-ribosomal peptide synthetase BGC in the genus *Rhodococcus* (Ceniceros *et al*., 2017; Undabarrena *et al*., 2021).

In the marine actinomycete genus *Salinispora*, the production of some specialized metabolites (Jensen *et al*., 2007) and BGC are conserved at the species level (Letezel *et al*., 2017; Chase *et al*., 2021). For example, the salinosporamide pathway is fixed in all *S. tropica* genomes examined to date (Chase *et al*., 2020), and one of the products, salinosporamide A, has been proposed as a species-level chemotaxonomic marker for this species (Jensen *et al*., 2007). Interestingly, although this BGC has on occasion been observed in *S. arenicola*, *S. tropica* strains consistently produce 10-fold more salinosporamide A [**1**]. While the ecological functions of this compound remain unknown, the high production rates in *S. tropica* may contribute to the selective maintenance of this BGC at the species level (Chase *et al*., 2021). In the case of *S. arenicola*, the

production of rifamycin [**2**] and saliniketal A [**3**], which originate from the same BGC, distinguish it from other species in the genus (Bose *et al*., 2014), with both compounds proposed as taxonomic markers (Jensen *et al*., 2007; Jensen *et al*., 2015; Letzel *et al*., 2017; Bauermeister *et al*., 2018). The high sequence identity between rifamycin biosynthetic genes in *A. mediterranei* and *S. arenicola* suggests a recent HGT acquisition event (Kim *et al*., 2006 & Bauermeister *et al*., 2018). This suggests that some horizontally acquired genes can subsequently become fixed at the species level and thus serve as taxonomic markers. The suggestion that rifamycin or its BGC can be used as taxonomic markers in *S. arenicola* does not preclude its detection on other species and thus must be considered on case by case basis. Notably, this compound has been observed in all *S. arenicola* strains tested regardless of media formulation (Hewavitharana *et al*., 2007; Matsuda *et al*., 2009; Duncan *et al*., 2015).

Other studies have shown phylogenetic patterns based on metabolite production. For example, Choudoir *et al*. (2019) studied volatile organic compounds in actinomycetes, finding a strong phylogenetic signal regardless of culture medium. Another example can be found in the MAR4 group of marine-derived *Streptomyces,* which is enriched in the production of hybrid isoprenoid specialized metabolites (Gallager *et al.*, 2010 & 2015). While these compounds are not commonly observed from bacteria, MAR4 hybrid isoprenoid production follows clade-specific patterns with marinone [**4, 5, 6**] production associated with the *S. synnematoformans* clade and naprydomicin [**7, 8, 9**] production associated with the *S. aculeolatus* clade (Gallagher *et al*., 2013 & Bauermeister *et al*., 2018). To the best of our knowledge, no MAR4 strains produce both compound classes (Gallagher *et al*., 2015). Moreover, multivariate analysis of metabolomic profiles from the *S. aculeolatus* and *S. synnematoformans* clades showed significant separation consistent with the MAR4 phylogeny (Bauermeister *et al*., 2019).

Myxobacteria, phylum *Myxococcia,* represent another group of chemically prolific bacteria where there is mounting evidence that specialized metabolites represent taxonomically informative markers. Metabolomic profiling and BGC distributions in the order *Myxococcales* revealed conservation within several genera including *Aetherobacter*. (Hoffmann *et al*., 2018 & Ahearne *et al*., 2021). In the case of *Myxococcus,* clading based on metabolomic profiles was found for two out of the four species studied (Hoffmann *et al*., 2018).

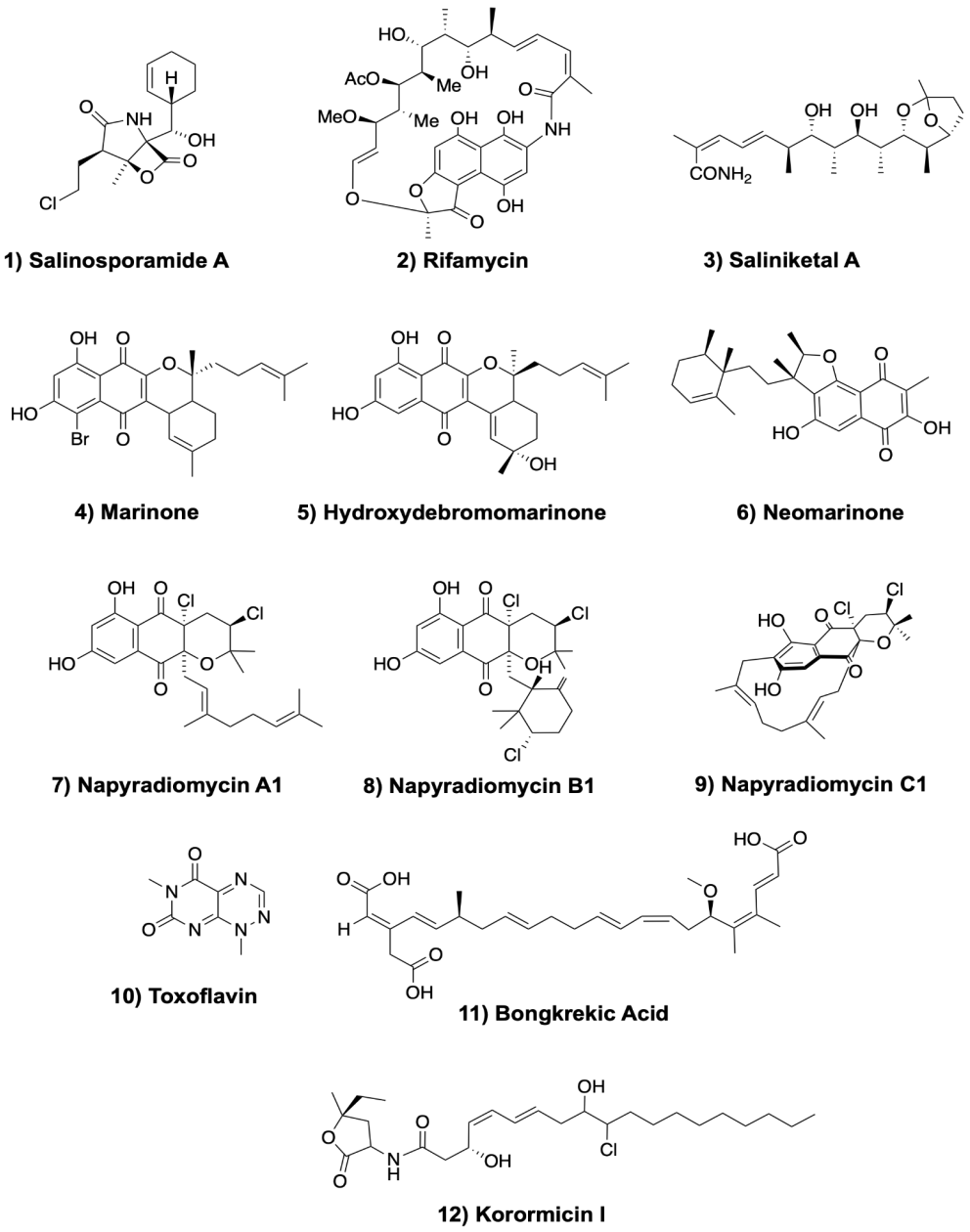

Figure 5.2 Species and clade-specific specialized metabolites in the *Actinomycetota* and *Pseudomonadota* phyla.

There are also examples of taxon-specific specialized metabolite production in the phylum *Pseudomonadota.* Among over 4000 strains in the clade *Burkholderia sensu latu*, which includes the genera *Burkholderia*, *Mycetohabitants*, *Trinickia, Paraburkholderia, Caballeronia, Robbsia,* and *Pararobbsia*, it was noted that each genus harbors around 15 distinct BGCs and that

approximately 24% of these to be species-specific. Unfortunately, many of these gene clusters are cryptic and thus cannot be linked to their small molecule products (Mullins & Mahenthiralingam 2021).

More specifically, the BGC for toxoflavin **[10]** was reported to be widely distributed among 206 strains of *Burkholderia gladioli* (Jones *et al*., 2021). To the best of our knowledge, this metabolite has not been identified in other *Burkholderia* species. The *Burkholderia* metabolite bongkrekic acid **[11]** has also been proposed as a pathogenicity marker for the *B. gladioli* subspecies 1 (Jones *et al*., 2021), providing further evidence that specialized metabolites can be used for fine-scale taxonomic resolution.

Another group from the *Pseudomonadota* phylum is the genus *Pseudoalteromonas,* a well-known producer of halogenated specialized metabolites including pentabromopseudilin and the brominated pyrrole monomer tetrabromopyrrole, both produced by the *bmp* biosynthetic pathway (Busch *et al*., 2019). Busch and collaborators (2019) studied 101 *Pseudoalteromonas* genomes and found a strong correlation between *bmp* gene composition, compound production, and taxonomy. For example, all *P. phenolic* strains analyzed maintained the same version of the gene cluster and produced the same metabolites. The gene cluster was also highly conserved in *P. luteoviolacea* but with only few gene loss events within the clade. Among other *Pseudoalteromonas* species, only *P. peptidolytica* is known to produce the chlorinated alkaloid korormicin I **[12]** (Offret *et al*., 2016 & Paulsen *et al*., 2020), which could be proposed as a chemotaxonomic marker for the species.

An interesting example of chemotaxomony is observed in the entomopathogenic bacterial genera *Xenorhabdus* and *Photorhabdus*, also in the *Pseudomonadota* phylum. These bacteria live in mutualistic symbiosis with nematodes of the genera *Steinernma* and *Heterorhabditis*,

respectively. Shi and collaborators (2022) studied the BGC diversity in 45 strains from these two genera. They found in total nine BGCs located in the core *Xenorhabdus* genome, including the NRPS that encodes xenoamicin biosynthesis. Additionally, they also found a highly conserved NRPS cluster observed in 28 of 29 genomes *Xenorhabdus* genomes. For the genus *Photorhabdus*, the glidobactin BGC, which encodes a potent eukaryotic proteasome inhibitor (Stein *et al*., 2012) was conserved in 15 of the 16 genomes studied.

5.7 Challenges of specialized metabolism in bacterial chemotaxonomy

The application of specialized metabolism to bacterial classification requires either the detection of compounds or BGCs that have been recognized as consistent features of a specific lineage. With both approaches, there are challenges that need to be considered. Given that compound production is highly dependent on growth conditions, (Huang *et al*., 2001; Machado *et al*., 2017), the cultivation and extraction methods must be standardized if the results are to be comparable. This challenge is magnified when comparisons are made across laboratories and also includes the standardization of extraction protocols. Similarly, the analytical methods employed can affect the compounds detected. In the case of mass spectrometry is used, the data acquisition protocols, ionization source, and detector can all affect what compounds are detected.

The creation of public data repositories for mass spectrometry data (Wang *et al*., 2016), nuclear magnetic resonance spectroscopy data (Reher *et al*., 2020), paired genomic and metabolomics data (Schorn *et al*., 2021), and more traditional databases such as the Natural Products Atlas (van Santen *et al*., 2021), the Collection of Open of Natural products database (Sorokina *et al*., 2021) and the LOTUS initiative for Open Natural Products database (Rutz *et al*., 2021) can all facilitate the use of specialized metabolites in chemotaxonomy. However, the effectiveness of these databases in many cases is dependent on community participation.

In addition to BGC distributions, gene regulation represents another level of phenotypic differentiation that must be considered (Konstantinidis *et al*., 2008; Buckley & Roberts, 2007). In the case of specialized metabolism, incongruences are often observed between BGC distributions and metabolite production (Nett *et al*., 2009; Rutledge & Challis, 2015). This can be due to subtle differences in regulatory elements that are not readily apparent when comparing gene clusters (Amos *et al*., 2017). A BGC that is differentially regulated across species could facilitate functional differentiation yet would not be apparent based simply on standard BGC comparisons. In addition to differential regulation, genetic variations within a gene cluster family, from the presence/absence of tailoring enzymes (Busch *et al*., 2019) to point mutations that affect substrate specificities, or differential enzymatic reactions can affect the structures of specialized metabolites (Miller *et al*., 2017; Crüsemann, 2021). For example, the biosynthetic pathway for the glycosylated polyketide rosamicin includes module-skipping, resulting in the production of the linear polyketides salinipyrones and pacificanones in addition to rosamicin (Oh *et al*., 2008; Awakawa *et al*., 2015). The effects of subtle differences in gene content and organization can be difficult to predict at the compound level thus creating challenges when using BGCs as taxonomic markers. This can be especially problematic if the products are unknown or if their biosynthesis is poorly understood.

Horizontal gene transfer must also be considered. This process is prevalent in microbial communities (Cohen, *et al*., 2011), can play a key role in speciation (Lawrence, 1999), and helps organisms adapt to new ecological niches by acquiring novel traits (Cohen, *et al*., 2011; Gogarten & Townsend, 2005). While some BGCs are conserved at the species level, others can be distributed among phylogenetically distant species (Fischbach *et al*, 2008). Given this, it's not surprising that similar compounds can be detected across diverse taxa. Thus, it's not simply the detection of a

BGC or compound that is informative in a chemotaxonomic sense but the consistency with which they are detected within a specific lineage. Even in cases where a BGC may have been acquired by HGT, there is evidence that these BGCs can subsequently become fixed at the species level (Chase *et al*., 2020). Following this logic, more recent horizontal gene transfer events would simply appear as noise among closely related strains if there was insufficient time for selection to drive their distribution. Alternatively, more ancient acquisition events can result in fixation as shown with rifamycin in *Salinispora arenicola* or napridiomycins and marinones in the Mar4 *Streptomyces* (Jensen *et al*., 2007; Gallagher *et al*., 2010; Bauermeister *et al*., 2018).

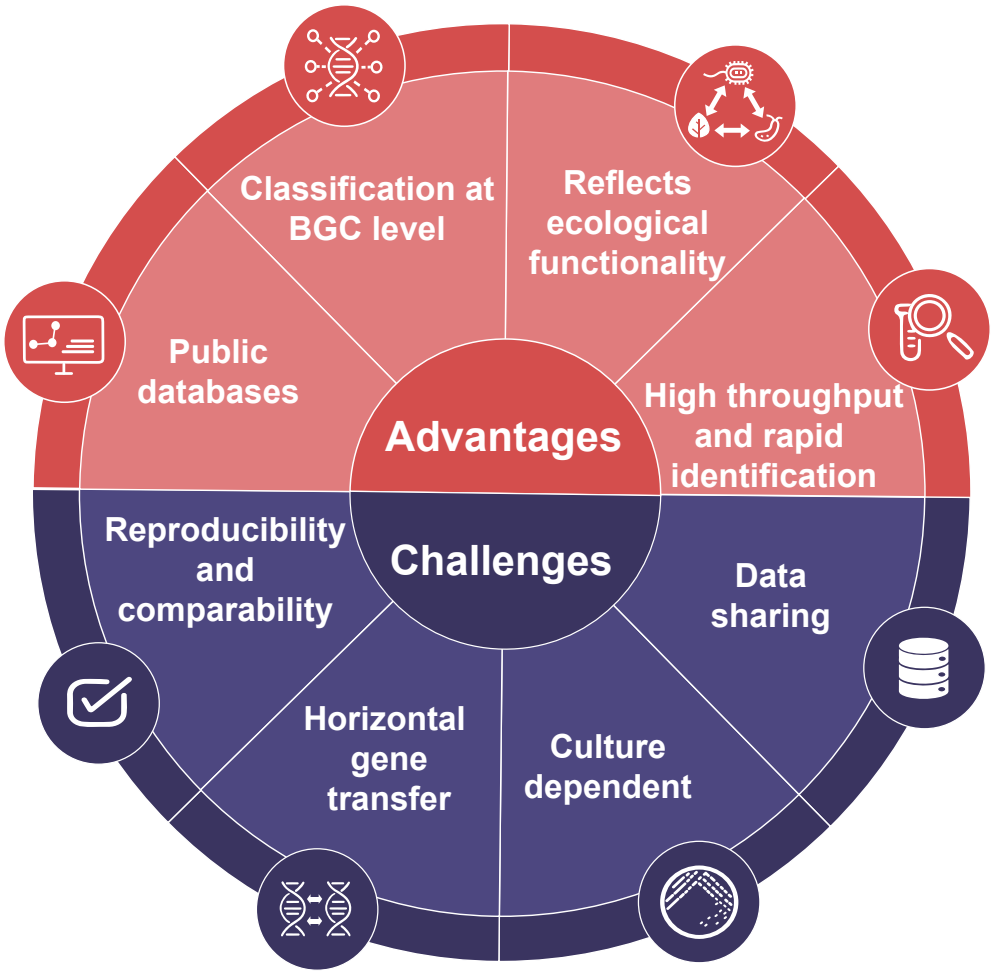

Figure 5.3 Advantages and challenges of implementing specialized metabolism in bacterial chemotaxonomy.

5.8 Summary

While specialized metabolites have played an important role in plant taxonomy, they have largely been studied for their biomedical properties in bacteria. Based on mounting evidence, there appear to be opportunities to expand the applications of chemotaxonomy to include bacterial specialized metabolism. While this may only apply to certain taxa, evidence that compounds or gene clusters are fixed within specific lineages indicates they could be useful taxonomic markers. The ability to recognize the phylogenetic distributions of these traits is highly dependent on the

number of strains studied, their relationships to each other, and the extent to which they use specialized metabolites to mediate ecological interactions.

While species concepts for bacteria remain controversial, the rapid expansion of genomic information has provided rapid methods to assess diversity. Identifying functionally distinct units of diversity from sequence data however remains challenging. If specialized metabolites represent niche defining traits, they could be used to help resolve ecologically meaningful groups that might be otherwise obscured when using sequence data alone. The aim of including specialized metabolism in bacterial classification is to use these features as complementary metrics for more robust species descriptions in the same way they have been used with plants, fungi and other organisms. A chemotaxonomy based on specialized metabolism offers an opportunity for an enhanced polyphasic approach for bacterial taxonomy, especially for taxa where species delimitations have been a challenge and added incentives to resolve the ecological functions mediated by these compounds.

5.9 Acknowledgements

Chapter 5 is coauthored with Alexander B. Chase and Paul R. Jensen. The dissertation author was the primary investigator and author of this chapter.

5.10 References

Abdel-Aziz, S., Elsoud, M.M., & Anise, A.A. (2017). Microbial Biosynthesis: A Repertory of Vital Natural Products. In: Mihai Grumezescu A. & A. M. Holban.Handbook of Food Bioengineering, Food Biosynthesis. Academic Press.Pages 25-54.

Adamek, M., Alanjary, M., Sales-Ortells, H., Goodfellow, M., Bull, A. T., Winkler, A., Wibberg, D., Kalinowski, J., & Ziemert, N. (2018). Comparative genomics reveals phylogenetic distribution patterns of secondary metabolites in Amycolatopsis species. *BMC Genomics*, *19*(1), 1–15.

Ahearne, A.T., Albataineh, H., Dowd, S.E., & Stevens, D.C. (2021). Assessment of Evolutionary Relationships for Prioritization of Myxobacteria for Natural Product Discovery. *Microorganisms, 9*.

Amos, G.C., Awakawa, T., Tuttle, R.N., Letzel, A., Kim, M.C., Kudo, Y., Fenical, W., S. Moore, B., & Jensen, P.R. (2017). Comparative transcriptomics as a guide to natural product discovery and biosynthetic gene cluster functionality. *Proceedings of the National Academy of Sciences, 114*, E11121 - E11130.

Andryukov, B.G., Mikhailov, V.V., & Besednova, N.N. (2019). The Biotechnological Potential of Secondary Metabolites from Marine Bacteria. *Journal of Marine Science and Engineering*.

Awakawa, T., Crüsemann, M., Munguia, J., Ziemert, N., Nizet, V., Fenical, W., & Moore, B.S. (2015). Salinipyrone and Pacificanone Are Biosynthetic By-products of the Rosamicin Polyketide Synthase. *ChemBioChem, 16*.

Bacon, C.W., Porter, J.K., Norred, W.P., & Leslie, J.F. (1996). Production of fusaric acid by Fusarium species. *Applied and Environmental Microbiology, 62*, 4039 - 4043.

Baltz, R.H. (2016). Gifted microbes for genome mining and natural product discovery. *Journal of Industrial Microbiology & Biotechnology, 44*, 573-588.

Bauermeister, A., Pereira F, Grilo IR, Godinho CC, Paulino M, Almeida V, Gobbo-Neto L, Prieto-Davó A, Sobral RG, Lopes NP, Gaudêncio SP (2019). Intra-clade metabolomic profiling of MAR4 Streptomyces from the Macaronesia Atlantic region reveals a source of anti-biofilm metabolites. *Environ Microbiol*.,21(3):1099-1112.

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., & Costa-Lotufo, L. V. (2018). Metabolomic Fingerprinting of Salinispora From Atlantic Oceanic Islands. *Frontiers in Microbiology*, *9*, 1–13.

Benedict, R.G. (1970). Chemotaxonomic Relationships Among the Basidiomycetes. *Advances in Applied Microbiology, 13*, 1-23.

Bergquist, P. R., & Wells, R. J. (1983). Chemotaxonomy of the Porifera: The Development and Current Status of the Field. In *Marine Natural Products* (pp. 1–50)

Bode, H. B., Bethe, B., Höfs, R., & Zeeck, A. (2002). Big effects from small changes: Possible ways to explore nature's chemical diversity. *ChemBioChem*, *3*(7), 619–627

Bose, U., Hewavitharana, A.K., Vidgen, M.E., Ng, Y.K., Shaw, P.N., Fuerst, J.A., & Hodson, M.P. (2014). Discovering the Recondite Secondary Metabolome Spectrum of Salinispora Species: A Study of Inter-Species Diversity. *PLoS ONE, 9*.

Buchanan, R. E. (1955). TAXONOMY. *Annu. Rev. Microbiol.*, *9*(1), 1–20

Buckley, M., R. J. Roberts (2007). Reconciling microbial systematics and genomics. Amer Acad Microbiol Rep.

Busch, J.C., Agarwal, V., Schorn, M.A., Machado, H., Moore, B.S., Rouse, G.W., Gram, L., & Jensen, P.R. (2019). Diversity and distribution of the bmp gene cluster and its Polybrominated products in the genus Pseudoalteromonas. *Environmental Microbiology, 21*, 1575–1585.

Ceniceros, A., Dijkhuizen, L., Petrusma, M., & Medema, M.H. (2017). Genome-based exploration of the specialized metabolic capacities of the genus Rhodococcus. *BMC Genomics, 18*

Chase, A.B., Sweeney, D., Muskat, M.N., Guillén Matus, D. G., & Jensen, P.R. (2020). Vertical inheritance governs biosynthetic gene cluster evolution and chemical diversification. *bioRxiv*.

Chevrette, M.G., & Handelsman, J. (2021). Needles in haystacks: reevaluating old paradigms for the discovery of bacterial secondary metabolites. *Natural product reports*.

Choudoir, M., Rossabi, S., Gebert, M., Helmig, D., & Fierer, N. (2019). A Phylogenetic and Functional Perspective on Volatile Organic Compound Production by Actinobacteria . *MSystems*, *4*(2), 1–15.

Cimermancic, P., Medema, M. H., Claesen, J., Kurita, K., Wieland Brown, L. C., Mavrommatis, K., Pati, A., Godfrey, P. A., Koehrsen, M., Clardy, J., Birren, B. W., Takano, E., Sali, A., Linington, R. G., & Fischbach, M. A. (2014). Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters. *Cell*, *158*(2), 412–421.

Cohan, F. (2002). What are bacterial species? *Annual review of microbiology, 56*, 457-87.

Crüsemann, M. (2021). Coupling Mass Spectral and Genomic Information to Improve Bacterial Natural Product Discovery Workflows. *Marine Drugs, 19*.

Doroghazi, J. R., Albright, J. C., Goering, A. W., Ju, K. S., Haines, R. R., Tchalukov, K. A., Labeda, D. P., Kelleher, N. L., & Metcalf, W. W. (2014). A roadmap for natural product discovery based on large-scale genomics and metabolomics. *Nature Chemical Biology*, *10*(11), 963–968.

Duncan, K. R., Dorrestein, P. C., Jensen, P. R., Duncan, K. R., Cru, M., Lechner, A., Sarkar, A., Li, J., Ziemert, N., & Wang, M. (2015). Molecular Networking and Pattern-Based Genome Mining Improves Discovery of Biosynthetic Gene Clusters and their Products from Salinispora Species. *Chemistry & Biology*, *22*(4), 460–471.

Embley T.M., E. Stackebrandt (1994). The molecular phylogeny and systematics of the actinomycetes. Annu Rev Microbiol 48:257–289

Erpenbeck, D., & Soest, R.W. (2006). Status and Perspective of Sponge Chemosystematics. *Marine Biotechnology, 9*, 2-19.

Figueiredo C., M. Baptista, I. C. Rosa, opes, A. R. Lopes, G. Dionísio, R.J.M. Rocha, I.C.S. Cruz, R.K.P. Kikuchi, N. Simões, M.C. Leal, I. Tojeira, N. Bandarra, R. Calado, R. Rosa.(2017). 3D chemoecology and chemotaxonomy of corals using fatty acid biomarkers: Latitude, longitude and depth. *Biochem Syst Ecol*. 70:35-42. doi:10.1016/j.bse.2016.10.016

Fischbach, M. A., Walsh, C. T., & Clardy, J. (2008). The evolution of gene collectives: How natural selection drives chemical innovation. *Proceedings of the National Academy of Sciences*, *105*(12), 4601–4608.

Freel, K. C., Nam, S.-J., Fenical, W., & Jensen, P. R. (2011). Evolution of secondary metabolite genes in three closely related marine actinomycete species. *Applied and Environmental Microbiology*, *77*(20), 7261–7270.

Frisvad, J. C., Andersen, B., & Thrane, U. (2007). The use of secondary metabolite profiling in chemotaxonomy of filamentous fungi. *Mycological Research*, *112*(2), 231–240.

Gallagher, K. A., & Jensen, P. R. (2015). Genomic insights into the evolution of hybrid isoprenoid biosynthetic gene clusters in the MAR4 marine streptomycete clade. *BMC Genomics*, *16*(1), 1–13.

Gallagher, K. A., & Jensen, P. R. (2015). Genomic insights into the evolution of hybrid isoprenoid biosynthetic gene clusters in the MAR4 marine streptomycete clade. *BMC Genomics*, *16*(1), 1–13.

Gallagher, K. A., Fenical, W., & Jensen, P. R. (2010). Hybrid isoprenoid secondary metabolite production in terrestrial and marine actinomycetes. *Current Opinion in Biotechnology*, *21*(6), 794–800.

Gallagher, K. A., Rauscher, K., Ioca, L. P., & Jensen, P. R. (2013). Phylogenetic and chemical diversity of a hybrid-isoprenoid-producing streptomycete lineage. *Applied and Environmental Microbiology*, *79*(22), 6894–6902.

Gevers, D., Cohan, F., Lawrence, J.G., Spratt, B.G., Coenye, T., Feil, E.J., Stackebrandt, E., Peer, Y.V., Vandamme, P.A., Thompson, F.L., & Swings, J. (2005). Re-evaluating prokaryotic species. *Nature Reviews Microbiology, 3*, 733-739.

Gogarten, J.P., & Townsend, J.P. (2005). Horizontal gene transfer, genome innovation and evolution. *Nature Reviews Microbiology, 3*, 679-687.

Griffin, W.J., & Lin, G.D. (2000). Chemotaxonomy and geographical distribution of tropane alkaloids. *Phytochemistry, 53 6*, 623-37.

Habener, L.J., Hooper, J.N., & Carroll, A.R. (2016). Chemical and Biological Aspects of Marine Sponges from the Family Mycalidae. *Planta medica, 82 9-10*, 816-31 .

Hamana, K., & Matsuzaki, S. (1992). Polyamines as a chemotaxonomic marker in bacterial systematics. *Critical Reviews in Microbiology*, *18*(4), 261–283.

Helaly, S.E., Thongbai, B., & Stadler, M. (2018). Diversity of biologically active secondary metabolites from endophytic and saprotrophic fungi of the ascomycete order Xylariales. *Natural product reports, 35 9*, 992-1014

Hewavitharana, A. K., Shaw, P. N., Kim, T. K., & Fuerst, J. a. (2007). Screening of rifamycin producing marine sponge bacteria by LC-MS-MS. *Journal of Chromatography. B, Analytical Technologies in the Biomedical and Life Sciences*, *852*(1–2), 362–366.

Hoffmann, T., Krug, D., Bozkurt, N., Duddela, S., Jansen, R., Garcia, R., Gerth, K., Steinmetz, H., & Müller, R. (2018). Correlating chemical diversity with taxonomic distance for discovery of natural products in myxobacteria. *Nature Communications*, *9*(1), 803.

Huang J, Lih C-J, Pan K-H, Cohen SN.(2001) Global analysis of growth phase responsive gene expression and regulation of antibiotic biosynthetic pathways in Streptomyces coelicolor using DNA microarrays. Genes Dev. ; 15:3183–3192.

Imbs, A.B., Demidkova, D.A., Latypov, Y.Y., & Pham, L.Q. (2007). Application of Fatty Acids for Chemotaxonomy of Reef-Building Corals. *Lipids, 42*, 1035-1046

Imhoff, J.F., Labes, A., & Wiese, J. (2011). Bio-mining the microbial treasures of the ocean: new natural products. *Biotechnology advances, 29 5*, 468-82

Jensen, P. R. (2016). Natural Products and the Gene Cluster Revolution. In *Trends in Microbiology* (Vol. 24, Issue 12, pp. 968–977). Elsevier Ltd.

Jensen, P. R., Moore, B. S., & Fenical, W. (2015). The marine actinomycete genus Salinispora: A model organism for secondary metabolite discovery. In *Natural Product Reports* (Vol. 32, Issue 5, pp. 738–751). Royal Society of Chemistry.

Jensen, P. R., Williams, P. G., Oh, D.C., Zeigler, L., & Fenical, W. (2007). Species-specific secondary metabolite production in marine actinomycetes of the genus Salinispora. *Applied and Environmental Microbiology*, *73*(4), 1146–1152.

Jensen, P.R. (2016). Natural Products and the Gene Cluster Revolution. *Trends in microbiology, 24 12*, 968-977

Jones, C., Webster, G., Mullins, A.J., Jenner, M., Bull, M.J., Dashti, Y., Spilker, T., Parkhill, J., Connor, T.R., Lipuma, J.J., Challis, G.L., & Mahenthiralingam, E. (2021). Kill and cure: genomic phylogeny and bioactivity of Burkholderia gladioli bacteria capable of pathogenic and beneficial lifestyles. *Microbial Genomics, 7*.

Juhas, M. (2019). Genomic Islands and the Evolution of Multidrug-Resistant Bacteria. In: Villa T. G. & M. Viñas (eds).Horizontal Gene Transfer: Breaking Borders Between Living Kingdoms Springer . Pages 143-154

Kautsar, S., Blin, K., Shaw, S., Navarro-Muñoz, J.C., Terlouw, B.R., van der Hooft, J.J., van Santen, J.A., Tracanna, V., Suarez Duran, H.G., Pascal Andreu, V., Sélem-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., & Medema, M.H. (2020). MIBiG 2.0: a repository for biosynthetic gene clusters of known function. *Nucleic Acids Research, 48*, D454 - D458.

Kim, T. K., Hewavitharana, A. K., Shaw, P. N., & Fuerst, J. A. (2006). Discovery of a new source of rifamycin antibiotics in marine sponge actinobacteria by phylogenetic prediction. *Applied and Environmental Microbiology*, *72*(3), 2118–2125.

Konstantinidis, K. T., & Tiedje, J. M. (2005). Genomic insights that advance the species definition for prokaryotes. *Proceedings of the National Academy of Sciences*, *102*(7), 2567–2572.

Lawrence J.G. (1999) Gene transfer, speciation, and the evolution of bacterial genomes. Curr Opin Microbiol 2:519–523

Lechevalier, M. P., De Bievre, C., & Lechevalier, H. (1977). Chemotaxonomy of aerobic Actinomycetes: Phospholipid composition. *Biochemical Systematics and Ecology*, *5*(4), 249–260.

Liang, L., Wang, J., Ying, X., Wang, K., Hu, Y., Tian, R., Yang, Bo., Lai, Q., Li, Y., Zhang, W., Shao, Z., Lam, H., Qian, P. Y. (2014). A High-Resolution LC-MS-Based Secondary Metabolite Fingerprint Database of Marine Bacteria. *Sci Rep.* 4, 6537

Lu, L., Wang, J., Xu, Y., Wang, K., Hu, Y., Tian, R., Yang, B., Lai, Q., Li, Y., Zhang, W., Shao, Z., Lam, H., & Qian, P. Y. (2014). A high-resolution LC-MS-based secondary metabolite fingerprint database of marine bacteria. *Scientific Reports*, *4*, 1–7.

Mao, D., Okada, B.K., Wu, Y., Xu, F., & Seyedsayamdost, M.R. (2018). Recent advances in activating silent biosynthetic gene clusters in bacteria. *Current opinion in microbiology, 45*, 156-163

Matsuda, S., Adachi, K., Matsuo, Y., Nukina, M., & Shizuri, Y. (2009). Salinisporamycin, a novel metabolite from salinispora arenicora. *Journal of Antibiotics*, *62*(9), 519–526.

Miller, I.J., Chevrette, M.G., & Kwan, J.C. (2017). Interpreting Microbial Biosynthesis in the Genomic Age: Biological and Practical Considerations. *Marine Drugs, 15*.

Miller, L.T. (1982). Single derivatization method for routine analysis of bacterial whole-cell fatty acid methyl esters, including hydroxy acids. *Journal of Clinical Microbiology, 16*, 584 - 586.

Mordarska, H., Mordarski, M., & Goodfellow, M. (1972). Chemotaxonomic characters and classification of some nocardioform bacteria. *Journal of General Microbiology*, *71*(1), 77–86.

Mosunova, O., Navarro-Muñoz, J.C., & Collemare, J. (2021). The Biosynthesis of Fungal Secondary Metabolites: From Fundamentals to Biotechnological Applications. In: Zaragoza O. & A. Casadevall (eds). Encyclopedia of Mycology. Elsevier. Pages 458-476

Mullins, A.J., & Mahenthiralingam, E. (2021). The Hidden Genomic Diversity, Specialized Metabolite Capacity, and Revised Taxonomy of Burkholderia Sensu Lato. *Frontiers in Microbiology, 12*.

Nett, M., Ikeda, H., & Moore, B.S. (2009). Genomic basis for natural product biosynthetic diversity in the actinomycetes. *Natural product reports, 26 11*, 1362-84

Offret, C., Desriac, F., Le Chevalier, P., Mounier, J., Jégou, C., & Fleury, Y. (2016). Spotlight on Antimicrobial Metabolites from the Marine Bacteria Pseudoalteromonas: Chemodiversity and Ecological Significance. *Marine Drugs, 14*.

Oh, D., Gontang, E.A., Kauffman, C.A., Jensen, P.R., & Fenical, W. (2008). Salinipyrones and pacificanones, mixed-precursor polyketides from the marine actinomycete Salinispora pacifica. *Journal of natural products, 71 4*, 570-5 .

Park, C.J., Smith, J.T., & Andam, C.P. (2019). Horizontal Gene Transfer and Genome Evolution in the Phylum Actinobacteria. In: Villa T. G. & M. Viñas (eds).Horizontal Gene Transfer: Breaking Borders Between Living Kingdoms Springer . Pages 155-174

Parks, D.H., Chuvochina, M., Waite, D.W., Rinke, C., Skarshewski, A., Chaumeil, P., & Hugenholtz, P.B. (2018). A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nature Biotechnology, 36*, 996-1004.

Paulsen, S.S., Isbrandt, T., Kirkegaard, M., Buijs, Y., Strube, M.L., Sonnenschein, E.C., Larsen, T.O., & Gram, L. (2020). Production of the antimicrobial compound tetrabromopyrrole and the Pseudomonas quinolone system precursor, 2-heptyl-4-quinolone, by a novel marine species Pseudoalteromonas galatheae sp. nov. *Scientific Reports, 10*.

Penn, K., Jenkins, C.A., Nett, M., Udwary, D.W., Gontang, E.A., McGlinchey, R.P., Foster, B., Lapidus, A.L., Podell, S., Allen, E.E., Moore, B.S., & Jensen, P.R. (2009). Genomic islands link secondary metabolism to functional adaptation in marine Actinobacteria. *The ISME journal, 3*, 1193 - 1203.

Pinhal S, Ropers D., Geiselmann J., de Jong H. (2019). Acetate metabolism and the inhibition of bacterial growth by acetate. J Bacteriol 201:e00147-19.

Polizzi, V., Adams, A., Malysheva, S.V., de Saeger, S., Van Peteghem, C.H., Moretti, A., Picco, A.M., & de Kimpe, N. (2012). Identification of volatile markers for indoor fungal growth and chemotaxonomic classification of Aspergillus species. *Fungal biology, 116 9*, 941-53

Reher, R., Kim, H.W., Zhang, C., Mao, H.H., Wang, M., Nothias, L., Caraballo-Rodríguez, A.M., Glukhov, E., Teke, B., Leão, T.F., Alexander, K.L., Duggan, B.M., Van Everbroeck, E.L., Dorrestein, P.C., Cottrell, G., & Gerwick, W.H. (2020). A Convolutional Neural Network-Based Approach for the Rapid Characterization of Molecularly Diverse Natural Products. *Journal of the American Chemical Society*.

Rohde S. & P. Schupp. (2018). Spatial and Temporal Variability in Sponge Chemical Defense. In: Puglisi M.P & Becerro M. A. (eds) Chemical Ecology. 373-396.

Rosselló-Móra, R. (2012). Towards a taxonomy of Bacteria and Archaea based on interactive and cumulative data repositories. *Environmental microbiology, 14 2*, 318-34 .

Rutledge, P.J., & Challis, G.L. (2015). Discovery of microbial natural products by activation of silent biosynthetic gene clusters. *Nature Reviews Microbiology, 13*, 509-523.

Rutz, A., Sorokina, M., Galgonek, J., Mietchen, D., Willighagen, E., Gaudry, A., Graham, J.G., Stephan, R., Page, R.D., Vondrášek, J., Steinbeck, C., Pauli, G.F., Wolfender, J., Bisson, J., & Allard, P. (2021). The LOTUS Initiative for Open Natural Products Research: Knowledge Management through Wikidata. *bioRxiv*.

Saini, V., Raghuvanshi, S., Talwar, G. P., Ahmed, N., Khurana, J. P., Hasnain, S. E., Tyagi, A. K., & Tyagi, A. K. (2009). Polyphasic Taxonomic Analysis Establishes Mycobacterium indicus pranii as a Distinct Species. *PLoS ONE*, *4*(7).

Schleifer, K. H. (2009). Classification of Bacteria and Archaea: Past, present and future. *Systematic and Applied Microbiology*, *32*(8)

Schorn, M.A., Verhoeven, S., Ridder, L., Huber, F., Acharya, 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.D., 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., 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., Hempel, B., Hur, J., Iorio, M., Jensen, P.R., Kang, K.B., Kaysser, L., Kelleher, N.L., Kim, C.S., Kim, K., Koester, I., König, G.M., Leão, T.F., Lee, S.R., Lee, 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.N., Nothias, L., O'Neill, E.C., Parkinson, E.I., Petráš, D., Piel, J., Pierce, E.C., Pires, K., Reher, R., Romero, D., Roper, M.C., Rust, M., Saad, H.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.,

& van der Hooft, J.J. (2021). A community resource for paired genomic and metabolomic data mining. *Nature Chemical Biology, 17*, 363 - 368.

Shi, Y., Hirschmann, M., Shi, Y., Ahmed, S., Abebew, D., Tobias, N.J., Grün, P., Crames, J.J., Pöschel, L., Kuttenlochner, W., Richter, C., Herrmann, J., Müller, R., Thanwisai, A., Pidot, S.J., Stinear, T.P., Groll, M., Kim, Y., & Bode, H. (2022). Global analysis of biosynthetic gene clusters reveals conserved and unique natural products in entomopathogenic nematode-symbiotic bacteria. *Nature Chemistry*, 1 - 12.

Singh, R., P. Geetanjali (2018). Chemotaxonomy of Medicinal Plants: Possibilities and Limitations. In: Subhash C., Mandal V., Konishi T. (eds)*Natural Products and Drug Discovery: An Integrated Approach.* Elsevier. pp. 119–136. Elsevier.

Singh, R., P. Geetanjali (2019). Chemotaxonomic Significance of Alkaloids in Plants. In: Ramawat K. (eds) Biodiversity and Chemotaxonomy. Sustainable Development and Biodiversity, vol 24. Springer pp 121-136

Sorokina, M., Merseburger, P., Rajan, K., Yirik, M.A., & Steinbeck, C. (2021). COCONUT online: Collection of Open Natural Products database. *Journal of Cheminformatics, 13*.

Stackebrandt E., P. Schumann. (2006). Introduction to the taxonomy of Actinobacteria. In: Dworkin M, Falkow S, Rosenberg E *et al* (eds) The prokaryotes, Archaea. Bacteria: Firmicutes, Actinomycetes. Springer. pp 297–321

Stein, M.L., Beck, P., Kaiser, M., Dudler, R., Becker, C.F., & Groll, M. (2012). One-shot NMR analysis of microbial secretions identifies highly potent proteasome inhibitor. *Proceedings of the National Academy of Sciences, 109*, 18367 - 18371.

Tang, H., Wang, X., Xu, L., Ran, X., Li, X., Chen, L., Zhao, X., Deng, H., & Liu, X. (2016). Establishment of local searching methods for orbitrap-based high throughput metabolomics analysis. *Talanta, 156-157*, 163-171.

Thompson, C.C., Amaral, G.R., Campeão, M.E., Edwards, R.A., Polz, M.F., Dutilh, B.E., Ussery, D.W., Sawabe, T., Swings, J., & Thompson, F.L. (2015). Microbial taxonomy in the post-genomic era: Rebuilding from scratch? *Archives of Microbiology, 197*, 359-370.

Tindall, B.J., Rosselló-Móra, R., Busse, H., Ludwig, W., & Kämpfer, P. (2010). Notes on the characterization of prokaryote strains for taxonomic purposes. *International journal of systematic and evolutionary microbiology, 60 Pt 1*, 249-66.

Tobias, N.J., Parra-Rojas, C., Shi, Y., Shi, Y., Simonyi, S., Thanwisai, A., Vitta, A., Chantratita, N., Hernández-Vargas, E.A., & Bode, H.B. (2019). Focused natural product elucidation by prioritizing high-throughput metabolomic studies with machine learning. *bioRxiv*.

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

Umoh, O. T. (2020). Chemotaxonomy: The Role of Phytochemicals in Chemotaxonomic Delineation of Taxa. *Asian Plant Research Journal*, *5*(1), 43–52.

Undabarrena, A., Valencia, R., Cumsille, A.G., Zamora-Leiva, L., Castro-Nallar, E., Barona-Gómez, F., & 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. *Microbial Genomics, 7*.

van Santen, J.A., Poynton, E.F., Iskakova, D., McMann, E., Alsup, T.A., Clark, T.N., Fergusson, C.H., Fewer, D.P., Hughes, A.H., McCadden, C.A., Parra, J., Soldatou, S., Rudolf, J.D., Janssen, E.M., Duncan, K.R., & Linington, R.G. (2021). The Natural Products Atlas 2.0: a database of microbially-derived natural products. *Nucleic Acids Research, 50*, D1317 - D1323.

Viljoen A.M., B.E. van Wyk, F.R. van Heerden. (1998). Distribution and chemotaxonomic significance of flavonoids in Aloe (Asphodelaceae), Plant Syst. Evol. 211 (1) 31-42.

Wang, M., Carver, J., Phelan, V.V., Sanchez, L.M., Garg, N., Peng, Y., Nguyen, D.D., Watrous, J.D., Kapono, C.A., Luzzatto-Knaan, T., Porto, C., Bouslimani, A., Melnik, A.V., Meehan, M.J., Liu, W., Crüsemann, M., Boudreau, P.D., Esquenazi, E., Sandoval-Calderón, M., Kersten, R.D., Pace, L.A., Quinn, R.A., Duncan, K.R., Hsu, C., Floros, D.J., Gavilán, R.G., Kleigrewe, K., Northen, T.R., Dutton, R.J., Parrot, D., Carlson, E.E., Aigle, B., Michelsen, C.F., Jelsbak, L., Sohlenkamp, C., Pevzner, P.A., Edlund, A., McLean, J.S., Piel, J., Murphy, B.T., Gerwick, L., Liaw, C., Yang, Y., Humpf, H., Maansson, M., Keyzers, R.A., Sims, A.C., Johnson, A.R., Sidebottom, A.M., Sedio, B.E., Klitgaard, A.K., Larson, C.B., P, C.A., Torres-Mendoza, D., Gonzalez, D.J., Silva, D.B., Marques, L.M., Demarque, D.P., Pociūtė, E., O'Neill, E.C., Briand, E., Helfrich, E.J., 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.S., Silva, R.R., Mascuch, S.J., Tomasi, S., Jenkins, S., Macherla, V.R., Hoffman, T., Agarwal, V., Williams, P.G., Dai, J., Neupane, R.P., Gurr, J.R., Rodriguez, A.M., Lamsa, A., Zhang, C., Dorrestein, K., Duggan, B.M., Almaliti, J., Allard, P., Phapale, P., Nothias, L., Alexandrov, T., Litaudon, M., Wolfender, J., Kyle, J.E., Metz, T.O., Peryea, T., Nguyen, D., Vanleer, D., Shinn, P., Jadhav, A., Müller, R., Waters, K.M., Shi, W., Liu, X., Zhang, L., Knight, R., Jensen, P.R., Palsson, B.O., Pogliano, K., Linington, R.G., Gutiérrez, M., Lopes, N.P., Gerwick, W.H., Moore, B.S., Dorrestein, P.C., & Bandeira, N. (2016). Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nature Biotechnology, 34*, 828-837.

Wei, W., Sung, P., Duh, C.Y., Chen, B., Sheu, J.H., & Yang, N.S. (2013). Anti-Inflammatory Activities of Natural Products Isolated from Soft Corals of Taiwan between 2008 and 2012. *Marine Drugs, 11*, 4083 - 4126.

Whalley, A.J., & Edwards, R.L. (1995). Secondary metabolites and systematic arrangement within the Xylariaceae. *Botany, 73*, 802-810.

Wink, M. (2003). Evolution of secondary metabolites from an ecological and molecular phylogenetic perspective. *Phytochemistry*, *64*(1), 3–19.

Wink, M., Botschen, F., Gosmann, C., Schäfer, H., & Waterman, P. G. (2010). Chemotaxonomy Seen from a Phylogenetic Perspective and Evolution of Secondary Metabolism. In *Annual Plant Reviews online* (Vol. 40).

Wu, Q., Sun, J., Chen, J., Zhang, H., Guo, Y., & Wang, H. (2018). Terpenoids from Marine Soft Coral of the Genus Lemnalia: Chemistry and Biological Activities. *Marine Drugs, 16*.

Zain, M. (2010). Biochemical markers in taxonomy of the genus Fusarium. *Research Journal of Agriculture and Biological Sciences, 6*, 1-7.

Zampieri, M., Sekar, K., Zamboni, N., & Sauer, U.H. (2017). Frontiers of high-throughput metabolomics. *Current opinion in chemical biology, 36*, 15-23.

Zervas, A., Aggerbeck, M.R., Allaga, H., Güzel, M., Hendriks, M., Jonuškienė, I., Kedves, O., Kupelı, A.E., Lamovšek, J., Mülner, P., Munday, D., Namli, S., Samut, H., Tomičić, R., Tomičić, Z., Yeni, F., Zghal, R.Z., Zhao, X., Sanchis-Borja, V., & Hendriksen, N.B. (2020). Identification and Characterization of 33 Bacillus cereus sensu lato Isolates from Agricultural Fields from Eleven Widely Distributed Countries by Whole Genome Sequencing. *Microorganisms, 8*.

Zidorn, C. (2019). Plant chemophenetics − A new term for plant chemosystematics/plant chemotaxonomy in the macro-molecular era. *Phytochemistry*, *163*(March), 147–148.

Ziemert, N., Lechner, A., Wietz, M., Millán-aguiñaga, N., Chavarria, K. L., & Jensen, P. R. (2014). *Diversity and evolution of secondary metabolism in the marine actinomycete genus Salinispora*. *2014*(3).

Chapter 6

The case of missing pigment: an investigation into color variation in *Salinispora*

6.1 Abstract

Colors in nature are vital. In bacteria, pigments can protect against ultraviolet radiation, oxidants, extreme temperature, and antimicrobial compounds. Further, they can aid in iron acquisition and pathogenesis. Carotenoid pigments are widespread among bacteria and undoubtedly play important ecological roles. Sioxanthin is a glycosylated carotenoid that gives *Salinispora* its characteristic bright orange pigmentation. Herein, I assessed the genetic basis for albinism in three strains of two *Salinispora* species that display rare albino phenotypes. In *S. arenicola* strains, a single point mutation in intergenic regions between genes *crtB* and *crtE*, was found, and in *S. vitiensis* several non-synonymous mutations in the *crtI* gene were associated with the loss of pigmentation in these strains. Although extremely rare, albinism in *Salinispora* doesn't seem to affect growth efficiency under laboratory conditions. The ecological outcomes of these mutations remain to be determined.

6.2 Introduction

Colors in nature are vital. From birds that exhibit brightly colored plumage to attract members of the opposite sex, to chameleons camouflaging by adapting to their surrounding's color, to the bright coloration of the poison dart frog warning potential predators to stay away (Liu & Nizet, 2009), colors play an essential role in survival. In bacteria, the production of pigments has attracted attention from the beginning, possibly because pigmentation is one of the most

striking characteristics of microorganisms (Sobin & Stahly, 1941). These hallmark phenotypes not only provide an easy way to describe bacteria, but special emphasis has been placed on their chemical nature and the metabolic roles they carry out within bacterial cells. For non-photosynthetic bacteria, the acquisition of pigments might be related to evolutionary selective pressures beyond the pigment's spectral properties such as light absorbance, reflection, or emission (Liu & Nizet, 2009).

Some functions proposed for microbial pigments are protection against ultraviolet radiation, oxidants, and extreme temperatures (Ramesh *et al*., 2019; Genç *et al.,* 2020) as well as protection from antimicrobial compounds and iron acquisition (Chatfield & Cianciotto, 2007). Particularly in pathogenic bacteria, it has been noted that pigment production enhances pathogenesis by interfering with host immune clearance mechanisms or by exhibiting pro-inflammatory or cytotoxic properties (Liu & Nizet, 2009; Liu et al., 2005).

Carotenoid pigments are widespread among bacteria and undoubtedly play important ecological roles; these pigments are ubiquitously synthesized by photosynthetic and non-photosynthetic bacterial lineages (Phadwal, 2005). Pigment production is a distinctive feature of *Salinispora sp.*; during vegetative growth the colonies exhibit a bright orange pigmentation that becomes darkened during sporulation (Maldonado *et al.,* 2005). The responsible for *Salinispora's* bright orange coloration is the recently discovered sioxanthin; a C40 carotenoid with a sugar at one end of the carbon chain and an aryl group at the other (Richter *et al.*, 2015). There are four different regions within the *Salinispora* genome that contribute to the production of Sioxanthin (fig 6.1). These regions correspond to two gene clusters (*terp1* and *terp2*) and two single genes (*strop_0241* and *strop_0248*) (Penn *et al.*, 2009; Richter *et al*., 2015).

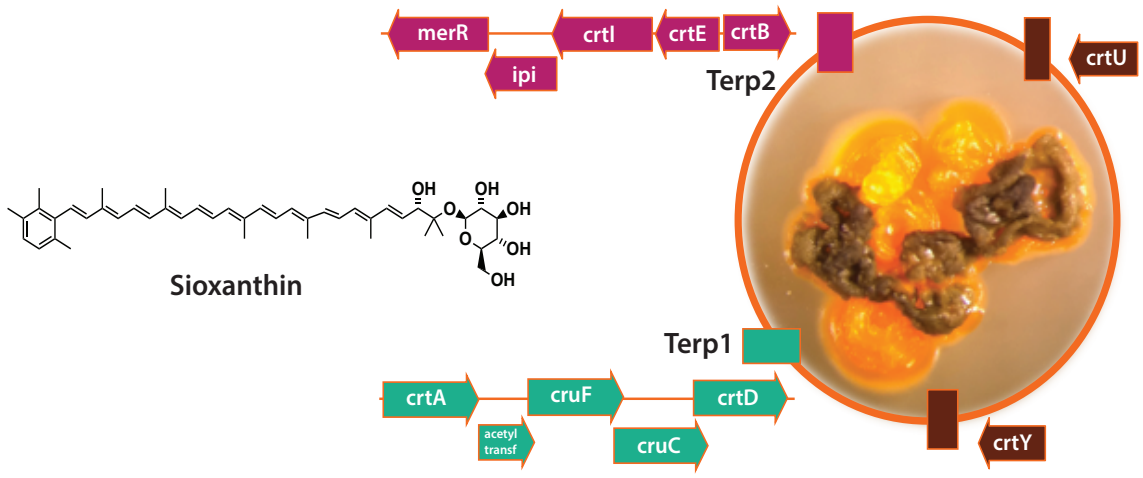

Figure 6.1 Genes involved in the biosynthesis of sioxanthin. The circle around a *Salinispora* colony represents the genome and the location of the two gene clusters Terp1 and Terp2 and the two individual genes *crtY* and *crtU* are indicated.

Throughout the years researching *Salinispora* specialized metabolism, the strains have been cultured under laboratory conditions with nutrient-rich fermentation media without apparent selective pressure. This has led to the isolation of stable populations of four non-pigmented mutant strains of *Salinispora* sp. In other bacteria, it has been observed that the loss in pigmentation equates to a loss in biological activity and can abate the production of specialized metabolites (Song *et al*., 2009; Kalan *et al*., 2013). To determine the causes for this phenotypic differentiation and explore if there are any effects on growth efficiency, comparative genomic analyses were performed on both wild and mutant types of several *Salinispora* strains.

6.3 Methods

6.3.1 Strain cultivation and growth curves

Frozen stocks of six *Salinispora* strains (table 6.1) were inoculated in A1 medium and subsequently inoculated at 1% v/v in 200 ml f A1 medium with the following formulation: soluble starch 10g/L; yeast 4 g/L; peptone 2 g/L and 22 g/L instant ocean mix ® in deionized water.

Fermentations were performed in duplicate using 500 mL Erlenmeyer flasks at 28ºC with shaking at 210 rpm. To avoid colony clumping, around 20 glass beads were added to each flask. During a period of 20 days, 3 mL aliquots were taken every 24 hours and optical density was measured in triplicate at 600 nm (OD600) using a spectrophotometer.

Table 6.1 Salinispora strains with pigmented (wild) and albino (mutant) phenotypes, used in these experiments.

Species	**Strain**	**Type**
Salinispora arenicola	CNH-996 A	Wild-type
	CNH-996 B	mutant
	CNY-861 A	Wild-type
	CNY-861 B	mutant
Salinispora vitiensis	CNS-055 A	Wild-type
	CNS-055 B	mutant

6.3.2 Genomic DNA extraction

Salinispora strains CNY-861 A and B, and CNS-055 B, were cultured in 250 mL Erlenmeyer flasks containing 150 mL of liquid A1 medium for seven days. Two 50 mL aliquots were centrifuged at 10,000 rpm for 5 minutes and the supernatant removed from the cell pellet and frozen at -20°C until DNA extraction. Genomic DNA (gDNA) was extracted following a phenol-chloroform method (Millán-Aguiñaga *et al.,* 2017). The cell pellet was re-suspended in TE Buffer (pH 8.0, 10 mM Tris-HCl, 1.0 mM EDTA) to an OD_{600} of 1.0; lysozyme (100 mg/mL; end concentration 3 mg/mL) and RNAse A (100 mg/mL; end concentration 100 μg/mL) was added and the mixture was incubated at least 80 min at 37°C. After incubation, 10% SDS and Proteinase K (20 mg/mL) was added, and the entire mixture was incubated overnight at 55°C. The next day, 1.5 mL of 5M NaCl and 1 mL 10% CTAB/NaCl 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 4

mL phenol/chloroform/isoamyl alcohol (25:24:1 ratio saturated with 10 mM Tris, pH 8.0, 1 mM EDTA) mixing, centrifuging for 10 min at 10,000 rpm, 4°C, followed by transferring the upper aqueous layer to 4 mL chloroform, twice. The samples were transferred using wide-bore tips to a new tube; DNA was precipitated by adding 0.6 volumes of molecular grade isopropanol, followed by centrifugation for 10 min at 10,000 rpm, 4°C. The precipitated DNA was rinsed with ice-cold 70% molecular grade EtOH and centrifuged for 10,000 rpm for 2 min after which the EtOH was removed and samples air dried; the genomic DNA was then dissolved overnight in TE buffer at 4°C. The gDNA purity was confirmed by measuring the 260/280 nm and 260/230 nm absorbance ratios using a NanoDrop 1000 spectrophotometer. The concentration of the DNA was measured using both a NanoDrop 1000 spectrophotometer and a Qubit 3.0 fluorometer according to the manufacturer's instructions. The gDNA molecular weight was analyzed by running 60 ng of the gDNA on a 1% agarose gel for 90 minutes at 90V alongside a 23kb DNA ladder. The gDNA samples were stored long-term at -80°C.

6.3.3 Genome sequencing, assembly, and annotation

The *Salinispora* strains were sequenced at the UC San Diego Institute for Genomic Medicine (IGM); libraries were prepared using KAPA-Biosystems KAPA Hyper Prep Kit. Samples were pooled and sequenced with paired end 100 base pair reads (PE100) on an Illumina HiSeq4000. Base calling was performed by Illumina Real-Time Analysis (RTA) v2.7.7, and the output of RNA was demultiplexed and converted to FASTQ format with bcl2fastq2 Conversion Software v2.18. The FASTQ files were not trimmed or filtered; the number of total reads for the three sequenced strains ranged from 13-19.9 million.

Genome assembly and annotation was made with the open-source software KBase (DOE Systems Biology Knowledgebase) from the US Department of Energy (DOE) (Arkin *et al.,* 2018).

Kbase enables *de novo* assembly of prokaryotic NGS reads from various sequencing platforms. Read assembly was made using SPAdes (Bankevich *et al.,* 2012). The assemblies were annotated with the Prokka annotation pipeline (Seemann, 2014).

6.3.4 Breseq pipeline

The reads from albino strains were mapped to the genomes of their respective wild-type strains. Mutations were predicted by breseq by comparing the sequences of the albino strains to that of their pigmented counterparts. In the case of CNY-861, where both the albino and pigmented phenotypes were *de novo* assembled, CNH-996 A was used as a reference. To visualize the assembly and annotations I used Geneious version 10.2.6 (http://www.geneious.com). The Breseq pipeline, developed by the Barrick lab at the University of Texas at Austin, allows users to find mutations relative to a reference sequence in short-read DNA re-sequence data. It reports single-nucleotide mutations, point insertions and deletions, large deletions, and new junctions supported by mosaic reads (such as those produced by new mobile element insertions) in an annotated HTML format (Deatherage & Barrick 2014).

6.4 Results and discussion

Color loss in bacteria in known to decrease biological activity. For instance, in the pathogenic bacterium *Staphylococcus aureus*, when the production of staphyloxanthin is inhibited, the virulence of unpigmented mutants is lower (Song *et al.,* 2009). Additionally, the composition of the culture medium may influence the formation of carotenoids, and in general, light exposure induces carotenogenesis in microorganisms (Bhosale, 2003). It was recently reported that sioxanthin production in crude extracts of *S. tropica* (CNB-440) was light and temperature dependent, with an optimal temperature of 28^{o}C and light exposure of 18.6 μE/m^2/s (Jezkova *et*

al., 2021). Further, in the same study, the protective activity of sioxanthin against oxidative radicals in mammalian cells was evaluated. Here, the authors report protective activity against hydrogen peroxide (2000-2500 μM) with the human prostate adenocarcinoma (LNCaP strain) and esophageal squamous carcinoma (KYSE-30) among other cell lines (Jezkova *et al.,* 2021).

In laboratory conditions, albino *Salinispora* phenotypes are a rare occurrence. Only a handful of strains have presented this phenotypic characteristic (figure 6.2B and C). Although albinism is a stable mutation in *Salinispora*, when an orange culture turned white it happened progressively and usually only a portion of the colonies would show reduced sioxanthin production (figure 6.2A). Thus, albinism was more easily identified on solid media, where a clear difference between orange and white colonies could be observed. This is not to say that pigment loss could also be occurring in liquid media and with many other *Salinispora* strains where it would not be noticed by the naked eye. Perhaps the albino phenotype is more common than we assume, and we simply have failed to isolate it from environmental samples.

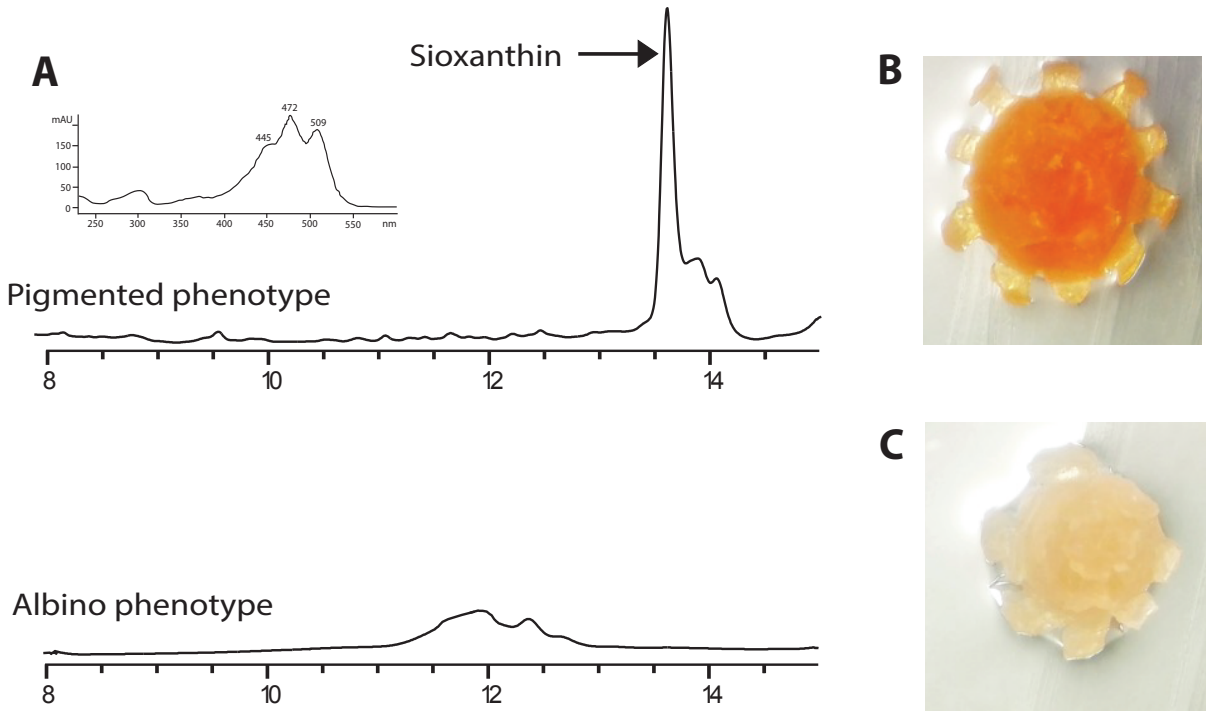

Figure 6.2 Variation in sioxanthin production in *Salinispora* strains. **A.** HPLC chromatogram showing the peak corresponding sioxanthin at 13.5 min and the lack of this peak in extracts of an albino strain. Inset: sioxanthin UV absorption spectrum. **B**. Orange colony (wild type) of *Salninispora arenicola* CNH-996. **C.** White colony (albino phenotype) of *Salninispora arenicola* CNH-996B.

Differences in growth between phenotypes was observed when cultures were grown on agar plates. Seemingly, the white cultures would grow slower than the pigmented strains. Therefore, I decided to compare growth rates between the two phenotypes. For this, I used optical density (600 nm, OD600) as a proxy for growth and cultivated the six *Salinispora* strains for 20 days taking aliquots every 24 hours. Unfortunately, I did not observe clear differences in growth between the albino and wild phenotypes. In some cases, the wild type had a faster growth rate and reached higher biomass than the mutant strain, such as in *S. arenicola* strain CNH-996A and B (figure 6.3); while in other cases, like in *S. vitiensis* CNS-055A and B, the albino phenotype would reach higher biomass with a faster growth rate, only to later stabilize with both the wild-type and

albino strains reaching similar biomasses. In the case of *S. arenicola* strains CNY0861 A and B, both grew similarly.

Salinispora form clumps when growing in liquid media making it very difficult to estimate growth rates in an accurate and simple way. Clumping in bacterial cultures has been attributed to to culture media, e.g., C:N ratio of the specific carbon source, pH, oxygen availability and nutrient depletion (Depass *et al.,* 2019; Brandl *et al.*, 2014; Bible *et al.,* 2015). Screening of carbon sources or the addition of buffers or anticlumping agents is commonly done to circumvent this issue. For this, I decided to repeat the experiment using Tween 80, which has been shown to help in biofilm disruption and would not interfere with growth rates since *Salinispora* spp. don't use tween 80 as a carbon source (Nielsen *et al.,* 2016; Román-Ponce *et al*., 2020). However, I observed similar growth rates, when adding Tween 80 to the growth media (data not shown). This suggest that both phenotypes have similar growth rates in laboratory conditions, and *Salinispora* spp. does not require to produce sioxanthin to grow.

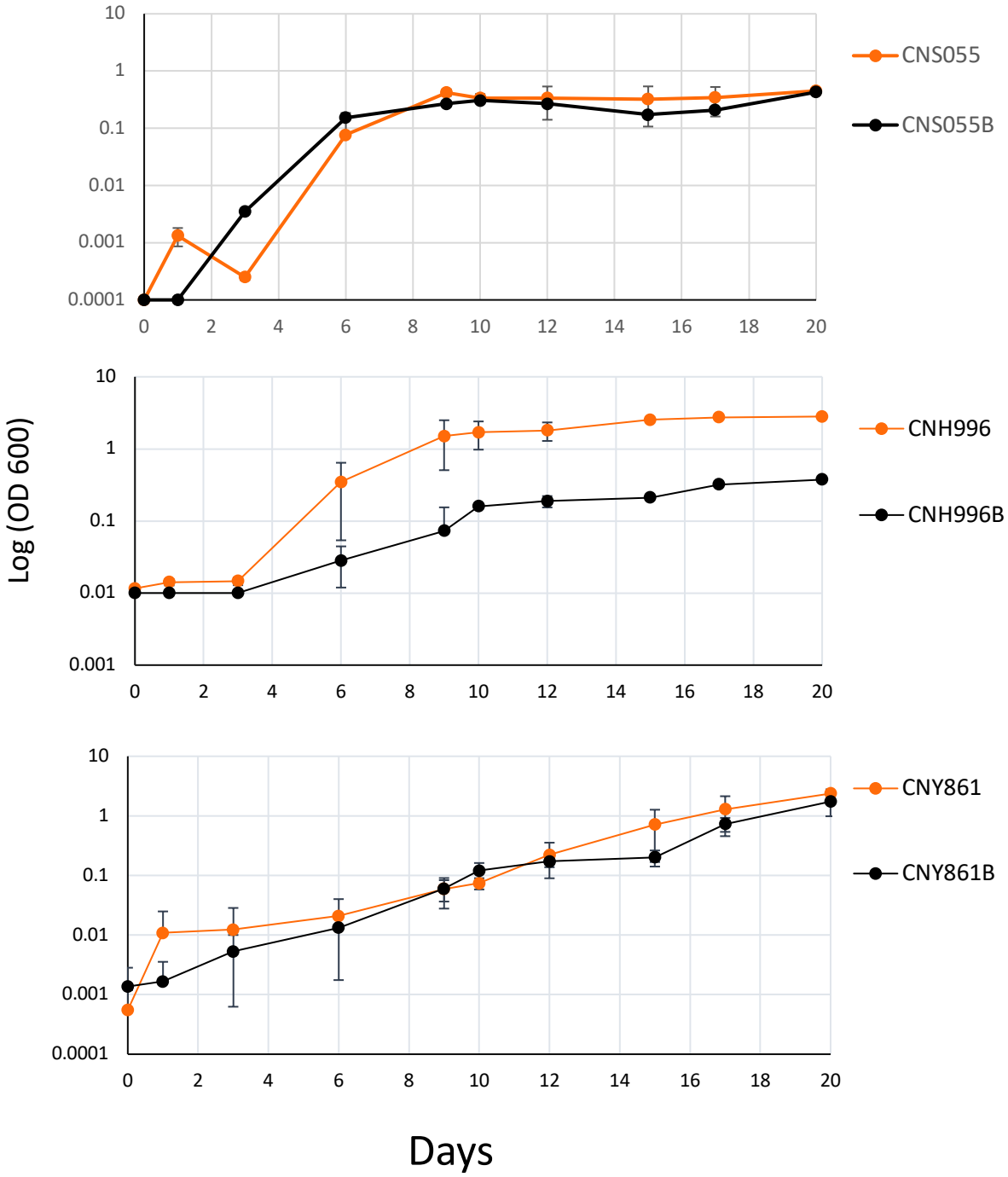

Figure 6.3 Growth curves of *Salnispora* albino (orange lines) and wild-type (black lines) phenotypes. The strain name is indicated at the right of each plot. OD600 is represented in log10 scale and the time scale is in days as indicated by the X axis.

Pigment production in actinomycetes has been linked to spore formation, with agents promoting pigment production simultaneously promoting sporulation (Onaka, 2017). However, I didn't observe a correlation between variations in sporulation and pigment production in the albino phenotype compared to the wild type, unless spores in the albino strains lacked their characteristic black coloration.

From previous and current sequencing projects, six genomes from three pairs of albino/wild-type strains phenotypes have been fully assessed facilitating the identification of mutations associated with pigment production. To identify the genomic differences between the albino and wild phenotypes, full genome comparisons were made using the breseq pipeline, which allows the detection of mutations relative to reference sequences in short-read DNA re-sequencing data (Deatherage *et al.,* 2015). Additionally, the mutations found with breseq were corroborated with sequence alignments from the genome of each strain (data not shown). After analyzing the breseq results, only a few differences were found related to genes responsible for pigment production (figure 6.4). Three SNPs (single nucleotide polymorphisms) and one nucleotide deletion were detected. In *S. vitiensis*, two SNPs are localized in the *crtI* gene, a phytoene desaturase, changing the coding sequence from glycine to glutamic acid in position 88 from the start codon and glutamic acid to a stop codon in position 557. One of the most surprising outcomes is the position of the additional two mutations, which are both located in the intergenic region between the *crtE* (Phytoene/squalene synthetase) and *crtB* (Geranylgeranyl pyrophosphate synthase) genes. The first one is observed in the two *S. arenicola* strains and is an adenine (A) to guanine (G) transition located upstream of the *crtB* and *crtE* genes at positions -54 and -12, respectively from the start codon of the genes. The second is observed in *S. vitiensis* and is a deletion of a G also in the intergenic region upstream from *crtB* and *crtE* genes at position -6 and

-363, respectively. The most striking observation is that none of the 118 *Salinispora* wild type genomes contain a guanine residue in the same position than the albino strains, on the contrary the guanine nucleotide seems to be highly conserved at that position across all genome strains from the wild types.

Richter *et al.* (2015) identified the genes responsible for sioxanthin biosynthesis in *S. tropica* by knocking out predicted genes for carotenoid production. In their study, the inactivation of *crtE* and *crtB* generated white strains similar to the albino phenotypes observed of this study. No other mutations displayed a similar phenotype in their study; however, variation of coloration was observed based on different mutations including pink, yellow, and light orange colonies.

The *crtE* and *crtB* genes are required in the early steps of the carotenoid biosynthesis. The *crtE* gene encodes for geranylgeranyl pyrophosphate synthase. This enzyme catalyzes the biosynthesis of a C_{20} precursor during carotenoid biosynthesis by the condensation of isopentenyl, a C_5 building block, and farnesyl pyrophosphate (Goodwin,1971). *CrtB* encodes a phytoene synthase that is responsible for the condensation of two molecules of GGPP to form a C_{40} phyotoene molecule that is subsequently, depending on the biosynthetic pathway, transformed into the vast diversity of carotenoids observed in bacteria (Phadwal, 2005). *CrtB* and *crtE* are well conserved genes among non-photosynthetic bacteria. GGPP synthases (*crtE*) are extensively associated with pathway branches, and thus, may have a greater effect on pathway flux than other genes in carotenoid biosynthesis (Phadwal, 2005; Crabtree & Newsholme, 1987). Although rare, the albino phenotype is very stable. Since sioxanthin production doesn't seem to affect growth under laboratory conditions, perhaps the colorless mutant is an adaptation to save metabolic energy or redirect it into the production of other metabolites. Bacterial cells continuously adjust gene expression in response to challenges from their environment, with SNPs providing an efficient mechanism to

disrupt metabolite production. It would be interesting to further analyze the albino strains following exposure to higher light intensities as was done with *S. tropica* in Jezkova *et al.,* (2021) and compare the effects on growth.

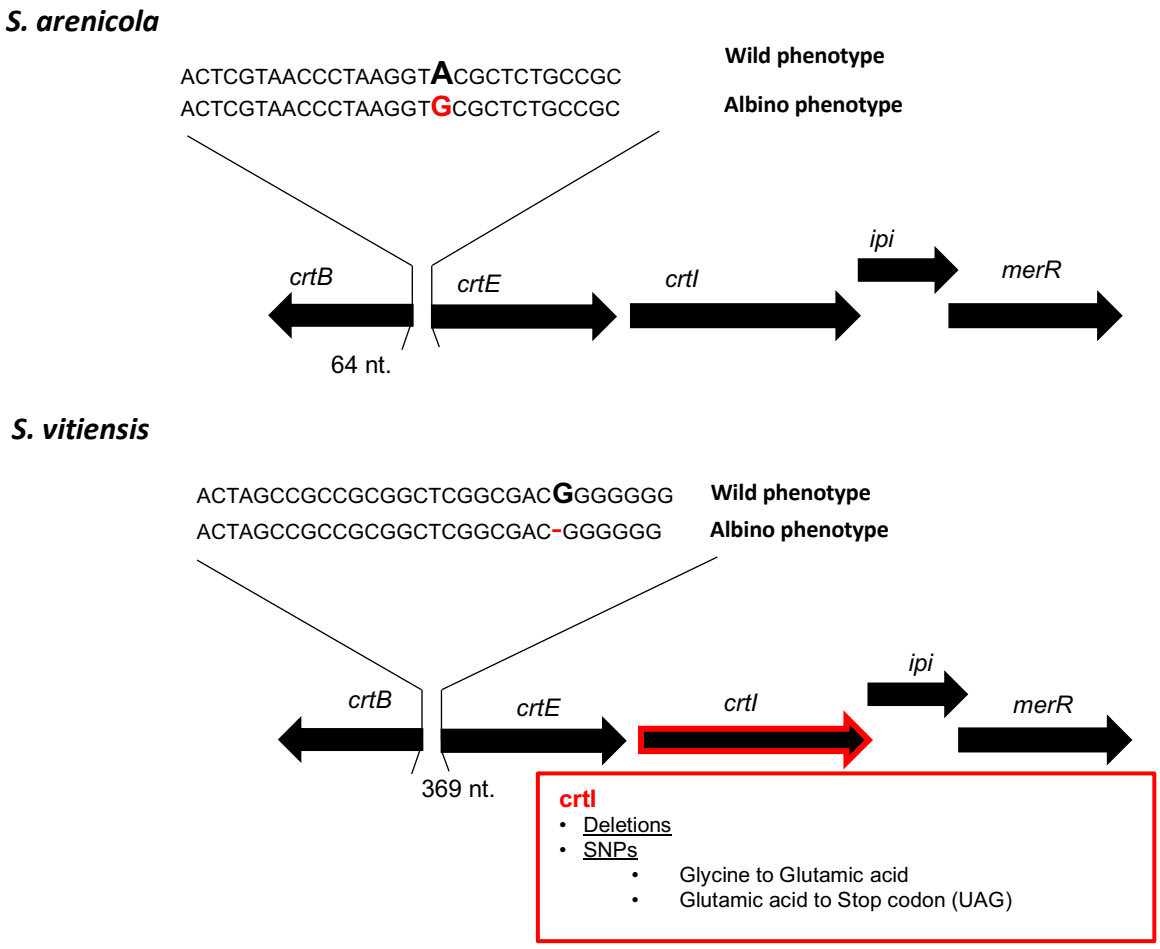

Figure 6.4 Mutations observed in genes involved in sioxanthin production between albino and wild phenotypes of Salinispora spp.; the red text box lists the mutations found in the *crtI* gene.

6.5 Conclusion

I was able to propose a genetic basis for a rare albino phenotype in *Salinispora* spp., finding several point mutations in the intergenic region between *crtB* and *crtE* and non-synonymous mutations in the *crtI* gene, particularly in *S. vitiensis.* It is interesting that single nucleotide mutations can change production of carotenoid in *Salinispora* spp., although these mutations are mostly related to *crtE*, which has previously been associated with controlling carotenoid pathway

flux. Although extremely rare, the albino phenotype is stable and seems to have similar growth rates as the wild types under laboratory conditions. While the ecological outcomes of pigment loss remain to be determined, other studies indicate that sioxanthin production is enhanced by increased light and temperature. It would be interesting to observe the growth efficiency of these albino mutants under intense light exposure and compare the results to the pigmented strains. Furthermore, evidence that sioxanthin provides antioxidant protection to ROS suggests pigment production is importance in natural environments, especially for *Salinispora* which is often found in tropical sediments where high light intensities and temperatures are common.

6.6 Acknowledgements

I would like to thank members of the Jensen Laboratory including Henrique Ramalho Machado and Kaitlin Creamer for their assistance with breseq analysis, and genome sequencing. Special thanks to Andrew Bradshaw and Kira Kawano for their help with the growth curves and data collection.

This research was supported by the NIH grant to R01GM085770 PRJ and BSM, the CONACYT and UC MEXUS doctoral fellowship, and the Shirley Boyd Memorial Fellowship.

6.7 References

Arkin, A.P., Cottingham, R.W., Henry, C.S., Harris, N.L., Stevens, R.L., Maslov, S., Dehal, P.S., Ware, D.H., Pérez, F., Canon, S., Sneddon, M.W., Henderson, M.L., Riehl, W.J., Murphy-Olson, D., Chan, S.Y., Kamimura, R.T., Kumari, S., Drake, M., Brettin, T.S., Glass, E.M., Chivian, D., Gunter, D., Weston, D.J., Allen, B.H., Baumohl, J.K., Best, A.A., Bowen, B., Brenner, S.E., Bun, C., Chandonia, J., Chia, J., Colasanti, R.L., Conrad, N., Davis, J.J., Davison, B.H., DeJongh, M., Devoid, S., Dietrich, E.M., Dubchak, I., Edirisinghe, J.N., Fang, G., Faria, J.P., Frybarger, P., Gerlach, W., Gerstein, M.B., Greiner, A., Gurtowski, J., Haun, H.L., He, F., Jain, R., joachimiak, M.P., Keegan, K.P., Kondo, S., Kumar, V., Land, M.L., Meyer, F., Mills, M., Novichkov, P.S., Oh, T., Olsen, G.J., Olson, R., Parrello, B.D., Pasternak, S., Pearson, E., Poon, S.S., Price, G.A., Ramakrishnan, S., Ranjan, P., Ronald, P.C., Schatz, M.C., Seaver, S.M., Shukla, M., Sutormin, R.A., Syed, M.H.,

Thomason, J., Tintle, N.L., Wang, D., Xia, F., Yoo, H.S., Yoo, S., & Yu, D. (2018). KBase: The United States Department of Energy Systems Biology Knowledgebase. *Nature Biotechnology, 36*, 566 - 569.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M., Nikolenko, S.I., Pham, S.K., Prjibelski, A.D., Pyshkin, A., Sirotkin, A., Vyahhi, N., Tesler, G., Alekseyev, M.A., & Pevzner, P.A. (2012). SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *Journal of computational biolog : a journal of computational molecular cell biology, 19 5*, 455-77

Bhosale, P. (2003). Environmental and cultural stimulants in the production of carotenoids from microorganisms. *Applied Microbiology and Biotechnology, 63*, 351-361.

Bible, A., Khalsa-Moyers, G.K., Mukherjee, T., Green, C., Mishra, P., Purcell, A.M., Aksenova, A., Hurst, G.B., & Alexandre, G. (2015). Metabolic adaptations of Azospirillum brasilense to oxygen stress by cell-to-cell clumping and flocculation. *Applied and environmental microbiology, 81 24*, 8346-57.

Brandl, M. T., & Huynh, S. (2014). Effect of the surfactant tween 80 on the detachment and dispersal of Salmonella enterica serovar Thompson single cells and aggregates from cilantro leaves as revealed by image analysis. *Applied and environmental microbiology*, *80*(16), 5037–5042. doi:10.1128/AEM.00795-14

Chatfield C. H. and Cianciotto N. P. (2007). The Secreted Pyomelanin Pigment of *Legionella pneumophila* Confers Ferric Reductase Activity. Infection and Immunity. ASM journal. 75(8): 4062–4070.

Crabtree, B., Newsholme, E.A., (1987). The derivation and interpretation of control coefficients. Biochem. J. 247, 113–120.

Deatherage, D.E., & Barrick, J.E. (2014). Identification of mutations in laboratory-evolved microbes from next-generation sequencing data using breseq. *Methods in molecular biology, 1151*, 165-88.

DePas, W.H., Bergkessel, M., & Newman, D.K. (2019). Aggregation of Nontuberculous Mycobacteria Is Regulated by Carbon-Nitrogen Balance. *mBio*.

Genç, Y., Bardakcı, H., Yücel, Ç., Karatoprak, G.Ş., Küpeli Akkol, E., Hakan Barak, T., & Sobarzo-Sánchez, E. (2020). Oxidative Stress and Marine Carotenoids: Application by Using Nanoformulations. *Marine Drugs, 18*.

Goodwin, T.W., 1971. Biosynthesis of carotenoids and plant triterpenes: the fifth CIBA Medal lecture. Cited in: Phadwal K. (2005). Carotenoid biosynthetic pathway: Molecular phylogenies and evolutionary behavior of *crt* genes in eubacteria. Gene 345:35-43.

Jezkova, Z., Binda, E., Potočár, T., Marinelli, F., Halecký, M., & Brányik, T. (2021). Laboratory scale cultivation of Salinispora tropica in shake flasks and mechanically stirred bioreactors. *Biotechnology letters*.

Jezkova, Z., Schulzova, V., Křížová, I., Karabin, M., & Brányik, 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. *Marine Drugs, 19*.

Kalan L., Gessner A., Thaker M. N., Waglechner N., Zhu X. M., Szawiola A., Bechthold A., Wright G. D. Zechel D. L. (2013). A Cryptic Polyene Biosynthetic Gene Cluster in *Streptomyces calvus* Is Expressed upon Complementation with a Functional *bldA* Gene. Chemistry & Biology 20, 1214–1224, 2013.

Liu G. U., Nizet V. (2009). Color me bad: microbial pigments as virulence factors. Trends Microbiol. 17(9): 406-413.

Liu G.Y., Essex A., Buchanan J. T., Datta V., Hoffman H. M., Bastian J. F., Fierer J., Nizet V. (2005) Staphylococcus aureus golden pigment impairs neutrophil killing and promotes virulence through its antioxidant activity. J Exp Med 202, 209–215.

Maldonado, L.A., Fenical, W., Jensen, P.R., Kauffman, C.A., Mincer, T.J., Ward, A.C., Bull, A.T., & Goodfellow, M. (2005). Salinispora arenicola gen. nov., sp. nov. and Salinispora tropica sp. nov., obligate marine actinomycetes belonging to the family Micromonosporaceae. *International journal of systematic and evolutionary microbiology, 55 Pt 5*, 1759-66.

Millán-Aguiñaga, N., Chavarria, K.L., Ugalde, J.A., Letzel, A., Rouse, G.W., & Jensen, P.R. (2017). Phylogenomic Insight into Salinispora (Bacteria, Actinobacteria) Species Designations. *Scientific Reports, 7*.

Nielsen, C.K., Kjems, J., Mygind, T., Snabe, T., & Meyer, R.L. (2016). Effects of Tween 80 on Growth and Biofilm Formation in Laboratory Media. *Frontiers in Microbiology, 7*.

Onaka, H. (2017). Novel antibiotic screening methods to awaken silent or cryptic secondary metabolic pathways in actinomycetes. *The Journal of Antibiotics, 70*, 865-870.

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

Phadwal K. (2005). Carotenoid biosynthetic pathway: Molecular phylogenies and evolutionary behavior of *crt* genes in eubacteria. Gene 345:35-43.

Ramesh, C., Vinithkumar, N.V., & Kirubagaran, R. (2019). Marine pigmented bacteria: A prospective source of antibacterial compounds. *Journal of Natural Science, Biology and Medicine, 10*, 104 - 113.

Richter T. K. S., Hughes C. C., Moore B. S. (2015). Sioxanthin, a novel glycosylated carotenoid, reveals an unusual subclustered biosynthetic pathway. Environ. Microbiol. 17(6), 2158–2171.

Román-Ponce, B., Millán-Aguiñaga, N., Guillen-Matus, D., Chase, A. B., Ginigini, J., Soapi, K., Feussner, K. D., Jensen, P. R., & 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., *Salinispora oceanensis* sp. nov. and *Salinispora vitiensis* sp. nov., and emended description of the genus *Salinispora*. *International journal of systematic and evolutionary microbiology*, *70*(8), 4668–4682.

Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. *Bioinformatics, 30 14*, 2068-9 .

Sobin B., Stahly G. L. (1942). The Isolation and Absorption Spectrum Maxima of bacterial carotenoid pigments. J. Bacteriol. 44(3):265-276.

Song Y., Liu C. I., Lin F. Y., No J. H., Hensler M., Liu Y. L., J W. Y., Low J., Liu G. Y., Wang A.H.J., Oldfield E. (2009). Inhibition of Staphyloxanthin Virulence Factor Biosynthesis in Staphylococcus aureus: In Vitro, In Vivo, and Crystallographic Results. J. Med. Chem. 52, 3869-3880.

Chapter 7

Final Remarks

"Bacteria are everywhere, some can make you sick and some are good for you" This is the most common answer people give when asked what they know about bacteria. But what does *everywhere* really mean? Bacteria are crucial for the health of virtually all ecosystems including soils, oceans, rivers, and even within other organisms as symbionts (van der Meij *et al.*, 2017). Further, bacteria are the source of many important medicines including antibiotic, anticancer, anti-inflammatory, and antiviral agents (Wright 2019). In the last decades, the increased access to bacterial genome sequences revealed a wealth of unexplored biosynthetic potential. In fact, only a small percentage of the specialized metabolites bacteria encode are produced under laboratory conditions. This discovery has revolutionized the way we search for natural products. Many approaches have been developed to "*turn on"* these seemingly silent biosynthetic genes, and the concept of genome mining has become well established as a method to access this newly discovered potential. One remarkably rich source of natural products is the marine actinomycete *Salinispora,* a prolific producer of many specialized metabolites with various biological activities and exceptional structures. Included among these is ketomemicin C, an extremely rare pseudopeptide containing a methylene carbonyl bond. To my knowledge, ketomemicins and arphamemines are the only two classes of naturally occurring pseudopeptides containing this type of moiety. My research on ketomemicin C resulted in the discovery of two new analogs of this natural product. I also confirmed cultured *Salinispora* spp. as sources of ketomemicins, which had only been isolated via heterologous expression until now. Additionally, I used genome mining to assess the distribution of ketomemicin in *Salinispora.* Pseudopeptides are valued in human health for their abilities to mimic biological substrates and inhibit important enzymes in the cell such as

proteases. I hypothesize ketomemicin C may have a similar activity, and further experiments are undertaken to prove this hypothesis.

The biosynthetic potential of *Salinispora* extends beyond the natural products isolated to date. In fact, most of the biosynthetic gene clusters within *Salinispora* genomes remain uncharacterized. Here, I report the discovery of biosynthetic gene clusters with unusual and recently discovered architectures such a trans-adenylation domain NRPS in tandem with a type I PKS and a PKS with a trans-acyltransferase domain. These types of biosynthetic domains remain generally understudied in actinomycetes. I propose targeting these two BGCs for heterologous expression, especially the trans-A NRPS, since it is likely they will yield novel natural products and expand our understanding of *Salinispora*'s natural product repertoire.

Surprisingly, specialized metabolism seems to play an important role in the speciation of *Salinispora*. My research supports the concept of species-specific patterns in *Salinispora* specialized metabolism. I discovered 15 unique BGCs distributed across seven different species including the first report of salinichelin as a species-specific metabolite in *S. pacifica*. Most species-specific BGCs remain cryptic and around 80% are silent. It remains to be determined to what extend these biosynthetic gene clusters are functional and under what conditions they are activated.

While the phylogenetic patterns observed in *Salinispora* specialized metabolism are relatively new, several other bacterial groups show similar patterns at various taxonomic levels. There is mounting evidence that compounds or biosynthetic gene clusters can become fixed within specific lineages. Many agree that specialized metabolites may represent niche defining traits and have important roles in evolution across all levels of life. In fact, in eukaryotic organisms such as plants, fungi, and marine sponges, specialized metabolism has aided in the delineation of species.

In bacteria, the species concept is controversial, and many suggest we should move to genomic information to assess bacterial diversity, although identifying functionally distinct units of diversity from sequence data remains challenging.

In chapter five, I present a literature review of the most compelling evidence for phylogenetic patterns in bacterial specialized metabolism and propose that, in some cases, they can be used as taxonomic markers. The aim of using specialized metabolism as a complement to other methods of bacterial classification is to provide more robust species descriptions as has been done with plants, fungi, and other organisms.

As I mentioned before, bacterial natural products are not only important for human health, but they also play important ecological roles. *Salinispora* is a marine actinomycete, isolated from shallow sediments in tropical areas. This type of environment is characterized by high temperatures and high light exposure. *Salinispora*, produces the carotenoid pigment sioxanthin, which has proven to provide antioxidant protection against ROS (Richter *et al*., 2015; Jezkova *et al*., 2021). Further, the production of sioxanthin is highly correlated with warm temperatures and high light intensities. During my time working with *Salinispora*, I discovered three albino strains. Surprisingly, this rare phenotype seems to not affect growth efficiency under standard laboratory conditions. I was able to correlate the lack of pigmentation with several point mutations related to early stages of sioxanthin biosynthesis, including a single nucleotide polymorphism within an intergenic region. The ecological implications of this mutation remain unclear; however, it seems that at random, some strains of *Salinispora* have stopped producing sioxanthin when they are not under thermo or photo stress.

Marine environments are some of the most diverse. For instance, coral reefs are thought to outnumber tropical rainforests in species diversity (Haefner 2003). Marine bacteria are under

intense competitive pressure for space, light, and nutrients, especially in locations with higher diversity and lower nutrient availability like tropical areas. We have observed that microbial natural products fulfill various tasks in the environment and seem to be important for maintaining and controlling many community functions and population dynamics (Petersen *et al*., 2020). However, our knowledge of the ecological roles of bacterial natural products is limited.

Indeed, bacteria are *everywhere*. We can find microbial communities in every environmental ecosystem described to date (Merino *et al.,* 2019). All these ecosystems harbor not only bacteria, but other microbes that play key roles in the regulation of many processes. In soils, they are important for the growth of crops and wildlife. In the oceans, they are pivotal for nutrient cycling, either directly as food sources, or indirectly by facilitating access to carbon, nitrogen, and other essential nutrients.

Almost all organisms including humans have symbiotic relationships with microbes (Gupta *et al.,* 2016). Our body has 10 times more microbial cells than human cells. Within the human body, microbes perform a plethora of functions: they help us digest our food, they educate our immune system, they help us resist diseases, and even can affect our behavior (Gilbert *et al*., 2018). We have much more to learn about microbes and the way they interact with us humans, the environment, and other organisms. Nonetheless, what we know so far, has made me realize how important these organisms are for our health and our ecosystem's health.

The more we learn about microbial communities the more we realize the important roles they have in our lives. Contradictorily, we are living in the era of sterile practices, of cleaning ourselves and our houses with antimicrobial soaps, of eating ultra-processed food, of overusing antibiotics, and all these practices are putting our microbial diversity in danger. When we think about endangered species, we rarely think about the species within microbial communities.

Nonetheless, we might be having an "invisible extinction" of our microbiota. Accordingly to recent studies of modern-day hunter-gatherers, who provided a window into what our microbiota was like for the approximately 180,000 years that humans exclusively foraged food, our modern guts harbor 50% less bacterial species and half as many bacterial genes than our ancestors (Schnorr *et al.*, 2014). In regard to global microbial diversity, we know even less, experts are unsure if the global microbial diversity is increasing, decreasing or staying the same. The consequences of microbial species becoming extinct are yet to be discovered, but the outcomes might be more sever of what we can imagine.

What can be done to protect these microbial communities? Perhaps a sterile environment is not as beneficial as we once thought. Of course, we precise of sterile operating rooms but, not our kitchen counters. We can choose soaps and cleaners without antimicrobial substances; we really don't need to kill 99% of microbes on surfaces. We can advocate for a responsible use of antibiotics; afterall the antibiotic crisis is a real thread. We can vote for policies supporting the research and protection of microbial communities across agriculture, healthcare, and consumer goods industries. And most importantly, we can tell our families, friends, and other people the important role microbes play in our lives and ask them to join us and support the protection of our microbial communities.

7.1 A letter for those pursuing a PhD in science.

Dear friends wanting to become doctors in science, please reconsider. I know it is hypocritical coming from me, and perhaps is not my true self talking but my sleep deprived self. Afterall, who can stay sane after working 20 hours a day for a full month and a half? Surely not me. And if me mentioning the extremes of dissertation writing and defense preparation doesn't scare you away, maybe the following lines will.

If someone had told me in detail how my life as a PhD student was going to be, I would have refused to join the program. In reality, I have no one else to blame but myself. I already have a Master's degree in science, I should have known better. Citing one of my favorite fictional scientists, Olive Smith, I was *"well aware that committing to years of unappreciated, underpaid eighty-hour workweeks might not going to be good for [my] mental health. That nights spent toiling away in front of a Bunsen burner* (or staring at mass spectrometry data, in my case) *to uncover a trivial slice of knowledge might not be the key to happiness… Academia takes a lot from you and gives back very little"* And still, here I am after six years, graduating. Becoming a doctor despite it all.

Because is all true, you will be working nonstop, you will be putting more hours into your research than in anything else and you will be getting very little back. You will spend endless hours trying to figure out the better way to design an experiment just to realize that was indeed not the best way, because your experiment would not work. And you will try again because we are not quitters! And you will fail again, and again… and again. You won't be able to sleep, eat, exercise, go out for a walk, meet a friend, grocery shop, or take a shower without thinking about troubleshooting your experiment. Why is it not working? What can you modify? What are you doing wrong? And one day after exhausting all options, it will work. You will have your answer, you will know a piece of knowledge no one else knows, and… you inevitable will have more questions. So, the process will start again, months and months of wrecking your head around something for a slice of information that 99% of the world population might not even be aware we need to know or are even interested to know. Because that's the reality of us scientists, our life is consumed by wanting to know something and thinking of ways to answer endless questions. And once we have answers, we will ask another question, we will look for another problem, and the

cycle will continue. Citing one of my favorite real-life scientists, Dr. Kaitlin Creamer, *"we are all perpetual students of nature".*

My dear friend, this all might sound appealing, and I get it, you love science, you are young, curious, and driven. You can do it all, you can take it. But this is not all, up to now I've only mentioned the good part. Spending long hours in the lab, solving unsolvable puzzles, and knowing something no one knows. Best job ever, I would say.

Of course, you will not be working in isolation. Unless you decide to study one of the most pathogenic strains of bacteria or radioactive materials, in that case yes, you might spend long hours all by yourself. Regardless, you will have lab mates. People that, like you, are overworked and consumed by the frustrations of failing experiments or in rare occasions, ecstatic for the success of their research. You will befriend them, and you will have amazing brainstorming sessions. You will feel invigorated and will share your best ideas. And sometimes, you will hear those, your ideas, presented back at you while in a group meeting session, by the friendliest lab mate of all. And you will be shocked. You will feel confused and frustrated. You won't know what to do. You might bring it up to them or your advisor, but don't expect an apology, you won't get it. They will tell you that you are not the only one thinking about those concepts; it doesn't matter no one else was interested in researching that before you joined the lab. And the friendliest lab mate will finish your project before you because they don't have to troubleshoot (you've already done that). They will publish their (your) research, and they will be praised for your ideas. So, if you are pursuing a PhD in science, please reconsider, unless you are ready for your ideas to be stolen or are willing to fight for them. Of course, you will also make good friends— the best ones. People like you that are driven, with endless ideas of their own, so they won't need to take yours. And you will love them and spend the little free time you have with them; you will form very strong bonds and, when

the time comes, saying goodbye will be hard. You will always cherish those times in lab with your good friends and you will learn to guard your best ideas.

Now, if you are pursuing science for the prestige, please reconsider. Think of your advisor, after long years of research, publishing in prestigious -and overly expensive- scientific journals, they are well known and respected— by the other few people within your field. Maybe someone has named a species after them out of respect. Maybe that someone was their best friend, one of those few scientists I just mentioned. Don't get me wrong, being a professor can be prestigious: awards, recognitions, invitations for seminars, and conferences. But remember, we the students, postdocs, and technicians, are the ones generating most of the results and ideas our advisors get recognized for. Some professors are aware of that, some others not so much, or choose not to. Think about which one you want to become.

Now, if you are pursuing a PhD in science because you are smart and think that is what smart people do, please reconsider. Yes, many scientists are very smart, some of the most intelligent people in history were scientists, but some others were artists, politicians, cooks, carpenters, teachers, etcetera (you get the gist). So, if you are pursuing a PhD just because everyone around you is telling you that smart people like you should become scientists, please reconsider. The academic path is rough and emotionally draining, more than you can imagine. I think we scientists are not only smart, but also stubborn. Research is for many of us a challenge. We need, crave to solve nature's phenomena and we would not give up easily. On the contrary, more than once we will find ourselves working in projects that are almost impossible to solve. High risk projects that are more about luck than hard work, and we never know when to give up, some can learn to reframe the story to be able to publish and move on. But I can assure you my friend, the failure will stick with us forever. Have you heard the phrase "luck is what happens when

preparation meets opportunity" by Seneca? Well, in science this doesn't apply, we are prepared, we have all the resources and still, sometimes we are not lucky enough to find our answer and we blame ourselves. We tell ourselves we were not smart enough or innovative enough. And that is devastating. We are on an emotional rollercoaster, and we refuse to get off. Many of us are scientists before anything else; friends, husbands, wives, daughters, sons, parents, and siblings they all come second after our craft. And it can be devastating for them as well. Please, remember this when you start working 80 hours a week and neglect your beloved ones; don't let your own curiosity and stubbornness consume you. Be better than what I've been, be you before being a scientist, put your health and your loved ones before your craft.

Lastly, if you are pursuing a PhD in science because you have a question, one well-thought question that you want to answer or an issue you want to solve. If you care more about learning than about prestige. If value people more than you value success. Then, I beg you, pursue science. We need you. Don't let anyone belittle you and don't give up. You have talent, you are capable, and you belong.

Maybe, with some luck we will run into each other in a conference, some day.

7.2 References

Gilbert, J.A., Blaser, M.J., Caporaso, J.G., Jansson, J.K., Lynch, S.V., & Knight, R. (2018). Current understanding of the human microbiome. *Nature medicine, 24*, 392 - 400.

Gupta, A., Gupta, R., & Singh, R.L. (2016). Microbes and Environment. *Principles and Applications of Environmental Biotechnology for a Sustainable Future*, 43 - 84.

Haefner B (2003) Drugs from the deep: marine natural products as drug candidates. Drug Discov Today 8:536–544

Jezkova, Z., Schulzova, V., Křížová, I., Karabin, M., & Brányik, 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. *Marine Drugs, 19*.

Merino, N., Aronson, H.S., Bojanova, D.P., Feyhl-Buska, J., Wong, M.L., Zhang, S., & Giovannelli, D. (2019). Living at the Extremes: Extremophiles and the Limits of Life in a Planetary Context. *Frontiers in Microbiology, 10*.

Petersen, L., Kellermann, M.Y., & Schupp, P.J. (2020). Secondary Metabolites of Marine Microbes: From Natural Products Chemistry to Chemical Ecology. *YOUMARES 9 - The Oceans: Our Research, Our Future*.

Richter T. K. S., Hughes C. C., Moore B. S. (2015). Sioxanthin, a novel glycosylated carotenoid, reveals an unusual subclustered biosynthetic pathway. Environ. Microbiol. 17(6), 2158–2171.

Schnorr, S.L., Candela, M., Rampelli, S., Centanni, M., Consolandi, C., Basaglia, G., Turroni, S., Biagi, E., Peano, C., Severgnini, M., Fiori, J., Gotti, R., De Bellis, G., Luiselli, D., Brigidi, P., Mabulla, A.Z., Marlowe, F.W., Henry, A.G., & Crittenden, A.N. (2014). Gut microbiome of the Hadza hunter-gatherers. *Nature Communications, 5*.

van der Meij, A., Worsley, S.F., Hutchings, M.I., & van Wezel, G.P. (2017). Chemical ecology of antibiotic production by actinomycetes. *FEMS microbiology reviews, 41 3*, 392-416 .

Wright, G.D. (2019). Unlocking the potential of natural products in drug discovery. *Microbial Biotechnology, 12*, 55 - 57.