UNIVERSITY OF CALIFORNIA SAN DIEGO

Diversity and distribution of polyketide synthase genes across biomes, taxa and abyssal sediments

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

in

Marine Biology

by

Hans Wu Singh

Committee in charge:

Professor Paul R. Jensen, Chair
Professor Eric E. Allen
Professor Lihini Aluwihare
Professor Linda W. Kelly
Professor Bradley S. Moore

2024

The Dissertation of Hans Wu Singh is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2024

DEDICATION

To my parents and grandparents.
I will forever be grateful for all your hard work and sacrifice,
along with the love and support you have given me.
I would not be here without you.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

First and foremost, I would like to thank Paul for giving me an opportunity to be a graduate student in your lab. Every day I am astonished at both the guidance and freedom you have given me as a scientist. From taking part in research cruises and traveling for conferences half a world away, to putting together and publishing my research projects piece by piece, I have experienced some of the highest highs as a member of your lab. In addition to my growth as a scientist, I truly believe you have made me a better person, both through your mentorship and your conduct as a role model.

Additionally, I would like to thank all of my committee members for their support, input and help along the way – Prof. Paul Jensen, Prof. Eric Allen, Prof. Lihini Aluwihare, Prof. Linda Wegley Kelly, and Prof. Bradley Moore. I have the utmost respect and admiration for each of you as both scientists and stewards of your respective labs.

Next, I owe a great deal of thanks to my co-authors and contributors along each of my research projects. Specifically, I would like to thank Dr. Kaitlin Creamer who is a co-author on my third chapter of this dissertation and who served as an exemplary role model during my time in the Jensen lab. I also would like to thank Dr. Sheila Podell for giving me advice in many fields but especially regarding the NaPDoS2 webtool, which was used in all three of my chapters. Next, I would like to thank Dr. Johanna Gutleben, who was amazingly helpful during my time in the lab and worked extensively to help me with the amplicon analyses in chapter 4. I would also like to thank Dr. Alexander Chase for his bioinformatic and metagenomic expertise that I leaned on in chapters 2 and 4. Finally, many thanks to Dr. Alex Bogdanov, who helped teach me the skills needed for metabolomics analyses I conducted in chapter 4. Additionally, I would like to

thank other members of the Jensen lab such as Dr. Gabriel Castro, Dr. Dulce Guillen-Matus, Dr. Douglas Sweeney, and Dr. Alyssa Demko who were instrumental for my development as a scientist due to the advice they gave me along the way. I also had the enormous privilege of mentoring many talented undergraduate and master's students during my time in the Jensen lab. As such, I would like to thank Zachary Daniel, Erien Cross, Sandra Martinez and Ivan Chavez for their enthusiasm and dedication.

Next, I want to thank many within the SIO administration for helping my time as a graduate student run smoothly: Maureen McGreevy, Shelley Weisel, Gilbert Bretado, Dana Jimenez, Adrielle Wai and Annamarie Bryson. I also want to thank SIO FM, particularly Dejan and Jackie for their attentiveness in maintaining a safe and clean lab. Finally, I would like to give a particular thanks to Keiara Auzenne, for the enormous efforts she has undertaken to improve equity, diversity and inclusion at SIO.

I have an enormous debt to the mentors that opened the doors of science for me. First, I would like to thank Dr. Alison Sherwood who gave me my first research experience as a REU student at UH Manoa, along with Dr. Rachael Wade who helped mentor me. Next, I am extremely grateful to Dr. Eoin Brodie, who took me on as a member of his lab at LBNL, along with Dr. Heejung Cho and Dr. Patrick Sorensen for their mentorship and patience with me as an undergraduate researcher.

I would like to thank my friends – Ziwa, Michael, Kevin, Sahana, Mary, Biz, Aizah, Namita, Gabe, Alma, Dulce, Kaitlin, Dani and Leonard. Thank you so much for being there for me always, through the highs and lows. I would not have made it through this thesis without your support.

I am also extremely grateful to my partner Monica for her energy, patience, support and love over the years. You have helped me be happy and steady throughout the stresses of my program and for that I am very thankful.

Lastly, and most importantly, I would like to thank my family for their unconditional love and support. I would not be where I am without my grandparents, aunts, uncles and cousins. Particularly, I would like to thank my cousin Leo, as only children I am glad to have you as a brother for life. Above all, I have the utmost gratitude to my parents, who have paved the way for all I accomplish in life. I have learned much from the values you live your life by – dedication, tenacity, empathy, confidence, kindness, loyalty, and a pursuit of knowledge. I love you and am forever grateful.

Chapter 2, in full, is a reprint of the materials as it was submitted to *mSystems*. Singh HW, Creamer KE, Chase AB, Klau LJ, Podell S, Jensen PR. 2023. Metagenomic data reveals type I polyketide synthase distributions across biomes. mSystems 8:e00012-23. The dissertation author was the primary investigator of this manuscript.

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

Chapter 4 is coauthored with Dr. Johanna Gutleben, Dr. Alexander Chase, Dr. Alex Bogdanov, and Dr. Paul R. Jensen. The dissertation author was the primary investigator and author of this chapter.

VITA

EDUCATION

08/2019 - Current **Scripps Institution of Oceanography, University of California, San Diego**

Doctor of Philosophy in Marine Biology

Master of Science in Marine Biology

08/2015-06/2019 **University of California, Berkeley**

B.A. in Molecular and Cell Biology: Genetics, Genomics and Development

B.S. in Environmental Science (honors)

RESEARCH TRAINING

2019 – Current Graduate Student, Scripps Institution of Oceanography, UC San Diego, Center for Marine Biotechnology and Medicine

Advisor: Dr. Paul Jensen

2017 – 2019 Undergraduate Researcher, Lawrence Berkeley National Laboratory

Advisor: Dr. Eoin Brodie

2017 Undergraduate Researcher, University of Hawai‘i at Mānoa

Advisor: Dr. Alison Sherwood

PUBLICATIONS

Singh, H.W., K.E. Creamer, A.B. Chase, L.J. Klau, S. Podell, P.R. Jensen. 2023. Metagenomic Data Reveals Type I Polyketide Synthase Distributions across Biomes. mSystems 8(3).

Singh, H.W., R.M. Wade and A.R. Sherwood. 2018. Diurnal patterns of airborne algae distribution in the Hawaiian Islands: a preliminary study. Aerobiologia 34: 363-373

Singh, H.W., J. Gutleben, A. Bogdanov, A.B. Chase, A.M. Demko, S. Podell, B. Haley, P. Jensen. In prep. A multi-omic assessment of biosynthetic potential across abyssal sediments.

Creamer, K.E., **H.W. Singh**, A.B. Chase, L.J. Klau, S. Podell, P.R. Jensen. In prep. Genomes across the tree of life highlight novel polyketide potential.

Klau, L.J., S. Podell, K.E. Creamer, A.M. Demko, **H.W. Singh**, E.E. Allen, B.S. Moore, N. Ziemert, A.C. Letzel, P.R. Jensen. 2022. The Natural Product Domain Seeker version 2 (NaPDoS2) webtool relates ketosynthase phylogeny to biosynthetic function. Journal of Biological Chemistry 298:102480.

Chadwick K.D., P.G. Brodrick, K. Grant, T. Goulden, A. Henderson, N. Falco, H. Wainwright, K.H. Williams, M. Bill, I. Breckheimer, E.L. Brodie, H. Steltzer, C.F.R. Williams, B. Blonder, J. Chen, B. Dafflon, J. Damerow, M. Hancher, A. Khurram, J. Lamb, C.R. Lawrence, M. McCormick, J. Musinsky, S. Pierce, A. Polussa, M. Hastings-Porro, A. Scott, **H.W. Singh**, P.O. Sorensen, C. Varadharajan, B. Whitney, K. Maher. 2020. Integrating airborne remote sensing and field campaigns for ecology and Earth system science. Methods Ecol Evol 11:1492–1508.

ABSTRACT OF THE DISSERTATION

Diversity and distribution of polyketide synthase genes across biomes, taxa and abyssal sediments

by

Hans Wu Singh

Doctor of Philosophy in Marine Biology

University of California San Diego, 2024

Professor Paul R. Jensen, Chair

Microorganisms interact with their surrounding environment via the production of metabolites, which they use to regulate growth, obtain nutrients, signal, and compete with other microbes. Polyketides are a diverse and bioactive class of microbially-produced specialized

metabolites. Microbial polyketide synthase (PKS) genes encode the biosynthesis of polyketides, although differences in PKS biosynthetic gene cluster (BGC) structure can lead to structural diversity in polyketides ranging from small aromatic compounds to large macrolides. With advances in DNA sequencing technologies, genome mining has become a valuable method for natural product discovery. The webtool NaPDoS2 detects ketosynthase (KS) domains from genomic, metagenomic, and amplicon query data and utilizes phylogenetic conservation to classify the type of polyketide synthase in which they reside. In this thesis, I employed a range of microbiological, genomic, and bioinformatic techniques to probe for untapped polyketide biosynthetic potential across Earth's microbiomes and the tree of life.

In my second chapter, I assessed PKS diversity and distributions by classifying KS domains across 137 metagenomes using NaPDoS2. From this, biomes were found to be differentially enriched in type I KS domains, providing a roadmap for future biodiscovery strategies. Furthermore, KS phylogenies reveal sediment-specific clades that do not include biochemically characterized PKSs, highlighting the biosynthetic potential of poorly explored environments. After exploring metagenomes, I was curious how polyketide biosynthetic potential varies across the tree of life. To investigate this, I analyzed the KS diversity from over 600,000 genomes across the tree of life in chapter 3. This work illuminated that underexplored taxonomic lineages carry KS domains that differ from known polyketide biosynthetic pathways. Finally, in chapter 4, I used multi-omics (KS amplicon sequencing, metagenomics and metabolomics) to assess the biosynthetic potential in 5000-meter deep abyssal sediments. KS amplicon sequencing showed that abyssal sediments have distinct KS sequence signatures compared to nearshore sediments. Environmental metabolomes and KS amplicon communities were also linked to distinct biogeochemical regimes across abyssal sites and sediment horizons,

raising the possibility that certain PKS pathways are crucial for navigating specific microenvironments. Together, my work suggests that abyssal sediments, poorly explored biomes, and understudied taxonomic lineages harbor unique opportunities for natural product discovery.

CHAPTER 1. Introduction

1.1 Diversity and function of microbes on Earth

For billions of years, microbes have been foundational to life in every biome of planet Earth. While microorganisms (bacteria, archaea, fungi, protists, and viruses) are usually invisible to the human eye, they play numerous roles that are integral from both an ecological and human perspective. Across ecosystems, microbes play essential roles in Earth's major elemental cycles (hydrogen, carbon, nitrogen, oxygen, sulfur, and phosphorus) by using thermodynamic reactions to recycle nutrients and biomass (Falkowski et al., 2008). Microbes are also critical and versatile parts of global food webs - in oceans cyanobacteria and diatoms serve as primary producers, and within biomes, microbes convert dissolved organic carbon into biomass that sustains higher trophic levels (Smetacek et al., 2002).

From a human lens, our bodies are host to trillions of microbial cells that combine to positively and negatively regulate human health. The composition of the human gut microbiome, for example, is essential to maintaining healthy digestion and nutrition within our bodies (Thursby et al., 2017), and fecal microbiomes have been transplanted from healthy to diseased patients with remarkable success (Park et al., 2021). On the other hand, each year more than 10 million people die from microbe-derived infectious diseases (Gray et al., 2022), so understanding how microbes interact with the human body is critical. Industrially, microbes are useful in food production, improving agricultural yields, as producers of natural ingredients used in cosmetics, as biofuel producers, in fermentation processes, and as bioremediation agents (Salwan et al., 2022). Additionally, perhaps one of the most beneficial applications of microbes is within drug

discovery, as bacterial and fungal-derived compounds are important sources of therapeutic agents (Rani et al., 2021).

The field of microbiology has advanced significantly since the first successful cultivation of microorganisms by Louis Pasteur in the 1860s. This is in part due to the astonishing advancement in sequencing capability since the first double-stranded DNA genome (bacteriophage lambda) was sequenced in 1982 (Sanger et al., 1982). In particular, high-throughput sequencing of the 16S rRNA gene has advanced our knowledge of microbial diversity quite remarkably. However, our understanding of microbial processes is still limited as Earth is home to an estimated 10^{30} total microbial cells, encompassing over 1 trillion microbial species, with over 99.99% of them yet to be cultivated (Locey et al., 2016). Even at broader taxonomic levels, it has been estimated that 81% of microbial cells belong to uncultured genera and 25% of microbial cells belong to uncultured phyla (Lloyd et al., 2018). It is important to note that the exact cutoff for estimating microbial species-level 16S rRNA diversity is a work in progress - early studies used 97% similarity thresholds, while many newer studies are shifting to unique amplicon sequence variants (ASVs) (Chiarello et al., 2022).

Comparisons across biomes have illustrated that free-living microbial communities are much more diverse than host-associated microbial communities and that saline and non-saline microbial communities differ in microbial composition (Thompson et al., 2017). Further, marine sediments are one of the most microbially species-rich environments (Thompson et al., 2017), while also containing the highest levels of microbial cell abundance and the most phylogenetically novel genera (Lloyd et al., 2018). Indeed, recent work showed that less than 3% of tropical sediment microbial communities are readily culturable (Demko et al., 2021). Remarkably, microbial density and diversity persist at smaller spatial scales - in just a single

sand grain, over 100,000 bacterial cells could be identified, belonging to more than 10,000 distinct operational taxonomic units (OTUs) (Probandt et al., 2018). While it is important for marine microbiologists to generate pure strains from uncultured lineages to understand microbial metabolism and traits, the sheer quantity of this microbial dark matter within sediments highlights the need to analyze marine microbiomes using culture-independent methods.

1.2 Metabolites are the language of microbial communication

Microorganisms interact with their surrounding environment through the production of secondary metabolites, which they use to regulate growth, obtain nutrients, signal, and compete with other microbes (Krautkramer et al., 2021). Microbes, like other organisms across the tree of life, can produce both primary and secondary metabolites, with the two differentiated by the roles they play for a given organism. Primary metabolites are used for basic metabolic functions that support the growth, development, and maintenance of a given microbe. Essential primary metabolites include nucleotides, amino acids, sugars, and fatty acids; these in turn, are building blocks for larger, more complex molecules such as DNA, RNA, proteins, mono and polysaccharides, and membrane phospholipids (Sanchez et al., 2008).

In contrast, secondary metabolites refer to small molecules that are not required for growth and general life functions, but instead provide a competitive edge for the producer (Davies et al., 2013). Also termed specialized metabolites or natural products, it can be difficult to gauge the exact ecological functions these compounds play, but they have been shown to have important roles in microbial defense, competition, symbiosis, signaling, development, pigmentation, and resource acquisition (Fouillaud et al., 2022). Microbially produced specialized

metabolites such as saxitoxin, tetrodotoxin, and domoic acid can even directly impact humans and other organisms at the top of the food chain (Vilariño et al., 2018).

While specialized metabolites can be produced across the tree of life and contain a vast diversity of chemical structures, there are a few common scaffolds that describe the majority of natural products. Peptidic natural products such as RiPPs (ribosomally synthesized and post-translationally modified peptides) and NRPs (non-ribosomal peptides) are formed via the condensation of amino acids, with the two differing in their modes of synthesis (Mordhorst et al., 2023). NRPs are synthesized via large assembly-line multienzyme complexes that incorporate both proteinogenic and non-proteinogenic amino acids, while RiPPs are generated in the ribosome from a precursor peptide and use post-translational modifications to convert proteinogenic to non-proteinogenic amino acids (Wenski et al., 2022). Terpenes are another widespread class of chemical scaffold - these compounds are built from combinations of five-carbon isoprene units. While seen across the tree of life, terpenes are most common in plants where they can function as antimicrobial defense agents (Masyita et al., 2022). Polyketides are a fourth large class of natural products. They are assembled from simple acyl building blocks into a wide diversity of carbon skeletons (Nivina et al. 2019). Other common natural product scaffolds include alkaloids (basic, nitrogen-containing organic compounds) and glycosides (a carbohydrate attached to another non-sugar compound via a glycosidic bond) (Davison et al., 2019).

1.3 Marine natural product discovery

Many specialized metabolites are important therapeutics, and the vast structural diversity seen across natural products make them useful as antibiotics, antitumor agents, antifungals, and

antivirals, among others (Davison et al., 2019). While the majority of natural products that have become FDA-approved drugs have been derived from plants (47%), bacteria (30%), and fungi (23%) (Patridge et al., 2016), recent work has highlighted that animals and protists are also promising sources of specialized metabolites (Torres et al. 2019). Humans have used natural products for thousands of years, with the earliest documented example recorded from ancient Mesopotamia around 2600 B.C. where oils from cypress trees were used to treat sickness (Dias et al., 2012). Another prominent natural therapeutic that has been used for more than two thousand years is salicylic acid, which is a compound found within the bark of willow trees that is used to ease pain and lower fevers (Dias et al., 2012). Today, salicylic acid forms the base of the commonly used drug Aspirin. Two other historically well-known natural products include morphine (extracted from the opium poppy) and penicillin (isolated from the fungus *Penicillium*), which are used for treating pain and as an antibiotic, respectively (Dias et al., 2012).

For millennia, natural product remedies have been intertwined with traditional knowledge possessed by indigenous communities across the globe, and many of the bioactive compounds from the sources are still yet to be identified (Dias et al., 2012). Within the industry of natural product discovery, many of these traditional knowledge bases or lands were exploited, with the yields of research findings kept separate from the indigenous communities (Vierros et al., 2016). In recent decades, global discussions have centered on addressing this inequitable extraction of genetic resources, with comparisons made to the way wealthy nations siphon off natural resources like oil and rare metals from abroad (Vierros et al., 2016). With the Nagoya protocol (passed in 2010) there are now guardrails in place to ensure that local communities are properly compensated for the use of their genetic resources.

While the majority of early natural products research occurred in terrestrial biomes, many marine natural products have been discovered in the past 50 years with important implications for human health. A few compounds that have been clinically approved include Ecteinascidin 743 (isolated from a tunicate, anti-cancer alkaloid, Yondelis®, PharmaMar), Ziconotide (isolated from cone snail, chronic pain treatment peptide, Prialt®, Elan), Eribulin (isolated from sponge, anti-cancer macrocyclic ketone, Halaven®, Eisai), Dolastatin (isolated from symbiotic bacteria within sea hare, anti-cancer peptide, Adcetris®, Pfizer), cytarabine (isolated from sponge, anti-cancer nucleoside, Cytosar-U®, Pfizer) and Vidarabine (isolated from sponge, antiviral nucleoside, Arasena A®, Mochida) (Dias et al., 2012). One unique feature that many marine natural products contain is halogenation, with brominated and chlorinated compounds seen much more often than in terrestrial systems (Wagner et al., 2009).

While a lot of marine natural products research has focused on invertebrate animal phyla such as Porifera and Cnidaria, microbes isolated from marine sediments have proven to be a fruitful source of specialized metabolites as well (Dias et al., 2012). In fact, certain obligate marine microbes such as the actinomycete *Salinispora* allocate up to 10% of their genome to the production of specialized metabolites, making it an ideal model organism to unearth novel natural products (Udwary et al., 2007; Jensen et al., 2015). Most notably, the compound salinosporamide A was discovered from a *Salinispora* strain (CNB-440) and showed strong anticancer activity (Fenical, 2020). After bioassay-guided fractionation, the structure of salinosporamide A was shown to contain an unusual β-lactone-γ-lactam ring system (Feling et al., 2003). Salinosporamide A is a potent proteasome inhibitor and has the unique ability to cross the blood-brain barrier (Di et al., 2016), making it an attractive target for the treatment of glioblastoma (brain tumor). The drug Marizomib (salinosporamide A) has progressed to phase

III clinical trials, although recent studies show that when used in addition to standard radiochemotherapy treatment, Marizomib does not significantly increase the overall survival rate of patients with glioblastoma (Roth et al., 2024). Nonetheless, the discoveries of salinosporamide A, along with other promising microbial natural products from marine sediment-derived bacteria such as cyclomarin A (anti-inflammatory), marinopyrrole (antibiotic), merochlorin A (antibacterial), and anthracimycin (antibacterial) (Fenical, 2020) have firmly established free-living microbial communities to be promising sources of novel therapeutics.

1.4 Biosynthetic gene clusters

The rise of genomic and metagenomic sequencing has allowed researchers to map out complete microbial genomes, and therefore visualize the genes responsible for specialized metabolite production (Udwary et al., 2007). In many organisms, particularly most bacteria and fungi, the enzymatic pathways responsible for the production of a specialized metabolite are physically clustered together in what is termed a biosynthetic gene cluster (BGC). The clustering of biosynthetic genes is predicted to help with the regulation of the pathway and the ability of pathways to move between microbes via horizontal gene transfer (Crits-Christoph et al., 2021). The various types of biosynthetic gene clusters contain distinct core genes that encode for the carbon skeletons associated with each class of natural products. For example, polyketide compounds are derived from polyketide synthases (PKSs), non-ribosomal peptides are linked to non-ribosomal peptide synthetases (NRPSs), and terpenes from terpene cyclases (TCs) (Blin et al., 2023). These conserved biosynthetic genes have been enormously helpful in the rapid identification of BGCs within genomic assemblies and form the foundation for genome mining

tools such as antiSMASH or NaPDoS2 (Blin et al., 2023; Klau, Podell, and Creamer et al., 2022).

In addition to genome mining tools, having repositories of BGCs that have been experimentally linked to their corresponding natural product, such as the MIBiG database, is important for comparative analyses and predicting the function of specific genes (Terlouw et al., 2023). By identifying the various enzymes used in nature's biosynthetic pathways, it also becomes easier to improve the rational design of novel compounds. While linking BGCs to their associated specialized metabolite remains the gold standard, there remain some pitfalls in this process. For one, microbial genomes can contain more than 50 BGCs, and if none show any similarity to previously characterized pathways, it can be difficult to ascertain which BGC corresponds to which metabolite (Chase et al., 2023). Additionally, not all BGCs are necessarily expressed by the microbe, and these so-called cryptic BGCs often require genetic manipulation, alternative culture conditions, or the introduction of constitutive promoters for activation (Hoskisson et al., 2020).

1.5 Polyketides

Polyketides are a diverse and bioactive group of natural products that utilize acyl-CoA precursors to form carbon skeletons via consecutive Claisen condensation reactions (Nivina et al., 2019). However, differences in PKS BGC structure can lead to polyketide structural diversity ranging from small aromatic compounds to large macrolides. Many polyketides are medicinally-relevant, such as antibiotics (erythromycin, oxytetracycline, doxycycline), antifungals (amphotericin), anticancer agents (epothilone and doxorubicin), antitumor agents (calicheamicin), and cholesterol-lowering drugs (lovastatin) (Korman et al., 2010) (Fig. 1.1).

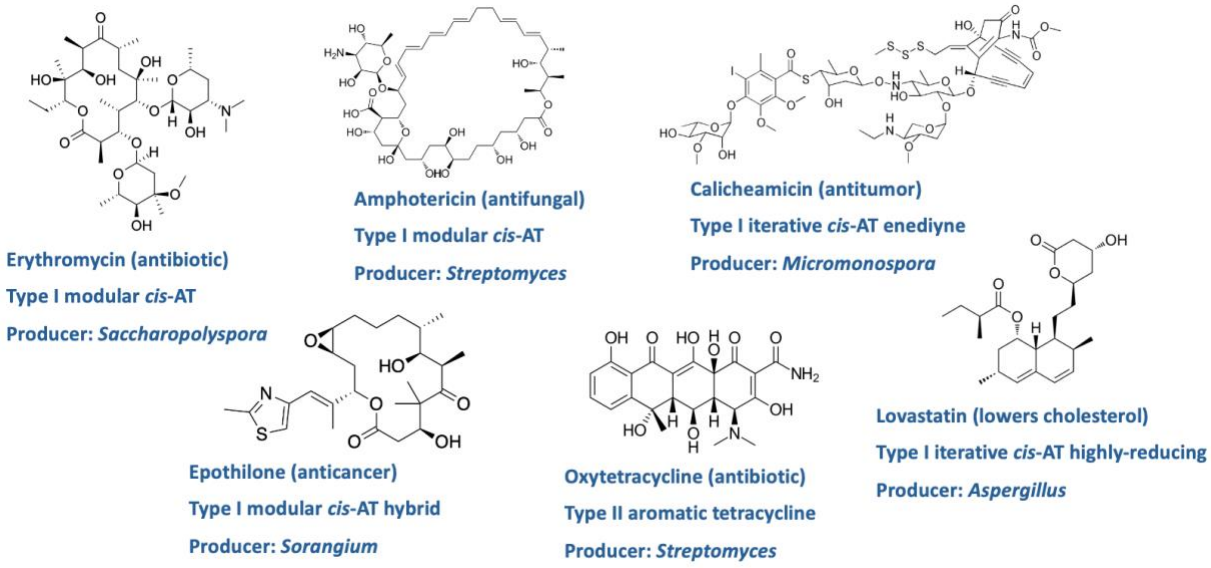

Figure 1.1. Medicinally-important polyketides. Compound name followed by type of bioactivity, NaPDoS2 classification of the PKS, and the name of the producing organism is listed for each compound.

Polyketide synthases (PKSs) are multifunctional enzymes responsible for producing polyketides and fall into three main types (I, II, III) (Hertweck et al., 2009). Type I PKSs are typically organized into modules and can either resemble an assembly line where multiple modules each catalyze a single elongation of the polyketide chain or act iteratively where a single module catalyzes multiple polyketide chain elongations (Shen, 2003). There are two main groups of assembly-line PKSs - *cis*-AT and *trans*-AT. Each module within both assembly-line and iterative PKSs typically contains at minimum an acyltransferase (AT) domain, an acyl carrier protein (ACP) domain, and a ketosynthase (KS) domain (Nivina et al., 2019). The AT domain selects the acyl-CoA chain extender unit and loads it onto an ACP domain. The ACP domain then transfers the extender unit to the KS domain, and the KS domain catalyzes the Claisen condensation reaction and extends the growing polyketide chain (Nivina et al., 2019). Trans-AT PKSs are a major exception to this rule - in contrast to *cis*-AT PKSs, *trans*-AT PKSs lack an AT domain within each module, and this function instead occurs via a free-standing AT

domain that can often be shared across multiple PKS modules (Helfrich et al. 2016). Type I PKS modules can also contain tailoring domains such as ketoreductase (KR), dehydratase (DH), and enoylreductase (ER) domains, which function to reduce the β-ketone to a hydroxyl, dehydrate the hydroxyl to form a double bond, and reduce the double bond to a methylene group, respectively (Nivina et al., 2019).

Canonical type II PKSs are mostly found in bacteria and linked to either the production of small polycyclic aromatic compounds or highly reduced polyenes. Typically, type II PKSs function iteratively and contain two subunits, with the first responsible for the elongation of the chain, and the second determining the chain length (Wang et al., 2020). In contrast to type II PKSs, type III PKSs are mainly found in plants and bacteria, do not employ carrier proteins, and resemble chalcone synthases to produce a variety of flavonoid, stilbene, and lipid-like structures via iteration (Katsuyama et al., 2012).

1.6 Genome mining to inform natural product discovery

Both natural product discovery and synthetic chemistry have been explored as methods to discover novel therapeutics. While the pharmaceutical industry has poured significant resources into the generation of large compound libraries via combinatorial chemistry, these processes are limited by the number of reactions chemists can perform, and the resulting compounds pale in comparison to the chemical diversity found in nature (Feher et al., 2003). Traditionally, specialized metabolites have been isolated through chemical extractions of either environmentally collected organisms or laboratory cultured cells. These chemical extracts can then be tested using bioassays and further fractionated to identify the specialized metabolite of interest (Atanasov et al., 2021). While this methodology was extremely successful during the

golden age of natural product discovery, there has been a noticeable decrease in the isolation of novel chemical structures due to high rediscovery rates (Pye et al. 2017). Nonetheless, advances in the sensitivity of tools used in analytical chemistry, along with creative approaches such as environmental metabolomics or in situ metabolite capture (Bogdanov et al., 2024), have ensured that natural product discovery remains important to finding new pharmaceuticals.

In the last few decades, the rapid development of sequencing technologies and the deposition of large amounts of sequencing data into online databases have supported another method for natural product discovery, genome mining. This approach utilizes bioinformatic tools such as antiSMASH 7.0, which detects and classifies biosynthetic gene clusters from genomic queries (Blin et al., 2023). Another useful genome mining web tool is NaPDoS2, which detects and classifies sequence tags (ketosynthase domains within PKSs and condensation domains within NRPSs) to give broader context about the types of PKSs and NRPSs found within metagenomic, genomic, and KS amplicon sequence data (Klau, Podell, and Creamer et al., 2022). The updated NaPDoS2 database supports the classification of KS domains into 41 distinct class and subclass assignments, allowing one to quickly gauge the polyketide biosynthetic potential of a given organism or biome (Singh et al., 2023).

To date, many genome mining studies have analyzed the biosynthetic potential within taxa of interest by running genomes through antiSMASH and analyzing the BGC outputs (Cimermancic et al., 2014; Wei et al., 2021). This has helped illuminate taxonomic lineages that could be targeted for natural product discovery. These types of genome analysis studies have been carried out across bacterial, fungal, protist, and animal lineages. In other instances, full biomes have been analyzed for BGC potential, such as the ocean microbiome, with certain notable BGCs within that biome then heterologously expressed to identify the specialized

metabolite of interest (Paoli et al., 2022). Recently, genome mining has been integrated with sequencing KS amplicons, with amplicons of interest traced back to metagenomic cosmid libraries for heterologous expression (Libis et al. 2019). These studies suggest that the future of genome mining is bright as only a small percentage of the microbial diversity on Earth has currently been sequenced.

1.7 Overview of the dissertation

The overall objectives of this dissertation is to determine how polyketide biosynthetic diversity differs across biomes, taxonomic lineages, and in understudied, abyssal sediment ecosystems. This was accomplished using a combination of metagenomic, targeted amplicon sequencing, metabolomic and bioinformatic techniques.

In Chapter 2, I evaluate whether PKS distribution and diversity vary across biomes by using NaPDoS2 to extract over 35,000 KS domains across 137 metagenomes. By analyzing datasets from soil, rhizosphere, peat soil, freshwater, seawater, freshwater sediment, marine sediment, and host-associated biomes, I established that biomes differ in their T1PKS repertoires, with soils enriched in modular *cis*-AT and hybrid *cis*-AT KSs and marine sediments enriched in PUFA and enediyne KSs. Further, I showed that marine sediment enediyne KSs can be linked to taxa that have yet to be reported to produce enediyne PKs. Finally, I utilize the metagenome-extracted KS domains to evaluate existing PCR primer sets and guide modifications used in the KS amplicon sequencing analyses I carry out in Chapter 4 of this dissertation. The work carried out in Chapter 2 has been published in the journal mBio (Singh et al., 2023).

With Chapter 3, I transition from metagenomic datasets to explore the polyketide biosynthetic potential across the genome-sequenced tree of life. In this chapter, my collaborator

(Kaitlin Creamer) and I analyzed over 600,000 genomes from all sequenced phyla, including bacterial, fungal, plasmid, animal, plant, protist, algae, archaeal, CPR, and viral lineages. NaPDoS2 allowed us to detect and classify KS distributions at the subclass level and identify taxonomic lineages that contain untapped biosynthetic potential. While biosynthetic potential has been evaluated at the BGC level within genomes before, this is the first time to our knowledge that polyketide diversity has been subclassified across the entire tree of life. This is notable, as the PKS subclass diversity reported in this chapter presents a unique opportunity to guide the discovery of specific types of polyketides. This chapter is being developed into a manuscript with the goal of publication in collaboration with co-first author Kaitlin Creamer.

In Chapter 4, I use multi-omics (16S amplicon sequencing, KS amplicon sequencing, metagenomics, and metabolomics) to evaluate the microbial communities and their biosynthetic potential within abyssal sediments. Little is known about the PKS diversity seen within abyssal sediments, which comprise more than 80% of the ocean floor. Here, targeted amplicons showed that abyssal sediments have distinct 16S and KS communities compared to nearshore sediments. Further, a transect across three abyssal sediment sites showed that amplicon community and metabolome differences mirrored the distinct biogeochemical regimes at different locations and sediment layer depths. Abyssal sediment KS phylogenies showed clades that are distinct from experimentally characterized PKS pathways, illustrating that abyssal plains carry significant biosynthetic potential. The work I have completed in Chapter 4 is in preparation for publication.

Finally, Chapter 5 highlights notable insights from this dissertation and discusses future directions that relate to this research.

1.8 References

Atanasov AG, Zotchev SB, Dirsch VM, Supuran CT. 2021. Natural products in drug discovery: advances and opportunities. Nat Rev Drug Discov 20:200–216.

Blin K, Shaw S, Augustijn HE, Reitz ZL, Biermann F, Alanjary M, Fetter A, Terlouw BR, Metcalf WW, Helfrich EJN, van Wezel GP, Medema MH, Weber T. 2023. antiSMASH 7.0: new and improved predictions for detection, regulation, chemical structures and visualisation. Nucleic Acids Research 51:W46–W50.

Bogdanov A, Salib MN, Chase AB, Hammerlindl H, Muskat MN, Luedtke S, da Silva EB, O'Donoghue AJ, Wu LF, Altschuler SJ, Molinski TF, Jensen PR. 2024. Small molecule in situ resin capture provides a compound first approach to natural product discovery. Nat Commun 15:5230.

Chase AB, Bogdanov A, Demko AM, Jensen PR. 2023. Biogeographic patterns of biosynthetic potential and specialized metabolites in marine sediments. The ISME Journal 17:976–983.

Chiarello M, McCauley M, Villéger S, Jackson CR. 2022. Ranking the biases: The choice of OTUs vs. ASVs in 16S rRNA amplicon data analysis has stronger effects on diversity measures than rarefaction and OTU identity threshold. PLoS ONE 17:e0264443.

Cimermancic P, Medema MH, Claesen J, Kurita K, Wieland Brown LC, Mavrommatis K, Pati A, Godfrey PA, Koehrsen M, Clardy J, Birren BW, Takano E, Sali A, Linington RG, Fischbach MA. 2014. Insights into Secondary Metabolism from a Global Analysis of Prokaryotic Biosynthetic Gene Clusters. Cell 158:412–421.

Crits-Christoph A, Bhattacharya N, Olm MR, Song YS, Banfield JF. 2021. Transporter genes in biosynthetic gene clusters predict metabolite characteristics and siderophore activity. Genome Res 31:239–250.

Davies J. 2013. Specialized microbial metabolites: functions and origins. J Antibiot 66:361–364.

Davison EK, Brimble MA. 2019. Natural product derived privileged scaffolds in drug discovery. Current Opinion in Chemical Biology 52:1–8.

Demko AM, Patin NV, Jensen PR. 2021. Microbial diversity in tropical marine sediments assessed using culture-dependent and culture-independent techniques. Environmental Microbiology 23:6859–6875.

Di K, Lloyd GK, Abraham V, MacLaren A, Burrows FJ, Desjardins A, Trikha M, Bota DA. 2016. Marizomib activity as a single agent in malignant gliomas: ability to cross the blood-brain barrier. Neuro Oncol 18:840–848.

Dias DA, Urban S, Roessner U. 2012. A Historical Overview of Natural Products in Drug Discovery. Metabolites 2:303–336.

Falkowski PG, Fenchel T, Delong EF. 2008. The Microbial Engines That Drive Earth's Biogeochemical Cycles. Science 320:1034–1039.

Feher M, Schmidt JM. 2003. Property Distributions: Differences between Drugs, Natural Products, and Molecules from Combinatorial Chemistry. J Chem Inf Comput Sci 43:218–227.

Feling RH, Buchanan GO, Mincer TJ, Kauffman CA, Jensen PR, Fenical W. 2003. Salinosporamide A: A Highly Cytotoxic Proteasome Inhibitor from a Novel Microbial Source, a Marine Bacterium of the New Genus Salinospora. Angew Chem Int Ed 42:355–357.

Fenical W. 2020. Marine microbial natural products: the evolution of a new field of science. J Antibiot 73:481–487.

Fouillaud M, Dufossé L. 2022. Microbial Secondary Metabolism and Biotechnology. Microorganisms 10.

Gray A, Sharara F. 2022. Global and regional sepsis and infectious syndrome mortality in 2019: a systematic analysis. The Lancet Global Health 10:S2.

Helfrich EJN, Piel J. 2016. Biosynthesis of polyketides by *trans*-AT polyketide synthases. Nat Prod Rep 33:231–316.

Hertweck C. 2009. The Biosynthetic Logic of Polyketide Diversity. Angew Chem Int Ed 48:4688–4716.

Hoskisson PA, Seipke RF. 2020. Cryptic or Silent? The Known Unknowns, Unknown Knowns, and Unknown Unknowns of Secondary Metabolism. mBio 11:e02642-20.

Jensen PR, Moore BS, Fenical W. 2015. The marine actinomycete genus *Salinispora*: a model organism for secondary metabolite discovery. Nat Prod Rep 32:738–751.

Katsuyama Y, Ohnishi Y. 2012. Type III Polyketide Synthases in Microorganisms, p. 359–377. In Methods in Enzymology. Elsevier.

Klau LJ, Podell S, Creamer KE, Demko AM, Singh HW, Allen EE, Moore BS, Ziemert N, Letzel AC, Jensen PR. 2022. The Natural Product Domain Seeker version 2 (NaPDoS2) webtool relates ketosynthase phylogeny to biosynthetic function. Journal of Biological Chemistry 298:102480.

Korman TP, Ames B, (Sheryl) Tsai S-C. 2010. Structural Enzymology of Polyketide Synthase: The Structure–Sequence–Function Correlation, p. 305–345. In Comprehensive Natural Products II. Elsevier.

Krautkramer KA, Fan J, Bäckhed F. 2021. Gut microbial metabolites as multi-kingdom intermediates. Nat Rev Microbiol 19:77–94.

Libis V, Antonovsky N, Zhang M, Shang Z, Montiel D, Maniko J, Ternei MA, Calle PY, Lemetre C, Owen JG, Brady SF. 2019. Uncovering the biosynthetic potential of rare metagenomic DNA using co-occurrence network analysis of targeted sequences. Nat Commun 10:3848.

Lloyd KG, Steen AD, Ladau J, Yin J, Crosby L. 2018. Phylogenetically Novel Uncultured Microbial Cells Dominate Earth Microbiomes. mSystems 3:10.1128/msystems.00055-18.

Locey KJ, Lennon JT. 2016. Scaling laws predict global microbial diversity. Proc Natl Acad Sci USA 113:5970–5975.

Masyita A, Mustika Sari R, Dwi Astuti A, Yasir B, Rahma Rumata N, Emran TB, Nainu F, Simal-Gandara J. 2022. Terpenes and terpenoids as main bioactive compounds of essential oils, their roles in human health and potential application as natural food preservatives. Food Chemistry: X 13:100217.

Mordhorst S, Ruijne F, Vagstad AL, Kuipers OP, Piel J. 2023. Emulating nonribosomal peptides with ribosomal biosynthetic strategies. RSC Chem Biol 4:7–36.

Nivina A, Yuet KP, Hsu J, Khosla C. 2019. Evolution and Diversity of Assembly-Line Polyketide Synthases: Focus Review. Chem Rev 119:12524–12547.

Paoli L, Ruscheweyh H-J, Forneris CC, Hubrich F, Kautsar S, Bhushan A, Lotti A, Clayssen Q, Salazar G, Milanese A, Carlström CI, Papadopoulou C, Gehrig D, Karasikov M, Mustafa H, Larralde M, Carroll LM, Sánchez P, Zayed AA, Cronin DR, Acinas SG, Bork P, Bowler C, Delmont TO, Gasol JM, Gossert AD, Kahles A, Sullivan MB, Wincker P, Zeller G, Robinson SL, Piel J, Sunagawa S. 2022. Biosynthetic potential of the global ocean microbiome. Nature 607:111–118.

Park S-Y, Seo GS. 2021. Fecal Microbiota Transplantation: Is It Safe? Clin Endosc 54:157–160.

Patridge E, Gareiss P, Kinch MS, Hoyer D. 2016. An analysis of FDA-approved drugs: natural products and their derivatives. Drug Discovery Today 21:204–207.

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

Pye CR, Bertin MJ, Lokey RS, Gerwick WH, Linington RG. 2017. Retrospective analysis of natural products provides insights for future discovery trends. Proc Natl Acad Sci USA 114:5601–5606.

Rani A, Saini KC, Bast F, Varjani S, Mehariya S, Bhatia SK, Sharma N, Funk C. 2021. A Review on Microbial Products and Their Perspective Application as Antimicrobial Agents. Biomolecules 11:1860.

Roth P, Gorlia T, Reijneveld JC, De Vos F, Idbaih A, Frenel J-S, Le Rhun E, Sepulveda JM, Perry J, Masucci GL, Freres P, Hirte H, Seidel C, Walenkamp A, Lukacova S, Meijnders P, Blais A, Ducray F, Verschaeve V, Nicholas G, Balana C, Bota DA, Preusser M, Nuyens S, Dhermain F, Van Den Bent M, O'Callaghan CJ, Vanlancker M, Mason W, Weller M. 2024. Marizomib for patients with newly diagnosed glioblastoma: A randomized phase 3 trial. Neuro-Oncology 26:1670–1682.

Salwan R, Sharma V. 2022. Plant beneficial microbes in mitigating the nutrient cycling for sustainable agriculture and food security, p. 483–512. In Plant Nutrition and Food Security in the Era of Climate Change. Elsevier.

Sanchez S, Demain AL. 2008. Metabolic regulation and overproduction of primary metabolites. Microbial Biotechnology 1:283–319.

Shen B. 2003. Polyketide biosynthesis beyond the type I, II and III polyketide synthase paradigms. Current Opinion in Chemical Biology 7:285–295.

Singh HW, Creamer KE, Chase AB, Klau LJ, Podell S, Jensen PR. 2023. Metagenomic data reveals type I polyketide synthase distributions across biomes. mSystems 8:e00012-23.

Smetacek V. 2002. Microbial food webs: The ocean's veil. Nature 419:565–565.

Terlouw BR, Blin K, Navarro-Muñoz JC, Avalon NE, Chevrette MG, Egbert S, Lee S, Meijer D, Recchia MJJ, Reitz ZL, van Santen JA, Selem-Mojica N, Tørring T, Zaroubi L, Alanjary M, Aleti G, Aguilar C, Al-Salihi SAA, Augustijn HE, Avelar-Rivas JA, Avitia-Domínguez LA, Barona-Gómez F, Bernaldo-Agüero J, Bielinski VA, Biermann F, Booth TJ, Carrion Bravo VJ, Castelo-Branco R, Chagas FO, Cruz-Morales P, Du C, Duncan KR, Gavriilidou A, Gayrard D, Gutiérrez-García K, Haslinger K, Helfrich EJN, van der Hooft JJJ, Jati AP, Kalkreuter E, Kalyvas N, Kang KB, Kautsar S, Kim W, Kunjapur AM, Li Y-X, Lin G-M, Loureiro C, Louwen JJR, Louwen NLL, Lund G, Parra J, Philmus B, Pourmohsenin B, Pronk LJU, Rego A, Rex DAB, Robinson S, Rosas-Becerra LR, Roxborough ET, Schorn MA, Scobie DJ, Singh KS, Sokolova N, Tang X, Udwary D, Vigneshwari A, Vind K, Vromans SPJM, Waschulin V, Williams SE, Winter JM, Witte TE, Xie H, Yang D, Yu J, Zdouc M, Zhong Z, Collemare J, Linington RG, Weber T, Medema MH. 2023. MIBiG 3.0: a community-driven effort to annotate experimentally validated biosynthetic gene clusters. Nucleic Acids Research 51:D603–D610.

Thompson LR, Sanders JG, McDonald D, Amir A, Ladau J, Locey KJ, Prill RJ, Tripathi A, Gibbons SM, Ackermann G, Navas-Molina JA, Janssen S, Kopylova E, Vázquez-Baeza Y, González A, Morton JT, Mirarab S, Zech Xu Z, Jiang L, Haroon MF, Kanbar J, Zhu Q, Jin Song S, Kosciolek T, Bokulich NA, Lefler J, Brislawn CJ, Humphrey G, Owens SM, Hampton-Marcell J, Berg-Lyons D, McKenzie V, Fierer N, Fuhrman JA, Clauset A, Stevens

RL, Shade A, Pollard KS, Goodwin KD, Jansson JK, Gilbert JA, Knight R, The Earth Microbiome Project Consortium, Rivera JLA, Al-Moosawi L, Alverdy J, Amato KR, Andras J, Angenent LT, Antonopoulos DA, Apprill A, Armitage D, Ballantine K, Bárta J, Baum JK, Berry A, Bhatnagar A, Bhatnagar M, Biddle JF, Bittner L, Boldgiv B, Bottos E, Boyer DM, Braun J, Brazelton W, Brearley FQ, Campbell AH, Caporaso JG, Cardona C, Carroll J, Cary SC, Casper BB, Charles TC, Chu H, Claar DC, Clark RG, Clayton JB, Clemente JC, Cochran A, Coleman ML, Collins G, Colwell RR, Contreras M, Crary BB, Creer S, Cristol DA, Crump BC, Cui D, Daly SE, Davalos L, Dawson RD, Defazio J, Delsuc F, Dionisi HM, Dominguez-Bello MG, Dowell R, Dubinsky EA, Dunn PO, Ercolini D, Espinoza RE, Ezenwa V, Fenner N, Findlay HS, Fleming ID, Fogliano V, Forsman A, Freeman C, Friedman ES, Galindo G, Garcia L, Garcia-Amado MA, Garshelis D, Gasser RB, Gerdts G, Gibson MK, Gifford I, Gill RT, Giray T, Gittel A, Golyshin P, Gong D, Grossart H-P, Guyton K, Haig S-J, Hale V, Hall RS, Hallam SJ, Handley KM, Hasan NA, Haydon SR, Hickman JE, Hidalgo G, Hofmockel KS, Hooker J, Hulth S, Hultman J, Hyde E, Ibáñez-Álamo JD, Jastrow JD, Jex AR, Johnson LS, Johnston ER, Joseph S, Jurburg SD, Jurelevicius D, Karlsson A, Karlsson R, Kauppinen S, Kellogg CTE, Kennedy SJ, Kerkhof LJ, King GM, Kling GW, Koehler AV, Krezalek M, Kueneman J, Lamendella R, Landon EM, Lane-deGraaf K, LaRoche J, Larsen P, Laverock B, Lax S, Lentino M, Levin II, Liancourt P, Liang W, Linz AM, Lipson DA, Liu Y, Lladser ME, Lozada M, Spirito CM, MacCormack WP, MacRae-Crerar A, Magris M, Martín-Platero AM, Martín-Vivaldi M, Martínez LM, Martínez-Bueno M, Marzinelli EM, Mason OU, Mayer GD, McDevitt-Irwin JM, McDonald JE, McGuire KL, McMahon KD, McMinds R, Medina M, Mendelson JR, Metcalf JL, Meyer F, Michelangeli F, Miller K, Mills DA, Minich J, Mocali S, Moitinho-Silva L, Moore A, Morgan-Kiss RM, Munroe P, Myrold D, Neufeld JD, Ni Y, Nicol GW, Nielsen S, Nissimov JI, Niu K, Nolan MJ, Noyce K, O'Brien SL, Okamoto N, Orlando L, Castellano YO, Osuolale O, Oswald W, Parnell J, Peralta-Sánchez JM, Petraitis P, Pfister C, Pilon-Smits E, Piombino P, Pointing SB, Pollock FJ, Potter C, Prithiviraj B, Quince C, Rani A, Ranjan R, Rao S, Rees AP, Richardson M, Riebesell U, Robinson C, Rockne KJ, Rodriguezl SM, Rohwer F, Roundstone W, Safran RJ, Sangwan N, Sanz V, Schrenk M, Schrenzel MD, Scott NM, Seger RL, Seguin-Orlando A, Seldin L, Seyler LM, Shakhsheer B, Sheets GM, Shen C, Shi Y, Shin H, Shogan BD, Shutler D, Siegel J, Simmons S, Sjöling S, Smith DP, Soler JJ, Sperling M, Steinberg PD, Stephens B, Stevens MA, Taghavi S, Tai V, Tait K, Tan CL, Tas¸ N, Taylor DL, Thomas T, Timling I, Turner BL, Urich T, Ursell LK, Van Der Lelie D, Van Treuren W, Van Zwieten L, Vargas-Robles D, Thurber RV, Vitaglione P, Walker DA, Walters WA, Wang S, Wang T, Weaver T, Webster NS, Wehrle B, Weisenhorn P, Weiss S, Werner JJ, West K, Whitehead A, Whitehead SR, Whittingham LA, Willerslev E, Williams AE, Wood SA, Woodhams DC, Yang Y, Zaneveld J, Zarraonaindia I, Zhang Q, Zhao H. 2017. A communal catalogue reveals Earth's multiscale microbial diversity. Nature 551:457–463.

Thursby E, Juge N. 2017. Introduction to the human gut microbiota. Biochemical Journal 474:1823–1836.

Torres JP, Schmidt EW. 2019. The biosynthetic diversity of the animal world. Journal of Biological Chemistry 294:17684–17692.

Udwary DW, Zeigler L, Asolkar RN, Singan V, Lapidus A, Fenical W, Jensen PR, Moore BS. 2007. Genome sequencing reveals complex secondary metabolome in the marine actinomycete *Salinispora* tropica. Proc Natl Acad Sci USA 104:10376–10381.

Vierros M, Suttle CA, Harden-Davies H, Burton G. 2016. Who Owns the Ocean? Policy Issues Surrounding Marine Genetic Resources. Limnology & Oceanogr Bull 25:29–35.

Vilariño N, Louzao M, Abal P, Cagide E, Carrera C, Vieytes M, Botana L. 2018. Human Poisoning from Marine Toxins: Unknowns for Optimal Consumer Protection. Toxins 10:324.

Wagner C, El Omari M, König GM. 2009. Biohalogenation: Nature's Way to Synthesize Halogenated Metabolites. J Nat Prod 72:540–553.

Wang J, Zhang R, Chen X, Sun X, Yan Y, Shen X, Yuan Q. 2020. Biosynthesis of aromatic polyketides in microorganisms using type II polyketide synthases. Microbial Cell Factories 19:110.

Wei B, Du A, Zhou Z, Lai C, Yu W, Yu J, Yu Y, Chen J, Zhang H, Xu X, Wang H. 2021. An atlas of bacterial secondary metabolite biosynthesis gene clusters. Environmental Microbiology 23:6981–6992.

Wenski SL, Thiengmag S, Helfrich EJN. 2022. Complex peptide natural products: Biosynthetic principles, challenges and opportunities for pathway engineering. Synthetic and Systems Biotechnology 7:631–647.

Chapter 2. Metagenomic data reveals type I polyketide synthase distributions across biomes

2.1 Abstract

Microbial polyketide synthase (PKS) genes encode the biosynthesis of many biomedically or otherwise commercially important natural products. Despite extensive discovery efforts, metagenomic analyses suggest that only a small fraction of nature's polyketide biosynthetic potential has been realized. Much of this potential originates from type I PKSs (T1PKSs), which can be further delineated based on their domain organization and the structural features of the compounds they encode. Notably, phylogenetic relationships among ketosynthase (KS) domains provide an effective method to classify the larger and more complex T1PKS genes in which they occur. Increased access to large metagenomic data sets from diverse habitats gives opportunities to assess T1PKS biosynthetic diversity and distributions through their smaller and more tractable KS domain sequences. Here, I used the web tool NaPDoS2 to detect and classify over 35,000 type I KS domains from 137 metagenomic data sets reported from eight diverse, globally distributed biomes. Biome-specific separation was seen, with soils enriched in KSs from modular *cis*-acetyltransferase (AT) and hybrid *cis*-AT KSs relative to other biomes and marine sediments enriched in KSs associated with polyunsaturated fatty acid and enediyne biosynthesis. The phylum Actinobacteria was linked to soil-derived enediyne and *cis*-AT KSs. In contrast, marine-derived KSs associated with enediyne and monomodular PKSs were linked to phyla from which the compounds produced by these biosynthetic enzymes have not been reported. These KSs were phylogenetically distinct from those linked with experimentally characterized PKSs suggesting they may yield novel structures or enzyme functions. Finally, I employed the metagenome-extracted KS domains to evaluate the PCR primers commonly used to amplify type I KSs and identified modifications that could increase the KS sequence diversity recovered from amplicon libraries.

2.2 Introduction

Microorganisms are a valuable source of structurally diverse specialized metabolites, including many with clinically relevant biological activities (Abdel-Razek et al., 2020; Pye et al., 2017). Recent advances in DNA sequencing technologies and molecular genetics have fostered new discovery paradigms based on the detection of natural product biosynthetic gene clusters (BGCs) in microbial genomes (Ziemert et al., 2016). Instrumental to this field are online tools such as antiSMASH (Medema et al., 2011; Blin et al., 2021) and PRISM (Skinnider et al., 2017) that detect and classify BGCs within query data. Additionally, the MIBiG repository (Kautsar et al., 2020), which lists BGCs that have been experimentally linked to compounds, and IMG-ABC (Palaniappan et al., 2020), which details BGCs within sequenced microbial genomes, serve as important comparison points for genome mining efforts.

Polyketides represent a major source of pharmaceutically relevant specialized metabolites (Nivina et al., 2019). Their biosynthesis is mediated by polyketide synthase (PKS) genes, which can be classified into types I–III, depending on their domain structure (Nivina et al., 2019). Type I PKSs (T1PKSs) are composed of multidomain proteins and represent the largest source of polyketide natural products within the MIBiG repository (Kautsar et al., 2020). A minimal T1PKS comprises an acetyltransferase (AT) domain, which selects the appropriate building block, an acyl carrier protein (ACP) domain, to which the building block is tethered, and a ketosynthase (KS) domain, which catalyzes chain elongation between the growing polyketide and the ACP-bound extender unit (Weissman et al., 2004; Fischbach et al., 2006; Shen et al., 2003). Based on the organization and function of these domains, type I PKS genes can be further delineated into two primary classes, the first of which encode enzymes that function as multimodular assembly lines (referred to here as modular *cis*-AT) where each KS domain

catalyzes one round of chain elongation. Trans-AT PKS genes represent another version of these multimodular systems in which the AT domain occurs outside of the PKS gene (Piel et al., 2010). The second major class of T1PKSs generally has only one module (monomodular) with the KS domain functioning iteratively to catalyze more than one round of chain elongation (Chen et al., 2016).

It has long been recognized that KS phylogenies can be used to distinguish sequences associated with type I modular *cis*-AT, iterative, and *trans*-AT PKSs and thus make broader predictions about the types of PKS genes in which they occur (Ziemert et al., 2012; Klau et al., 2022; Wang et al., 2020; Miyanaga et al., 2018). Type I KS phylogenies can further provide insight into the types of compounds produced (e.g., enediynes, polyunsaturated fatty acids [PUFAs], polyketide–peptide hybrids) and the functional roles of KSs in polyketide assembly (e.g., loading vs extension). The web tool NaPDoS2 (Klau et al., 2022) uses DIAMOND (Buchfink et al., 2021) and a well-curated reference database to detect KS domains in genomic, metagenomic, or amplicon query data. It further classifies these domains based on their top database match, which can be used to make broader predictions about PKS diversity and distributions. In this way, NaPDoS2 circumvents the need for complete PKS gene or BGC assembly, which can be particularly challenging for highly repetitive, multimodular T1PKSs, thus making it ideal for assessing biosynthetic potential within poorly assembled metagenomic data sets.

While metagenomic data have provided important new insights into natural product biosynthetic gene diversity, PCR-amplified KS domains allow for the detection of low-frequency sequences within complex assemblages. This approach has enabled the large-scale comparison of KS diversity across environmental samples (Moffitt et al., 2003; Ayuso-Sacido et al., 2005;

Wawrik et al., 2005; Rascher et al., 2003; Shulse et al., 2011; Charlop-Powers et al., 2014; Charlop-Powers et al., 2015; Libis et al., 2019; Bech et al., 2020; Rego et al., 2020; Elfeki et al., 2021; Hochmuth et al., 2009; Della et al., 2014) and guided the discovery of novel natural products (Libis et al., 2019). The primers used to amplify KS sequences were originally designed based on modular *cis*-AT KSs detected in the phyla Actinobacteria, Cyanobacteria, and Deltaproteobacteria (Moffitt et al., 2003; Ayuso-Sacido et al., 2005; Rascher et al., 2003). While this primer set has been modified over the years (Charlop-Powers et al., 2014; Charlop-Powers et al., 2015; Libis et al., 2019), it is unclear how well it conforms to the KS diversity now being observed in metagenomic data sets. Recent evidence that this primer set would amplify relatively few of the KS domains detected in the poorly studied phyla Acidobacteria, Verrucomicrobia, and Gemmatimonadetes suggests that modifications are warranted (Crits-Christoph et al., 2018).

Assessing biosynthetic diversity using metagenomic data sets carries distinct advantages over amplicons in that complete BGCs can be captured and PCR biases avoided. For example, work to date using metagenome-assembled genomes (MAGs) has identified previously unknown or poorly studied microbial taxa, including Acidobacteria and Candidatus Eremiobacteraeota from soils (Crits-Christoph et al., 2018) and seawater (Paoli et al., 2022), respectively, that are enriched in uncharacterized BGCs and thus could be targeted for natural product discovery. However, rare community members, which are an important source of natural products (Crits-Christoph et al., 2018), will be poorly represented among the MAGs assembled from complex communities, with only 5.3 MAGs binned on average per metagenome in a recent analysis of 1,500 metagenomes (Parks et al., 2017). To date, metagenomes have largely been used to analyze the biosynthetic potential of individual biomes, with the aim of finding new products (Crits-Christoph et al., 2018; Paoli et al., 2022; Sugimoto et al., 2019; Carrion et al., 2019;

Storey et al., 2020). For example, direct cloning of metagenomic DNA from the human microbiome led to the discovery of new polyketide antibiotics including two that potentially play a role in microbe-microbe competition (Sugimoto et al., 2019). In other studies, metagenomic analyses of root endophyte microbiomes led to the identification of a non-ribosomal peptide synthetase (NRPS)-PKS BGC that played a key role in disease suppression (Carrion et al., 2019), while a sponge metagenome was used to link compounds to the microbes that produce them (Storey et al., 2020). Comparisons of PKS diversity across biomes are less common, although it was recently suggested, based on BGC distributions in MAGs, that specific chemistry is not limited or amplified by environment (Nayfach et al., 2021).

In this study, I used NaPDoS2 to detect and classify type I KS domains from eight environmental biomes. Using KS phylogenies, biome-specific clades were detected that are distinct from those associated with experimentally characterized BGCs. Additionally, on average less than 3% of the KS domains in each metagenome are associated with MAGs, supporting their value as a proxy to assess biosynthetic diversity. Finally, access to environmental KS sequences provided an opportunity to evaluate the effectiveness of a widely used type I KS primer set.

2.3 Methods

KS domain identification

One hundred and thirty-seven shotgun metagenomes representing eight biomes (agricultural/forest soil, rhizosphere, peat soil, marine sediment, freshwater sediment, seawater, host-associated, and freshwater) were selected from the JGI IMG database (Palaniappan et al., 2020) and filtered to exclude contigs <600 nucleotides using a custom script (https://github.com/spodell/NaPDoS2_website/data_management_scripts/size_limit_seqs.pl),

resulting in a combined size of 240 Gbp. NaPDoS2 (Klau et al., 2022) was used to identify KS domains using a minimum match length of 200 amino acids and a minimum E-value of 10−30. KS domains were similarly extracted from the MIBiG 2.0 database (Kautsar et al., 2020), and MAGs listed on the JGI IMG database (which they assembled using MetaBAT) (Palaniappan et al., 2020).

KS domain amino acid sequences associated with T1PKSs and FASs were identified in the NaPDoS2 output. These included the following classifications: modular *cis*-AT, *cis*-loading module, olefin synthase, iterative aromatic, iterative PTM, *trans*-AT, hybrid *trans*-AT, hybrid *cis*-AT, PUFA, enediyne, and FAS. NaPDoS2 KS classifications were verified for a randomly selected subset of metagenomic sequences across the range of KS domain types by running the associated contigs through antiSMASH 6.0 (Medema et al., 2011; Blin et al., 2021) and comparing the output. The relative abundance of each KS domain type was calculated for each metagenome as the number of KS domains/Gbp of metagenomic data and compared using a one-way analysis of variance (ANOVA) followed by Tukey's HSD test. KS domain classifications were rarified to 100 sequences per metagenome using the average from 1,000 permutations and transformed into a Bray–Curtis dissimilarity matrix using the Vegan R program (Oksanen et al., 2020). This matrix was used to perform a PCoA with significant differences between biomes identified using a permutational ANOVA with the Vegan R program.

Full-length KS domain diversity and taxonomic assignments

Full-length KS domains were filtered from the total type I metagenomic KS pool and the MIBiG 2.0 database (Kautsar et al., 2020). All type I KS domains within the NaPDoS2 reference database contain the start residues IAIVG and end residues GTNAH (with some degeneracy at

these positions). As such, metagenomic sequences were categorized as full-length if they spanned the entirety of these regions. Geneious ver. 2020.2 (Kearse et al., 2012) was used for alignments. KS richness comparisons were made by randomly selecting 580 (the minimum number in any one biome) full-length KS domains from each biome using Geneious ver. 2020.2 (Kearse et al., 2012) and clustering them into OBUs at 95%, 90%, 80%, and 70% amino acid sequence identity using UCLUST (Edgar et al., 2010). KS richness was estimated at each sequence identity level using the Chao1 index based on the average of 10 replicate analyses and compared using a one-way ANOVA followed by Tukey's HSD test. The number of KS domains in OBUs shared between biomes was calculated using pairwise comparisons between all biome combinations. Taxonomic affiliations were assigned to full-length KS domains based on the phylum of the closest NCBI (O'Leary et al., 2016) Blastp ver. 2.11.0 (Camacho et al., 2009) match (based on E-value).

KS phylogeny

An SSN of all full-length sequences was constructed using EFI (Zallot et al., 2019) with an E-value edge calculation of 100 and visualized using Cytoscape (Shannon et al., 2003). Phylogenetic trees were constructed individually for the *trans*-AT, hybrid *cis*-AT, iterative PUFA, iterative enediyne, and *cis*-AT/iterative KS clusters identified in the SSN using FastME on the ngphylogeny.fr website (Lemoine et al., 2019) with default settings and visualized using iTOL (Letunic et al., 2021). Due to the large number of KS domains identified in the *cis*-AT/iterative group (n = 3,162), they were grouped by biome and clustered into 70% OBUs using UCLUST (Edgar et al., 2010). KS sequences from the MIBiG 2.0 database that were classified in the *cis*-AT/iterative group (n = 2,149) were similarly clustered into 70% OBUs. Centroid

representatives for each OBU were used to construct FastME trees using ngphylogeny.fr (Lemoine et al., 2019) under default parameters and visualized using iTOL version 6 (Letunic et al., 2021). Similarly, for the enediyne phylogeny, context was added to the metagenomic enediyne KSs (n = 210) by extracting and clustering (70% OBUs) all enediyne KS domains from the RefSeq select genomes (n = 271) and MIBiG 2.0 (n = 11) databases. No clustering was needed prior to generating the *trans*-AT (n = 831), hybrid *cis*-AT (n = 1,746), or PUFA (n = 1,996) phylogenies.

To visualize the genomic context of select KS domains within an uncharacterized clade in the *cis*-AT/iterative phylogeny, the relevant metagenomic contigs were analyzed using antiSMASH 5.0 (Blin et al., 2021). Since only two of what appeared to be complete BGCs were detected, related KSs that clustered into 70% OBUs with the metagenome-extracted KS domains were extracted from the NCBI RefSeq (O'Leary et al., 2016) protein database (release number 200) using Blastp version 2.11.0 (Camacho et al., 2009). RefSeq genomes that contained these were then analyzed using antiSMASH 5.0 to identify the relevant BGCs. A multilocus phylogeny was constructed with the metagenome-extracted BGCs, RefSeq genome-extracted BGCs, and the closest related BGCs from the MIBiG 2.0 reference database using CORASON (Navarro-Munoz et al., 2020).

Evaluation of KS primers

The commonly used KS2F/KS2R (Shulse et al., 2011; Charlop-Powers et al., 2014; Charlop-Powers et al., 2015; Libis et al., 2019; Bech et al., 2020; Rego et al., 2020; Elfeki et al., 2021; Hochmuth et al., 2009) primer set is composed of the forward primer 5′-GCNATGGAYCCNCARCARMGNVT-3′ (translated to AMDPQQ(RS) (LIMV)) and the

reverse primer 5′-GTNCNNGTNCCRTGNSCYTCNAC-3′ (*trans*lated to VE(AG)HGT(CWRSG)T). I aligned this primer set with the metagenome-extracted type I KS domains (amino acids), and the percent matching at each amino acid residue was calculated using Geneious ver. 2020.2 (Kearse et al., 2012). The PUFA pfaA-specific primer set (Shulse et al., 2011) was similarly analyzed using the metagenome-extracted KS domains that were classified as PUFA KS01 or pfaA KS domains using NaPDoS2.

2.4 Results

Type I PKS distributions across biomes

NaPDoS2 was used to identify KS domains associated with T1PKSs in 137 shotgun metagenomes comprising 240 Gbps of data. The metagenomes represent the following eight environmental biomes: forest/agricultural soil, rhizosphere, peat soil, freshwater, seawater, freshwater sediment, marine sediment, and host-associated (Table 2.S1). In total, 35,116 KS domains were assigned to T1PKSs and an additional 409 to type I fatty acid synthases (FASs) using a minimum alignment length >200 aa. The NaPDoS2 output further delineated the non-FAS KS domains into three groups (*cis*-AT, *trans*-AT, and iterative *cis*-AT) and eight subgroups (hybrid *cis*-AT, *cis*-loading module, olefin synthase, PUFA, enediyne, aromatic, polycyclic tetramate macrolactam [PTM], and hybrid *trans*-AT) (Table 2.S2). To validate the NaPDoS2 KS classifications, representative sequences across the range of KS domain types were analyzed by running the associated metagenomic contigs through antiSMASH 6.0 (Medema et al., 2011; Blin et al., 2021). In each case, the KSs were associated with PKS genes that matched the NaPDoS2 classification (Fig. 2.S1).

The majority of metagenome-extracted type I KSs (37.5%) were classified by NaPDoS2 as *cis*-AT with no further subgroup designation. Following that, the iterative *cis*-AT PUFA (20.9%) and hybrid *cis*-AT (18.9%) designations were the next most abundant (Table 2.S2). I also analyzed the MAGs binned from each metagenome through the Joint Genome Institute (JGI) Integrated Microbial Genome (IMG) pipeline finding that, on average, only 2.7% of the type I KS domains within a given metagenome were located within MAGs (Fig. 2.S2). This highlights the fragmented nature of the metagenomic assemblies and the utility of targeting KS sequences when assessing biosynthetic potential in complex communities.

A principal coordinates analysis (PCoA) based on a Bray–Curtis dissimilarity matrix showed a significant separation of biomes based on type I KS composition [permutational multivariate analysis of variance (ANOVA), $P < 0.001$, $R2 = 0.499$] with PUFA, *cis*-AT, and hybrid *cis*-AT KS domains representing major drivers of biome separation between marine and non-marine samples (Fig. 2.1a). To further address differences in KS composition across biomes, I determined the frequency of KSs per gigabase pair (Gbp) and found that marine sediments had significantly more PUFA and enediyne sequences (Fig. 2.1b) than other biomes (Tukey's honestly significant difference [HSD], $P < 0.01$). Likewise, forest/agricultural soil and rhizosphere metagenomes encoded significantly more hybrid *cis*-AT KS domains per Gbp ($P < 0.01$), and forest/agricultural soil metagenomes encoded more *cis*-AT KS domains ($P < 0.01$) than non-soil biomes (Fig. 2.1b).

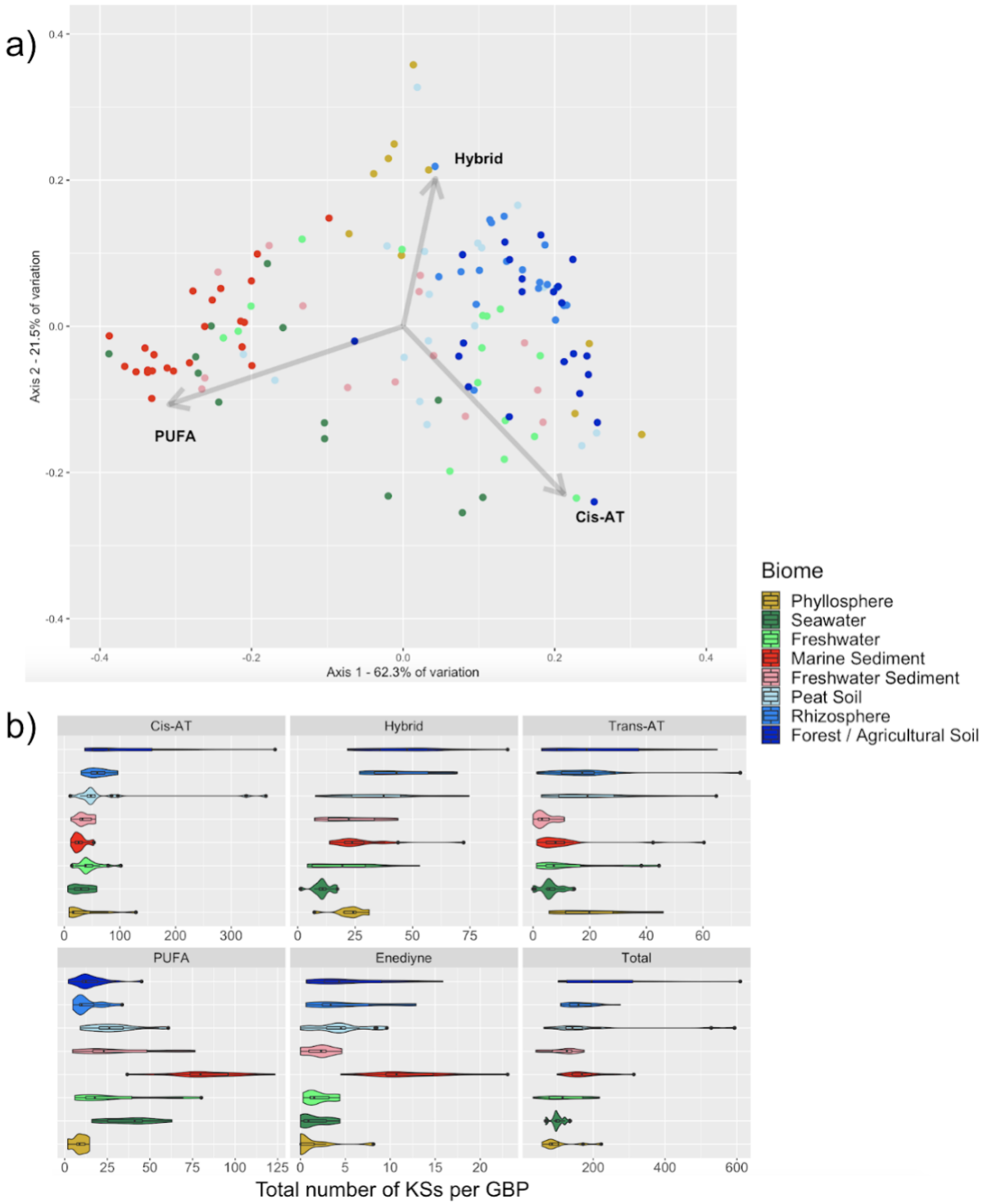

Figure 2.1. Biome-specific type I KS diversity and abundance. (a) PCoA of type I KS domain distributions after transformation using a Bray–Curtis dissimilarity matrix. Each point represents a metagenomic data set (colored by biome). Arrows indicate the three KS domain types driving the most variation. (b) Violin plots showing the number of type I KS domains per Gbp of metagenomic data across the eight biomes.

KS diversity across biomes

To evaluate KS diversity across biomes, the 7,945 full-length KS domains that could be extracted across the metagenomes were used, as they provide a standardized framework for comparison. To assess KS richness, these were clustered sequences into operational biosynthetic units (OBUs) (40) over a range of 70%–95% amino acid sequence identity and alpha diversity was calculated using Chao1 index values, a predictive measure typically applied to measure taxonomic (or operational taxonomic unit) diversity that gives more weight to rare taxa (Kim et al., 2017). Applying this approach to biosynthetic diversity, I found that soil and freshwater sediment biomes consistently carried greater OBU richness than marine sediment and seawater biomes at all clustering levels (Fig. 2.S3). However, the only significant differences in richness were observed in forest/agricultural and peat soils, which were more diverse than non-soil biomes at the 90% and 95% clustering thresholds (Tukey's HSD, $P < 0.01$).

Next, I asked how the full-length KS sequences compared with those associated with experimentally characterized PKSs in the MIBiG 2.0 database. Regardless of biome, most sequences shared little similarity with the database, with only 1% overall sharing >90% amino acid sequence identity (Fig. 2.S4). The number of KS matches dropped precipitously with increasing sequence identity across all KS types, most noticeably for hybrid and iterative PUFA KSs where 97% of the sequences had matches of <70% sequence identity. While relatively few BGCs have been experimentally characterized, this nonetheless supports the concept that considerable new polyketide diversity remains to be discovered from Earth's microbiomes. Among the few KSs that shared >90% sequence identity with the MIBiG 2.0 database, the associated BGCs represent six different biosynthetic types that account for a diverse range of natural products, including siderophores (yersiniabactin), antibiotics (e.g., rifamycin), and

cyanobacterial toxins (e.g., microcystin) (Fig. 2.S5). In previous experiences, KS sequence identity matches of >90% to characterized BGCs are good predictors of compound production (Freel et al., 2011; Gontang et al., 2010). The largest numbers of >90% MIBiG 2.0 sequence identity matches were observed in forest/agricultural soil, rhizosphere, host-associated, and freshwater biomes, while seawater and marine sediment had relatively few matches at this level, and peat soil and freshwater sediment biomes had none (Fig. 2.S5). This may provide insight into biomes that remain underexplored in terms of novel polyketide discovery.

Next, the number of KS domains in OBUs shared between biomes was identified by rarefying each biome to 580 full-length KS sequences (the lowest number in any one biome), clustering the sequences into OBUs over a range of 70%–95% amino acid sequence identity, and performing pairwise comparisons (Fig. 2.2). Overall, marine sediment and seawater biomes had the greatest number of KS sequences within shared OBUs at all clustering levels except 70%. Shared OBUs were also commonly identified in pairwise comparisons among forest/agricultural soil, peat soil, rhizosphere, and freshwater sediment biomes, and these always ranked among the top 10 in the number of KS domains within the shared OBUs. Surprisingly, at 95% clustering, no OBUs were shared between freshwater and freshwater sediments, and only the seawater and marine sediment biomes had more than 10 KS sequences within shared OBUs. In fact, at both the 95% and 90% clustering levels, many biome combinations (64% and 46%, respectively) contained no shared OBUs, with maxima of 42 and 96 KS sequences, respectively, within the OBUs shared at these clustering levels. In contrast, at the lower clustering thresholds of 80% and 70%, all biome combinations shared at least two KS sequences, and the maximum number of shared KS sequences was 195 and 416, respectively (Fig. 2.2). The lack of shared OBUs

between biomes at the higher sequence identity levels (90%–95%) suggests little overlap in the polyketides produced.

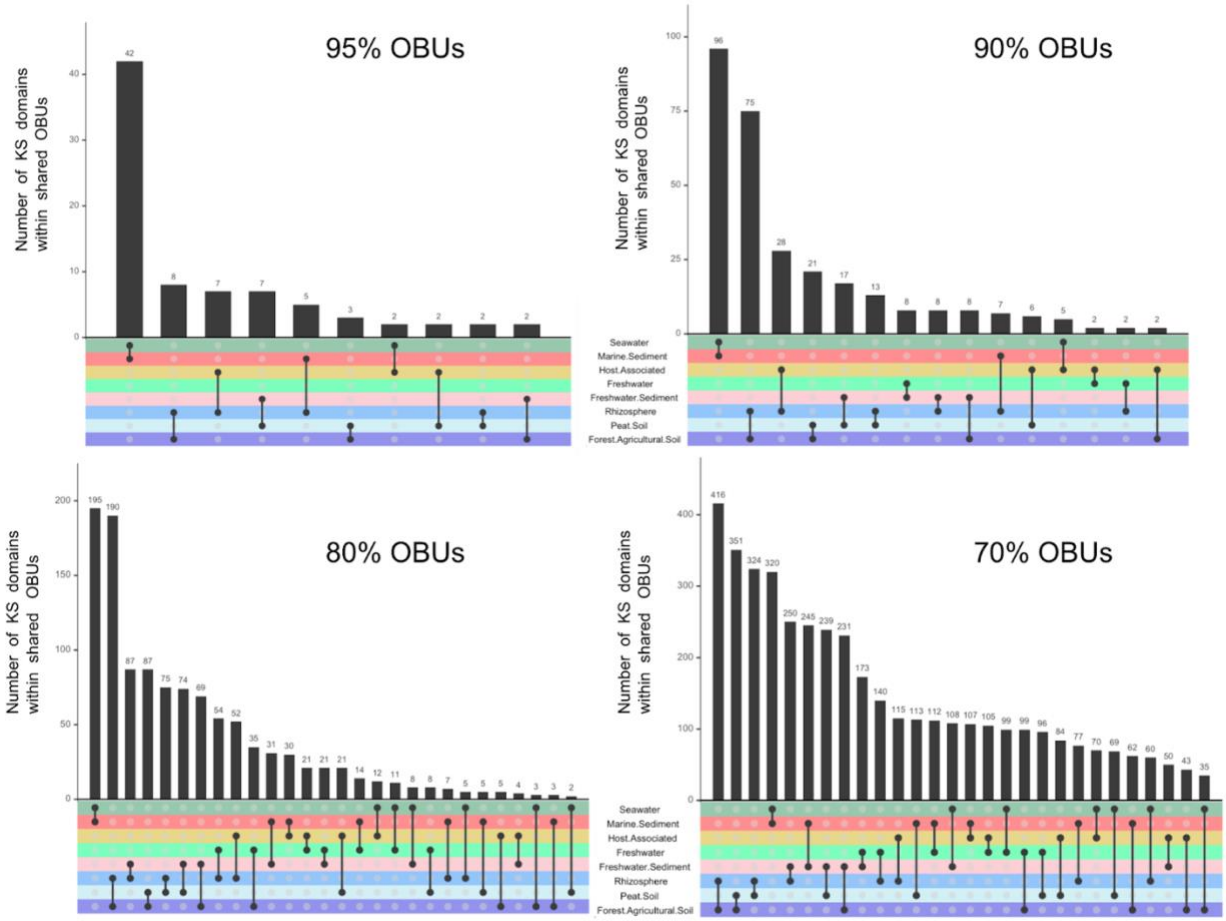

Figure 2.2. KSs shared between biomes. Each biome was rarefied to 580 full-length KS sequences (the smallest number in any one biome). These sequences were clustered into operational biosynthetic units over a range (70%–95%) of amino acid sequence identities. Biome pairwise comparisons were then made and the number of KS domains within OBUs shared between biomes determined (y-axis). Black dots connected with a line indicate biome pairs in which shared OBUs were identified.

Type I KS domains form five major groups

A sequence similarity network (SSN) was used to visualize relatedness among the 7,945 full-length metagenome-extracted KS domains in the context of their NaPDoS2 classification (Fig. 2.S6). Additionally, 3,040 full-length KS domains identified in the MIBiG 2.0 database were included (2,149 *cis*-AT/iterative, 725 *trans*-AT, 126 hybrid *cis*-AT, 29 PUFA, and 11

enediyne KS domains). The hybrid *cis*-AT (n = 1,746), *trans*-AT (n = 831), PUFA (n = 1,996), and enediyne (n = 210) KS domains are clearly distinguished within the SSN. The *cis*-AT/iterative cluster (n = 3,162) represents the fifth group, and includes KSs classified by NaPDoS2 as modular (assembly line) *cis*-AT (including olefin synthase and loading module subgroups) and iteratively acting *cis*-AT (including iterative aromatic and iterative PTM subgroups). The three PUFA KS clusters correlate with the three KS domains that are usually found in PUFA PKSs (Chen et al., 2016; Shulse et al., 2011). Next, I generated KS phylogenies to ask if biome-specific clades could be detected within each of the five major groups identified in the SSN.

Biome-specific and uncharacterized clades within the *cis*-AT/iterative group

Since *cis*-AT KS domains were a major driver of the separation among biomes (Fig. 2.1), a phylogeny was constructed using the 3,162 metagenomic and 2,149 MIBiG 2.0 sequences that comprised the *cis*-AT/iterative cluster in the SSN. This phylogeny revealed a large clade (923 sequences) in which 98.0% of the sequences mapped to soil biomes (Fig. 2.3a, yellow inner ring). This clade includes more than 50% of the MIBiG 2.0 reference sequences, all of which originate from multimodular, assembly-line *cis*-AT PKSs. Outside of this soil-dominant clade, the remaining 2,239 metagenome-extracted KS domains within the *cis*-AT/iterative group were more evenly spread between soil (50.5%) and non-soil (49.5%) biomes. Additionally, the phylogeny revealed a small clade of 89 sequences that originated from sponge metagenomes (Fig. 2.3a, brown inner ring) and was exclusive of any MIBiG 2.0 sequences. This is consistent with previous work describing an unusual clade of T1PKSs associated with sponge symbionts

(Fieseler et al., 2007) and KS domain sequences that predominate in sponge KS amplicon libraries (Hochmuth et al., 2009; Della et al., 2014).

The taxonomic affiliations of the metagenome-extracted KS domains, assessed using the closest NCBI Blastp match, revealed that 96.2% of the *cis*-AT/iterative sequences within the soil-dominant clade could be assigned to the phylum Actinobacteria (Fig. 2.3b). This is not surprising given that soil-derived Actinobacteria belonging to genera, such as Streptomyces, have been a rich source of polyketide natural products. The sequences outside of this clade had a wider taxonomic distribution (Fig. 2.3c), with 23.5% assigned to Actinobacteria, 18.2% to Cyanobacteria, and 3%–9% to low abundance phyla. In contrast with the soil-dominant clade, the sponge-specific clade displayed greater taxonomic diversity, with most sequences mapping to the phyla Gemmatimonadetes (24.7%) and Alphaproteobacteria (20.2%).

The KS phylogeny of the *cis*-AT/iterative group also revealed a large clade (379 sequences) that did not group with any MIBiG 2.0 sequences (Fig. 2.3a, pink inner ring), indicating a lack of functional characterization. The sequences in this clade were classified by NaPDoS2 as iterative PTMs, which have been reported from Actinobacteria and Proteobacteria (6) and produce macrolactams with fused carbocyclic systems that possess wide-ranging biological activities (Blodgett et al., 2010) making them of interest for natural product discovery efforts. PTM biosynthesis proceeds by an unusual hybrid iterative PKS-NRPS system that contains a single module (monomodular) with the unique hybrid domain architecture of KS-AT-DH-KR-ACP-C-A-PCP-TE (45). From searching the metagenomic assemblies, two contigs of sufficient length for antiSMASH 5.0 analysis were found, which showed that the KSs were associated with PTM-like domain architectures (Fig. 2.S7). BLAST and NCBI RefSeq were also used to identify related KSs (>70% sequence identity) in four phyla (Proteobacteria,

Verrucomicrobia, Planctomycetes, and Bacteroidetes), which determined they were all associated with monomodular PKSs that possessed PTM-like architectures (Fig. 2.S7). A multilocus phylogeny of these BGCs showed that the metagenome-extracted sequences were most closely related to RefSeq BGCs observed in Verrucomicrobia and Proteobacteria and distinct from the MIBiG 2.0 reference PTM BGCs (Fig. 2.S7). The diversity and lack of MIBiG 2.0 matches among these monomodular PKSs suggest that new PTMs await discovery.

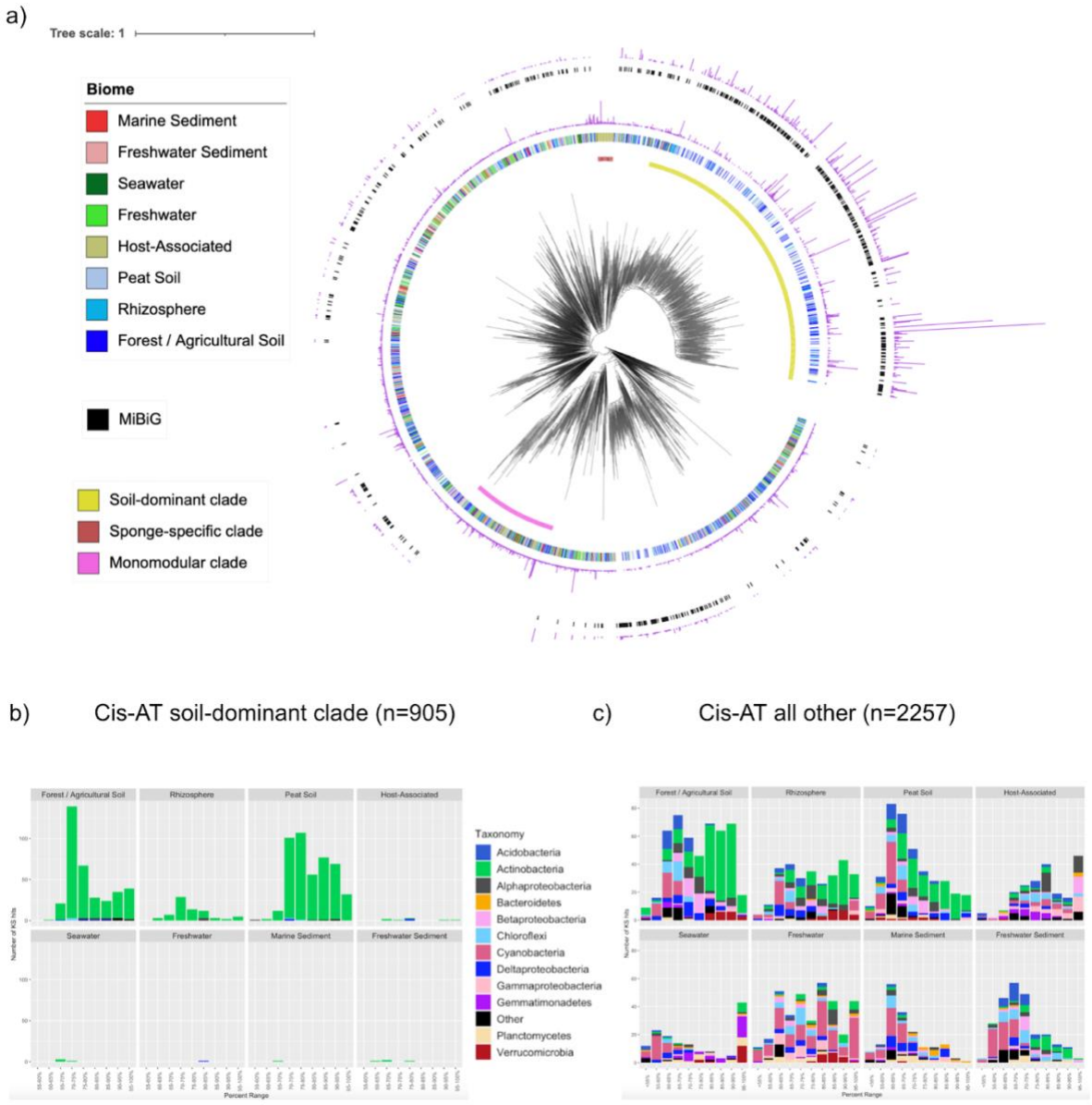

Figure 2.3. Phylogeny and taxonomic distribution of KS domains from the *cis*-AT/iterative group across biomes. (a) FastME phylogeny of full-length, metagenome- extracted, *cis*-AT/iterative group KS OBUs (70% sequence identity). The innermost ring denotes the soil-dominant clade (n = 905, yellow), a sponge-specific clade (n = 89, brown), and a monomodular clade that does not include any MIBiG 2.0 sequences (n = 379, pink). The second ring indicates the biome from which the KS was derived, and the third ring (purple) indicates the number of metagenome-extracted KS domains in each OBU. The fourth ring (black) depicts the MIBiG 2.0 database *cis*-AT/iterative group KS domains grouped into 70% OBUs, and the fifth ring (purple) shows the number of MIBiG 2.0-extracted KS sequences in each OBU. (b) Closest Blastp taxonomic match across eight biomes for the soil-dominant clade, with the x-axis denoting the range of percent similarity to the closest match and the y-axis denoting the number of KS domains. (c) Taxonomic distributions across eight biomes for all *cis*-AT/iterative group KSs other than the soil-dominant clade with the same x- and y-axis denotations.

Enediyne KS diversity across biomes

Given that marine sediments had significantly more enediyne KS sequences than other biomes (Fig. 2.1b), I assessed their novelty in comparison with enediyne KSs from the MIBiG 2.0 database (n = 11) and the NCBI RefSeq genome database (n = 271) after clustering into 70% OBUs. Enediynes represent a rare class of natural products that contain two acetylenic groups conjugated to a double bond within either a 9- or 10-membered ring. They are highly cytotoxic, have been developed into effective cancer drugs (Shen et al., 2015), and to date have only been reported from the phylum Actinobacteria and marine ascidian extracts (Rudolf et al., 2016). A phylogeny generated using the 210 full-length, metagenome-extracted enediyne KS sequences revealed a soil-specific lineage (n = 50) that co-localized with the genome-derived (NCBI RefSeq) sequences from Actinobacteria. This lineage encompassed all 11 MIBiG 2.0 enediyne KS domains (Fig. 2.4). The remaining 160 metagenome-extracted enediyne KS domains were linked to a range of phyla, including Cyanobacteria, Proteobacteria, Firmicutes, Bacteroidetes, Chloroflexi, and Spirochaetes based on their associations with the NCBI RefSeq sequences, with only two sequences of Actinobacterial origin. To further assess the non-Actinobacterial enediyne KS domains detected in RefSeq, I analyzed the respective genomes using antiSMASH 5.0 and found that all of the KS domains were associated with enediyne-like T1PKSs. Notably, 40% of the metagenome-extracted enediyne KS domains in the non-Actinobacterial portion of the phylogeny originated from marine sediments, with many (n = 31) observed in a large, sediment-specific clade. Sequences in this clade shared 85% or greater amino acid identity (NCBI BlastP) with a KS domain observed in the Deltaproteobacteria (Fig. 2.S8), suggesting a potential new source of enediyne natural products. Interestingly, four of the RefSeq enediyne KS domains were

observed in anaerobes (three from Deltaproteobacteria and one from Spirochaetes), which are also not known to produce enediyne compounds.

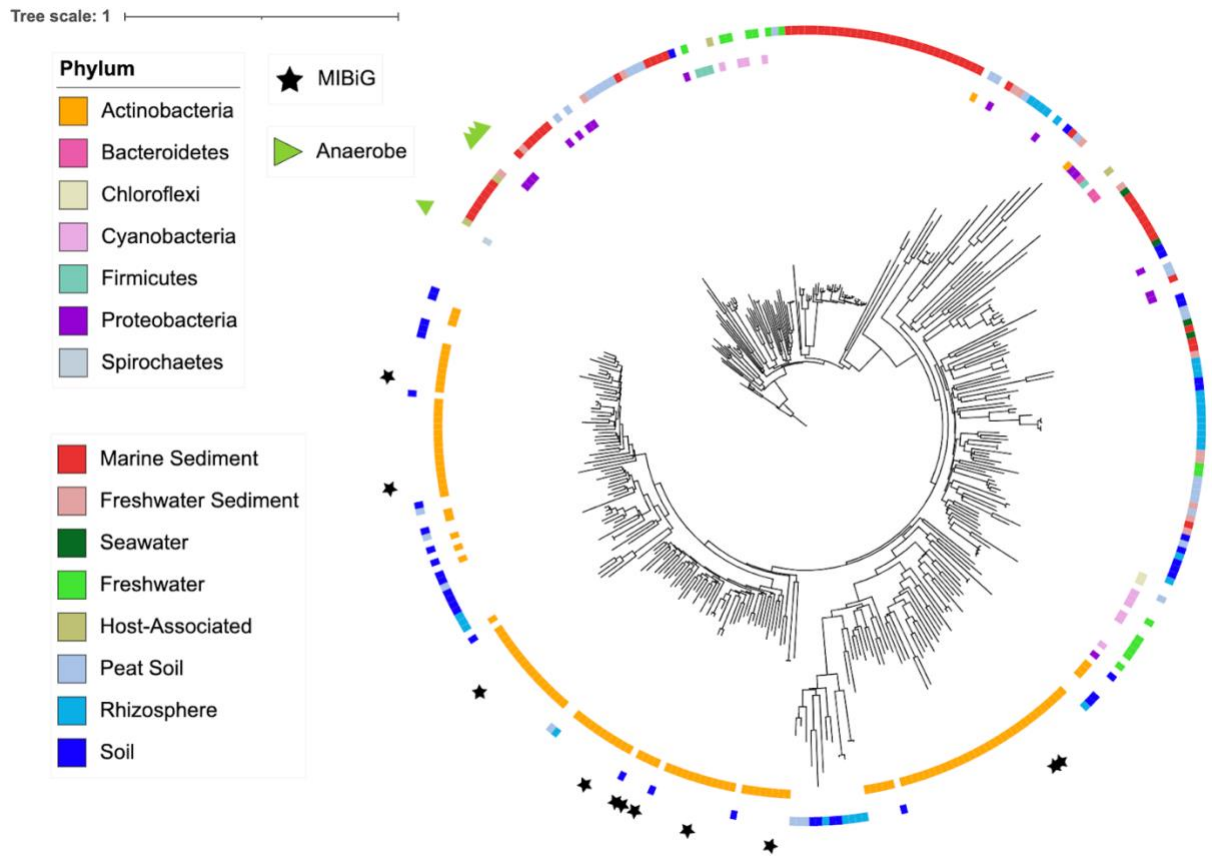

Figure 2.4. Distribution of enediyne KS domains across biomes and taxa. A FastME phylogeny was built using full-length enediyne KS domains obtained from metagenomes (n = 210, outer ring, colored by biome), the NCBI RefSeq database (inner ring, colored by taxonomy), and the MIBiG 2.0 database (stars).

Hybrid *cis*-AT, *trans*-AT, and PUFA KS domains largely lack biome specificity

The taxonomic distributions and biome specificities of the metagenome-extracted hybrid *cis*-AT, *trans*-AT, and PUFA KS clusters identified in the SSN (Fig. 2.S6) were analyzed next. Hybrid *cis*-AT KS domains catalyze the condensation of an acyl group onto a PCP-tethered intermediate. Trans-AT KS domains occur in PKSs in which the AT domain acts in trans as a stand-alone AT (Piel et al., 2010). PUFA KSs occur across a wide range of bacterial phyla and contribute to the biosynthesis of linear carbon chains with multiple *cis* double bonds (Shulse et

al., 2011). Little biome-specific clustering was seen across the hybrid *cis*-AT (n = 1,746), *trans*-AT (n = 831), and PUFA KS domains (n = 1,996) (Fig. 2.S9 to 2.S11). Based on their top NCBI Blastp matches, the hybrid *cis*-AT KS domains were most often assigned to Cyanobacteria (24.7%), the *trans*-AT KS domains to Gammaproteobacteria (27.0%) and Firmicutes (21.7%), and the PUFA KS domains to Deltaproteobacteria (27.7%). From the metagenome-extracted PUFA KS domains (n = 1,170), over 92% fell outside of those associated with known products (Fig. 2.S12), suggesting significant potential for the discovery and characterization of new PUFAs.

Metagenomic KS diversity allows for the evaluation of KS PCR primers

The metagenome-extracted KS domains analyzed here are diverse in terms of their taxonomic affiliations and biome of origin. Importantly, they are not biased toward cultured strains. As such, I saw an opportunity to evaluate the commonly used type I KS primer set (KS2F/R), which has been shown to amplify *cis*-AT, *trans*-AT, and hybrid *cis*-AT KSs across diverse bacterial phyla (Charlop-Powers et al., 2014; Charlop-Powers et al., 2015; Libis et al., 2019; Bech et al., 2020; Rego et al., 2020; Elfeki et al., 2021; Hochmuth et al., 2009; Della et al., 2014). When comparing the amino acid specificity of the forward primer (KS2F) to the metagenome-extracted KS domains, it aligned best with the *cis*-AT/iterative group, with >80% of the sequences matching the amino acids targeted by the primer (Fig. 2.S14). In contrast, only 46.0% and 40.0% of the hybrid *cis*-AT (n = 1,746) and *trans*-AT (n = 831) KS domains, respectively, matched at the third-codon position (glutamine) from the 3′ end of the KS2F primer (Fig. 2.S14). If this primer were modified to also target histidine (H) and glutamic acid (E) at this position, these percentages would go up to 88.0% for hybrid *cis*-AT and 88.9% for *trans*-AT

KSs. The KS2R reverse primer matched best with the *cis*-AT/iterative soil-dominant clade, with >90% of the sequences matching all five of the amino acids targeted by the primers. However, the other KS sequence types matched poorly with the 3′ amino acid (valine) targeted by the primer. If this 3′ position was modified to include isoleucine (I), the percentages for all KS sequence types would exceed 80%.

To date, enediyne KSs have not been reported, and PUFA KSs have rarely been reported when the KS2F/R primer set is used (Rego et al., 2020). This is consistent with my analyses, as <7% of the metagenome-extracted enediyne (n = 210) and PUFA (n = 1,996) KSs matched the amino acids targeted by the second (arginine/serine) and fourth (glutamine) positions of the forward primer (Fig. 2.S14). I also evaluated the PUFA-specific primer set pfaA, which has been used to amplify KSs from marine sediment and seawater samples (Shulse et al., 2011). The 1,170 metagenome-extracted PUFA KS domains identified by NaPDoS2 as pfaA matched well with the first five amino acids targeted by the reverse pfaA primer (glutamic acid, alanine, histidine, glycine, and threonine; Fig. 2.S13). However, <50% of these sequences matched the amino acids targeted by the forward primer at three of the four residues closest to the 3′ end (Fig. 2.S15).

Finally, given that PCR-generated KS amplicons are likely to be shorter than full-length sequences, I asked how this might affect the NaPDoS2 output by re-analyzing the full-length (~420 amino acids) metagenome-extracted *cis*-AT/iterative (n = 3,162), hybrid *cis*-AT (n = 1,746), and *trans*-AT (n = 831) KS domains after trimming them to an amplicon length typical of next-generation sequencing technologies (~138 amino acids). Overall, 94.6% of the shortened sequences yielded the same NaPDoS2 classification as their full-length counterparts (95.8% for *cis*-AT/iterative, 94.6% for hybrid *cis*-AT, and 90.5% for *trans*-AT). An SSN of the amplicon-

length KS domains showed the same clustering pattern observed in the full-length KS domain SSN (Fig. 2.S6b), further supporting the applications of NaPDoS2 for amplicon analysis.

2.5 Discussion

NaPDoS2 provides a rapid method to assess biosynthetic diversity in complex data sets by using KS and C domains to make broader predictions about PKS and NRPS genes and their small molecule products. Using the latest update to this publicly available web tool (Klau et al., 2022), which features an expanded KS classification scheme and greater capacity for large data sets, I performed an in-depth assessment of T1PKS diversity and distributions in 240 Gbp of metagenomic data representing eight environmental biomes. Significant differences in KS domain composition were observed, driven by PUFA KSs in marine biomes and modular *cis*-AT and hybrid *cis*-AT KSs in soil biomes (Fig. 2.1a). PUFAs have been suggested to aid in homeoviscous adaptation (Shulse et al., 2011), which could explain why these PKSs are enriched in marine biomes. My analyses also showed that similar biomes shared similar KS diversity, which could reflect biogeographical patterns among KS-containing microbes or environmental selection based on the functional roles of the products they encode. While a recent study of MAGs found no clear skew in relative BGC family content across Earth's microbiomes (Nayfach et al., 2020), I detected biome-specific variations when focusing on diversity within type I KSs, indicating that broad surveys can obscure more subtle but potentially important environmental differences in gene content. Furthermore, I found that only a small subset (<3% on average) of the type I KS domains occurred within MAGs, highlighting the value of using KS sequences to assess biosynthetic diversity in complex communities. While the drivers of these environmental differences cannot be distinguished here, the KS diversity discerned among

biomes can inform natural product discovery efforts and provide insights into the ecological roles of microbial natural products.

The majority of full-length, type I KS domains were classified as *cis*-AT with no further subclassification (Table 2.S2), which aligns with previous genomic explorations of type I KS diversity (O'Brien et al., 2014). The search for biome-specific clades within the *cis*-AT/iterative group revealed a soil-dominant clade that mapped almost exclusively to Actinobacteria and grouped with *cis*-AT KS domains associated with experimentally characterized assembly-line PKSs (Fig. 2.3a). While experimental characterization is biased toward certain taxa, select groups of Actinobacteria that possess large genomes may be uniquely suited for assembly-line megasynthases whose polyketide products may provide a competitive advantage in soil microbiomes. Notably, these results are consistent with previous studies that have found soil communities to be enriched in Actinobacteria compared with other biomes (Delgado-Baquerizo et al., 2018; Hoshino et al., 2020). Also aligning with previous work (Hochmuth et al., 2009; Della et al., 2014), the KS phylogeny revealed a sponge-specific clade that was distinct from all KS domains in the MIBiG 2.0 database (Fig. 2.3a), thus illustrating the potential for continued natural product discovery from sponge microbiomes. A monomodular clade that was distinct from functionally characterized sequences was also detected among the sequences classified as *cis*-AT/iterative (Fig. 2.3a; Fig. 2.S5). The affiliation of these sequences with Verrucomicrobia, Planctomycetes, and Proteobacteria complements previous studies in Streptomyces (Wang et al., 2020) and suggests that monomodular PKSs are more widely distributed than previously recognized.

Enediyne natural products are rare and of considerable importance as anticancer drugs due to their potent cytotoxicity (Shen et al., 2015). My analyses revealed that enediyne KSs were

enriched in marine sediments relative to other biomes (Fig. 2.1b). A phylogeny generated from full-length KS sequences revealed affiliations with diverse phyla such as Proteobacteria, Cyanobacteria, Firmicutes, Bacteroidetes, Chloroflexi, and Spirochaetes (Fig. 2.4), all taxa from which this class of compounds has yet to be reported. The large marine sediment enediyne lineage is most closely related to a KS domain identified in the Deltaproteobacteria. Searching for enediyne compounds from this taxon could yield new structural diversity in a biomedically important compound class. Additionally, I report the potential for enediyne PKSs in anaerobes based on the analysis of NCBI RefSeq genomes. The phylogeny also reveals a soil-specific lineage that mapped exclusively to Actinobacteria and included all the MIBiG 2.0-derived enediyne KS domains (Fig. 2.4). This agrees with previous work showing that Actinobacteria account for most of the enediyne natural products described to date (Shen et al., 2015; Rudolf et al., 2016).

Several patterns in the taxonomic distribution of KSs among bacteria were stood out. Actinobacteria were the most common taxonomic match for *cis*-AT (44%) and enediyne (28%) KS domains, whereas hybrid *cis*-AT, *trans*-AT, and PUFA domains mostly mapped to Cyanobacteria (24%), Gammaproteobacteria (28%), and Deltaproteobacteria (27%), respectively (Fig. 2.3; Fig. 2.S6 to 2.S9). In addition, 23% of the *trans*-AT KS domains mapped to Firmicutes while <4% of the other KS types mapped to this phylum. This tracks with previous reports of *trans*-AT PKSs mostly occurring in the Firmicutes and Proteobacteria phyla (Nguyen et al., 2008). While microbial gene databases are biased toward cultured representatives, these results suggest that bacterial phyla are differentially enriched in the types of polyketide genes they carry. Noting that 51.1% of the metagenome-extracted KS domains shared a sequence identity of less than 75% with the closest NCBI match (Fig. 2.S13), these taxonomic assignments can be

considered tentative and hint at the potential for natural product discovery from poorly studied taxa. Notably, this trend also extended when considering experimentally characterized PKSs within the MIBiG 2.0 database, as over 92% of metagenome-extracted KS domains shared less than 75% amino acid sequence identity with the closest database match, indicating that a wealth of PKS biosynthetic diversity remains to be characterized.

Our analyses revealed that soils and freshwater sediments held greater KS richness than marine sediments and seawater, which contrasts with previous KS amplicon work that reported marine sediments to have greater KS richness than soils (Bech et al., 2020). While the optimal clustering thresholds to group KS amplicons into meaningful biosynthetic units remain unknown, the KS richness trends observed in the metagenomes were consistent across thresholds from 70% to 95%. I also showed that when reducing full-length KS domains to next-generation amplicon lengths, 94.6% maintained the same classification (Fig. 2.S6b), thus supporting the use of KS amplicons obtained using next-generation sequencing as proxies for full-length type I KS domains.

While metagenomes are not biased toward any specific gene, they are limited in the coverage that can be obtained when complex communities are assessed. Conversely, the targeted nature of PCR can result in a more comprehensive coverage of KS sequence diversity within a given sample, while being limited to the diversity that can be amplified by the primers. While early studies using the KS2F/R primer set revealed that soil, sponge, and sediment biomes contained significant KS diversity (Charlop-Powers et al., 2014; Charlop-Powers et al., 2015; Libis et al., 2019; Bech et al., 2020; Rego et al., 2020; Elfeki et al., 2021; Hochmuth et al., 2009; Della et al., 2014), a recent study found that these primers matched poorly with KSs detected in novel soil bacteria (Crits-Christoph et al., 2018). Capitalizing on my metagenome-derived KS

data set, I found that the KS2F/R primer set aligned best with sequences in the soil-dominant clade within the *cis*-AT/iterative group (Fig. 2.S14). While these primers can recover some hybrid *cis*-AT and *trans*-AT KS sequences, their efficiency for these KS types could be improved with further primer design. Furthermore, the KS2F/R primer set matches poorly with both PUFA and enediyne KS domains (Fig. 2.S14), as did a PUFA-specific primer set (Shulse et al., 2011) (Fig. 2.S15), suggesting the need for primer modifications that maximize sequence detection within these KS types.

2.6 Conclusion

An analysis of KS domains in metagenomic data sets using NaPDoS2 revealed linkages between biosynthetic potential and environmental biomes. By processing 240 Gbps of metagenomic data, I highlight biome-specific differences in type I KS composition, with PUFA KSs driving the separation in marine biomes, and *cis*-AT and hybrid *cis*-AT KS domains driving the separation in soils. Furthermore, I show that similar biomes share more KS diversity than dissimilar biomes. Phylogenetic analyses of the metagenome-extracted KS domains revealed monomodular and enediyne clades that remain unexplored in terms of natural product discovery. Finally, my work showed that the commonly used KS2F/R primer set is biased toward modular *cis*-AT KSs and is not well designed to amplify iterative *cis*-AT enediyne and PUFA KSs. This study illustrates the applications of KS sequence tags to assess PKS diversity within complex metagenomic data sets.

2.7 Funding sources

This research was supported by the National Science Foundation Graduate Research Fellowship Program (grant no. DGE-2038238 to H.W.S.). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

2.8 Acknowledgements

Chapter 2, in full, is a reprint of the materials as it was submitted to mSystems. Singh HW, Creamer KE, Chase AB, Klau LJ, Podell S, Jensen PR. 2023. Metagenomic data reveals type I polyketide synthase distributions across biomes. mSystems 8:e00012-23.

2.9 Supplementary Figures and Tables

Table 2.S1. Summary table of all metagenomes analyzed using NaPDoS2 . All metagenomes are listed along with the biome type they fall under and the total size of the metagenome (base pairs).

Study Name	Metagenome Size (Base Pairs)	Biome	Accession number(s)
Hardwood forest soil microbial communities from Morgan-Monroe State Forest, Indiana, United States	3,204,058,601	Forest / Agriculture Soil	3300032180
Hardwood forest soil microbial communities from Morgan-Monroe State Forest, Indiana, United States	1,885,256,956	Forest / Agriculture Soil	3300032205
Hardwood forest soil microbial communities from Morgan-Monroe State Forest, Indiana, United States	1,298,045,371	Forest / Agriculture Soil	3300031715
Hardwood forest soil microbial communities from Morgan-Monroe State Forest, Indiana, United States	1,459,061,319	Forest / Agriculture Soil	3300031754
Hardwood forest soil microbial communities from Morgan-Monroe State Forest, Indiana, United States	1,474,821,223	Forest / Agriculture Soil	3300031718
Soil microbial communities from agricultural site in Penn Yan, New York, United States	1,390,993,318	Forest / Agriculture Soil	3300033551
Soil microbial communities from agricultural site in Penn Yan, New York, United States	1,285,422,822	Forest / Agriculture Soil	3300030336
Soil microbial communities from agricultural site in Penn Yan, New York, United States	1,522,946,930	Forest / Agriculture Soil	3300033550
Forest soil microbial communities from Eldorado National Forest, California, USA	1,084,380,358	Forest / Agriculture Soil	3300035687
Forest soil microbial communities from Eldorado National Forest, California, USA	999,898,556	Forest / Agriculture Soil	3300034268
Forest soil microbial communities from Eldorado National Forest, California, USA	1,071,649,480	Forest / Agriculture Soil	3300035667
Forest soil microbial communities from Barre Woods Harvard Forest LTER site, Petersham, Massachusetts, United States	1,455,646,258	Forest / Agriculture Soil	3300020579
Forest soil microbial communities from Barre Woods Harvard Forest LTER site, Petersham, Massachusetts, United States	1,494,582,847	Forest / Agriculture Soil	3300021171
Forest soil microbial communities from Barre Woods Harvard Forest LTER site, Petersham, Massachusetts, United States	1,558,207,418	Forest / Agriculture Soil	3300021168
Forest soil microbial communities from Barre Woods Harvard Forest LTER site, Petersham, Massachusetts, United States	1,363,538,788	Forest / Agriculture Soil	3300020582
Soil microbial communities from LAMPS site, Iowa State University, Ames, IA, USA	1,347,980,442	Forest / Agriculture Soil	3300037444
Soil microbial communities from LAMPS site, Iowa State University, Ames, IA, USA	1,401,762,085	Forest / Agriculture Soil	3300037529
Soil microbial communities from LAMPS site, Iowa State University, Ames, IA, USA	1,527,241,072	Forest / Agriculture Soil	3300037523
Soil microbial communities from LAMPS site, Iowa State University, Ames, IA, USA	1,490,249,658	Forest / Agriculture Soil	3300037803
Soil microbial communities from Everglades Agricultural Area, Florida, United States	1,179,311,592	Forest / Agriculture Soil	3300036759
Soil microbial communities from Everglades Agricultural Area, Florida, United States	1,305,585,406	Forest / Agriculture Soil	3300036838
Soil microbial communities from Everglades Agricultural Area, Florida, United States	1,152,188,752	Forest / Agriculture Soil	3300036840
Wetland soil microbial communities from Old Woman Creek delta, Ohio, United States	2,115,715,718	Peat Soil	3300033419
Wetland soil microbial communities from Old Woman Creek delta, Ohio, United States	1,217,355,322	Peat Soil	3300033483
Wetland soil microbial communities from Old Woman Creek delta, Ohio, United States	1,690,667,107	Peat Soil	3300033485
Wetland soil microbial communities from Old Woman Creek delta, Ohio, United States	1,546,155,278	Peat Soil	3300033414
Wetland soil microbial communities from Old Woman Creek delta, Ohio, United States	1,099,733,367	Peat Soil	3300033488
Peat permafrost microbial communities from Stordalen Mire near Abisko, Sweden	1,299,338,055	Peat Soil	3300030739
Peat permafrost microbial communities from Stordalen Mire near Abisko, Sweden	1,607,780,293	Peat Soil	3300031261
Peat permafrost microbial communities from Stordalen Mire near Abisko, Sweden	2,638,818,371	Peat Soil	3300030906
Peat permafrost microbial communities from Stordalen Mire near Abisko, Sweden	2,731,153,271	Peat Soil	3300031788
Peat permafrost microbial communities from Stordalen Mire near Abisko, Sweden	1,367,811,048	Peat Soil	3300030019
Arctic peat soil microbial communities from the Barrow Environmental Observatory site, Barrow, Alaska, USA	2,085,521,650	Peat Soil	3300006642, 3300025888, 3300006638, 3300006950
Rhizosphere microbial communities from Carex aquatilis grown in University of Washington, Seatle, WA, United States	1,337,824,726	Peat Soil	3300031344
Rhizosphere microbial communities from Carex aquatilis grown in University of Washington, Seatle, WA, United States	1,456,417,748	Peat Soil	3300028800
Rhizosphere microbial communities from Carex aquatilis grown in University of Washington, Seatle, WA, United States	1,716,094,089	Peat Soil	3300031711, 3300031712
Tropical peat soil microbial communities from peatlands in Loreto, Peru	2,435,655,096	Peat Soil	3300035643, 3300033805, 3300033806, 3300033977
Soil microbial communities from Populus trichocarpa stands in riparian zone in the Pacific Northwest, United States	1,733,731,890	Peat Soil	3300028802, 3300031226, 3300031455
Peat soil microbial communities from wetlands in Alaska, United States	1,390,497,852	Peat Soil	3300034163, 3300034196, 3300035209
Populus rhizosphere microbial communities from soil in Oregon, United States	2,047,615,555	Rhizosphere	3300036401
Populus rhizosphere microbial communities from soil in Oregon, United States	1,243,199,825	Rhizosphere	3300035692
Populus rhizosphere microbial communities from soil in Oregon, United States	1,094,358,263	Rhizosphere	3300035695
Populus rhizosphere microbial communities from soil in Oregon, United States	1,089,247,126	Rhizosphere	3300035725
Populus rhizosphere microbial communities from soil in Oregon, United States	1,597,279,552	Rhizosphere	3300037068
Corn rhizosphere microbial communities from Kellogg Biological Station, Michigan, USA	1,516,673,119	Rhizosphere	3300025917
Corn rhizosphere microbial communities from Kellogg Biological Station, Michigan, USA	1,561,591,177	Rhizosphere	3300025919
Switchgrass rhizosphere microbial communities from Kellogg Biological Station, Michigan, USA	1,567,954,971	Rhizosphere	3300025923
Switchgrass rhizosphere microbial communities from Kellogg Biological Station, Michigan, USA	1,595,709,255	Rhizosphere	3300025931
Switchgrass rhizosphere microbial communities from Kellogg Biological Station, Michigan, USA	2,410,994,453	Rhizosphere	3300005719
Switchgrass rhizosphere microbial communities from Kellogg Biological Station, Michigan, USA	1,861,559,781	Rhizosphere	3300025986
Miscanthus rhizosphere microbial communities from Kellogg Biological Station, Michigan, USA	1,638,079,859	Rhizosphere	3300025926
Miscanthus rhizosphere microbial communities from Kellogg Biological Station, Michigan, USA	1,485,183,434	Rhizosphere	3300025935
Miscanthus rhizosphere microbial communities from Kellogg Biological Station, Michigan, USA	1,136,041,079	Rhizosphere	3300005328
Miscanthus rhizosphere microbial communities from Kellogg Biological Station, Michigan, USA	1,786,668,392	Rhizosphere	3300005338
Rhizosphere microbial communities from Vellozia epidendroides in rupestrian grasslands, the National Park of Serra do Cipo, Br	957,257,553	Rhizosphere	3300021358, 3300021361, 3300021441
Maize rhizosphere microbial communities from greenhouse at UC Davis, California, United States	1,250,286,332	Rhizosphere	3300031903
Maize rhizosphere microbial communities from greenhouse at UC Davis, California, United States	1,611,017,323	Rhizosphere	3300031852
Arabidopsis thaliana rhizosphere microbial communities from the Joint Genome Institute, USA, that affect carbon cycling	967,377,510	Rhizosphere	3300027665, 3300027682, 3300027695, 3300027717, 3300027876
Populus root and rhizosphere microbial communities from Tennessee, USA	1,372,540,525	Rhizosphere	3300027907
Sediment microbial communities from Yellowstone Lake	3,482,940,763	Freshwater Sediment	3300032516
Sediment microbial communities from Yellowstone Lake	1,558,866,242	Freshwater Sediment	3300035688
Sediment microbial communities from Yellowstone Lake	1,301,812,343	Freshwater Sediment	3300031885
Sediment microbial communities from Yellowstone Lake	2,118,061,919	Freshwater Sediment	3300032046
Sediment microbial communities from Loxahatchee National Wildlife Refuge, Florida, United States	1,356,943,436	Freshwater Sediment	3300038410
Sediment microbial communities from Loxahatchee National Wildlife Refuge, Florida, United States	1,274,909,464	Freshwater Sediment	3300038550
Sediment microbial communities from Loxahatchee National Wildlife Refuge, Florida, United States	1,953,865,168	Freshwater Sediment	3300038408
Sediment microbial communities from Loxahatchee National Wildlife Refuge, Florida, United States	1,686,876,318	Freshwater Sediment	3300038455
Sediment microbial communities from Loxahatchee National Wildlife Refuge, Florida, United States	3,028,796,679	Freshwater Sediment	3300038552
Freshwater lake sediment microbial communities from the University of Notre Dame, USA, for methane emissions studies	1,235,723,351	Freshwater Sediment	3300027902
Freshwater lake sediment microbial communities from the University of Notre Dame, USA, for methane emissions studies	1,356,447,832	Freshwater Sediment	3300027900
Freshwater lake sediment microbial communities from the University of Notre Dame, USA, for methane emissions studies	1,474,716,571	Freshwater Sediment	3300027896
Sediment microbial communities from wetlands near Prado wetlands, California, USA	1,906,967,011	Freshwater Sediment	3300037458
Sediment microbial communities from wetlands near Prado wetlands, California, USA	2,119,703,235	Freshwater Sediment	3300037524
Sediment microbial communities from wetlands near Prado wetlands, California, USA	2,289,580,168	Freshwater Sediment	3300037461
Sediment microbial communities from wetlands near Prado wetlands, California, USA	1,646,841,975	Freshwater Sediment	3300037408
Freshwater microbial communities from Lake Tanganyika, Tanzania	2,767,785,519	Freshwater	3300020109, 3300020220
Freshwater microbial communities from Lake Tanganyika, Tanzania	1,302,498,413	Freshwater	3300020222
Freshwater microbial communities from Lake Tanganyika, Tanzania	1,273,692,525	Freshwater	3300020084
Freshwater microbial communities from Lake Tanganyika, Tanzania	1,460,997,943	Freshwater	3300020074
Freshwater microbial communities from Lake Fryxell liftoff mats and glacier meltwater in Antarctica	1,609,619,801	Freshwater	3300009032
Freshwater microbial communities from Lake Fryxell liftoff mats and glacier meltwater in Antarctica	1,606,851,264	Freshwater	3300009084
Freshwater microbial communities from Lake Fryxell liftoff mats and glacier meltwater in Antarctica	1,628,106,102	Freshwater	3300009083
Freshwater microbial communities from Lake Fryxell liftoff mats and glacier meltwater in Antarctica	1,278,739,541	Freshwater	3300007521
Freshwater microbial communities from Lake Bonney liftoff mats and glacier meltwater in Antarctica	1,474,819,565	Freshwater	3300007519
Freshwater microbial communities from Lake Mendota, Madison, Wisconsin, United States	1,326,969,174	Freshwater	3300035666
Freshwater microbial communities from Lake Mendota, Madison, Wisconsin, United States	1,596,426,758	Freshwater	3300036719
Freshwater microbial communities from Lake Mendota, Madison, Wisconsin, United States	2,099,022,913	Freshwater	3300034101, 3300034284
Freshwater microbial communities from Lake Mendota, Madison, Wisconsin, United States	2,061,414,356	Freshwater	3300034272, 3300034280
Freshwater microbial communities from meromictic Lake La Cruz, Castile	3,391,824,087	Freshwater	3300029286, 3300028581, 3300029268
Freshwater microbial communities from meromictic Lake La Cruz, Castile	1,486,589,259	Freshwater	3300027970, 3300028569
Freshwater microbial communities from Lake Lanier, Atlanta, Georgia, United States	4,054,488,353	Freshwater	3300023184, 3300023174, 3300022752, 3300023179
Marine sediment microbial communities from subtidal zone of North Sea	1,535,122,492	Marine Sediment	3300028600
Marine sediment microbial communities from subtidal zone of North Sea	1,029,147,879	Marine Sediment	3300028599
Marine sediment microbial communities off the coast of San Francisco, CA, United States	1,647,702,068	Marine Sediment	3300037489

Table 2.S1. (Continued) Summary table of all metagenomes analyzed using NaPDoS2 . All metagenomes are listed along with the biome type they fall under and the total size of the metagenome (base pairs).

Marine sediment microbial communities off the coast of San Francisco, CA, United States	2,382,593,820	Marine Sediment	3300038629
Marine sediment microbial communities off the coast of San Francisco, CA, United States	1,117,689,729	Marine Sediment	3300037835
Marine sediment microbial communities off the coast of San Francisco, CA, United States	1,323,119,854	Marine Sediment	3300037521
Marine sediment microbial communities off the coast of San Francisco, CA, United States	1,109,151,697	Marine Sediment	3300037450
Marine sediment microbial communities off the coast of San Francisco, CA, United States	2,555,594,725	Marine Sediment	3300038548
Marine sediment microbial communities off the coast of San Francisco, CA, United States	1,286,123,733	Marine Sediment	3300037459
Marine sediment microbial communities off the coast of San Francisco, CA, United States	1,241,427,685	Marine Sediment	3300037460
Marine sediment microbial communities off the coast of San Francisco, CA, United States	1,687,513,841	Marine Sediment	3300037540
Marine sediment microbial communities off the coast of San Francisco, CA, United States	1,507,997,651	Marine Sediment	3300037245
Marine sediment microbial communities off the coast of San Francisco, CA, United States	1,376,019,208	Marine Sediment	3300037247
Marine sediment microbial communities off the coast of San Francisco, CA, United States	2,233,392,739	Marine Sediment	3300037246, 3300022413
Coastal sediment microbial communities from Delaware Bay, Delaware, United States	1,697,423,932	Marine Sediment	3300032136
Coastal sediment microbial communities from Oude Bieten Haven, Netherlands	801,797,914	Marine Sediment	3300032251
Coastal sediment microbial communities from Maine, United States	1,509,772,824	Marine Sediment	3300033429
Coastal sediment microbial communities from Maine, United States	1,508,653,967	Marine Sediment	3300032272
Coastal sediment microbial communities from Maine, United States	1,396,560,300	Marine Sediment	3300032231
Coastal sediment microbial communities from Maine, United States	1,298,677,993	Marine Sediment	3300032258
Coastal sediment microbial communities from Maine, United States	1,135,204,813	Marine Sediment	3300032259
Coastal sediment microbial communities from Maine, United States	1,016,401,552	Marine Sediment	3300032262
Coastal sediment microbial communities from Maine, United States	1,157,046,311	Marine Sediment	3300032260
Marine microbial communities from station ALOHA, North Pacific Subtropical Gyre	2,957,439,220	Seawater	3300032820
Marine microbial communities from station ALOHA, North Pacific Subtropical Gyre	2,217,368,920	Seawater	3300032278
Marine microbial communities from station ALOHA, North Pacific Subtropical Gyre	1,666,929,232	Seawater	3300032006
Marine microbial communities from station ALOHA, North Pacific Subtropical Gyre	1,247,814,780	Seawater	3300031785
Marine microbial communities from western Arctic Ocean	2,003,936,505	Seawater	3300009786, 3300009173
Marine microbial communities from western Arctic Ocean	1,801,075,760	Seawater	3300009409, 3300009706
Marine microbial communities from western Arctic Ocean	2,592,698,426	Seawater	3300009420, 3300009526, 3300009705, 3300009785
Marine microbial communities from western Arctic Ocean	3,100,549,849	Seawater	3300031802, 3300031804, 3300036808
Seawater microbial communities from Jervis Inlet, British Columbia, Canada	2,779,312,094	Seawater	3300024518, 3300024520, 3300027865, 3300027881
Seawater microbial communities from Amundsen Gulf, Northwest Territories, Canada	1,727,509,430	Seawater	3300024521, 3300027872, 3300027997
Seawater microbial communities from Saanich Inlet, British Columbia, Canada	1,587,034,760	Seawater	3300023210, 3300027861, 3300028045
Seawater microbial communities from eastern tropical North Pacific Ocean	2,882,053,374	Seawater	3300035157, 3300035202, 3300035203, 3300036767
Seawater microbial communities from, Arabian Sea, Indian Ocean	3,094,561,910	Seawater	3300035204, 3300035205, 3300035206, 3300035250
Agave microbial communities from Guanajuato, Mexico	1,044,017,924	Host-Associated (Phyllosphere)	3300005562, 3300006020, 3300030499, 3300009144, 3300027809, 3300030505, 3300030497
Agave microbial communities from Guanajuato, Mexico	1,110,787,892	Host-Associated (Phyllosphere)	3300005661, 3300010395, 3300030515
Agave microbial communities from Guanajuato, Mexico	738,289,986	Host-Associated (Phyllosphere)	3300027761, 3300030512, 3300030516
Phyllosphere microbial communities from UC Gill Tract Community Farm, Albany, California, United States	1,300,189,725	Host-Associated (Phyllosphere)	3300031088, 3300031419
Phyllosphere microbial communities from UC Gill Tract Community Farm, Albany, California, United States	935,592,278	Host-Associated (Phyllosphere)	3300031132, 3300031370, 3300031418
Phyllosphere microbial communities from UC Gill Tract Community Farm, Albany, California, United States	1,578,246,309	Host-Associated (Phyllosphere)	3300031134, 3300031413
Phyllosphere microbial communities from UC Gill Tract Community Farm, Albany, California, United States	1,890,101,419	Host-Associated (Phyllosphere)	3300031087, 3300031110, 3300031133
Phyllosphere microbial communities from UC Gill Tract Community Farm, Albany, California, United States	1,739,354,604	Host-Associated (Phyllosphere)	3300031084, 3300031112, 3300031420
Phyllosphere microbial communities from UC Gill Tract Community Farm, Albany, California, United States	1,931,146,294	Host-Associated (Phyllosphere)	3300031372, 3300031414, 3300031416, 3300031417
Host-associated microbial community of the marine sponge Aplysina aerophoba from Gulf of Piran, Adriatic Sea	1,000,911,300	Host-Associated	3300002222, 3300002150, 3300027328, 3300027951
Host-associated microbial community of the marine sponge Aplysina aerophoba from Gulf of Piran, Adriatic Sea	1,111,906,834	Host-Associated	3300002159, 3300002448, 3300002147, 3300027327, 3300027386
Host-associated microbial community of the marine sponge Aplysina aerophoba from Gulf of Piran, Adriatic Sea	1,324,284,378	Host-Associated	3300002160, 3300002151, 3300027391, 3300027532
Marine algal microbial communities from Sidmouth, United Kingdom	1,003,807,699	Host-Associated	3300009417, 3300009439, 3300009446
Marine algal microbial communities from Sidmouth, United Kingdom	1,031,151,631	Host-Associated	3300009073, 3300009415, 3300009421, 3300027028
Tube worm associated microbial communities from hydrothermal vent at the East Pacific Rise, Pacific Ocean	1,278,100,677	Host-Associated	3300028026, 3300028029, 3300028042, 3300028534
Rumen microbial communities from sheep, dairy cows and beef cattle from various locations	2,488,892,422	Host-Associated	3300028805
Rumen microbial communities from sheep, dairy cows and beef cattle from various locations	1,559,553,778	Host-Associated	3300037453

Table 2.S2 - Type I KS hits classified by NaPDoS2. Type I KS domains listed by NaPDoS2 class and subclass across eight biomes, with the total KS hits across all biomes listed in the third column. The total number of Type I KS hits across all classes is listed in the second to last row. Type I FAS hits were not included in these totals and are listed in the last row.

Class	Subclass	All biomes	Forest / Agricultural soil	Peat Soil	Rhizosphere	Marine Sediment	Freshwater Sediment	Host-associated	Seawater	Freshwater
Modular *cis*-AT	No subclass	13186	3381	2539	1790	951	1070	1275	951	1229
	Hybrid KS	6655	1452	1105	1316	907	687	326	318	544
	Loading module	864	215	139	152	129	54	98	21	56
	Olefin synthase	719	135	79	121	98	58	27	99	102
Iterative *cis*-AT	PUFA	7346	460	791	386	2750	899	134	1141	785
	No subclass	1523	264	178	218	204	229	104	204	122
	Enediyne	1044	173	132	153	403	66	12	49	56
	Aromatic	355	140	48	79	4	34	10	7	33
	PTM-type	207	44	39	36	21	14	6	10	37
trans-AT	No subclass	3061	683	625	473	345	113	291	177	354
	Hybrid KS	156	26	35	27	12	10	15	16	15
	Total KSs	35116	6973	5710	4751	5824	3234	2298	2993	3333
Type I FAS		409	93	36	58	10	35	94	45	38

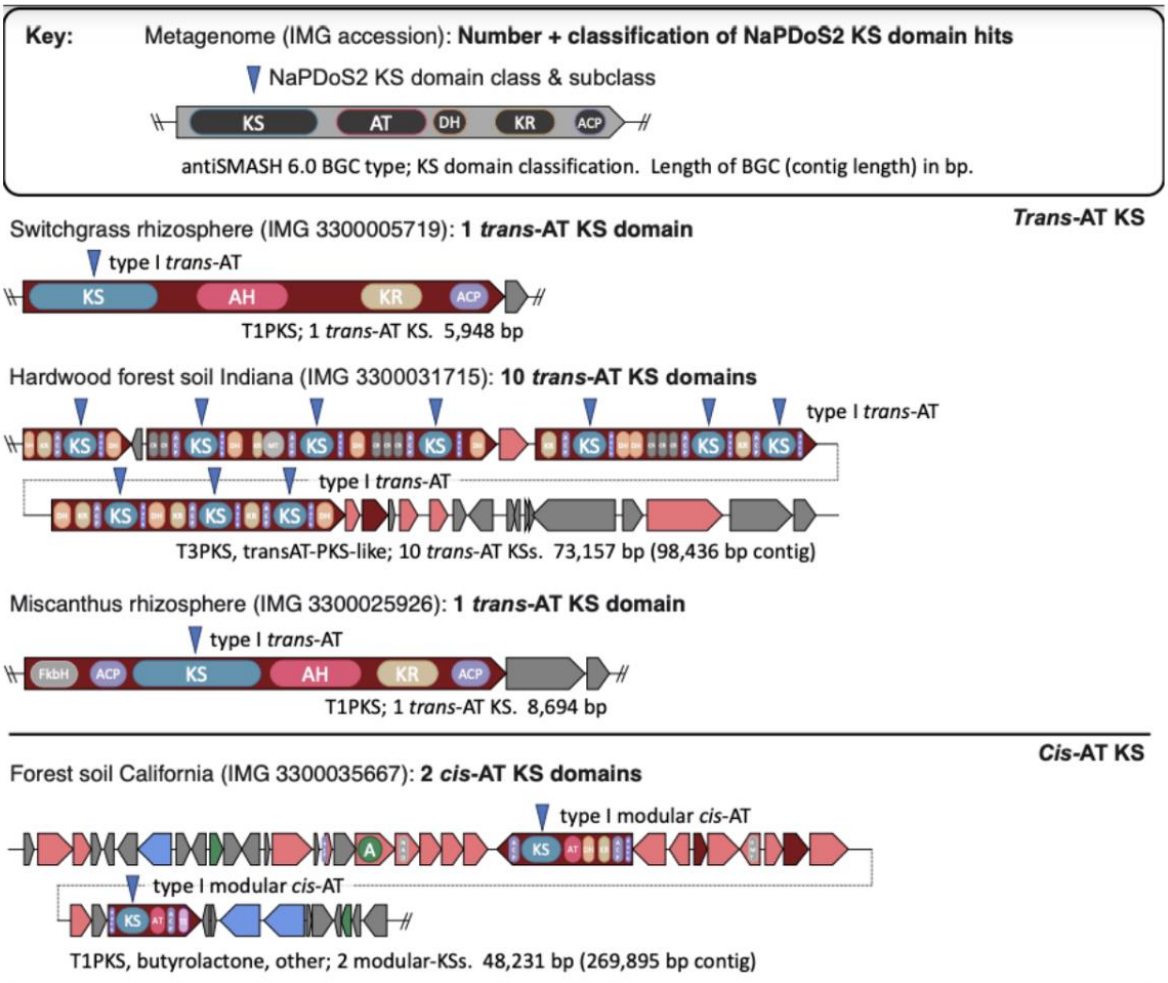

Figure 2.S1. Examples of BGC and gene neighborhood context for NaPDoS2 KS hits. Random KS hits of various subclasses including *trans*-AT, modular *cis*-AT, hybrid *cis*-AT, mixed hybrid *trans*/*cis*-AT, PUFA, *cis*-AT PUFA, and *cis*-AT enediyne) identified by NaPDoS2 were located in the metagenomes they were detected/extracted from by using blastP of the KS domains against the metagenome in the JGI IMG interface. The entire metagenome scaffold/contig with the KS hit was extracted, and run through antiSMASH 6 to identify BGC regions and important biosynthetic genes; transATor, and "PKS/NRPS Analysis Web-site" (http://nrps.igs.umaryland.edu/) for relevant BGC domain detection. The BGCs were drawn and colored as determined by antiSMASH6 (maroon, core biosynthetic gene; pink, additional biosynthetic gene; blue, transport-related genes; green, regulatory genes, gray, other genes); domain position and function were drawn and colored according to antiSMASH, transATor, and NRPS.IGS (blue, KS ketosynthase; pink: AH acyl hydrolase, AT acyl transferase; sand, KR ketoreductase; pale purple, ACP Phosphopantetheine acyl carrier protein; orange, DH dehydratase; dark gray, CR crotonase; light gray: MT methyltransferase, FkbH domain, CAL Co-enzyme A ligase domain, NAD Male sterility protein, oMT Oxygen methyltransferase, AmT aminotransferase; light pink, TE thioesterase; dark blue, dock PKS terminal docking domain; light green, C condensation domain; dark green: A adenylation domain, A-OX Adenylation domain with integrated oxidase). Blue arrows point to KS hits that NaPDoS2 detected and classified from each metagenome shown in the BGC context; arrows are labeled with the NaPDoS2 classification.

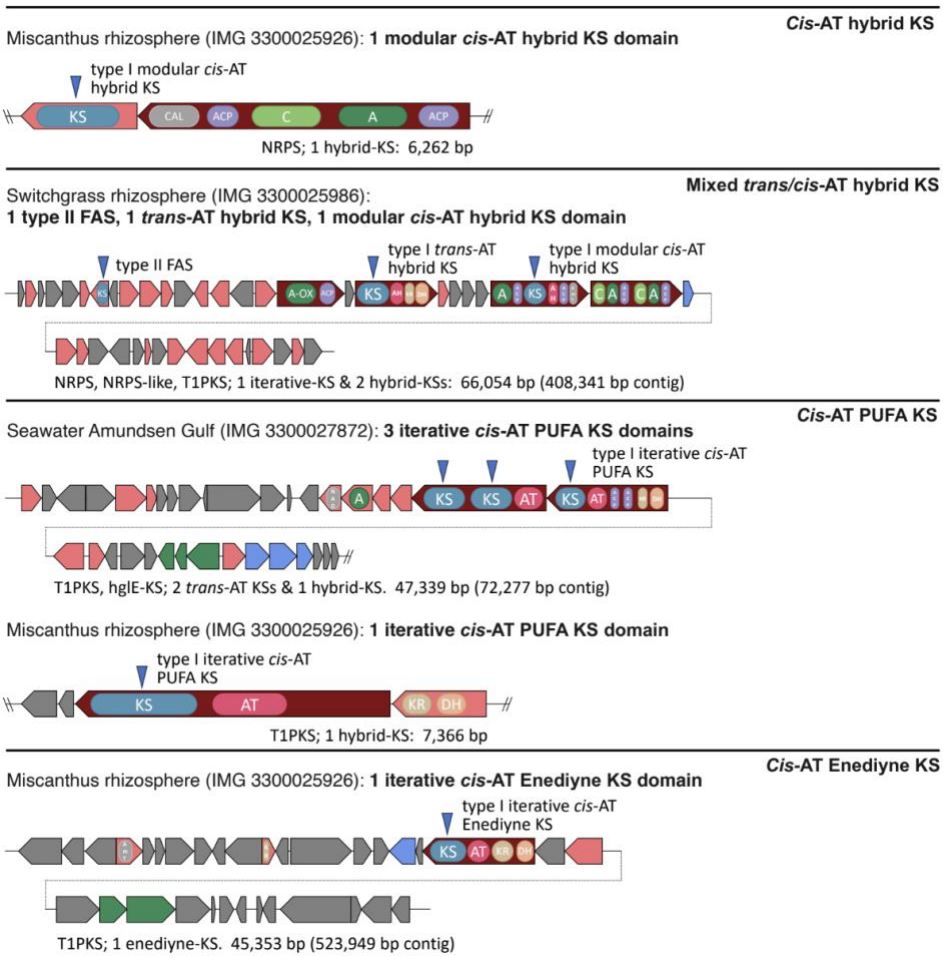

Figure 2.S1. (Continued) Examples of BGC and gene neighborhood context for NaPDoS2 KS hits. Random KS hits of various subclasses including *trans*-AT, modular *cis*-AT, hybrid *cis*-AT, mixed hybrid *trans*/*cis*-AT, PUFA, *cis*-AT PUFA, and *cis*-AT enediyne) identified by NaPDoS2 were located in the metagenomes they were detected/extracted from by using blastP of the KS domains against the metagenome in the JGI IMG interface. The entire metagenome scaffold/contig with the KS hit was extracted, and run through antiSMASH 6 to identify BGC regions and important biosynthetic genes; transATor, and "PKS/NRPS Analysis Web-site" (http://nrps.igs.umaryland.edu/) for relevant BGC domain detection. The BGCs were drawn and colored as determined by antiSMASH6 (maroon, core biosynthetic gene; pink, additional biosynthetic gene; blue, transport-related genes; green, regulatory genes, gray, other genes); domain position and function were drawn and colored according to antiSMASH, transATor, and NRPS.IGS (blue, KS ketosynthase; pink: AH acyl hydrolase, AT acyl transferase; sand, KR ketoreductase; pale purple, ACP Phosphopantetheine acyl carrier protein; orange, DH dehydratase; dark gray, CR crotonase; light gray: MT methyltransferase, FkbH domain, CAL Co-enzyme A ligase domain, NAD Male sterility protein, oMT Oxygen methyltransferase, AmT aminotransferase; light pink, TE thioesterase; dark blue, dock PKS terminal docking domain; light green, C condensation domain; dark green: A adenylation domain, A-OX Adenylation domain with integrated oxidase). Blue arrows point to KS hits that NaPDoS2 detected and classified from each metagenome shown in the BGC context; arrows are labeled with the NaPDoS2 classification.

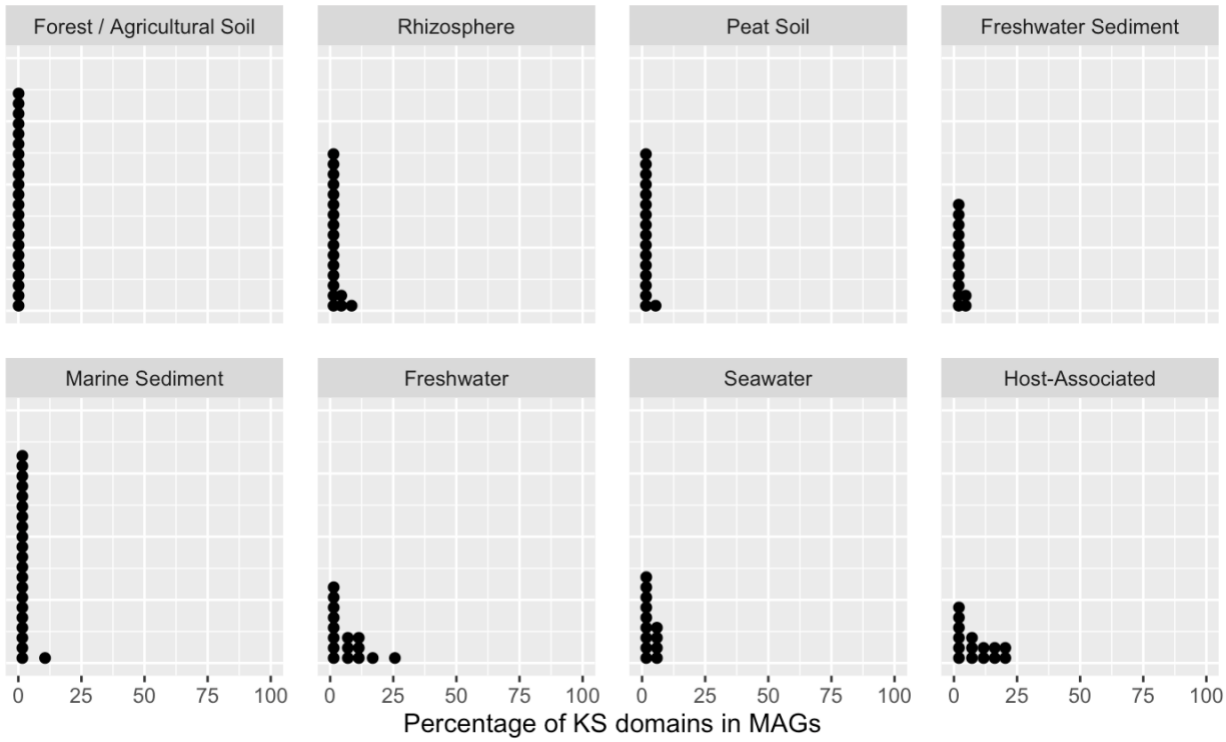

Figure 2.S2. Percentage of metagenomic KS domains detected within MAGs. Individual metagenomes (137 in total from IMG/M) are represented by black circles. MAGs were binned for each metagenome using MetaBAT according to the JGI IMG automated pipeline. The metagenomes and MAGs from those metagenomes were analyzed independently using NaPDoS2, and the percentage of KS domains in MAGs compared to the entire metagenome was plotted.

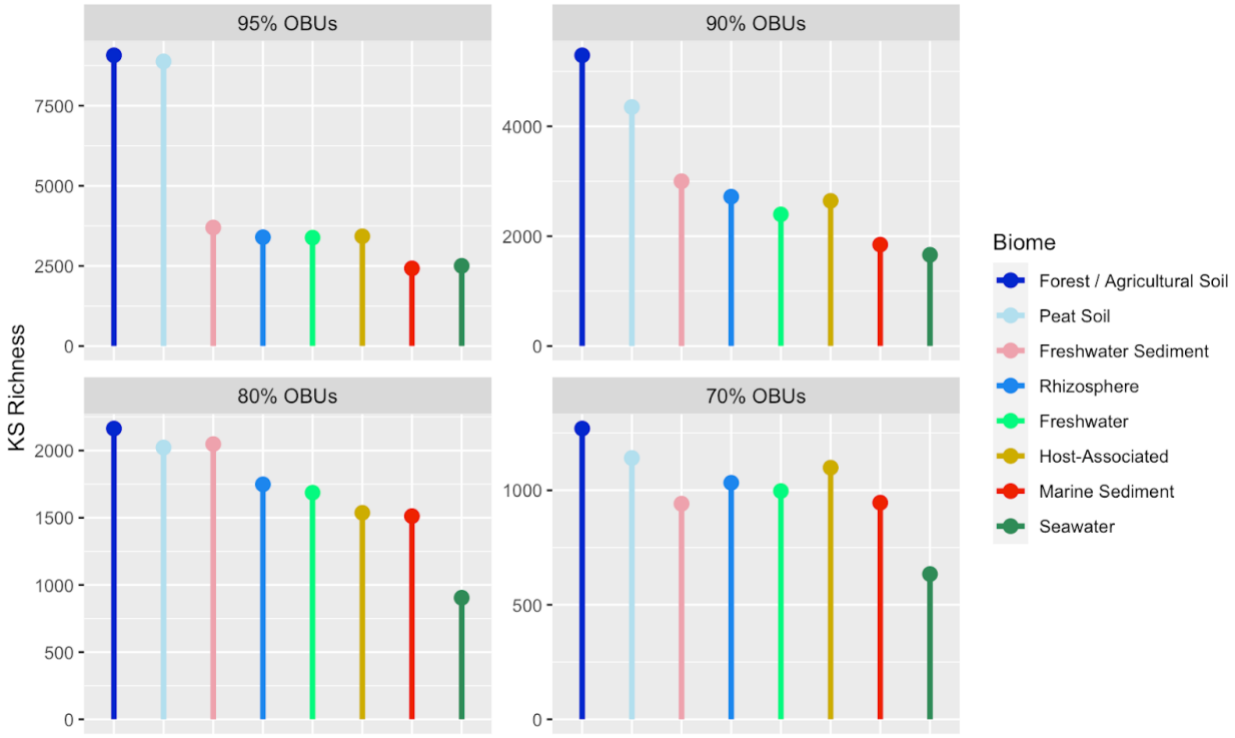

Figure 2.S3. KS richness across biomes. Bar plot showing KS OBU richness, calculated using the Chao1 richness index across eight biomes. For each biome, 580 full-length KS domains were randomly selected and clustered at four different OBU thresholds (panels) ranging from 70% to 95%. The resulting Chao1 richness values (y-axis) represent the average of 10 analyses per biome.

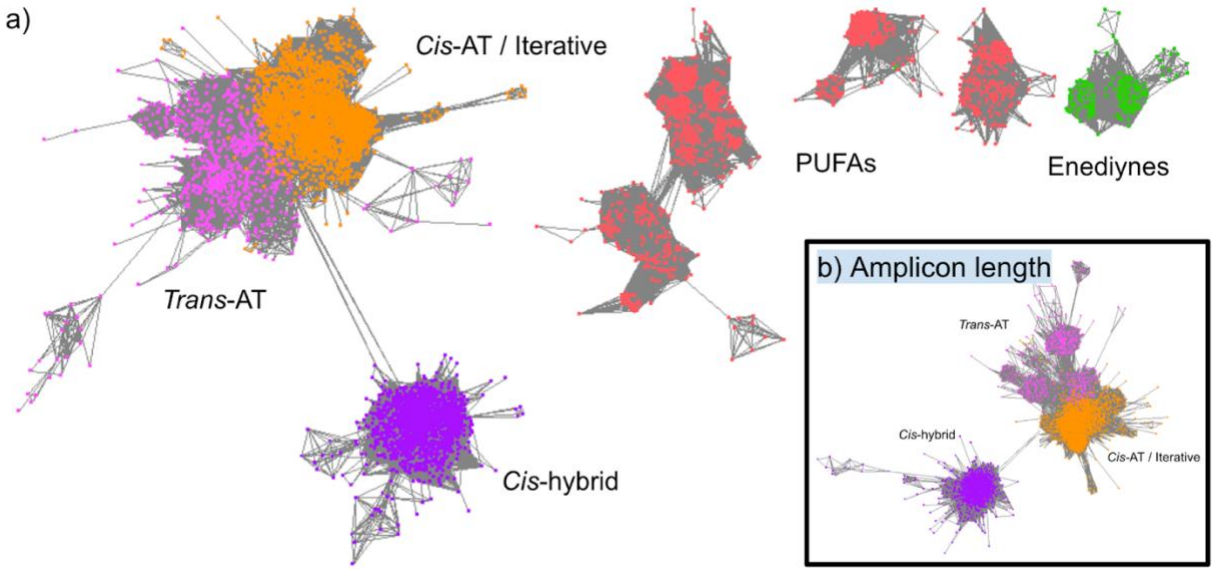

Figure 2.S4. Type I KS domain SSN colored by NaPDoS2 classification. A) An SSN of all full length metagenome-extracted KS domains (main) was constructed using EFI and visualized using Cytoscape. KS domain sequences (amino acids, average length = 420 aa) were colored according to their NaPDoS2 classification, with five clear groups appearing. KS domains from hybrid *cis*-AT (purple), *trans*-AT (pink), PUFAs (red) and enediynes (green) formed four subclass specific groups. Additionally, KS domains from the modular *cis*-AT, *cis*-loading modules, OLS, iterative aromatics and iterative PTMs classes formed a fifth group that was termed the *cis* AT/iterative group (orange). B) All KS domains from the *cis*-AT/iterative, *trans*-AT and hybrid *cis*-AT groups were shortened to amplicon lengths (amino acids, average length = 138 aa) and used to create a SSN. The same separation of these three clades was observed at amplicon lengths (inset,B) as for full-length KS domains (main,A).

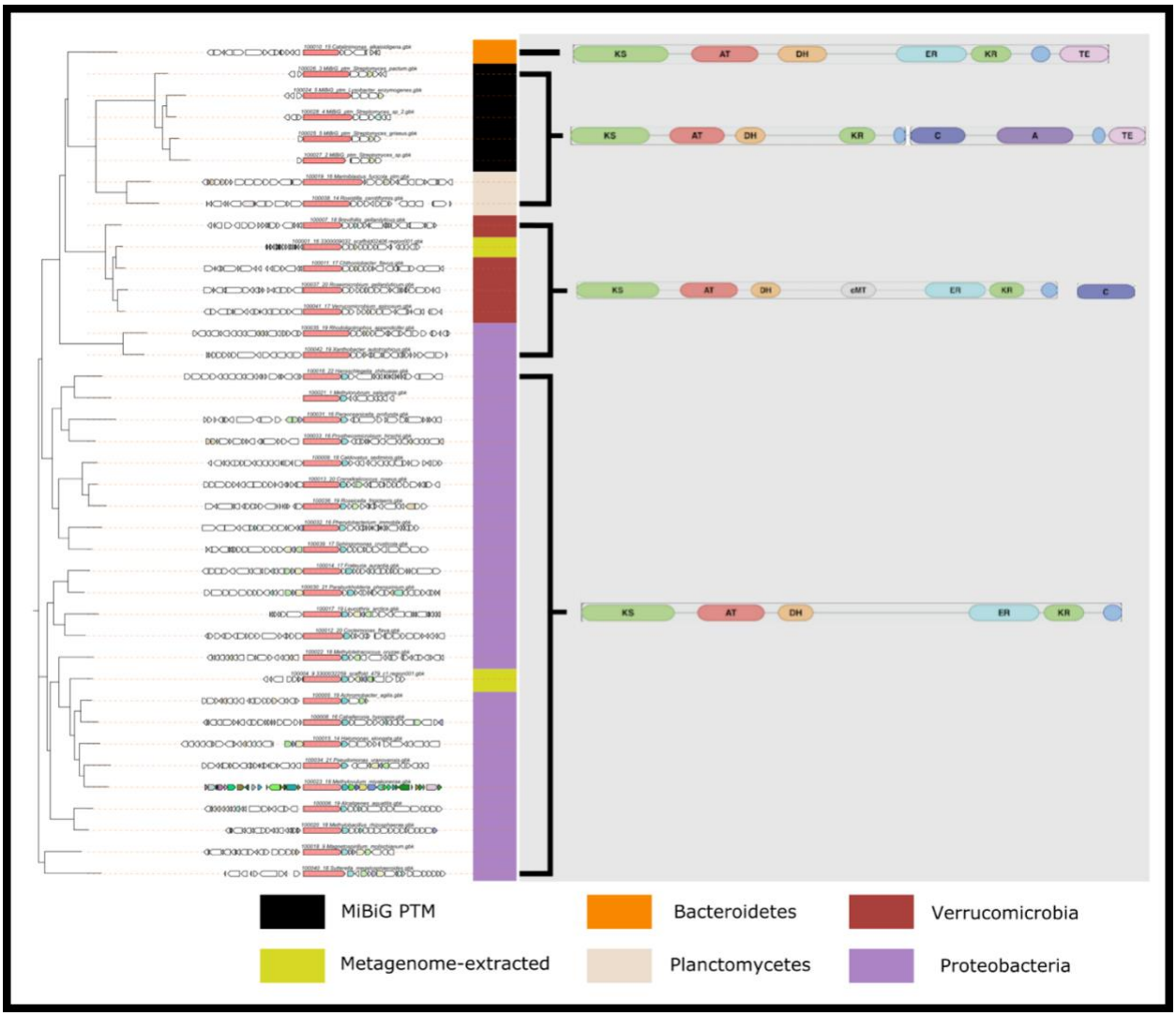

Figure 2.S5. Multilocus phylogeny of Monomodular clade adjacent to MIBiG PTMs. To contextualize a KS clade from the *cis*-AT/iterative phylogeny (Fig. 2a, pink) that was exclusive of KS sequences from MiBIG, a multilocus phylogeny was used. Two full-length BGCs were pulled from metagenomic contigs from this KS clade using antiSMASH (yellow), and BGCs from RefSeq genomes that grouped into KS OBUs (70%) from this clade were also extracted (colored by taxa of RefSeq genome). Included in this phylogeny as reference points are full length BGCs corresponding to the closest MIBiG representatives to this clade (Iterative PTM BGCs in black). To the right of the phylogeny, the architecture of the polyketide synthase gene as visualized in the AntiSMASH outputs are shown.

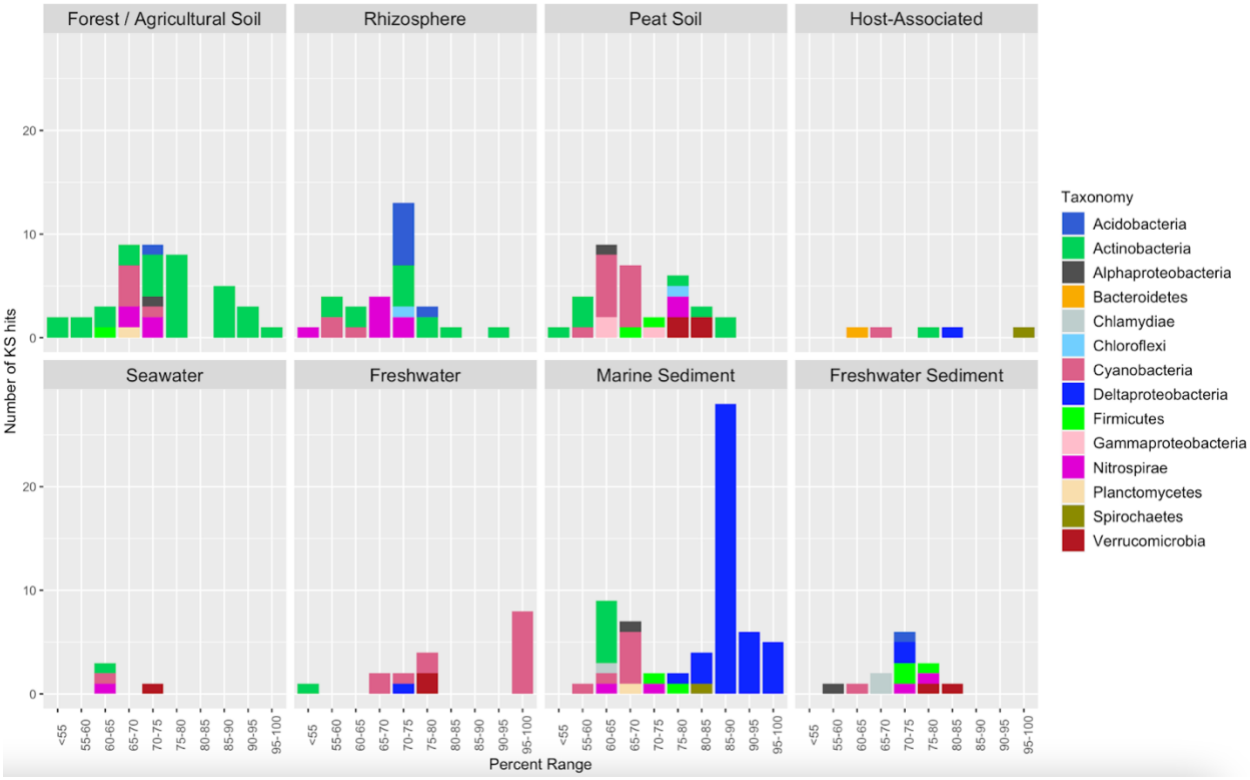

Figure 2.S6. Enediyne KS domain distributions across biomes. Stacked bar charts indicate the phylum-level taxonomic composition and abundance (y-axis) of full-length metagenome extracted enediyne KS domains for eight biomes grouped based on percent identity with the closest NCBI BLASTp database match (x-axis).

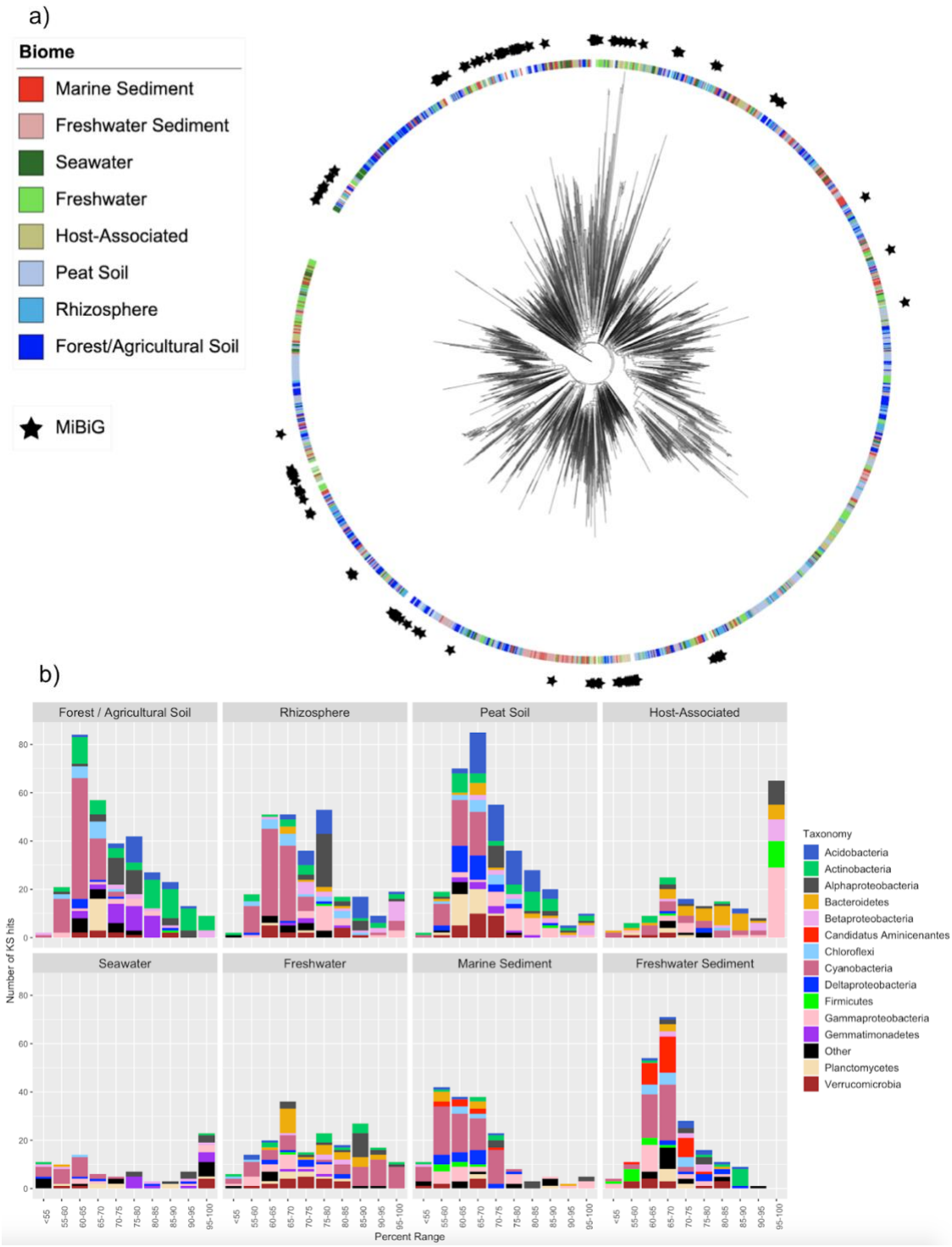

Figure 2.S7. Hybrid *cis*-AT KS domain phylogeny and distributions across biomes. A) A FastME phylogeny was generated from full-length metagenome-extracted hybrid *cis*-AT KS domains (n=1746, colored by biome) with the position of MiBiG-extracted hybrid *cis*-AT KS domains shown as black stars. B) Stacked bar charts indicate the phylum-level taxonomic composition and abundance (y-axis) of full-length metagenome-extracted *cis*-AT KS domains for eight biomes grouped based on percent identity with the closest NCBI BLASTp database match (x-axis).

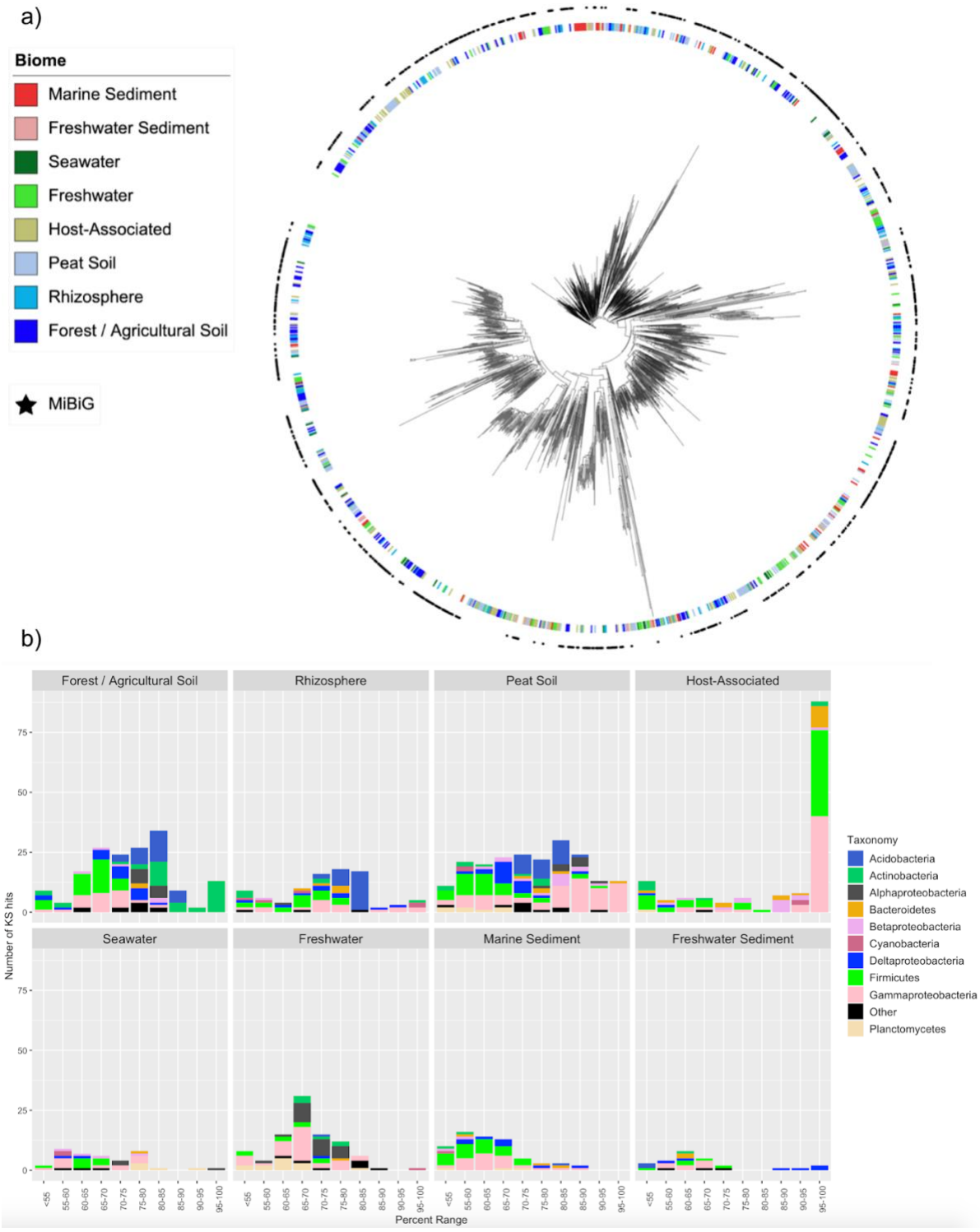

Figure 2.S8. Trans-AT KSs domain phylogeny and distributions across biomes. A) A FastME phylogeny generated from all full-length, metagenome-extracted *trans*-AT KS domains (n=831, colored by biome) with the position of MiBiG-extracted *trans*-AT KS domains shown as black stars). B) Stacked bar charts indicate the phylum-level taxonomic composition and abundance (y-axis) of full-length metagenome-extracted *trans*-AT KS domains for eight biomes grouped based on percent identity with the closest NCBI BLASTp database match (x-axis).

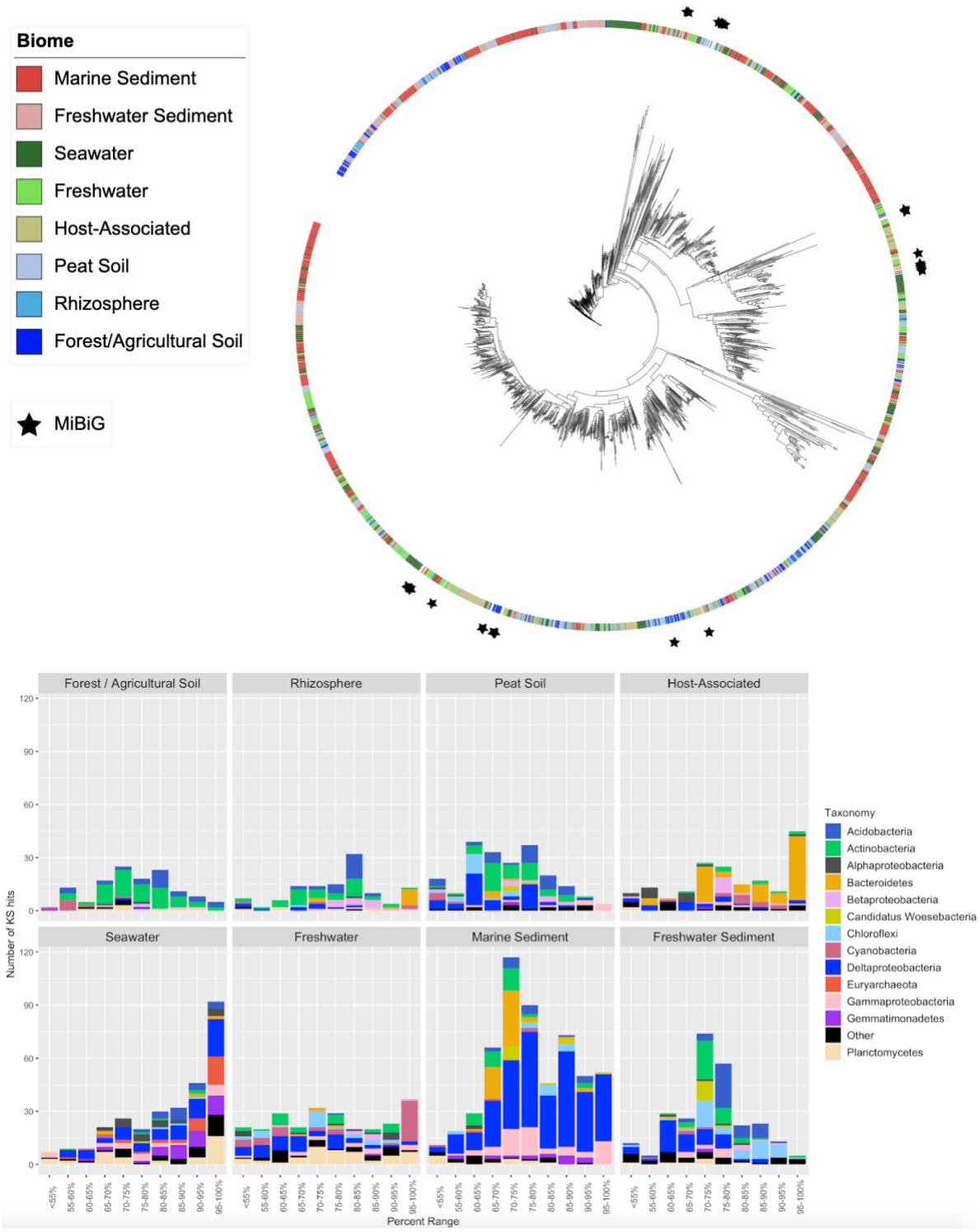

Figure 2.S9. PUFA KS domain phylogeny and distributions across biomes. A) FastME phylogeny generated from full-length metagenome-extracted PUFA KS domains (n=1996, colored by biome) with the position of MIBiG PUFA KS domains shown as black stars. B) Stacked bar charts indicate the phylum-level taxonomic composition and abundance (y-axis) of full-length metagenome-extracted PUFA KS domains for eight biomes grouped based on percent identity with the closest NCBI BLASTp database match (x-axis).

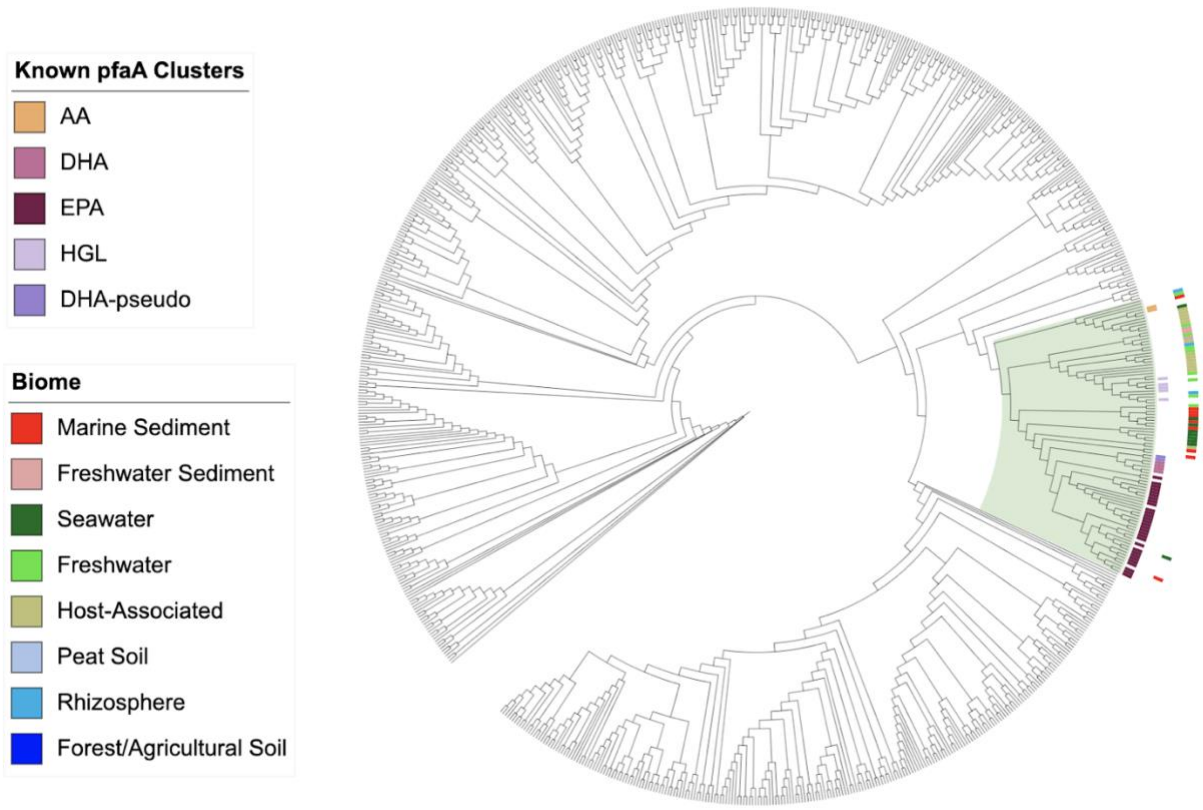

Figure 2.S10. PfaA KS domain phylogeny. FastME phylogeny generated from all full-length metagenome-extracted PUFA KS domains that were further classified as belonging to a pfaA module (n=1170) compared against previously characterized pfaA clusters (22). The clade in green includes all pfaA KS domains from these five previously characterized clusters - AA, DHA, EPA, HGL and pseudo-DHA, with 92% of all pfaA KS domains falling outside this clade.

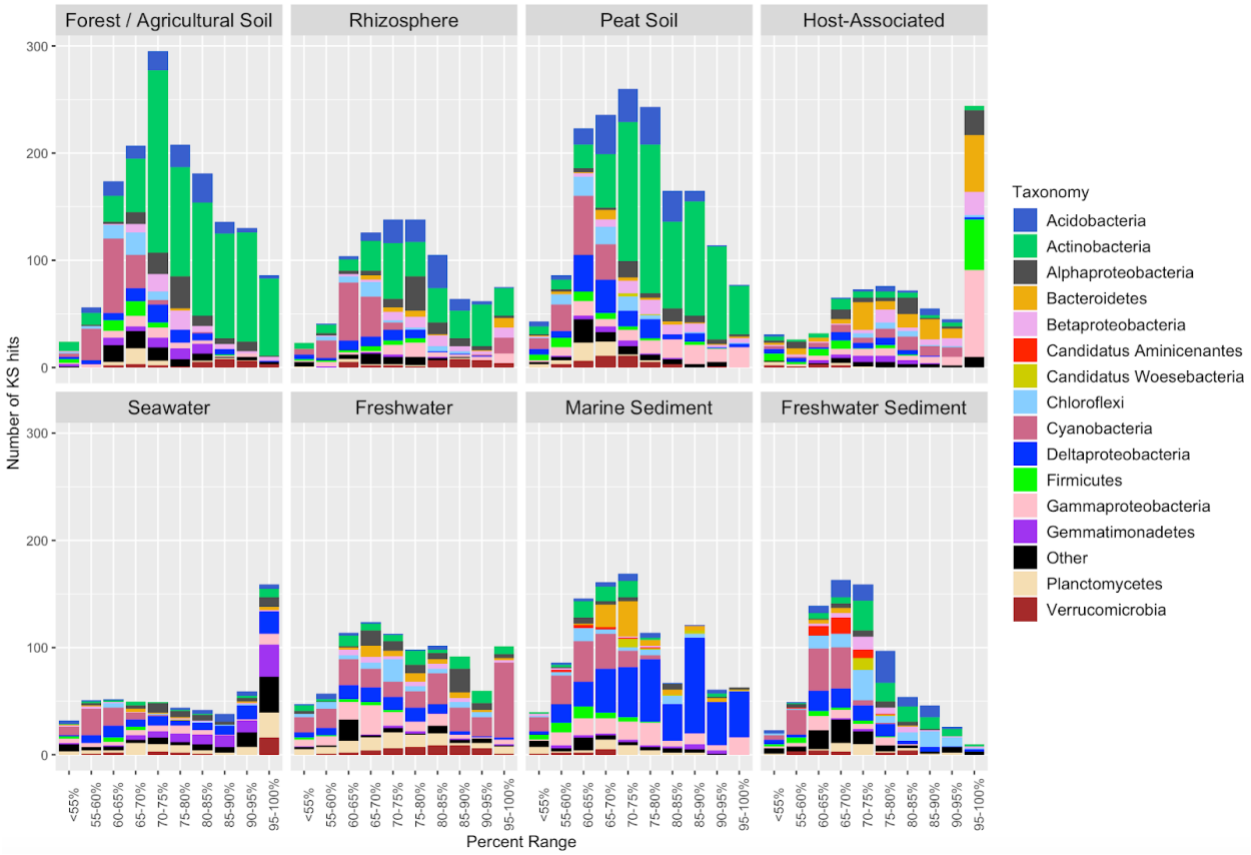

Figure 2.S11. Full-length KS domin distributions across biomes. Stacked bar charts indicate the phylum-level taxonomic composition and abundance (y-axis) of full-length metagenome extracted KS domains for eight biomes grouped based on percent identity with the closest NCBI BLASTp database match (x-axis).

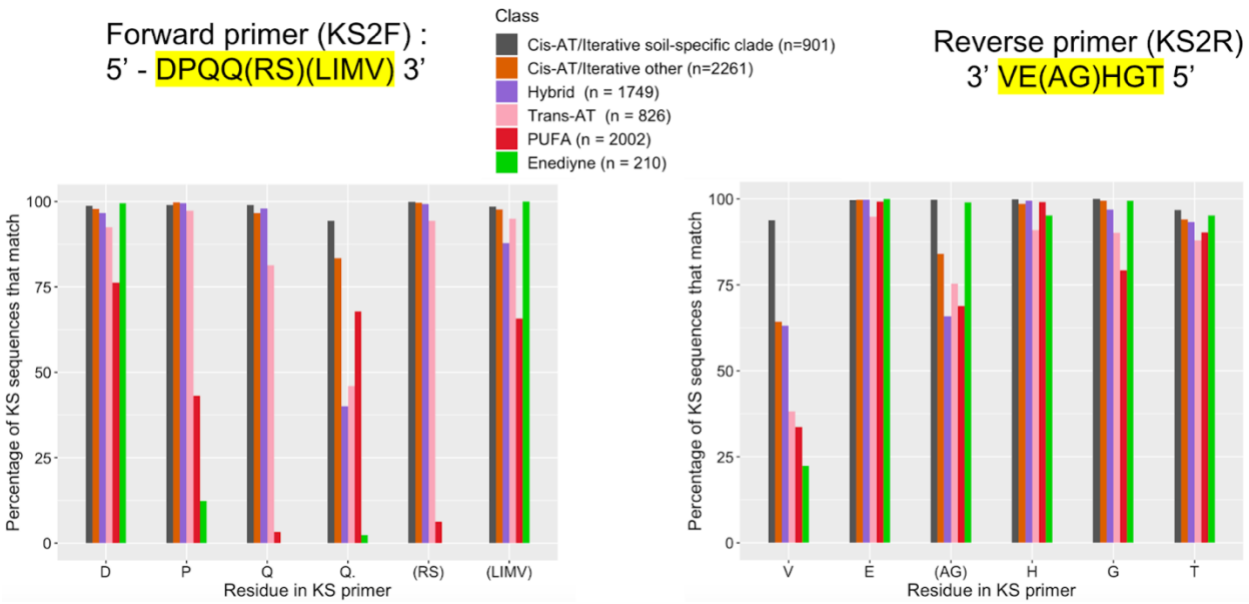

Figure 2.S12. Evaluation of the KS2F/R primer set. Percentage of metagenome extracted KS sequences (y-axis) that match the KS2F/R primers at each amino acid position. KS sequences are grouped by their NaPDoS2 classification. The *cis*-AT/iterative soil-dominant KSs (black) were analyzed separately from all other *cis*-AT/iterative KSs (orange).

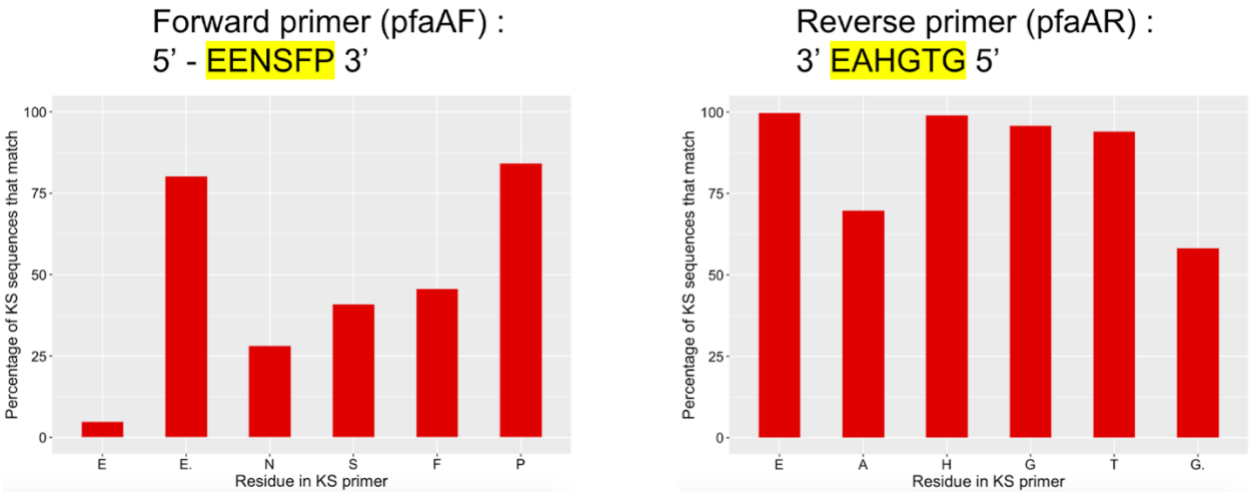

Figure 2.S13. Evaluation of the pfaA primer set. Bar charts were used to visualize the percentages (y-axis) with which the amino acid residues (x-axis) from the PUFA-specific pfaAF/R primer set matched with metagenome-extracted PUFA KS domains (n=1170) identified by NaPDoS2.

2.10 References

Abdel-Razek AS, El-Naggar ME, Allam A, Morsy OM, Othman SI. 2020. Microbial natural products in drug discovery. Processes 8:470.

Ayuso-Sacido A, Genilloud O. 2005. New PCR primers for the screening of NRPS and PKS-I systems in actinomycetes: detection and distribution of these biosynthetic gene sequences in major taxonomic groups. Microb Ecol 49:10–24.

Bech PK, Lysdal KL, Gram L, Bentzon-Tilia M, Strube ML. 2020. Marine sediments hold an untapped potential for novel taxonomic and bioactive bacterial diversity. mSystems 5:e00782-20.

Blin K, Shaw S, Kloosterman AM, Charlop-Powers Z, van Wezel GP, Medema MH, Weber T. 2021. AntiSMASH 6.0: improving cluster detection and comparison capabilities. Nucleic Acids Res 49:W29–W35.

Buchfink B, Reuter K, Drost H-G. 2021. Sensitive protein alignments at tree-of-life scale using DIAMOND. Nat Methods 18:366–368.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. BMC Bioinformatics 10:421.

Carrión VJ, Perez-Jaramillo J, Cordovez V, Tracanna V, de Hollander M, Ruiz-Buck D, Mendes LW, van Ijcken WFJ, Gomez-Exposito R, Elsayed SS, Mohanraju P, Arifah A, van der Oost J, Paulson JN, Mendes R, van Wezel 53. GP, Medema MH, Raaijmakers JM. 2019. Pathogen-induced activation of disease-suppressive functions in the endophytic root microbiome. Science 366:606–612.

Charlop-Powers Z, Owen JG, Reddy BVB, Ternei MA, Brady SF. 2014. Chemical-biogeographic survey of secondary metabolism in soil. Proc Natl Acad Sci U S A 111:3757–3762.

Charlop-Powers Z, Owen JG, Reddy BVB, Ternei MA, Guimarães DO, de 41. Frias UA, Pupo MT, Seepe P, Feng Z, Brady SF. 2015. Global biogeographic sampling of bacterial secondary metabolism. Elife 4:e05048.

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

Crits-Christoph A, Diamond S, Butterfield CN, Thomas BC, Banfield JF. 2018. Novel soil bacteria possess diverse genes for secondary metabolite 48. biosynthesis. Nature 558:440–444.

Delgado-Baquerizo M, Oliverio AM, Brewer TE, Benavent-González A, Eldridge DJ, Bardgett RD, Maestre FT, Singh BK, Fierer N. 2018. A global atlas of the dominant bacteria found in soil. Science 359:320–325.

Della Sala G, Hochmuth T, Teta R, Costantino V, Mangoni A. 2014. Polyketide synthases in the microbiome of the marine sponge plakortis 47. halichondrioides: a metagenomic update. Mar Drugs 12:5425–5440.

Edgar RC. 2010. Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26:2460–2461.

Elfeki M, Mantri S, Clark CM, Green SJ, Ziemert N, Murphy BT. 2021. Evaluating the distribution of bacterial natural product biosynthetic 45. genes across lake huron sediment. ACS Chem Biol 16:2623–2631.

Fieseler L, Hentschel U, Grozdanov L, Schirmer A, Wen G, Platzer M, Hrvatin S, Butzke D, Zimmermann K, Piel J. 2007. Widespread occurrence and genomic context of unusually small polyketide synthase genes in microbial consortia associated with marine sponges. Appl Environ Microbiol 73:2144–2155. https://doi.org/10.1128/AEM.02260-06 Blodgett JAV, Oh D-C, Cao S, Currie CR, Kolter R, Clardy J. 2010. Common biosynthetic origins for polycyclic tetramate macrolactams from phylogenetically diverse bacteria. Proc Natl Acad Sci U S A 107:11692– 11697.

Fischbach MA, Walsh CT. 2006. Assembly-line enzymology for polyketide and nonribosomal peptide antibiotics: logic, machinery, and mecha- nisms. Chem Rev 106:3468–3496.

Freel KC, Nam SJ, Fenical W, Jensen PR. 2011. Evolution of secondary metabolite genes in three closely related marine actinomycete species. Appl Environ Microbiol 77:7261–7270.

Gontang EA, Gaudêncio SP, Fenical W, Jensen PR. 2010. Sequence-based analysis of secondary-metabolite biosynthesis in marine actinobacteria. Appl Environ Microbiol 76:2487–2499.

Hochmuth T, Piel J. 2009. Polyketide synthases of bacterial symbionts in sponges -- evolution-based applications in natural products research. 46. Phytochemistry 70:1841–1849.

Hoshino T, Doi H, Uramoto G-I, Wörmer L, Adhikari RR, Xiao N, Morono Y, D'Hondt S, Hinrichs K-U, Inagaki F. 2020. Global diversity of microbial communities in marine sediment. Proc Natl Acad Sci U S A 117:27587– 27597.

Kautsar SA, Blin K, Shaw S, Navarro-Muñoz JC, Terlouw BR, van der Hooft JJJ, van Santen JA, Tracanna V, Suarez Duran HG, Pascal Andreu V, Selem- Mojica N, Alanjary M, Robinson SL, Lund G, Epstein SC, Sisto AC, Charkoudian LK, Collemare J, Linington RG, Weber T, Medema MH. 2020. MIBiG 2.0: a repository for biosynthetic gene clusters of known function. Nucleic Acids Res 48:D454–D458.

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

Kim BR, Shin J, Guevarra RB, Lee JH, Kim DW, Seol KH, Lee JH, Kim HB, Isaacson RE. 2017. Deciphering diversity indices for a better understand- ing of microbial communities. Journal of Microbiology and Biotechnol- ogy 27:2089–2093.

Klau LJ, Podell S, Creamer KE, Demko AM, Singh HW, Allen EE, Moore BS, Ziemert N, Letzel AC, Jensen PR. 2022. The natural product domain seeker version 2 (NaPDoS2) webtool relates ketosynthase phylogeny to biosynthetic function. J Biol Chem 298:102480.

Lemoine F, Correia D, Lefort V, Doppelt-Azeroual O, Mareuil F, Cohen- Boulakia S, Gascuel O. 2019. NGPhylogeny.fr: new generation phyloge- netic services for non-specialists. Nucleic Acids Res 47:W260–W265.

Letunic I, Bork P. 2021. Interactive tree of life (iTOL) v5: an online tool for phylogenetic tree display and annotation. Nucleic Acids Res 49:W293– W296.

Libis V, Antonovsky N, Zhang M, Shang Z, Montiel D, Maniko J, Ternei 42. MA, Calle PY, Lemetre C, Owen JG, Brady SF. 2019. Uncovering the biosynthetic potential of rare metagenomic DNA using co-occurrence network analysis of targeted sequences. Nat Commun 10:3848.

Medema MH, Blin K, Cimermancic P, de Jager V, Zakrzewski P, Fischbach MA, Weber T, Takano E, Breitling R. 2011. AntiSMASH: rapid identifica- tion, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. Nucleic Acids Res 39:W339–W346.

Miyanaga A, Kudo F, Eguchi T. 2018. Protein-protein interactions in polyketide synthase-nonribosomal peptide synthetase hybrid assembly lines. Nat Prod Rep 35:1185–1209.

Moffitt MC, Neilan BA. 2003. Evolutionary affiliations within the superfamily of ketosynthases reflect complex pathway associations. J Mol Evol 56:446–457.

Navarro-Muñoz JC, Selem-Mojica N, Mullowney MW, Kautsar SA, Tryon JH, Parkinson EI, De Los Santos ELC, Yeong M, Cruz-Morales P, Abubucker S, Roeters A, Lokhorst W, Fernandez-Guerra A, Cappelini LTD, Goering AW, Thomson RJ, Metcalf WW, Kelleher NL, Barona-Gomez F, Medema MH. 2020. A computational framework to explore large-scale biosyn- thetic diversity. Nat Chem Biol 16:60–68.

Nayfach S, Roux S, Seshadri R, Udwary D, Varghese N, Schulz F, Wu D, Paez-Espino D, Chen I-M, Huntemann M, Palaniappan K, Ladau J, Mukherjee S, Reddy TBK, Nielsen T,

Kirton E, Faria JP, Edirisinghe JN, Henry CS, Jungbluth SP, Chivian D, Dehal P, Wood-Charlson EM, Arkin AP, Tringe SG, Visel A, Woyke T, Mouncey NJ, Ivanova NN, Kyrpides NC, Eloe- Fadrosh EA, IMG/M Data Consortium. 2021. A genomic catalog of earth's microbiomes. Nat Biotechnol 39:499–509.

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

Nivina A, Yuet KP, Hsu J, Khosla C. 2019. Evolution and diversity of assembly-line polyketide synthases. Chem Rev 119:12524–12547.

O'Brien RV, Davis RW, Khosla C, Hillenmeyer ME. 2014. Computational identification and analysis of orphan assembly-line polyketide synthases. J Antibiot (Tokyo) 67:89–97.

O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D, Astashyn A, Badretdin A, Bao Y, Blinkova O, Brover V, Chetvernin V, Choi J, Cox E, Ermolaeva O, Farrell CM, Goldfarb T, Gupta T, Haft D, Hatcher E, Hlavina W, Joardar VS, Kodali VK, Li W, Maglott D, Masterson P, McGarvey KM, Murphy MR, O'Neill K, Pujar S, Rangwala SH, Rausch D, Riddick LD, Schoch C, Shkeda A, Storz SS, Sun H, Thibaud-Nissen F, Tolstoy I, Tully RE, Vatsan AR, Wallin C, Webb D, Wu W, Landrum MJ, Kimchi A, Tatusova T, DiCuccio M, Kitts P, Murphy TD, Pruitt KD. 2016. Reference sequence (RefSeq) database at NCBI: Current status, Taxonomic expansion, and functional Annotation. Nucleic Acids Res 44:D733–D745.

Oksanen J, Blanchet FG, Friendly M, Kindt R, Legendre P, McGlinn D, Minchin PR, O'Hara RB, Simpson GL, Solymos P, Henry M, Stevens H, Szoecs E, Wagner H. 2020. Vegan: community ecology package, p 5–7. In R package version 2.

Palaniappan K, Chen I-M, Chu K, Ratner A, Seshadri R, Kyrpides NC, Ivanova NN, Mouncey NJ. 2020. IMG-ABC v.5.0: an update to the IMG/ Atlas of biosynthetic gene clusters knowledgebase. Nucleic Acids Res 48:D422–D430.

Paoli L, Ruscheweyh H-J, Forneris CC, Hubrich F, Kautsar S, Bhushan A, Lotti A, Clayssen Q, Salazar G, Milanese A, Carlström CI, Papadopoulou C, 49. Gehrig D, Karasikov M, Mustafa H, Larralde M, Carroll LM, Sánchez P, Zayed AA, Cronin DR, Acinas SG, Bork P, Bowler C, Delmont TO, Gasol JM, Gossert AD, Kahles A, Sullivan MB, Wincker P, Zeller G, Robinson SL, Piel J, Sunagawa S. 2022. Biosynthetic potential of the global ocean micro- 50. biome. Nature 607:111–118.

Parks DH, Rinke C, Chuvochina M, Chaumeil P-A, Woodcroft BJ, Evans PN, Hugenholtz P, Tyson GW. 2017. Recovery of nearly 8,000 metagenome- 51. assembled genomes substantially expands the tree of life. Nat Microbiol 2:1533–1542.

Piel J. 2010. Biosynthesis of polyketides by *trans*-at polyketide synthases. Nat Prod Rep 27:996–1047.

Pye CR, Bertin MJ, Lokey RS, Gerwick WH, Linington RG. 2017. Retrospec- tive analysis of natural products provides insights for future discovery trends. Proc Natl Acad Sci U S A 114:5601–5606.

Rascher A, Hu Z, Viswanathan N, Schirmer A, Reid R, Nierman WC, Lewis M, Hutchinson CR. 2003. Cloning and characterization of a gene cluster for geldanamycin production in Streptomyces hygroscopicus NRRL 3602. FEMS Microbiol Lett 218:223–230.

Rego A, Sousa AGG, Santos JP, Pascoal F, Canário J, Leão PN, Magalhães C. 2020. Diversity of bacterial biosynthetic genes in maritime Antarctica. Microorganisms 8:279.

Rudolf JD, Yan X, Shen B. 2016. Genome neighborhood network reveals insights into enediyne biosynthesis and facilitates prediction and prioritization for discovery. J Ind Microbiol Biotechnol 43:261–276.

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13:2498–2504.

Shen B, Yan X, Huang T, Ge H, Yang D, Teng Q, Rudolf JD, Lohman JR. 2015. Enediynes: exploration of microbial genomics to discover new anticancer drug leads. Bioorg Med Chem Lett 25:9–15.

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

Shulse CN, Allen EE. 2011. Diversity and distribution of microbial long- chain fatty acid biosynthetic genes in the marine environment. Environ Microbiol 13:684–695.

Skinnider MA, Merwin NJ, Johnston CW, Magarvey NA. 2017. PRISM 3: expanded prediction of natural product chemical structures from microbial genomes. Nucleic Acids Res 45:W49–W54.

Storey MA, Andreassend SK, Bracegirdle J, Brown A, Keyzers RA, Ackerley DF, Northcote PT, Owen JG. 2020. Metagenomic exploration of the marine sponge Mycale hentscheli uncovers multiple polyketide- 54. producing bacterial symbionts. mBio 11:e02997-19.

Sugimoto Y, Camacho FR, Wang S, Chankhamjon P, Odabas A, Biswas A, Jeffrey PD, Donia MS. 2019. A metagenomic strategy for harnessing the chemical repertoire of the human microbiome. Science 366:eaax9176. 52.

Wang B, Guo F, Huang C, Zhao H. 2020. Unraveling the iterative type I polyketide synthases hidden in Streptomyces. Proc Natl Acad Sci U S A 117:8449–8454.

Wawrik B, Kerkhof L, Zylstra GJ, Kukor JJ. 2005. Identification of unique type II polyketide synthase genes in soil. Appl Environ Microbiol 71:2232–2238.

Weissman KJ. 2004. Polyketide biosynthesis: understanding and exploiting modularity. Philos Trans A Math Phys Eng Sci 362:2671–2690.

Zallot R, Oberg N, Gerlt JA. 2019. The EFI web resource for genomic enzymology tools: leveraging protein, genome, and metagenome databases to discover novel enzymes and metabolic pathways. Biochemistry 58:4169–4182.

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

Ziemert N, Lechner A, Wietz M, Millán-Aguiñaga N, Chavarria KL, Jensen PR. 2014. Diversity and evolution of secondary metabolism in the marine actinomycete genus *Salinispora*. Proc Natl Acad Sci U S A 111:E1130– E1139.

Ziemert N, Podell S, Penn K, Badger JH, Allen E, Jensen PR. 2012. The natural product domain seeker NaPDos: a Phylogeny based Bioinfor- matic tool to classify secondary metabolite gene diversity. PLoS One 7:e34064.

CHAPTER 3. Genomes across the tree of life highlight novel polyketide potential

3.1 Abstract

With the rise of genomic sequencing and accessibility of sequence data in online repositories, there lies an opportunity to characterize biosynthetic potential across lineages of life. Polyketides are important compounds from a pharmaceutical perspective, with select polyketides serving as important antibiotic, anti-fungal, and anti-cancer compounds. Polyketides are assembled by the products of polyketide synthase (PKS) gene clusters (BGCs). PKSs contain ketosynthase (KS) domains, which perform condensation reactions critical to compound assembly. Phylogenetic analysis of KS domains gives insight into PKS structure and function and, in some cases, structural insights into the compounds produced. To comprehensively map polyketide diversity across living organisms, NaPDoS2 was used to extract 53,000 KS domains across ~620,000 genomes that spanned the three domains of life—Bacteria, Archaea, and Eukaryota. From this, I was able to show that while well-examined phyla such as Actinobacteria, Myxococcota, Cyanobacteria, and Ascomycota contain the largest amounts of KS domains, a staggering amount of KS diversity remains to be realized in understudied taxa.

3.2 Introduction

Specialized metabolites are produced across the tree of life and play crucial ecological roles for the producer such as defense, development, communication, and nutrient acquisition (Davies, 2013). From a human perspective, natural products have been the basis of traditional indigenous medicines for thousands of years, and the inspiration for many pharmaceuticals, drugs, and therapeutics in the last century (Dias et al., 2012). Of the many classes of natural products, polyketides are one of the largest and are responsible for a staggering diversity of compounds - many of which have become important therapeutics (Nivina et al., 2019). While the

analysis of chemical extracts has been a productive approach to finding novel polyketide compounds, this methodology is now plagued by high rediscovery rates, leading to the rise of alternative natural product discovery strategies such as genome mining (Pye et al., 2017).

The genes encoding for polyketides are usually clustered within polyketide synthase (PKS) biosynthetic gene clusters (BGCs), which are evolutionarily related to fatty acid synthases (FASs) (Herbst et al., 2018). Both polyketides and fatty acids form through repeated decarboxylative Claisen condensation reactions, catalyzed by ketosynthase (KS) domains, until a defined chain length is reached and the substrate is released from the enzyme complex (Hertweck et al., 2009). However, PKSs notably differ from FASs in that the reduction/dehydration steps that can occur after each chain elongation (catalyzed by ketoreductase, dehydratase, and enoyl reductase enzymes) can be partly or fully omitted, leading to greater structural diversity (Nivina et al., 2019). And yet, the lines between PKSs and FASs are still being distinguished, as evidenced by the recent discovery of fatty acid synthase- (FAS)-like polyketide synthase (PKS) proteins in sacoglossans (Torres et al. 2019).

Initiatives to sequence animal genomes have expanded since the first genome (*Caenorhabditis elegans*) was sequenced in 1998. Yet, bacterial genomes still account for >75% of the genomes within the National Center for Biotechnology Information (NCBI) database (O'Leary et al., 2016), largely due to the comparative ease of sequencing microbial genomes. Nonetheless, the wealth of genomes sequenced in the last two decades has enabled the rise of genome mining as a technique to identify novel natural BGCs. As sequencing technologies have improved, so too have genome mining tools such as antiSMASH 7.0, which identify biosynthetic gene clusters (BGCs) based on the co-localization of core biosynthetic genes and their tailoring enzymes (Blin et al., 2023). Equally instrumental to the field of genome mining is the MIBiG

database, which is an online repository of all BGCs that have been experimentally linked to an associated natural product (Terlouw et al., 2023).

However, there remain difficulties in applying genome mining algorithms to non-microbial genomes, as these tools were largely trained on bacterial and fungal datasets (Blin et al., 2023). For example, certain biosynthetic genes may not co-localize in the same region of the eukaryotic chromosome, assemblies can be too fragmented and miss the core genes needed for calling matches, and eukaryotic splice-variant gene calling may cause difficulties (Kwon et al., 2023; Meleshko et al., 2019). As described in Chapter 2, the NaPDoS2 webtool also fills an important niche in genome mining in that it extracts KS domains from PKS BGCs and phylogenetically classifies sequences into one of 42 classes/subclasses (Klau, Podell, and Creamer et al., 2022). In bypassing the need for BGC assembly, this sequence tag approach leverages the powerful conservation of KS functional diversity to predict PKS biosynthetic outputs with scale, speed, and specificity.

In addition to the exploration of polyketide diversity across microbial metagenomes described in chapter 2 (Singh et al., 2023), recent studies have utilized genome mining tools to uncover biosynthetic potential within bacterial genomes, highlighting taxa-specific diversity and lineages that could be targeted for novelty (Cimermancic et al., 2014; Wei et al., 2021; Chen et al., 2022). In particular, one recent study found that only 3% of the BGC families recovered across more than 200,000 genomes were linked to experimentally characterized compounds, hinting at significant, yet to be discovered biosynthetic novelty (Gavriilidou et al., 2022). Similarly, work utilizing over 1,000 fungal genomes revealed taxa-specific BGC richness (Robey et al., 2021), and a combination of genome mining with “Hex” synthetic biology illustrated a systematic way to isolate novel polyketide compounds from challenging fungal

scaffolds (Harvey et al., 2018). While often overlooked in favor of their microbial counterparts, genome mining has shown that animals have significant untapped potential to produce polyketides, with a notable example being the recent discovery of animal FAS-like PKS (AFPK) proteins in Sacoglossans (Torres et al. 2019). Follow-up studies using a custom hidden Markov model (HMM) highlighted more than 6000 distinct AFPK sequences across mollusks and arthropods (Lin et al., 2024).

In this work, the NaPDoS2 webtool was used to detect and classify polyketide biosynthetic potential across 617,968 representative bacterial, fungal, animal, algae, plant, protist, plasmid, archaeal, and CPR (candidate phyla radiation) genomes. The study represents a collaboration between myself and former PhD student Kaitlin Creamer. In doing so, I uncovered taxa-specific KS diversity and novel KS phylogenetic clades across the tree of life.

3.3 Methods

Genomic dataset selection

To analyze KS diversity across the tree of life, carefully curated genomic datasets that maximized phylogenetic diversity were selected. For kingdoms that lacked carefully constructed datasets, genomes were selected to minimize duplication when possible, which was accomplished by using representative genomes. For bacteria and archaea, all 10,575 genomes within the "Web of Life Reference Phylogeny for Microbes" were selected, which is a comprehensive dataset of genomes from both cultured isolates and metagenome-assembled genomes, or MAGs (Zhu *et al.*, 2019). For four understudied bacterial phyla (Acidobacteria, Chloroflexi, Planctomycetes, and Verrucomicrobia), another 2,852 high-quality genomes from

the JGI IMG/MER repository were used. Plasmids within the COMPASS database (12,084 plasmids across 1,571 species) were also analyzed (Douarre *et al.*, 2020).

Viral genomes were collected from four different databases: the Reference Viral Database (RVDB protein database, downloaded February 2021) (Bigot *et al.*, 2020), the Virus Pathogen Resource (ViPR) database (downloaded June 2021) (Pickett *et al.*, 2012), the PATRIC virus and phage database (version 3.6.9) (Davis *et al.*, 2020), and the CheckV (version v1.0) curated database of viral genomes (Nayfach *et al.*, 2021).

Since introns have been observed in eukaryotic PKS genes, predicted amino acid coding sequences were used to remove the possibility that NaPDoS2 would recover false positive KS domains. For fungi, all 1,644 fungal genomes used in a recent, well-constructed phylogenetic analysis of fungal diversity were downloaded (Li *et al.*, 2021). However, through manual inspection of these genomes, many were revealed to be nucleic acid genome sequences, so the "ncbi-genome-download" script was used to download the amino acid genome sequences. For protist and algal genomes, all 149 genomes from the JGI PhycoCosm repository (Grigoriev *et al.*, 2021) were selected, along with the 411 plant genomes within the JGI Phytozome database (Goodstein *et al.*, 2012).

For animal genomes, I was unable to find a previous study or well-curated dataset that spanned the tree of life. As such, the NCBI Taxonomy browser and the NCBI Beta Genomes Datasets tool (O'Leary et al., 2016) were used to manually inspect the phylogenetic tree within every family in the kingdom Metazoa (~22,632 available amino acid protein genome assemblies). Annotated genomes were selected (containing amino acid protein CDS predictions), with at least one genome per genus downloaded using the "ncbi-genome-download" script, resulting in 1,125 Metazoan protein genomes.

Finally, KS domains within the NaPDoS2 (Klau *et al.*, 2022) database (1,877 KS domains), and the MIBiG 3.0 database (Terlouw *et al.,* 2023) of experimentally characterized BGCs (5,134 KS domains) were also extracted. These served as comparison points to analyze how genome-extracted KS domains compared to those encoding for known polyketides.

Extraction and classification of KS domains

All genomes (Table 3.1) were analyzed with NaPDoS2 (Klau *et al.*, 2022) at a minimum alignment length of 200 amino acids and an E-value of 1e-8. For each genome, hit results and trimmed KS domain sequences were saved. Furthermore, custom scripts were used to separate all KS class and subclass hits into individual FASTA sequence files. Similarly, all KS domains were grouped separately by taxonomy using custom scripts for downstream analyses. RawGraphs (Mauri *et al.,* 2017) and RStudio (RStudio team, 2021) were used to visualize the distribution of KS hits across taxa with alluvial diagrams and bar charts, respectively.

Phylogenetic distribution and diversity

KS domains were clustered into 80% operational biosynthetic units (OBUs) using UCLUST (Edgar, 2010) and phylogenies constructed using the FastTree (Price *et al.*, 2010) workflow implemented on NGPhylogeny (Lemoine *et al.,* 2019) and visualized using iTol (Letunic and Bork, 2019). Sequence similarity networks (SSNs) were calculated with the EFI-EST Enzyme Similarity Tool (Zallot *et al.,* 2021) and visualized using Cytoscape (version 3.7.2) (Carlin *et al.,* 2017).

3.4 Results

KS domains across the tree of life

To classify polyketide and fatty acid biosynthetic potential across the tree of life, 617,968 genomes across Bacteria, Archaea, Plasmids, Viruses, Fungi, Algae, Protists, Plants, and Animals were selected and analyzed for their KS diversity using NaPDoS2 (Table 3.1). When possible, geomes connected to well-curated datasets and rigorous phylogenetic analyses were selected. While the majority of genomes analyzed were viral (>590,000), KS domains were relatively rare in viruses (<0.01 per genome), and low rates were also seen in archaea and CPR bacteria (<0.01 per genome) and plasmids (0.06 per plasmid) (Fig. 3.1). In contrast, KS domains were most abundant in protists/algae (13.82 per genome), fungi (10.73 per genome), plants (6.53 per genome), animals (5.78 per genome) and bacteria (3.84 per genome) (Fig 3.2-3.3).

Table 3.1. Number of genomes in each taxa dataset.

Dataset	**Number of Genomes**	**Reference**
Bacteria, Archaea	13,427	(Zhu *et al.*, 2019), JGI IMG/MER
Plasmids	12,084	(Douarre *et al.*, 2020)
Viruses	591,387	(Pickett *et al.*, 2012; Bigot *et al.*, 2020; Davis *et al.*, 2020; Nayfach *et al.*, 2021)
Fungi	1,149	(Li *et al.*, 2021)
PhycoCosm (Green Algae, SAR)	140	(Grigoriev *et al.*, 2021)
Phytozome (plants)	105	(Goodstein *et al.*, 2012)
Animals	1,125	NCBI Datasets - Genomes
Total:	***617,968***	
MIBiG 3.0 (*reference*) - KSs	5,134 KSs	(Terlouw *et al.*, 2023)
NaPDoS2 (*reference*) – KSs	1,877 KSs	(Klau *et al.*, 2022)

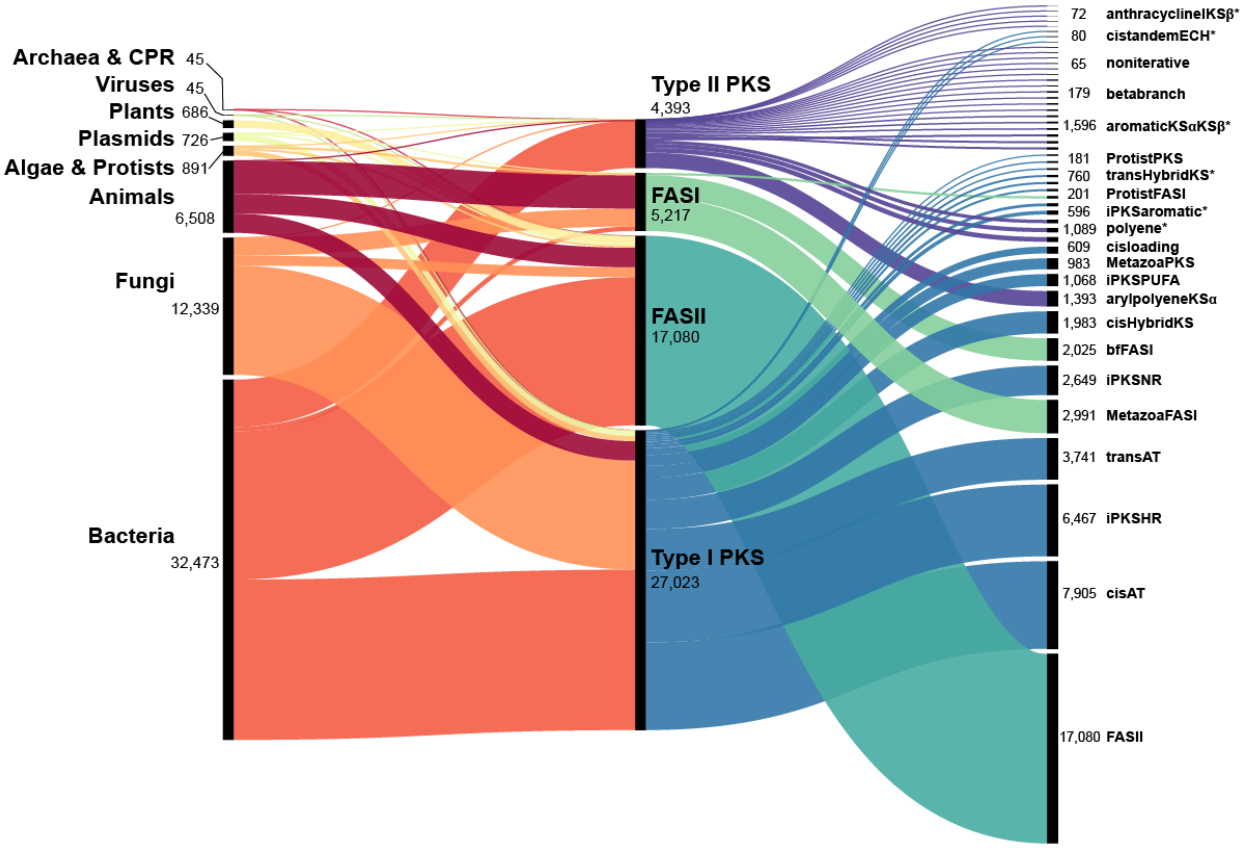

Figure 3.1. KS domains identified across Kingdoms and their corresponding NaPDoS2 classifications.

Of the 10,575 bacterial genomes within the WoL database, 72.3% came from cultured organisms, with the remaining from environmental samples. The majority of genomes from this dataset were from the Pseudomonadota (n=2,925), Bacillota (n=2,016), candidate phyla radiation (n=1454), Actinobacteria (n=1,103) and Bacteroidetes (n=833) phyla. Across taxonomic lineages, the number of KSs generally increased with genome size, although there was significant variability and notable outliers (Fig. 3.S1). For example, while 3.84 KSs were seen per genome on average across the entire WoL reference dataset, five genomes within the phylum Actinobacteria carried more than 100 KSs (four *Streptomyces* and one *Actinoplanes* strains) (Fig. 3.S1). Similar outliers were also seen in other phyla, with more than 50 KS domains seen in three Bacillota genomes (two *Bacillus* and one *Clostridium* strain), four Myxococcota genomes (one

Myxococcus, one Chondromyces, one Sorangium and one Enhygromyxa strain), and two Pseudomonadota genomes (one Gynuella and one Candidatus Magnetomorum strain) (Fig. 3.S1).

Four taxonomic lineages that have been found in culture-independent work to contain polyketide potential contained a low number of genomes within the WoL database: Acidobacteria (n=55), Chloroflexi (n=151), Planctomycetes (n=63) and Verrucomicrobia (n=55). Collectively, only three polyketide BGCs from the MIBiG database are from these four phyla - lasonolide A (Verrucomicrobia), palmerolide (Verrucomicrobia) and aurantoside A (Chloroflexi). To investigate the polyketide potential of these four understudied phyla in more detail, I extracted and analyzed high-quality genomes (both cultured isolates and metagenome-assembled genomes or MAGs) from the JGI IMG/MER database using NaPDoS2, which brought the total to 886 Acidobacteria genomes, 407 Chloroflexi genomes, 810 Planctomycetes genomes, and 1073 Verrucomicrobia genomes.

Distribution of type I KS domains across taxonomic lineages

Across bacterial genomes, the majority of type I KS domains were classified as modular *cis*-AT (46.5%), *trans*-AT (22.1%), and hybrid *cis*-AT (13.1%). The phyla with the greatest average number of type I KS domains per genome was Myxococcota (23.2), with Actinobacteria (7.0) and Cyanobacteria (3.9) being the next highest (Fig. 3.2). The next seven highest average abundances were observed in the Planctomycetes (1.58), Chloroflexi (1.02), Bacillota (1.01), Pseudmonadota (0.87), Acidobacteria (0.69), Bacteroidetes (0.54), and Verrucomicrobia (0.28) (Fig. 3.2).

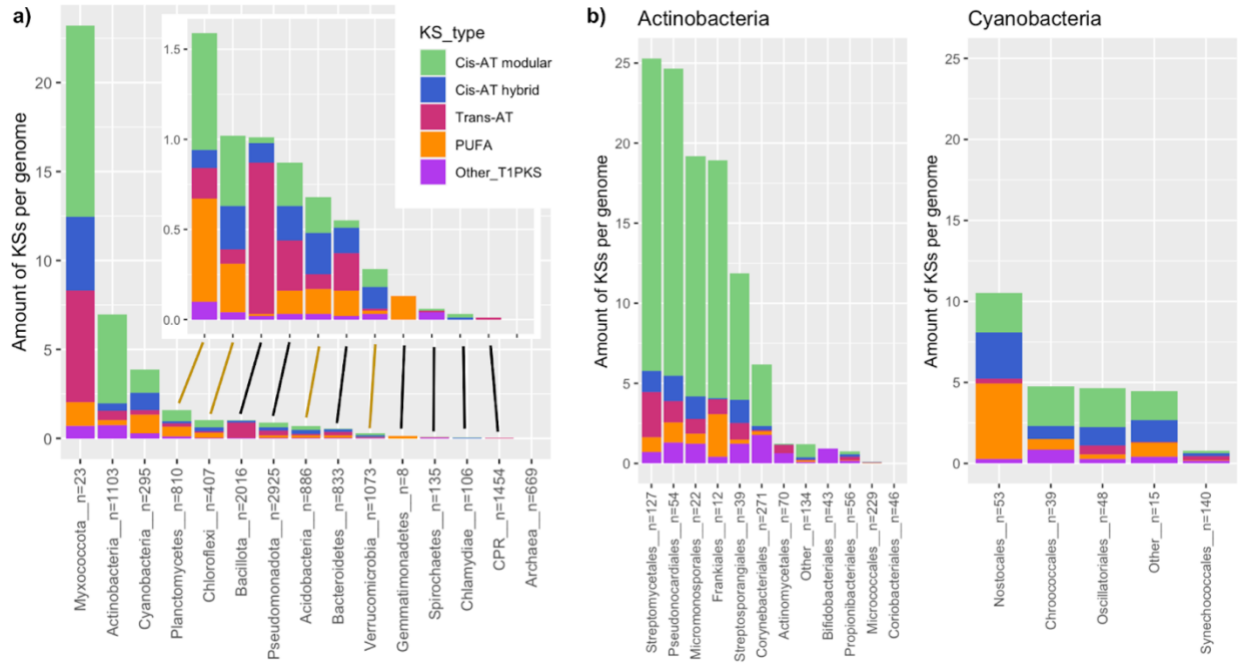

Figure 3.2. Type I KS composition in bacteria. (a) Average number of type I KS domains per genome (y-axis) across phyla (x-axis), with colors denoting the NaPDoS2 KS classification. n = the number of genomes analyzed per phylum. Inset: phyla with less than 2 KS domains per genome with brown lines indicating the four understudied phyla that were supplemented with additional genomes from the JGI IMG/MER database. (b) Average number and type of type I KS domains per genome (y-axis) in the phyla Actinobacteria and Cyanobacteria.

While the average number of KSs per genome in the phylum Myxococcota (23.22) is larger than other phyla, the top Actinobacterial orders, Streptomycetales (25.29), Pseudonocardiales (24.65), Micromonosporales (19.18) and Frankiales (18.92) carried comparable numbers of type I KS domains (Fig. 3.2). Their classification varied however, as 78% of the type I KS domains from these four Actinobacteria orders were modular *cis*-AT, whereas in the Myxococcota genomes, type I KS domains were split more evenly between modular *cis*-AT (46.4%), *trans*-AT (27.0%), and hybrid *cis*-AT (17.8%) (Fig. 3.2). Notably, the least diverse taxon was the Bacillota, with 83.2% of the type I KS domains extracted from these genomes belonged to the *trans*-AT subclass (Fig. 3.S2). When looking at the relative abundance of rare type I KS subclasses, I found enediynes and olefin synthase KSs to be most common in

Cyanobacteria (0.10 and 0.17 per genome, respectively), iterative aromatic KSs to be most common in Actinobacteria (0.27 per genome) and iterative PTMs to be most common in Myxococcota (0.26 per genome) (Fig. 3.S3).

In contrast, type I KS domains in fungal genomes were mostly classified as iterative highly reducing (56.1%), iterative non-reducing (23.2%), and bacteria/fungi type I FAS (13.9%) (Fig. 3.3). Type I KS domains were most abundant in Ascomycota (12.7 per genome) but also observed in the Chytridiomycota (5.3 per genome), Zoopagomycota (3.7 per genome), Basidiomycota (3.7 per genome), and Mucoromycota (1.5 per genome) (Fig. 3.3). Again however, there was variation in the subclassification of type I KS domains across different fungal phyla. Ascomycota genomes mostly contained iterative highly reducing (7.82 per genome) and iterative non-reducing (2.96 per genome) KS domains (Fig. 3.3). This was in sharp contrast to other phyla where iterative highly reducing KS domains were extremely rare (<0.15 per genome in all other fungal phyla). Iterative non-reducing KS domains were only seen in Basidiomycota (0.96 per genome) (Fig. 3.3). In contrast, modular *cis*-AT KS domains were rare in Ascomycota (0.17 per genome), whereas the Chytridiomycota and Basidiomycota phyla were much richer in modular *cis*-AT KS domains (3.38 and 1.18 per genome respectively) (Fig. 3.3). Finally, most type I KS domains in Zoopagomycota (98.4%) and Mucoromycota (68.5%) were from type I FAS bacterial/fungal pathways (Fig. 3.3).

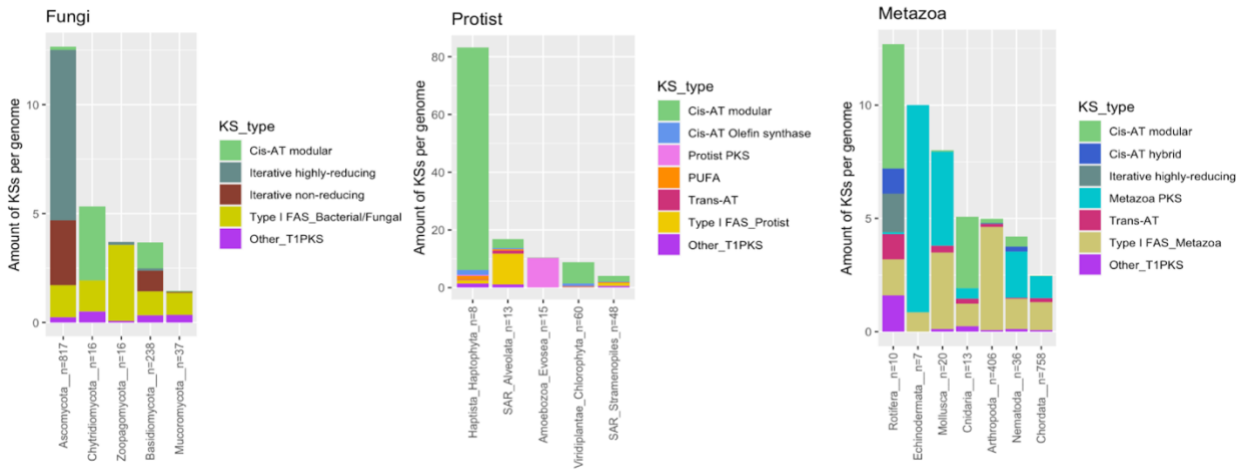

Figure 3.3. Type I KS composition in Fungi, Protists, and Metazoa. Average number of type I KS domains per genome (y-axis) across phyla (x-axis), with colors denoting the NaPDoS2 KS classification. n = the number of genomes analyzed per phylum.

Within the protist and animal genomes, many of the phyla had low numbers of genomes available. Nonetheless, Haptophytes displayed extraordinary KS diversity, averaging over 80 type I KS domains across the eight genomes analyzed, while Alveolata (16.9 KS domains per genome), Evosea (10.4 KS domains per genome), Chlorophyta (8.8 KS domains per genome) and Stramenopiles (4.0 KS domains per genome) also contained noticeable KS richness (Fig. 3.3). Modular *cis*-AT KS domains were abundant in Haptophytes (92.6%) and Chlorophyta (85.1%), whereas in the phylum Alveolata, type I FAS protist KS domains were most common (63.6%) and in the phylum Evosea, protist KS domains were dominant (99.3%) (Fig. 3.3). Most animal genomes analyzed were in the phyla Chordata (n=758) and Arthropoda (n=406), and they contained 2.48 and 4.99 type I KS domains respectively (Fig. 3.3). In Arthropoda, these type I KS domains mostly belonged to the type I FAS metazoa class (91.5%), while in Chordata, they belonged to both the type I FAS metazoa class (50.3%) and metazoa PKS class (39.3%) (Fig. 3.3). While only 10 Rotifera genomes were analyzed, this phylum held the greatest abundance of KS domains within the Metazoa (12.7 per genome), with these surprisingly spread out across the

modular *cis*-AT (5.5 per genome), iterative highly-reducing (1.7 per genome), type I FAS metazoa (1.6 per genome), hybrid *cis*-AT (1.1 per genome) and *trans*-AT (1.1 per genome) KS classes (Fig. 3.3).

Distribution of type II KS domains across taxonomic lineages

Most KS domains from type II PKS pathways were seen in bacterial genomes (98%), although their frequency was much lower than type I KS domains. Six phyla contained more than 0.5 type II KS domains per genome: Myxococcota (1.61), Bacteroidetes (1.16), Planctomycetes (1.05), Actinobacteria (1.05), Verrucomicrobia (0.81), and Pseudomonadota (0.52) (Fig. 3.S4). Abundances of the type II PKS subclasses varied, with aryl polyene KSs enriched in Bacteroidetes, Myxococcota, and Pseudomonadota (0.65, 0.57 and 0.37 per genome, respectively), while type II aromatic KSs were common in Planctomycetes, Actinobacteria, Myxococcota, and Verrucomicrobia (0.87, 0.66, 0.65 and 0.50 per genome, respectively) (Fig. 3.S3). To date, type II aromatic polyketides have only been observed in Actinobacteria, suggesting that type II KS diversity awaits in understudied taxonomic lineages.

Diversity of type I KS domains across the tree of life

To assess the diversity of type I KS domains, rarefaction curves were created by clustering full-length KS domains into operational biosynthetic units (OBUs) that shared 80% sequence similarity. I found 13 phyla with more than 400 full-length type I KS domains, with eight from bacteria (Acidobacteriota, Actinomycetota, Bacillota, Bacteroidota, Cyanobacteriota, Myxococcota, Planctomycetota, and Pseudomonadota), two from fungi (Ascomycota and Basidiomycota), one protist (Haptophyta) and two from metazoa (Arthropoda and Chordata). Of

these, Actinobacteria, Myxococcota, Pseudomonadota, and Ascomycota had the most diversity (clustering into 358, 341, 321, and 310 KS OBUs, respectively from 400 KS domains sampled), while Haptophyta and Chordata had the least diversity, clustering into 108 and 40 KS OBUs, respectively (Fig. 3.4). Within the metazoa, Arthropoda contained six times as many KS OBUs as Chordata, and at smaller KS sample sizes, Nematoda, Mollusca, and Rotifera showed elevated OBU diversity compared to Chordata as well (Fig. 3.4). While Haptophyta was found to contain high KS abundance (over 80 KS domains per genome), upon clustering these into OBUs, it had noticeably lower KS diversity compared to other protists and algae, with Alveolata and Chlorophyta containing the greatest KS diversity (Fig. 3.4).

To compare the KS diversity in genomes with experimentally characterized PKS BGCs, the genome-extracted KS domains were clustered with KS sequences from the MIBiG database. From this, Mycoxoccota (40.3%), Bacillota (33.1), Actinobacteria (27.7%), Cyanobacteria (25.1%), and Pseudomonadota (17.9%) all contained a high proportion of KS domains clustering with KS domains from the MIBiG database (Fig. 3.S5). These trends largely align with the composition of the 813 experimentally characterized bacterial polyketide BGCs within the MIBiG database, which mostly belong to Actinobacteria (65.3%), Pseudomonadota (11.7%), Myxococcota (10.1%), and Cyanobacteria (7.4%). In contrast, KSs from the understudied taxonomic lineages of Acidobacteria, Planctomycetes, and Verrucomicrobia had 0% clustering overlap with MiBIG sequences, and Chloroflexi held less than 1% overlap. Within fungi, KS domains from the phylum Ascomycota clustered at an elevated rate with KSs in the MIBiG database (14%) when compared to all other fungal phyla (<5%) (Fig. 3.S5).

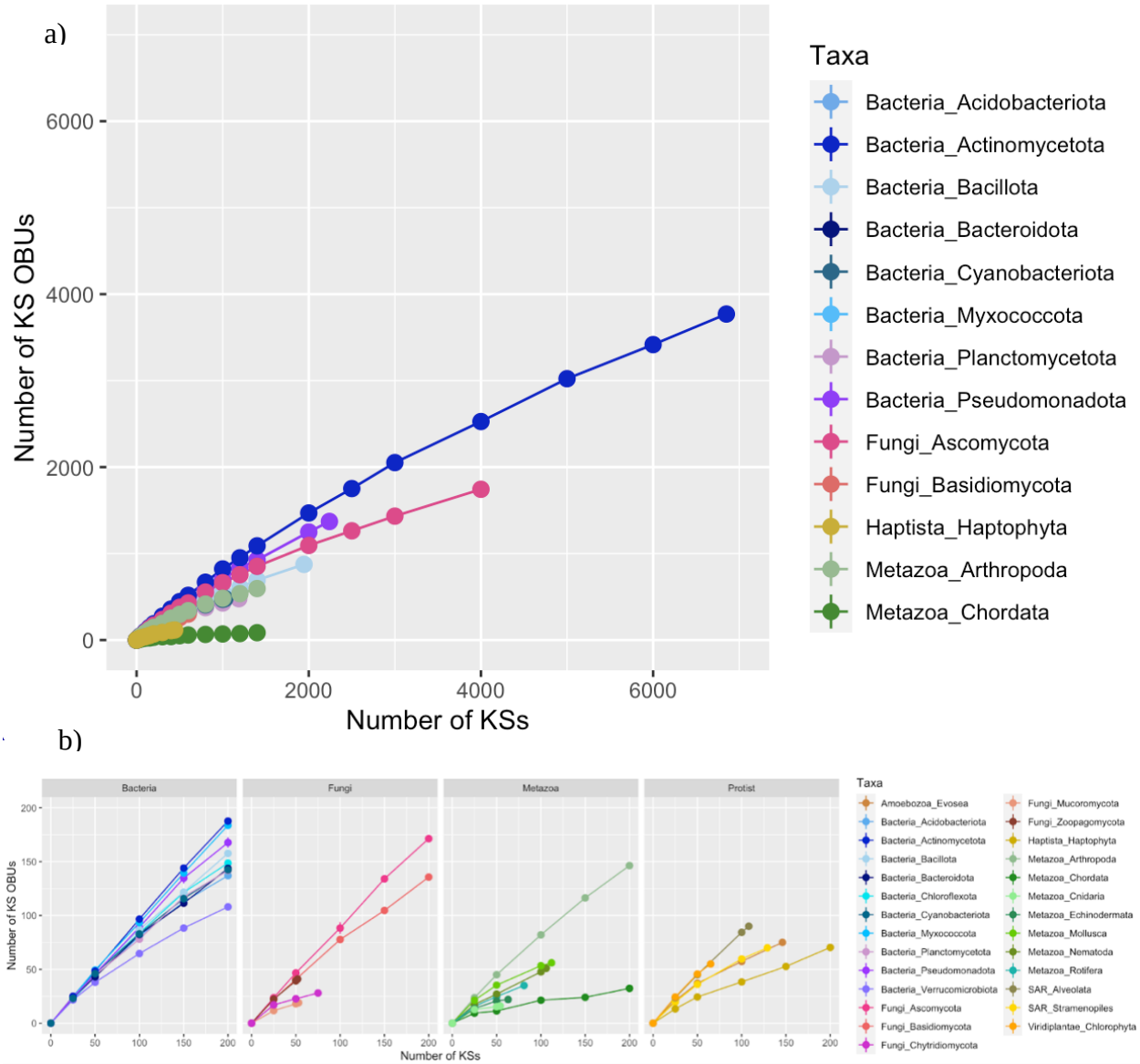

Figure 3.4. Type I KS diversity across phyla. (a) Rarefaction curve showing number of KS OBUs (80% similarity) across phyla with greater than 400 full-length type I KS domains. X-axis denotes the number of KS domains and the Y-axis denotes the number of KS OBUs. (b) All phyla with more than 50 full-length type I KS domains. Each panel represents a Kingdom.

Fungal type I KS diversity

To assess unexplored fungal KS diversity, a phylogeny using the 3,242 fungal KS OBUs was constructed, with the 192 KS OBUs from fungal polyketide pathways in MIBiG used as reference points (Fig. 3.5). From this, an interesting clade of 26 KS OBUs that was classified by

NaPDoS2 as hybrid *cis*-AT was observed. Notably, across all the fungal KS pathways in the MIBiG database, none contain a hybrid *cis*-AT KS (Fig. 3.5), suggesting that *cis*-hybrid polyketides have yet to be linked to a fungal producer. To follow up on this clade the genomes associated with these 26 KS OBUs were run through antiSMASH to understand their biosynthetic context. Notably, eight of these KSs were not identified by antiSMASH, highlighting the advantage of using NaPDoS2 to recover KS domains from fragmented assemblies or poorly understood BGCs. The other 18 KS OBUs within this clade were identified by antiSMASH as NRPS BGCs, and not as NRPS-PKS hybrids (Fig. 3.S6). This points to potentially novel hybrid peptide-polyketide natural product structural diversity within the fungal Kingdom.

The majority of the 3,242 fungal KS OBUs were classified by NaPDoS2 as iterative highly-reducing (65.9%) and iterative non-reducing (24.3%). Notably, these KS OBUs were mostly traced back to the phylum Ascomycota (98.6% and 85.8%, respectively). In contrast, 6.4% of the total fungal KS OBUs were classified by NaPDoS2 as modular *cis*-AT, with the majority seen in conserved clades tracing to Basidiomycota (62.7%) (Fig. 3.5). Some within this small subset may have been misclassified, as closer analysis of the Basidiomycota PKS pathway for the production of strobilurin A showed that NaPDoS2 classified the KS domain as modular *cis*-AT, but the percent similarity to the closest database match was only 36%. In contrast, analysis of the strobilurin A BGC showed that it functioned as a highly-reducing KS. In the MIBiG database to date, 292 PKS BGCs have been experimentally linked to a metabolite from the phylum Ascomycota, while only five PKS BGCs have been experimentally characterized in the phylum Basidiomycota, which could be the reason for certain KS misclassifications.

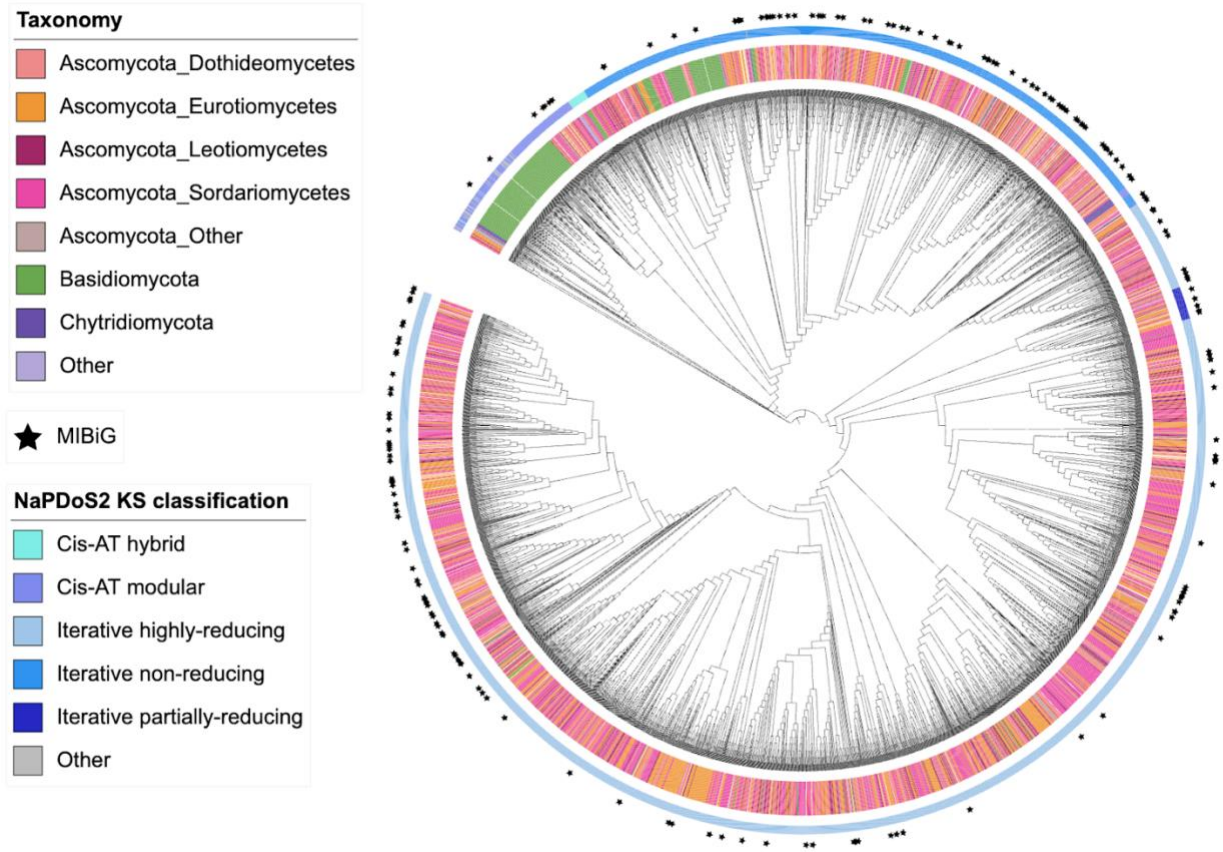

Figure 3.5. Type I KS diversity across fungi. KS OBUs (80% similarity) from fungi were used to construct a FastME phylogeny. The innermost ring denotes the taxonomy of the fungal genome containing each KS OBU. The second ring indicates the NaPDoS2 classification of each KS OBU. The outermost ring indicates all KS OBUs from fungal PKS pathways within the MIBiG database.

Comparison of type I KS diversity between cultured and uncultured bacteria

Most type I KS domains recovered from bacterial genomes were traced to strains that have been cultured (86.4%), with the remaining (13.6%) belonging to MAGs. To compare KS diversity identified using culture-dependent and culture-independent methods, the metagenome-extracted KS domains recovered in chapter 2 were compared to the genomic KSs. From this incorporation, over 20,000 bacterial KS domains extracted from chapters 2 and 3 were analyzed, with 57.1% obtained through culture-dependent methods and 42.9% through culture-independent

methods. To assess differences in polyketide diversity, a sequence similarity network (SSN) was constructed using KS OBUs (80%), with clustering of the SSN set to 70% similarity (Fig. 3.6). From this, 90 clusters with at least 10 KS OBUs were recovered, with these clusters spread across different KS classes: 16 KS clusters from hybrid *cis*-AT pathways, 17 KS clusters from *trans*-AT pathways, 23 KS clusters from iterative enediyne or PUFA pathways, and 34 KS clusters from modular *cis*-AT, olefin synthase, iterative PTM or iterative aromatic pathways (Fig. 3.6). Of these clusters, 36.6% contained a composition that was significantly different from the overall frequency across all KS OBUs (Fisher exact test, $p<0.05$). A few notable examples include a PUFA cluster of 93 KS OBUs with 96.7% collected from metagenomic data and a modular *cis*-AT cluster of 881 KS OBUs, with 79.5% from cultured isolates (Fig. 3.6).

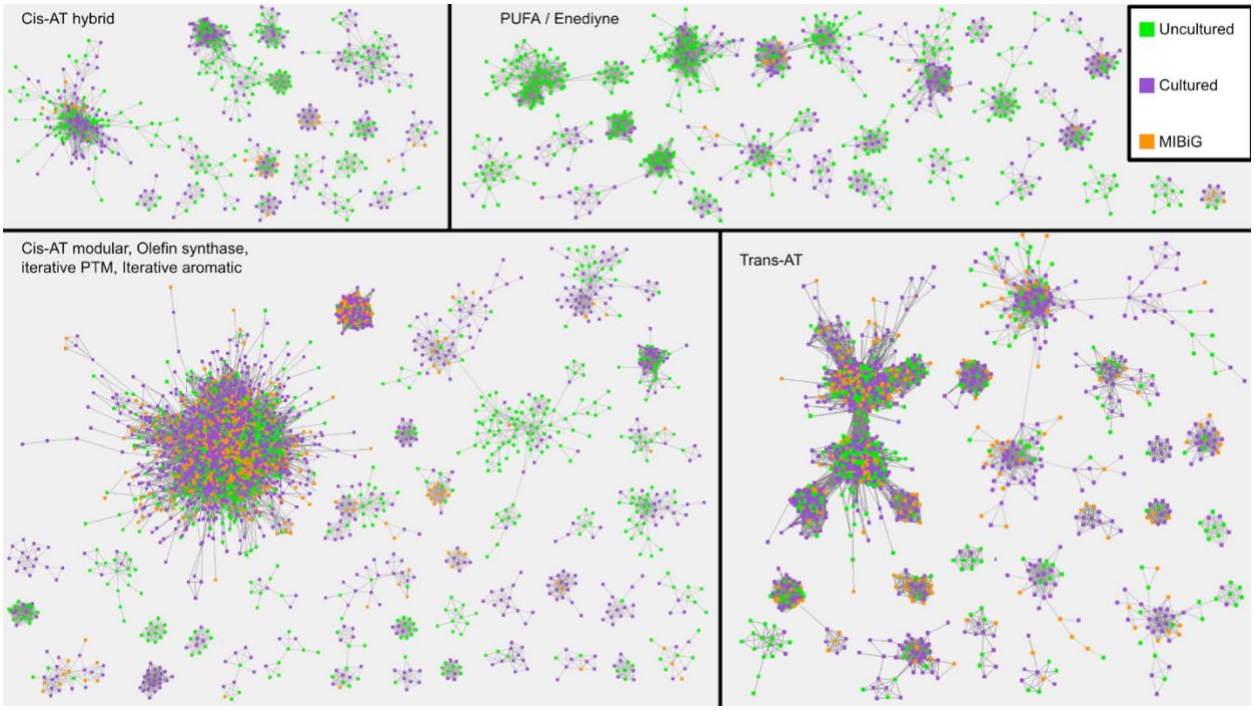

Figure 3.6. Type I KS diversity compared between cultured and uncultured taxa. SSN of KS OBUs from cultured strains (purple) and metagenomic data (green). MIBiG KS OBUs (orange) were used as reference points. Each node represents a single OBU (80% similarity), with nodes connected if they shared greater than 70% similarity. Panels represent different type I KS classes.

3.5 Discussion

In this work, I assessed the polyketide biosynthetic potential across the tree of life (Bacteria, Archaea, Plasmids, Viruses, Fungi, Algae, Protists, Plants, and Animals) using well-developed genomic datasets. This study was facilitated by the recently updated NaPDoS2 webtool, which allowed for the rapid extraction and classification of 53,713 KS domains into more than 40 KS subclasses. As the traditional natural product discovery pipeline of cultivation, chemical extraction, and bioassay-guided isolation has been challenged by high rediscovery rates in the past decade, genome mining has gained importance as a way to identify and prioritize BGCs of interest.

Recently, analyses of biosynthetic diversity across the bacterial kingdom have been carried out using tools like antiSMASH and PRISM to identify and compare full-length BGCs (Cimermancic et al., 2014; Wei et al., 2021; Chen et al., 2022; Gavriilidou et al., 2022). These studies have found widely distributed BGC families that lack characterized members, and that only 3% of natural product BGCs have been experimentally characterized. However, the analysis here using NaPDoS2 carries some distinct advantages. First, a phylogenetic classification scheme that focuses solely on polyketide diversity provides higher resolution (>40 KS subclasses) than pipelines such as antiSMASH, which require complete BGCs. Additionally, by using this KS sequence tag approach, the level of computation time decreases significantly; for example, an average fungal genome takes 22 seconds to run through NaPDoS2, while taking 13 minutes to run through antiSMASH. Additionally, NaPDoS2 can handle fragmented genomes (such as MAGs) and identify KS domains that could fall outside of BGCs if genomic co-localization is lacking (as in animals). Further, antiSMASH is designed for bacterial and fungal genomes, making NaPDoS2 the only webtool that also assesses PKS diversity within animals

and protists. While PKSs are best known from bacteria and fungi, animal KSs have recently been seen in birds, fish, (Ganley and Derbyshire, 2020), nematodes (Feng et al., 2021), Sacoglossans (Torres and Schmidt, 2019; Torres et al., 2020), and Echinoderms (Li et al., 2022).

Similar to previous genome mining studies (Cimermancic et al., 2014; Wei et al., 2021; Chen et al., 2022; Gavriilidou et al., 2022), Actinobacteria, Myxococcota, and Cyanobacteria were found to be biosynthetically rich. By using NaPDoS2 to explore PKS diversity in detail, I was able to show that Actinobacteria are enriched in modular *cis*-AT KS domains, while Myxococcota and Cyanobacteria are enriched in hybrid *cis*-AT KS domains. The phylum Bacillota was also found to be the most biased in terms of their KS composition, with over 80% of the type I KS domains in their genomes belonging to the *trans*-AT subclass. This aligns with previous studies that have shown Bacillota to contain large numbers of *trans*-AT PKS pathways (Nguyen et al., 2008), although the relative absence of other type I PKS pathways is still surprising.

Here, I establish that the presence of type I polyketides is widespread across phyla, with ten different phyla containing at least one type I KS domain per 4 genomes analyzed (Myxococcota, Actinobacteria, Cyanobacteria, Planctomycetes, Chloroflexi, Bacillota, Pseudomonadota, Acidobacteria, Bacteroidetes, and Verrucomicrobia). Of the more than 800 bacterial-derived polyketides in the MIBiG database, over 94% were isolated from Actinobacteria, Pseudomonadota, Myxococcota, and Cyanobacteria. These four phyla are among the most cultured, which is likely one of the reasons behind this skewed distribution. In contrast, no polyketides within the MIBiG database were isolated from the phyla Acidobacteria and Planctomycetes. The phyla Chloroflexi (one polyketide, aurantoside A) and Verrucomicrobia (two polyketides, lasonolide A and palmerolide) are also rare in terms of polyketide discovery.

The compound aurantoside A was originally discovered from the sponge *Theonella* and is a dichlorinated antifungal and cytotoxic metabolite that contains a polyene-tetramic acid core. This reinforces that understudied taxonomic lineages can be valuable sources of novel polyketides. While previous work has hinted that the phyla Verrucomicrobia and Acidobacteria contain PKS gene clusters, the detection of 1,172 type I KS OBUs (80% clustering) across these four understudied taxonomic lineages (Acidobacteria, Chloroflexi, Planctomycetes, and Verrucomicrobia) suggests that prioritization of these phyla could lead to novel and structurally diverse compounds.

A recent analysis of polyketide diversity across 1,000 fungal genomes found type I iterative highly reducing and non-reducing PKSs to be enriched in Ascomycota, while the Basidiomycota contained fewer PKS BGCs (Robey et al., 2021). Here, I show that Basidiomycota contains phylogenetically distinct clades of KS domains, many of which are classified by NaPDoS2 as modular *cis*-AT. This perhaps could be a misclassification, as the strobilurin A BGC (one of the five Basidiomycota PKSs within the MIBiG database) also contains a KS domain classified by NaPDoS2 as modular *cis*-AT, but experimental characterization has shown it to operate as an iterative highly reducing KS. However, this KS contains a different BGC architecture compared to typical highly reducing PKS pathways in that it has partial hydrolase and C-MeT (C-terminal methyltransferase) domains downstream of the terminal ACP, which is more characteristic of iterative non-reducing PKSs. This allows for the formation of the very unusual E,Z,E triene moiety that is seen within the compound strobilurin A, which is well-known as an agricultural fungicide. While most of these modular *cis*-AT classifications for Basidiomycota fungal KS domains are likely incorrect, the phylogeny suggests the existence of distinct KS functionality in Basidiomycota, which is validated by strobilurin A

having unique structural motifs. This indicates the value of placing KS sequences in a phylogenetic context to aid the discovery of new compounds, even if the NaPDoS2 classification is wrong. The discovery of more polyketide compounds from Basidiomycota and other non-Ascomycota fungal lineages, along with populating the NaPDoS2 database with the known Basidiomycota PKS pathways could improve KS resolution.

Hybrid *cis*-AT KSs are seen in a diverse set of fungi, yet to date, no fungal compounds have been found that contain a hybrid *cis*-AT KS domain. To validate these hits, I ran the genomes containing a hybrid *cis*-AT KS domain through antiSMASH, finding 80% of them to reside within pathways that antiSMASH classified as NRPSs. The remaining 20% were not identified by antiSMASH, highlighting the advantage of the NaPDoS2 sequence tag approach. The discovery of compounds that correspond to fungal hybrid *cis*-AT KSs could be important in uncovering new polyketide structures.

In the future, this dataset of 53,713 genome-extracted KS domains will be a useful reference for PKS genome-mining studies. This well-curated reference dataset of KSs with known genomic origins can serve as a taxonomic marker and provide phylogenetic context for taxonomy-independent surveys such as those utilizing KS amplicon or metagenomic data. Additionally, this collection of KSs could be useful in creating class and subclass-specific primers for high-throughput amplicon analyses. The analyses presented in this chapter show where targeted discovery efforts could focus, especially in taxa that have not been explored before, thus illustrating the power KS domains contain as search hooks for uncovering novel polyketides.

3.6 Funding sources

This research was supported by the National Science Foundation Graduate Research Fellowship Program (grant no. DGE-2038238 to H.W.S. and grant no. DGE-1650112 to K.E.C.). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

3.7 Acknowledgements

Chapter 3, in full, is not a reprint of any materials that have been submitted for publication. The dissertation author was one of two equal contributors to this work.

3.8 Supplementary Figures and Tables

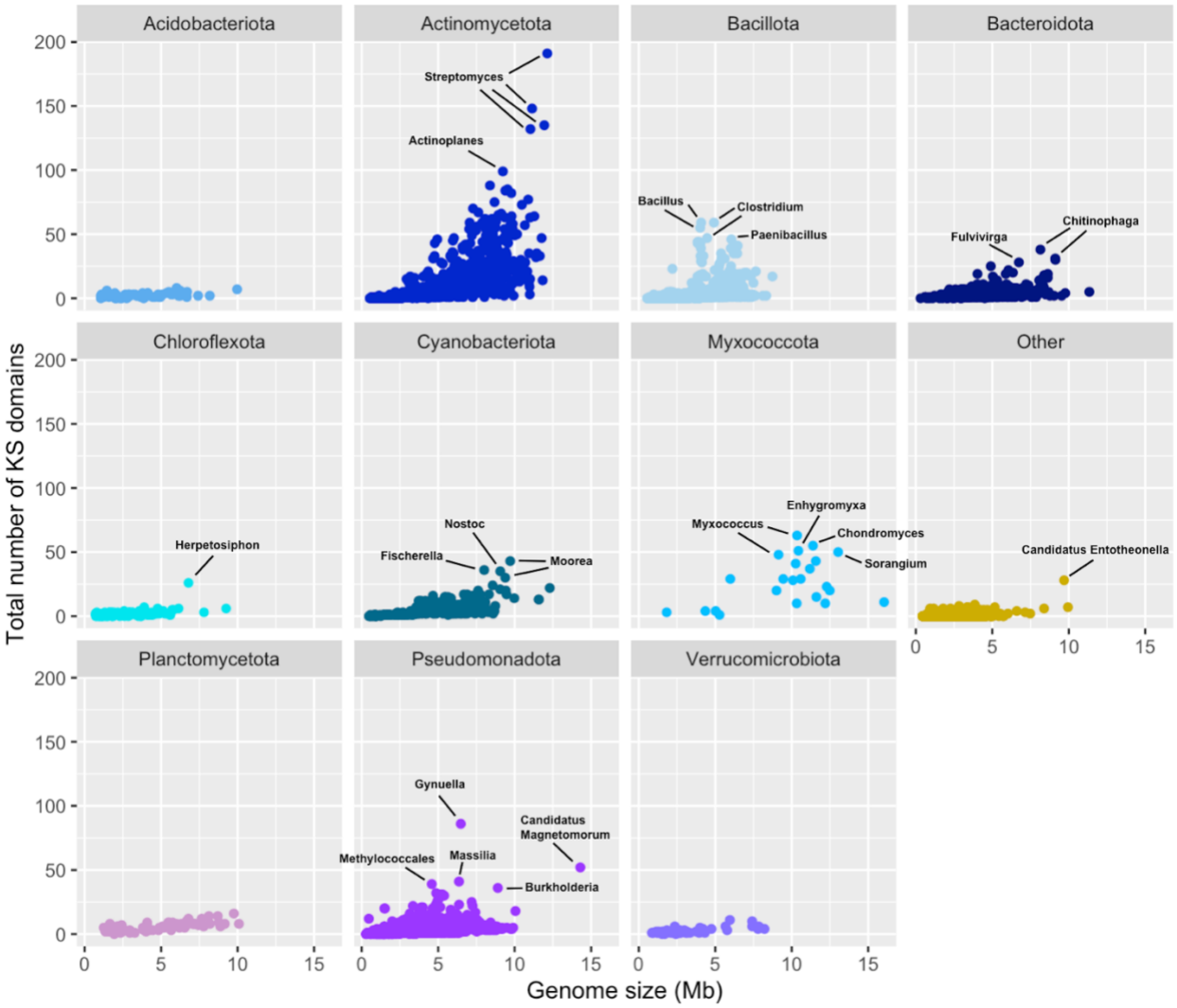

Figure 3.S1. KS abundance across bacteria. The graph shows the number of KS domains within a given genome (y-axis) relative to genome size (x-axis). Each dot represents a genome within the WoL database, divided into phyla across the 11 panels.

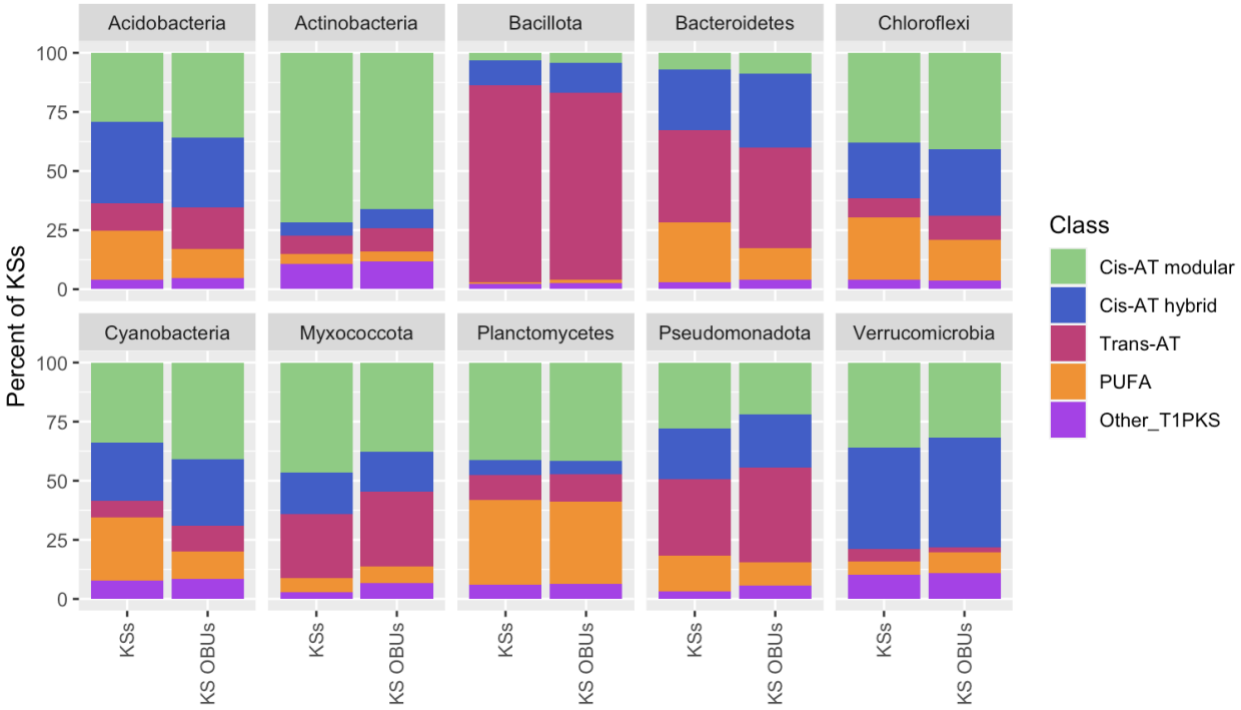

Figure 3.S2. Classification of total KSs and KS OBUs. The graph shows the NaPDoS2 classification of all KS domains within a given phylum (KSs) compared to the NaPDoS2 classification of all KS OBUs (80% similarity). Each panel represents a phylum, with the y-axis corresponding to the percentage of total KSs or KS OBUs in each category.

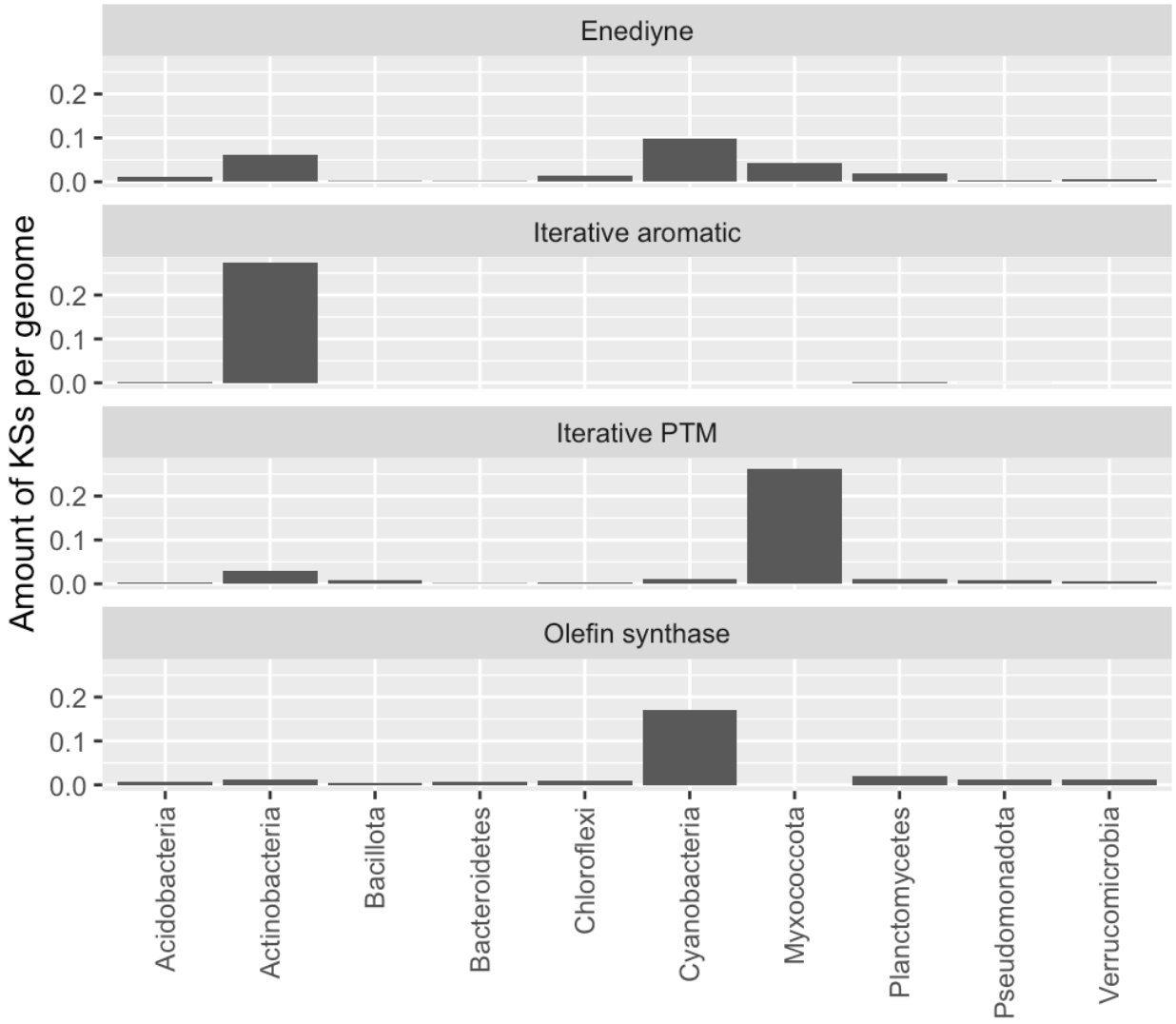

Figure 3.S3. Abundance of type I KSs from rare subclasses. Average number of type I KS domains per genome (y-axis) across bacterial phyla, with each panel displaying abundances for a different type I KS subclass (enediyne, iterative aromatic, iterative PTM, and olefin synthase).

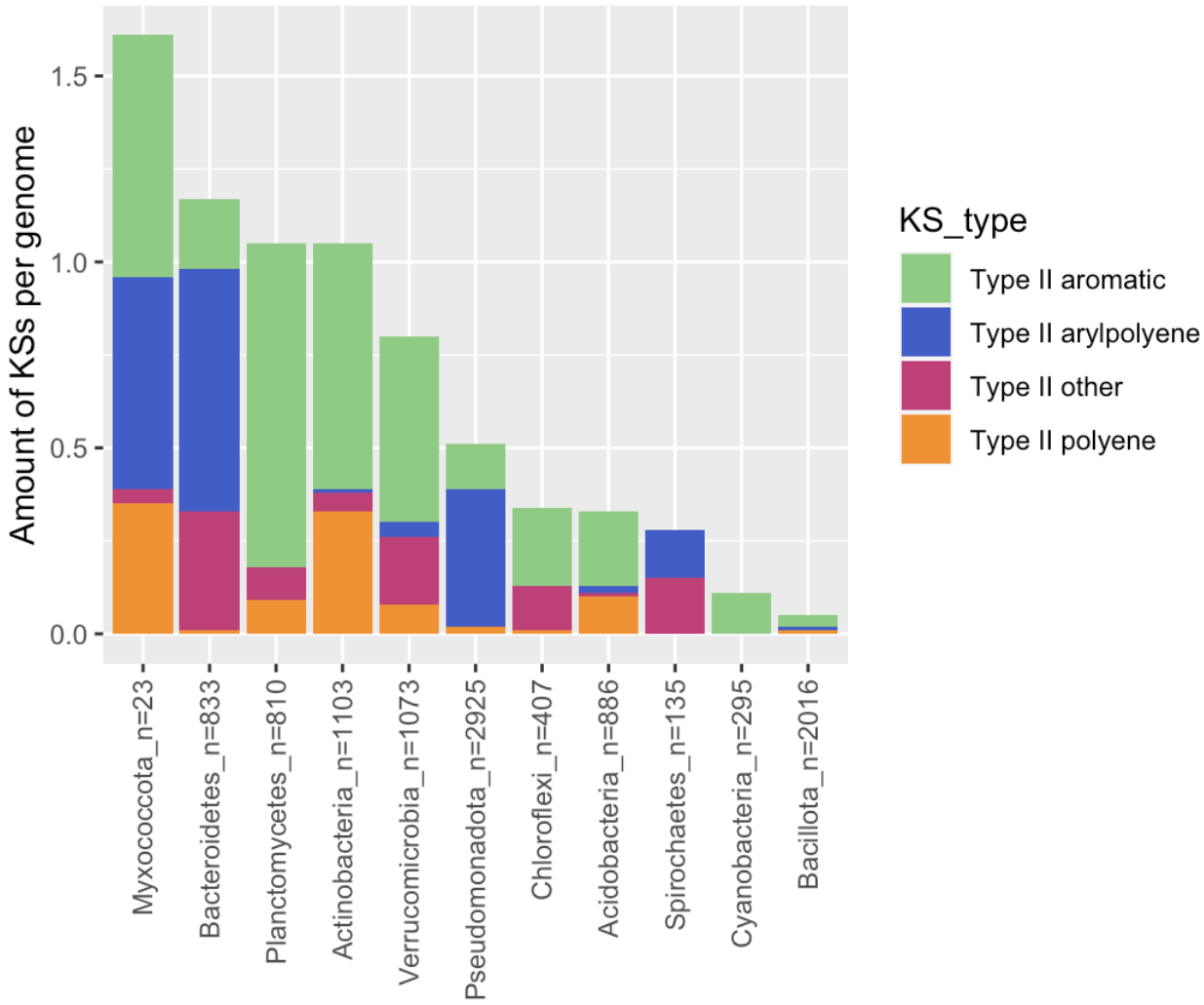

Figure 3.S4. Type II KS composition in bacteria. Average number of type II KS domains per genome (y-axis) across phyla, with colors denoting the NaPDoS2 KS classification. X-axis denotes the phyla and the amount of genomes analyzed in that phyla.

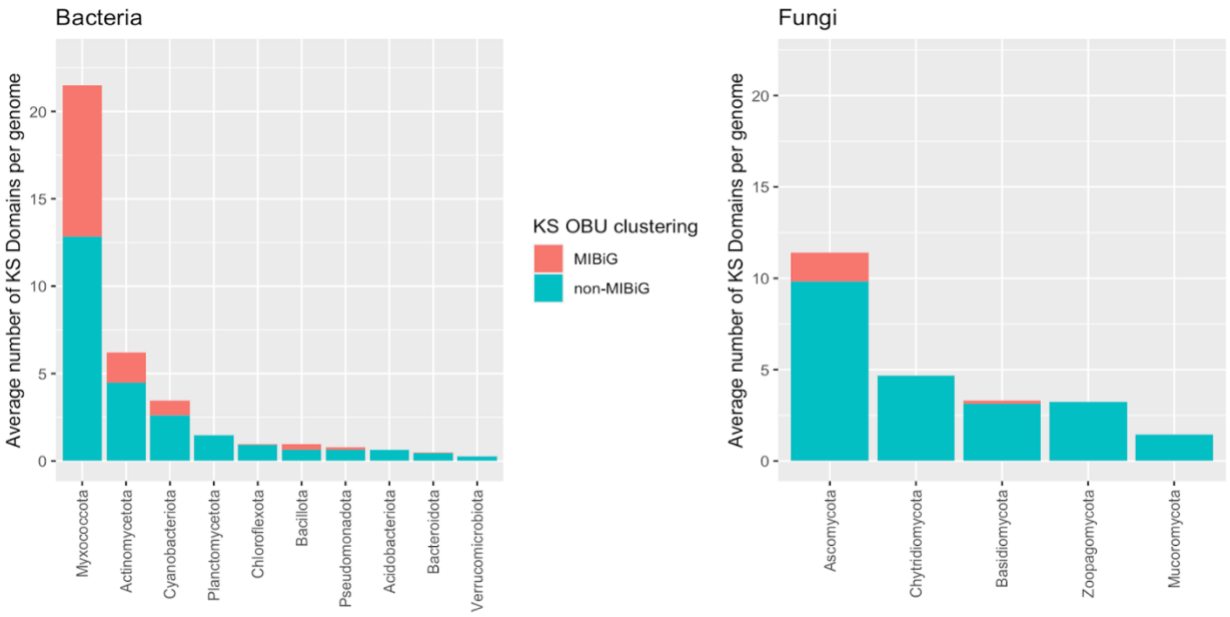

Figure 3.S5. Type I KS novelty in bacteria and fungi. Average number of type I KS domains per genome (y-axis) across phyla. Colors denote relationship relationships to MIBiG database KSs (80% OBUs). X-axis denotes phylum and the two panels represent bacteria and fungi.

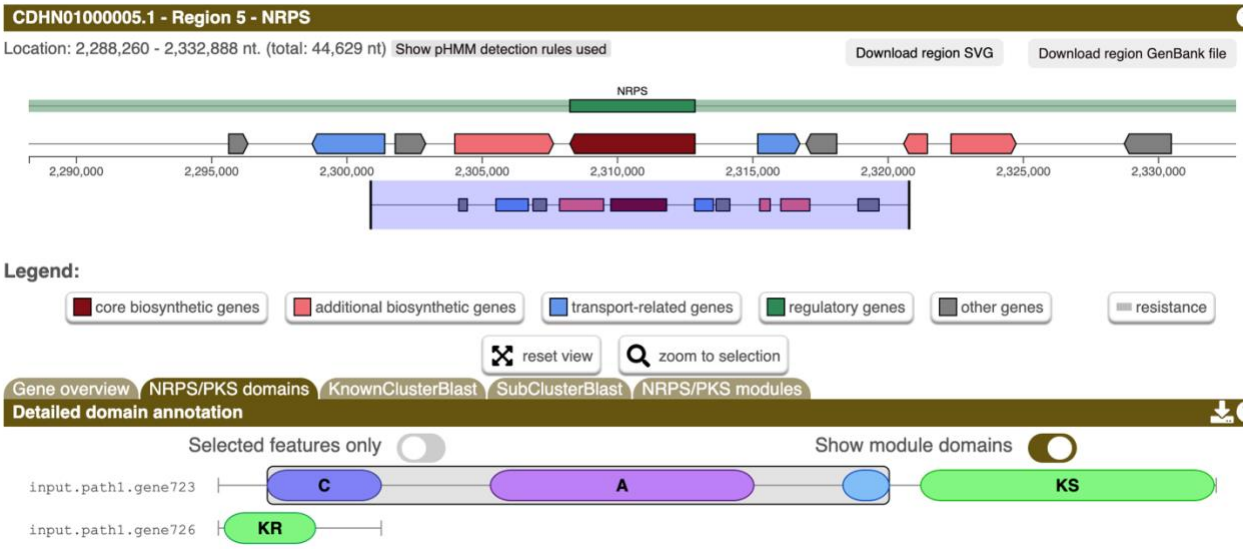

Figure 3.S6. BGC linked to fungal hybrid *cis*-AT KS domain. To follow up on a hybrid *cis*-AT clade in the fungal KS phylogeny, genomes with hybrid *cis*-AT KSs were run through antiSMASH. This is the antiSMASH output associated with the corresponding hybrid *cis*-AT KS domain from a *Torrubiella* genome. Those from other genomes had similar BGC architectures.

3.9 References

Bigot T, Temmam S, Pérot P, Eloit M. 2019. RVDB-prot, a reference viral protein database and its HMM profiles. F1000Res 8:530.

Blin K, Shaw S, Augustijn HE, Reitz ZL, Biermann F, Alanjary M, Fetter A, Terlouw BR, Metcalf WW, Helfrich EJN, van Wezel GP, Medema MH, Weber T. 2023. antiSMASH 7.0: new and improved predictions for detection, regulation, chemical structures and visualisation. Nucleic Acids Research 51:W46–W50.

Brinkac LM, Davidsen T, Beck E, Ganapathy A, Caler E, Dodson RJ, Durkin AS, Harkins DM, Lorenzi H, Madupu R, Sebastian Y, Shrivastava S, Thiagarajan M, Orvis J, Sundaram JP, Crabtree J, Galens K, Zhao Y, Inman JM, Montgomery R, Schobel S, Galinsky K, Tanenbaum DM, Resnick A, Zafar N, White O, Sutton G. 2010. Pathema: a clade-specific bioinformatics resource center for pathogen research. Nucleic Acids Research 38:D408–D414.

Carlin DE, Demchak B, Pratt D, Sage E, Ideker T. 2017. Network propagation in the cytoscape cyberinfrastructure. PLoS Comput Biol 13:e1005598.

Chen S, Zhang C, Zhang L. 2022. Investigation of the Molecular Landscape of Bacterial Aromatic Polyketides by Global Analysis of Type II Polyketide Synthases. Angew Chem Int Ed 61:e202202286.

Cimermancic P, Medema MH, Claesen J, Kurita K, Wieland Brown LC, Mavrommatis K, Pati A, Godfrey PA, Koehrsen M, Clardy J, Birren BW, Takano E, Sali A, Linington RG, Fischbach MA. 2014. Insights into Secondary Metabolism from a Global Analysis of Prokaryotic Biosynthetic Gene Clusters. Cell 158:412–421.

Davies J. 2013. Specialized microbial metabolites: functions and origins. J Antibiot 66:361–364.

Dias DA, Urban S, Roessner U. 2012. A Historical Overview of Natural Products in Drug Discovery. Metabolites 2:303–336.

Douarre P-E, Mallet L, Radomski N, Felten A, Mistou M-Y. 2020. Analysis of COMPASS, a New Comprehensive Plasmid Database Revealed Prevalence of Multireplicon and Extensive Diversity of IncF Plasmids. Front Microbiol 11:483.

Edgar RC. 2010. Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26:2460–2461.

Gavriilidou A, Kautsar SA, Zaburannyi N, Krug D, Müller R, Medema MH, Ziemert N. 2022. Compendium of specialized metabolite biosynthetic diversity encoded in bacterial genomes. Nat Microbiol 7:1324–1324.

Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, Mitros T, Dirks W, Hellsten U, Putnam N, Rokhsar DS. 2012. Phytozome: a comparative platform for green plant genomics. Nucleic Acids Research 40:D1178–D1186.

Grigoriev IV, Nikitin R, Haridas S, Kuo A, Ohm R, Otillar R, Riley R, Salamov A, Zhao X, Korzeniewski F, Smirnova T, Nordberg H, Dubchak I, Shabalov I. 2014. MycoCosm portal: gearing up for 1000 fungal genomes. Nucl Acids Res 42:D699–D704.

Harvey CJB, Tang M, Schlecht U, Horecka J, Fischer CR, Lin H-C, Li J, Naughton B, Cherry J, Miranda M, Li YF, Chu AM, Hennessy JR, Vandova GA, Inglis D, Aiyar RS, Steinmetz LM, Davis RW, Medema MH, Sattely E, Khosla C, St. Onge RP, Tang Y, Hillenmeyer ME. 2018. HEx: A heterologous expression platform for the discovery of fungal natural products. Sci Adv 4:eaar5459.

Herbst DA, Townsend CA, Maier T. 2018. The architectures of iterative type I PKS and FAS. Nat Prod Rep 35:1046–1069.

Hertweck C. 2009. The Biosynthetic Logic of Polyketide Diversity. Angew Chem Int Ed 48:4688–4716.

Klau LJ, Podell S, Creamer KE, Demko AM, Singh HW, Allen EE, Moore BS, Ziemert N, Letzel AC, Jensen PR. 2022. The Natural Product Domain Seeker version 2 (NaPDoS2) webtool relates ketosynthase phylogeny to biosynthetic function. Journal of Biological Chemistry 298:102480.

Kwon T, Hanschen ER, Hovde BT. 2023. Addressing the pervasive scarcity of structural annotation in eukaryotic algae. Sci Rep 13:1687.

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

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

Li Y, Steenwyk JL, Chang Y, Wang Y, James TY, Stajich JE, Spatafora JW, Groenewald M, Dunn CW, Hittinger CT, Shen X-X, Rokas A. 2021. A genome-scale phylogeny of the kingdom Fungi. Current Biology 31:1653-1665.e5.

Lin Z, Li F, Krug PJ, Schmidt EW. 2024. The polyketide to fatty acid transition in the evolution of animal lipid metabolism. Nat Commun 15:236.

Mauri M, Elli T, Caviglia G, Uboldi G, Azzi M. 2017. RAWGraphs: A Visualisation Platform to Create Open Outputs, p. 1–5. *In* Proceedings of the 12th Biannual Conference on Italian SIGCHI Chapter. ACM, Cagliari Italy.

Meleshko D, Mohimani H, Tracanna V, Hajirasouliha I, Medema MH, Korobeynikov A, Pevzner PA. 2019. BiosyntheticSPAdes: reconstructing biosynthetic gene clusters from assembly graphs. Genome Res 29:1352–1362.

Nayfach S, Camargo AP, Schulz F, Eloe-Fadrosh E, Roux S, Kyrpides NC. 2021. CheckV assesses the quality and completeness of metagenome-assembled viral genomes. Nat Biotechnol 39:578–585.

Nivina A, Yuet KP, Hsu J, Khosla C. 2019. Evolution and diversity of assembly-line polyketide synthases. Chem Rev 119:12524–12547.

O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D, Astashyn A, Badretdin A, Bao Y, Blinkova O, Brover V, Chetvernin V, Choi J, Cox E, Ermolaeva O, Farrell CM, Goldfarb T, Gupta T, Haft D, Hatcher E, Hlavina W, Joardar VS, Kodali VK, Li W, Maglott D, Masterson P, McGarvey KM, Murphy MR, O'Neill K, Pujar S, Rangwala SH, Rausch D, Riddick LD, Schoch C, Shkeda A, Storz SS, Sun H, Thibaud-Nissen F, Tolstoy I, Tully RE, Vatsan AR, Wallin C, Webb D, Wu W, Landrum MJ, Kimchi A, Tatusova T, DiCuccio M, Kitts P, Murphy TD, Pruitt KD. 2016. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res 44:D733–D745.

Pickett BE, Sadat EL, Zhang Y, Noronha JM, Squires RB, Hunt V, Liu M, Kumar S, Zaremba S, Gu Z, Zhou L, Larson CN, Dietrich J, Klem EB, Scheuermann RH. 2012. ViPR: an open bioinformatics database and analysis resource for virology research. Nucleic Acids Research 40:D593–D598.

Price MN, Dehal PS, Arkin AP. 2010. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. PLoS ONE 5:e9490.

Pye CR, Bertin MJ, Lokey RS, Gerwick WH, Linington RG. 2017. Retrospec- tive analysis of natural products provides insights for future discovery trends. Proc Natl Acad Sci U S A 114:5601–5606.

Robey MT, Caesar LK, Drott MT, Keller NP, Kelleher NL. 2021. An interpreted atlas of biosynthetic gene clusters from 1,000 fungal genomes. Proc Natl Acad Sci USA 118:e2020230118.

Singh HW, Creamer KE, Chase AB, Klau LJ, Podell S, Jensen PR. 2023. Metagenomic data reveals type I polyketide synthase distributions across biomes. mSystems 8:e00012-23.

Terlouw BR, Blin K, Navarro-Muñoz JC, Avalon NE, Chevrette MG, Egbert S, Lee S, Meijer D, Recchia MJJ, Reitz ZL, van Santen JA, Selem-Mojica N, Tørring T, Zaroubi L, Alanjary M, Aleti G, Aguilar C, Al-Salihi SAA, Augustijn HE, Avelar-Rivas JA, Avitia-Domínguez LA, Barona-Gómez F, Bernaldo-Agüero J, Bielinski VA, Biermann F, Booth TJ, Carrion Bravo VJ, Castelo-Branco R, Chagas FO, Cruz-Morales P, Du C, Duncan

KR, Gavriilidou A, Gayrard D, Gutiérrez-García K, Haslinger K, Helfrich EJN, van der Hooft JJJ, Jati AP, Kalkreuter E, Kalyvas N, Kang KB, Kautsar S, Kim W, Kunjapur AM, Li Y-X, Lin G-M, Loureiro C, Louwen JJR, Louwen NLL, Lund G, Parra J, Philmus B, Pourmohsenin B, Pronk LJU, Rego A, Rex DAB, Robinson S, Rosas-Becerra LR, Roxborough ET, Schorn MA, Scobie DJ, Singh KS, Sokolova N, Tang X, Udwary D, Vigneshwari A, Vind K, Vromans SPJM, Waschulin V, Williams SE, Winter JM, Witte TE, Xie H, Yang D, Yu J, Zdouc M, Zhong Z, Collemare J, Linington RG, Weber T, Medema MH. 2023. MIBiG 3.0: a community-driven effort to annotate experimentally validated biosynthetic gene clusters. Nucleic Acids Research 51:D603–D610.

Torres JP, Schmidt EW. 2019. The biosynthetic diversity of the animal world. Journal of Biological Chemistry 294:17684–17692.

Wei B, Du A, Zhou Z, Lai C, Yu W, Yu J, Yu Y, Chen J, Zhang H, Xu X, Wang H. 2021. An atlas of bacterial secondary metabolite biosynthesis gene clusters. Environmental Microbiology 23:6981–6992.

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

Zhu Q, Mai U, Pfeiffer W, Janssen S, Asnicar F, Sanders JG, Belda-Ferre P, Al-Ghalith GA, Kopylova E, McDonald D, Kosciolek T, Yin JB, Huang S, Salam N, Jiao J-Y, Wu Z, Xu ZZ, Cantrell K, Yang Y, Sayyari E, Rabiee M, Morton JT, Podell S, Knights D, Li W-J, Huttenhower C, Segata N, Smarr L, Mirarab S, Knight R. 2019. Phylogenomics of 10,575 genomes reveals evolutionary proximity between domains Bacteria and Archaea. Nat Commun 10:5477.

Chapter 4. Multi-omic assessment of polyketide biosynthetic potential across abyssal sediments

4.1 Abstract

Microbially-derived polyketides represent a diverse class of natural products that account for some of today's most useful medicines. Yet the discovery of novel polyketides has become challenging with high rediscovery rates from well-studied taxa posing a significant roadblock. To address this challenge, poorly explored biomes such as marine sediments have been targeted and the strains isolated have proven to be a rich source of novel polyketides. Yet these strains largely originate from nearshore sediments leaving abyssal (4000-6000 m) marine sediments, which comprise >80% of the ocean floor, poorly explored. Furthermore, the bacteria obtained in culture represent only a small component of the Earth's microbiome leaving much to be learned about the taxonomic composition and biosynthetic potential of abyssal sediment communities. Using methods such as shotgun metagenomics or PCR amplicon sequencing, it is now possible to use culture-independent approaches to explore taxonomic and biosynthetic diversity across Earth's biomes in greater depth. Here, I used culture-independent approaches to compare taxonomic and biosynthetic diversity across sediments from three abyssal locations spanning 880 km, with each containing distinct geochemical regimes. The sediment communities had distinct taxonomic (16S rRNA) and biosynthetic (ketosynthase domain) signatures across abyssal sites and sediment horizons when compared to nearshore sediments, suggesting they harbor unique opportunities for natural product discovery. KS phylogenies revealed clades that are largely distinct from those observed in experimentally characterized PKS pathways, and over 90% of our metabolomic features did not match spectra in the GNPS library, further supporting the discovery potential from abyssal sediments. Deep-sea metagenome-assembled genomes linked type I KS domains to the poorly studied phylum Gemmatimonadetes, providing a connection between biosynthetic potential and candidate polyketide producers. Taken together, these

findings suggest that abyssal sediments represent a reservoir of novel polyketide biosynthetic diversity.

4.2 Introduction

The oceans cover 70% of the earth's surface with 80% of the ocean floor corresponding to abyssal depths (4000-6000 m). Abyssal sediments represent one of the planet's most underexplored habitats with diverse microbial communities playing key roles in global biogeochemical cycles and marine food webs (Orsi et al., 2018; Orcutt et al., 2011; Baker et al., 2021; Hoshino et al., 2020). These sediments receive less organic input compared to continental margins, resulting in higher oxygen concentrations due to reduced microbial activity (D'Hondt et al., 2015). Abyssal sediment communities are less diverse than those observed in continental margin sediments, decrease in diversity with increasing sediment horizon depth (Kallmeyer et al., 2012), and are typically enriched in the phyla Proteobacteria and Firmicutes (Hoshino et al., 2020). Due to the presence of polymetallic nodules, some abyssal plains are being targeted for deep-sea mining operations (Stratmann et al., 2021), which will have unforeseen consequences on sediment microbial communities and their potential to yield useful new natural products.

Microorganisms have proven to be a key source of natural products, including many important pharmaceuticals (Abdel-Razek et al., 2020; Pye et al., 2017). Polyketides represent an important class of microbial natural products (Nivina et al., 2019) and are encoded by polyketide synthase (PKS) biosynthetic gene clusters (BGCs). These BGCs can exceed 100 kb and be difficult to assemble from genomic or metagenomic sequence data due to their highly repetitive sequences (Gao et al., 2021). PKS genes have been delineated into three types (I-III), with most experimentally characterized PKSs belonging to type I (Kautsar et al., 2020; Weissman et al.,

2004). A minimal T1PKS contains an acetyltransferase (AT) domain, which selects the appropriate building block, an acyl carrier protein (ACP) domain, to which the building block is tethered, and a ketosynthase (KS) domain, which catalyzes chain elongation between the growing polyketide and the ACP-bound extender unit (Shen et al., 2003; Fischbach et al., 2006). Type I PKSs can be further divided into three groups (*cis*-AT, *trans*-AT, and iterative *cis*-AT), with *cis*-AT and *trans*-AT PKSs functioning as multimodular assembly lines where each KS domain catalyzes one round of chain elongation (Piel et al., 2010; Chen et al., 2016; Lewis et al., 2020). Type I iterative *cis*-AT PKSs generally differ in having only one module (monomodular) with the single KS domain catalyzing multiple rounds of chain elongation (Lewis et al., 2020).

One major roadblock to microbial natural product discovery is the reliance on cultured bacteria, which represent a small percentage of the diversity observed in nature (Lloyd et al., 2018). Additionally, for the strains that can be cultured, many BGCs are not expressed under laboratory growth conditions (Amos et al., 2017). Genome mining has provided an alternative route to natural product discovery, with bioinformatic tools such as antiSMASH (Blin et al., 2021) and MIBiG (Kautsar et al., 2020) providing rapid methods to annotate BGCs and link them to their small molecule products, respectively. The webtool NaPDoS2 can be used probe polyketide biosynthetic diversity with greater specificity based on PKS associated KS domain sequences (Klau et al., 2022). Notably, this tool uses type I KS domain phylogenies to distinguish among *cis*-AT, iterative, and *trans*-AT PKSs and eight additional KS subgroups observed within these broader categories (hybrid *cis*-AT, *cis*-loading module, olefin synthase, PUFA, enediyne, aromatic, polycyclic tetramate macrolactam, and hybrid *trans*-AT) (Klau et al., 2022; Ziemert et al., 2012; Wang et al., 2020; Miyanaga et al., 2018). These KS categories are linked to distinct enzyme features or biosynthetic mechanisms that can inform discovery efforts

by allowing for rapid assessments of biosynthetic diversity across strains or biomes. A major advantage of this "sequence tag" approach is that it does not require BGC assembly, making it ideal for assessing type I PKS diversity associated with KS amplicons or poorly assembled metagenomic data (Klau et al., 2022).

PCR-amplified KS sequences have been used to document extensive type I PKS diversity across diverse environments (Charlop-Powers et al., 2014; Charlop-Powers et al., 2015; Libis et al., 2019; Bech et al., 2020). Notably, soils have been found to differ in KS diversity and richness based on nutrient content, geographic distance, and biome type (Charlop-Powers et al., 2014; Charlop-Powers et al., 2015). KS amplicon sequencing allows for the recovery of low-frequency sequences, giving an in-depth snapshot of the biosynthetic potential of a given sample. The prioritization of targeted KS types has been used to facilitate polyketide discovery through the screening of large-insert clone libraries (Libis et al., 2019). The most common KS primer set used to assess type I PKS diversity was designed based on a ~750 bp region within modular *cis*-AT KSs detected in the Actinobacteria, Cyanobacteria, and Deltaproteobacteria phyla (Moffitt et al., 2003; Ayuso-Sacido et al., 2005; Rascher et al., 2003). However, in our recent assessment of metagenomic KS diversity, primer modifications to improve the amplification of hybrid *cis*-AT and *trans*-AT KS domains were suggested (Singh et al., 2023). This new primer set is designed to target all type I KS subclasses except those associated with iterative PUFA and enediyne PKSs (Singh et al., 2023).

Microorganisms cultured from marine sediments have led to the discovery of novel polyketides, many of which carry promising bioactivity (Cappello et al., 2021). And yet, recent culture-independent work has shown that the majority of PKS potential in marine sediments remains unrealized (Bech et al., 2020; Chase et al., 2023). For example, type I KS amplicons

from shallow marine sediments collected off the coast of Denmark carried greater KS richness than soils or seawater samples and were largely distinct from the KSs detected in cultures from the same sediments (Bech et al., 2020). Sediment-derived amplicons also had low similarity to reference KSs, a finding also seen in type I KS domains amplified from shallow Yellow Sea marine sediments (Wei et al., 2018). Recent metagenomes from shallow marine sediments collected in Moorea found geographic variation in PKS potential across spatial scales, with most of the PKSs recovered being unrelated to sequences in the MIBiG database (Chase et al., 2023). While both culture- and culture-independent studies have targeted microbes associated with continental margin sediments, the biosynthetic potential of deep-sea marine sediments has yet to be assessed. Here, I investigate type I polyketide biosynthetic potential across sediment samples taken from three biogeochemically unique abyssal locations. The results from multi-omic analyses indicate that abyssal sediments contain distinct KS communities compared to continental margin sediments and experimentally characterized PKS pathways, and thus provide opportunities for polyketide natural product discovery.

4.3 Methods

Sediment collection and processing

Marine sediment samples were collected in October 2020 using a gravity core from three abyssal sites (3°N, 7°N, 11°N) along a transect at 152°W (Fig. 1) aboard the R/V Kilo Moana. At each of the three sites, three replicate sediment cores were sampled, with each analyzed for the 0-1 cm and 4-5 cm sediment horizons (3 sites, 3 replicates, 2 depths, n=18 sediment samples) and placed into Whirl-Pak® bags. Samples were frozen (-80°C) on board the vessel and shipped on dry ice to Scripps Institution of Oceanography where they were stored at −80°C until

processing. Inventoried (-40°C) near-shore sediments collected from Fiesta Island (<2 m), Point Loma (<2 m), Fiji (<20 m), Belize (<20 m), and the Southern California borderland (300 m and 900 m) were also processed (Demko et al., 2021; Bogdanov et al., 2024, Guraieb et al., 2024). DNA was extracted from 1 g of the abyssal and near-shore (n=18) sediments using a combination of physical (bead beating) and chemical (phenol-chloroform) steps (Patin et al., 2013).

PCR, amplicon sequencing, and amplicon processing

The v4 region of the 16S rRNA gene was PCR amplified using the Earth Microbiome Project primers 515F (GTGYCAGCMGCCGCGGTAA) and 806R (GGACTACNVGGGTWTCTAAT) (Caporaso et al., 2018). Ketosynthase domains were amplified using primers KSF (GCNATGGAYCCNCARSANMGNNT) and KSR (GTNSCNGTNCCRTGNGYYTCNAY) which were designed based on previous analyses (Singh et al., 2023). The Phusion Green Hot Start II High-Fidelity PCR Master Mix was used for the products of both primer sets with the following cycles: 98°C for 30 s followed by 30 cycles of 98°C for 10 s, 60°C for 30 s, and 72°C for 30 s followed by a final 10 min extension at 72°C. Gel electrophoresis was used to confirm the size of the PCR products. PCR products were then cleaned using AMPure XP beads and normalized to the same DNA concentration. Purified amplicon libraries were sent to the Institute for Genomics and Bioinformatics (IGB) at UC Irvine for sequencing on an Illumina MiSeq v2 500 cycle at a depth of 130k reads per sample.

Raw 16S rRNA sequences were imported into QIIME2-2020.2 (Estaki et al., 2020) and denoised using the DADA2 (Callahan et al., 2016) denoise-paired pipeline. Taxonomy was assigned using the SILVA v132 database (Quast et al., 2013). KS amplicons were verified and

classified using the NaPDoS2 webtool (Klau et al., 2022). Both sequence types were then transferred into the Geneious ver. 2020.2 (Kearse et al., 2012) platform for downstream analyses. Phylogenies of the KS amplicons were constructed using the FastTree (Price et al., 2010) workflow implemented on NGPhylogeny (Lemoine et al., 2019) and visualized using iTol (Letunic and Bork, 2019).

Metagenomic sequencing

For the 0-1 cm horizon of each replicate across the three abyssal sites (n = 9), shotgun metagenomic libraries were constructed using a Nextera XT DNA library preparation kit (Illumina) at the CMI. All samples were sequenced on a NovaSeq 6000 system (Illumina) with 150 bp paired-end reads at the UC Davis Genome Center. Raw reads were quality trimmed, and adapters were removed using the BBMap aligner (Bushnell et al., 2014). Metagenomes were annotated using PROKKA (Seemann et al., 2014) and assembled using IDBA (Peng et al., 2011) as previously described (Chase et al., 2023). Metagenomes and MAGs were analyzed using NaPDoS2 to assess KS composition and diversity (Klau et al., 2022) and by antiSMASH 6.0 (Blin et al., 2021) to extract BGCs.

Sediment metabolomics

Metabolomes were generated for each 0-1 and 4-5 cm sediment horizon (3 replicates across 3 sites, (n = 18) by extracting 50 grams of sediment with 100 mL 1:1 methanol dichloromethane. The extract was filtered, evaporated using a rotary evaporator, and the dried extracts resuspended in HPLC-grade methanol to 1 mg/mL. LC-HRMS analysis was performed by injecting 5 μL into a 6530 Accurate-Mass QToF with ESI-source (Agilent) coupled with a

1260 Infinity HPLC (Agilent) equipped with a 150×4.6 mm Kinetex C18 5 μm column (Phenomenex, USA). Runs started with 20:80 acetonitrile:water with 0.1 % formic acid (FA) for 2 min followed by an 18 min gradient using 95% acetonitrile, which was then held for 2 min and increased to 100 % acetonitrile over 1 min and held for 2 min at a flow of 1 mL/min. MS/MS data was taken in positive mode over a range of 135–1700 m/z and analyzed using the GNPS2 platform (Aron et al., 2020). Metabolic networks were visualized in Cytoscape (Carlin *et al.*, 2017).

4.4 Results

Sediment collection and characteristics

Sediment cores were collected from three abyssal (>5000 m) sites (3°N, 7°N, 11°N latitude) along a north to south transect at 152°W longitude (Fig. 4.1a) aboard the R/V Kilo Moana in October 2020. These locations are among the most remote habitats sampled for marine natural products research, with the closest land mass being the Kiribati Atoll >600 km away. From each site, three cores were collected, and sub-samples from the 0-1 cm and the 4-5 cm horizons were analyzed (18 total sediment samples). The sites differ in their biogeochemical profiles with the 3°N site associated with relatively high biological productivity due to equatorial upwelling and greater sediment organic deposition (Dai et al., 2023). In contrast, strong density gradients at the 7°N and 11°N sites limit water layer mixing thus reducing biological productivity and sediment organic content (Dai et al., 2023). These differences were borne out in the biogeochemical data, with the 3°N site displaying greater DOC, $NO_3^- + NO_2^-$, and PO_4^{3-} concentrations and lower oxygen concentrations due to greater biological activity (Fig. 4.1b).

DOC and oxygen concentrations decreased from the 0-1 cm to 4-5 cm horizons across all sites whereas $NO_3^- + NO_2^-$ and PO_4^{3-} increased with depth (Fig. 4.1b).

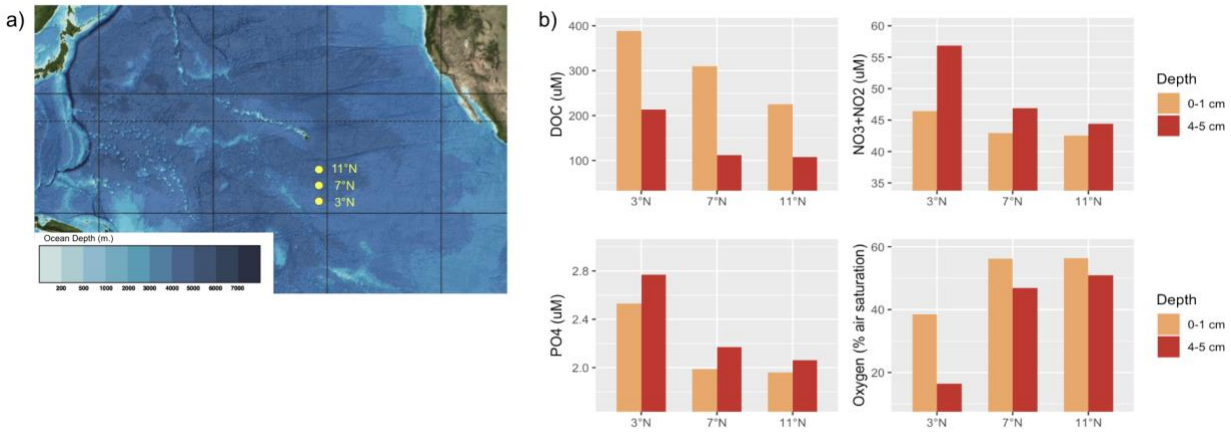

Figure 4.1. Abyssal sediment collection locations and biogeochemical signatures (a) Three abyssal sediment sampling sites mapped onto GEBCO bathymetric map. (b) Sediment biogeochemical measurements (DOC, $NO_3^- + NO_2^-$, PO_4^{3-}, oxygen % air saturation) for the three sites.

Microbial community diversity

The phylum-level taxonomic composition and relative abundance of microorganisms in the abyssal sediments (n=17, one sample yielded poor quality sequence data) varied little among sites based on 16S rRNA amplicon analysis (Fig. 4.S1). For comparative purposes, sediments from Fiesta Island, CA (<2 m), Point Loma, CA (<2 m), Fiji (<20 m), Belize (<20 m), and mesopelagic sediments from the Southern California borderland (300 m and 900 m) were also analyzed and collectively referred to as nearshore sediments. In contrast to the abyssal sites, the near-shore sediment communities (n=16) had greater variation, which was reflective of the greater geographic and habitat variability (Fig. 4.S1). A principal coordinates analysis (PCoA) based on a Bray–Curtis 16S rRNA gene sequence dissimilarity matrix revealed clustering patterns based on location, with the abyssal sediment communities showing dense clustering and significant separation (permutational multivariate analysis of variance (ANOVA), P <0.01) from

the near-shore sites (Fig. 4.2a). The primary phyla driving the separation of abyssal sediment communities from the nearshore sites were Acidobacteria, Gemmatimonadetes, Alphaproteobacteria, and Gammaproteobacteria (Fig. 4.2a).

Given the separation of the abyssal sites from better studied near-shore communities, the next aim was to assess variation among the three abyssal sites given their different biogeochemical profiles. A PCoA analysis based on 16S rRNA amplicon sequence variants (ASVs) revealed a significant difference between the 3°N site (ANOVA, $p< 0.01$) and the other two locations (Fig. 4.2b). These differences were also observed when the 0-1 and 4-5 cm horizons were compared (ANOVA, $p<0.01$). To explore these differences in a taxonomic context, I first mapped the 87 most abundant ASVs (i.e., those representing >0.25% of the community) detected across all samples to the closest NCBI RefSeq genomes (Fig. 4.S2) and found that the phylum Gammaproteobacteria was the most abundant (40/87 ASVs) and, within this phylum, *Woeseia* was the most abundant genus (16/40 ASVs). Seventy-seven percent of the most abundant ASVs shared less than 95% 16S rRNA similarity to their closest NCBI database match while 45% shared less than 90% similarity, providing evidence that most of the species-level taxa detected in these samples have not previously been observed. Average Shannon diversity index values were significantly greater for the 0-1 cm horizon (Tukey's HSD, $p>0.01$) but were not significantly different across sites (Tukey's HSD, $p=0.37$) (Fig. 4.S3).

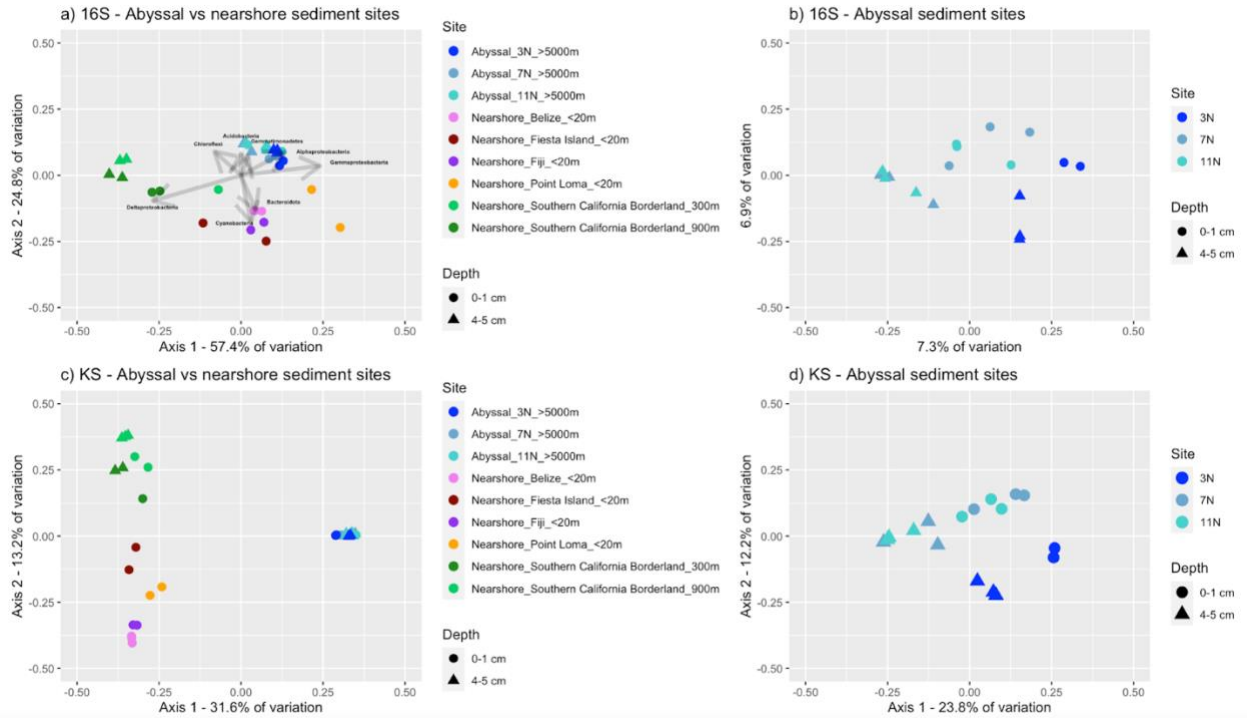

Figure 4.2. Sediment 16S rRNA and KS amplicon diversity. (a) PCoA of abyssal and nearshore sediment 16S communities based on a Bray–Curtis dissimilarity matrix. Arrows indicate the main taxa driving separation (b) PCoA of abyssal sediment 16S communities. (c) PCoA of abyssal and nearshore sediment KS communities. (d) PCoA of abyssal sediment KS communities.

KS amplicon diversity

Type I PKS biosynthetic gene diversity was estimated in abyssal (n=17) and nearshore (n=17) sediments based on amplified KS domain sequences. KS amplicons were classified using NaPDoS2 (22) after off-target sequences (20.1% of the total) were removed. Among the verified KS amplicons, a subset of 4000 was randomly selected from each sample and clustered into 80% sequence identity operational biosynthetic units (OBUs). Sediment KS composition strongly separated by location (Fig 4.2c), emphasizing the value of collecting from diverse sites. The greatest separation was between the abyssal and near-shore sediments (PERMANOVA, $p<0.01$, Fig. 4.2c). The majority of abyssal KSs were classified as modular *cis*-AT (42.9%), hybrid *cis*-AT (38.5%), and *trans*-AT (8.6%). Differences between abyssal and nearshore sediments (Fig.

4.S4) included the relative abundance of type I hybrid *cis*-AT KSs (38.5% vs. 20.4% relative abundance) and increased protist-type I FAS KSs in tropical nearshore samples from Belize and Fiji.

Despite the extreme environmental conditions, the abyssal KS amplicons revealed extraordinary levels of diversity, averaging 1209 OBUs (clustered at 80% similarity, Fig. 4.S5) across all samples. In aggregate, 6,426 OBUs were recovered from 68,000 abyssal sediment KS amplicons, more than twice the number of OBUs within the MiBIG database (Fig. 4.S6). A rarefaction curve suggests more diversity would be recovered with additional sequencing (Fig. 4.S6). The 0-1 cm horizon possessed greater KS diversity, both in number of OBUs and Shannon diversity index values, compared to the 4-5 cm horizon (Tukey's HSD, $p<0.01$) (Figs. 4.S5). While there were no significant differences (Tukey's HSD, $p>0.05$) in KS diversity across the three sites (Figs. 4.S5), KS composition within the 3°N site was significantly different from the 7°N and 11°N sites, with added separation between the 0-1 and 4-5 cm horizons (PERMANOVA, $p<0.01$) (Fig. 4.2d). Notably, this separation mirrored the differences between sites and sediment horizons observed in the biogeochemical data, establishing a linkage between environmental factors and biosynthetic potential.

KS amplicon novelty

Next, abyssal (n=68,000) and nearshore sediment (n=68,000) KS sequences were combined with KSs from the MIBiG and RefSeq databases (n=34,782) to assess their relationships. After clustering into 80% OBUs, only 0.8% of the 6,426 abyssal sediment OBUs contained a MIBiG or NCBI RefSeq sequence (Fig. 4.3a). This percentage was consistently low across the three largest KS subclasses (0.8% for modular *cis*-AT, 0.4% for hybrid *cis*-AT, and

1.9% for *trans*-AT). Further, only 18.0% of the abyssal sediment OBUs included KSs from the nearshore marine sediments, once again highlighting the biosynthetic gene novelty of the abyssal environments (Fig. 4.3a). In total, only 550 of 68,000 abyssal KSs shared >80% sequence similarity with a MIBiG PKS pathway (Fig 4.3b). Surprisingly, 28 of these shared >95% similarity, suggesting they could be associated with the production of similar compounds (Fig. 4.3b). These matches include KSs linked with the biosynthesis of salinilactam (20 sequences), salinosporamide (1 sequence), rifamycin (4 sequences), and quinolidomicin (3 sequences), all of which have been reported from the marine actinomycete genus *Salinispora* (20). While these close matches represent a small subset of the data, they were all detected in sediments from the 7°N and 11°N sites, where oxygen levels were the highest (Fig. 4.1) and could support *Salinispora* growth (38). Notably, the two most abundant MIBiG matches (both <80% sequence identity) were to KSs associated with the biosynthesis of phormidolide (n=288) and leptolyngbyalide (n=214), which are both produced by *trans*-AT PKSs found in Cyanobacteria (Fig. 4.3b).

The abyssal sediment KS sequences were also distinct from those observed in sequenced bacteria genomes, with most sharing <70% similarity with the NCBI RefSeq database (Fig. 4.S7). The most common taxonomic assignments were to Proteobacteria (27.3%), Cyanobacteria (26.0%), Actinobacteria (19.4%), and Bacillota (13.1%), although there were differences across KS subclasses. Notably, abyssal *trans*-AT (34.5%) and hybrid *cis*-AT (23.6%) KSs commonly matched to Bacillota, whereas modular *cis*-AT KSs rarely matched to this taxon (1.1%) (Fig. 4.S8). In contrast, 30.9% of modular *cis*-AT KS amplicons matched to Actinobacteria while relatively few hybrid *cis*-AT and *trans*-AT KS amplicons matched to Actinobacteria (7.0% and 12.4%) (Fig. 4.S8).

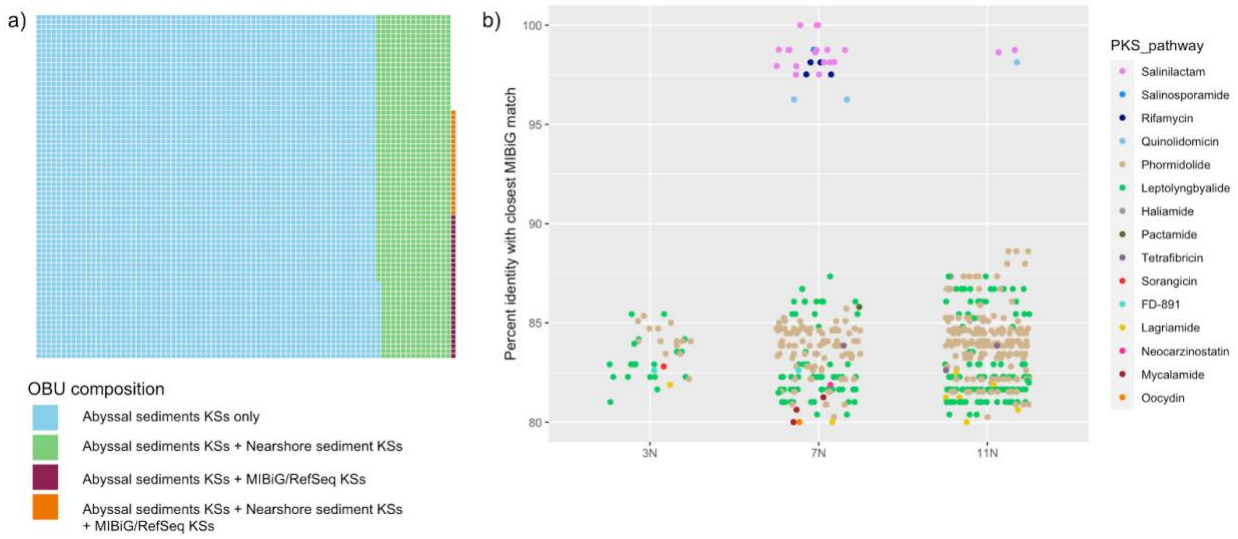

Figure 4.3. Sediment KS amplicon novelty. (a) Abyssal KS amplicon OBUs (80% clustering) were largely distinct from those observed in nearshore marine sediments and rarely clustered with MIBiG sequences (each dot represents an OBU). (b) MIBiG annotations for 550 OBUs (80% clustering) that contained abyssal sediment and MIBiG KSs. Sequences are grouped by sample location (x-axis) and colored by the MIBiG annotation.

Phylogenetic analysis of KS amplicons

Phylogenies were generated for the two largest KS subclasses detected in the abyssal sediment samples (modular *cis*-AT and hybrid *cis*-AT) to assess their evolutionary relationships to MIBiG and NCBI RefSeq sequences. The modular *cis*-AT phylogeny included 24,126 KS amplicons distributed among 509 OBUs. OBUs from the iterative polycyclic tetramate macrolactam (n=19 OBUs, 511 KSs), iterative aromatic (n=21 OBUs, 415 KSs), and olefin synthase (n=33 OBUs, 2,523 KSs) subclasses in the phylogeny (Fig. 4.4a) were added for completeness. Large clades that largely or entirely lacked MIBiG and RefSeq KSs were identified including clade 1, which included 2,374 KS amplicons across seven 80% sequence identity OBUs (Fig. 4.4a). Of these, 1,278 amplicons were classified as modular *cis*-at, 1096 as olefin synthase, and 22 as iterative aromatic. The closest MIBiG KSs are outside of this clade and belong to the dutomycin (BGC0001409) and bisenarsan (BGC0001283) pathways. All of the

5,295 KS sequences in clade 2 were classified as modular *cis*-AT and delineated into 112 OBUs that included 5 RefSeq KSs. Four MIBig KSs associated with the biosynthesis of psymberin (BGC0001110), chondrochloren (BGC0000970), guadinominic acid (BGC0000998), and cylindrospermopsin (BGC0000978) were observed in this clade. A third noteworthy clade contained 3,830 KS amplicons across 50 OBUs and no RefSeq or MIBiG KSs. The closest MIBiG KS was observed in the phenylnannolone (BGC0000122.5) pathway (Fig. 4.4a).

The hybrid *cis*-AT KS phylogeny (23,352 KS amplicons in 309 OBUs) revealed two large clades that lacked MIBiG representatives (Fig. 4.4b). Clade 4 was comprised of 9,468 KS amplicons across 123 OBUs and nine RefSeq KSs. The closest MIBiG matches were to sequences in the anachelin (BGC0002532) and DKxanthene (BGC0000986) pathways. Clade 5 was comprised of 2,977 KS amplicons across 22 OBUs. It contained no RefSeq KSs and the closest MIBiG KS was observed in the butyrolactol (BGC0001537) pathway. Together, these results provide evidence for large KS clades that have not been observed in experimentally characterized PKS pathways and are largely distinct from those observed in sequenced genomes.

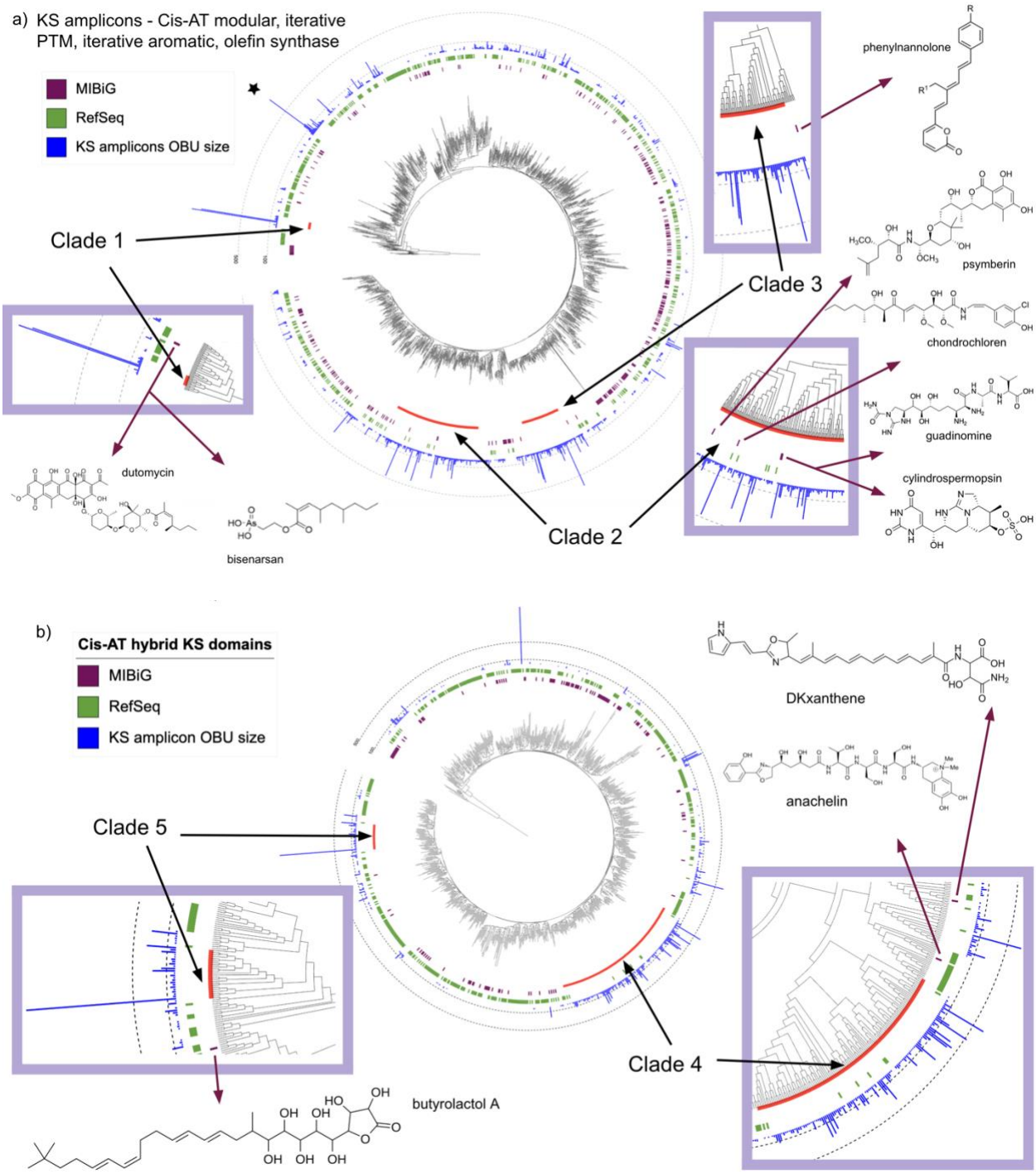

Figure 4.4. Abyssal sediment KS amplicon phylogenies. (a) FastME phylogeny of modular *cis*-AT (n=509), olefin synthase (n=33), iterative PTM (n=19), and iterative aromatic (n=18) KS amplicon OBUs (80% similarity). Included as reference are the closest RefSeq matches and MIBiG KSs. (b) FastME phylogeny of hybrid *cis*-AT KS amplicon OBUs (n=309), the closest RefSeq matches, and hybrid *cis*-AT MIBiG KSs. Large KS amplicon clades lacking MIBiG representatives are enlarged (purple boxes), and the structures associated with the closest MIBiG KS to each clade shown (gray). The star indicates the KS amplicon OBU that clustered with the KS domain extracted from the metagenomic data.

Metagenomic KS diversity

Metagenomics avoids PCR biases and thus provides an alternative approach to assess PKS diversity. Metagenomes were generated from the 0-1 cm horizons from all nine abyssal sediment cores (three replicates from each of the three sites) and polyketide diversity was evaluated using NaPDoS2. Despite the combined metagenome size of 790 Mbp, only 41 type I KS sequences were detected in the unassembled metagenomes, with the majority (78%) associated with polyunsaturated fatty acid pathways (PUFAs. Four of the KSs were classified as modular *cis*-AT and shared >90% similarity with a large OBU (947 KS sequences) identified in the amplicon analyses. In addition, six enediyne KSs were detected (Fig. 4.S9). The primers used in the amplicon study were not designed to detect the enediyne and PUFA KS classes, highlighting one of the drawbacks of this targeted approach. On the other hand, no hybrid *cis*-AT or *trans*-AT KS domains were recovered from the metagenomes despite the staggering diversity detected in the KS amplicon libraries. This emphasizes the value of amplicon-based approaches to assess biosynthetic diversity in marine sediments.

The metagenomic contigs were assembled into 25 high-quality and 22 medium-quality MAGs in an effort to resolve the genomic context of the KSs. Most MAGs were identified as Alphaproteobacteria (23.4%) and Archaea (21.2%) (Fig. 4.5a). Of note, the four metagenomic modular *cis*-AT BGCs (Fig. 4.S9) binned separately into four MAGs from the Gemmatimonadetes phylum (Fig. 4.5c), which is understudied for natural product discovery. To probe this further, the four KS sequences were included in a phylogeny with KS OBUs (80% similarity, n=93) extracted from publicly available Gemmatimonadetes genomes (n=240 JGI genomes, n=26 NCBI genomes). Here, the Gemmatimonadetes KSs are clearly distinct from all MIBiG KS, with the abyssal MAG KSs adding a new clade to this poorly studied group (Fig.

4.S10). T1PKSs comprised only 4.4% of the 922 metagenomic BGCs recovered using antiSMASH 6.0 (Fig. 4.5b), with the majority classified as terpenes (28.6%), RiPPs (23.1%), and NRPSs (17.0%).

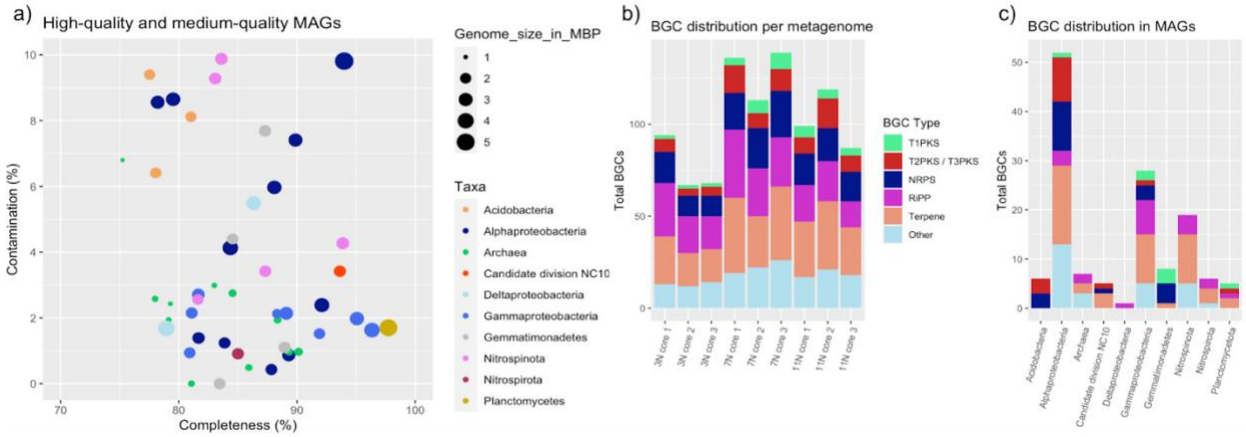

Figure 4.5. Abyssal MAGs and metagenomic biosynthetic diversity. (a) MAG dotplot showing taxonomy (color) and genome size (dot size) in the context of completeness (x-axis) and contamination (y-axis). (b) BGCs identified in nine abyssal sediment metagenomes. Colors indicate BGC type. (c) BGCs identified in 45 high and medium-quality MAGs grouped by taxonomy and colored by BGC type.

Abyssal sediment metabolomes

Given the high levels of sequence identity between some abyssal sediment KS amplicons and those detected in experimentally validated PKS pathways, I assessed if the products of those pathways could be detected in metabolomes generated from the sediments. Using untargeted liquid chromatography–high-resolution mass spectrometry (LC-HRMS), extracts were analyzed from 18 sediment samples (3 replicates from each of the 3 sites across 2 depth horizons). Although none of the molecular features aligned with the predicted BGC products, 746 unique molecular features were detected and the metabolomes largely separated between the 0-1 and 4-5 cm horizons (Fig. 4.S11) as was observed with the community analyses. Unlike the community (16S) and biosynthetic (KS) data sets, there was little site variation among samples, except for

two outliers from the 3°N site (Fig. 4.S11), which was the same site that held significantly different 16S and KS amplicon communities.

Previous metabolomes generated from nearshore sediments in Moorea (35) recovered 2164 unique molecular features across 36 samples (average = 60.1) compared to an average of 29.7 for the abyssal metabolomes. Despite displaying less complexity, only 47 molecular features (8.8%) revealed matching MS/MS spectra to the GNPS online platform compared to 17.9% for the nearshore Moorea sediments (35). Feature-based molecular networking was used to visualize the abyssal metabolomes in chemical space, revealing mostly singleton (58.0%) nodes (Fig. 4.S12). No molecular families were unique to the two outlier communities at the 3°N site despite their separation in the PCoA of molecular features (Fig. 4.S12).

4.5 Discussion

Greater than 80% of the ocean floor occurs at abyssal depths (>5000 m), yet little is known about the composition of the microbial communities associated with this vast biome and their natural product biosynthetic potential. This is due in part to the challenges associated with collecting samples at depth and the focus of most deep-sea research on anomalous bottom features such as hydrothermal vents. Here, I used multi-omic approaches to assess microbial community composition and natural product biosynthetic potential in sediments collected from three geographically remote abyssal plains and found them to be distinct from those detected in continental margin sediments. A closer examination of the abyssal sediment communities revealed significantly different taxonomic (16S) and biosynthetic (KS) diversity between the more equatorial 3°N site, which displayed higher DOC, NO_3^- + NO_2^-, and PO_4^{3-} concentrations and lower oxygen concentrations, and the 7°N and 11°N sites, which are less

biologically productive regions. This provides evidence that the geochemical properties associated with abyssal sediments affect community composition and biosynthetic potential as previously reported for shallow marine sediments (Hoshino et al., 2020) and soils (Charlop-Powers et al., 2014), respectively. These differences could point to environmental conditions shifting the composition of KS-containing microbes or, perhaps, selection for certain PKS pathways in distinct biogeochemical regimes.

From a microbial community perspective, 18% of the most abundant 16S ASVs fell within the Gammaproteobacterial order Woeseiales. This order is known to be widespread in both shallow and deep-sea sediments, comprising between 1-22% of the overall microbial community (Hoffman et al., 2020). Recent studies have reported Woeseiales MAGs and SAGs from deep-sea sediments, highlighting the adaptations of this taxonomic Order to grow on proteinaceous matter (Hoffman et al., 2020). Besides Gammaproteobacteria, three lineages that were also enriched in the abyssal sediments were Acidobacteria, Gemmatimonadetes, and Alphaproteobacteria. Notably, Acidobacteria (7.33%) and Gemmatimonadetes (4.85%) were less than 10% of the microbial communities in abyssal sediments, highlighting that rarer microbiome members often carry significant power when it comes to separating microbiomes (Jousset et al., 2017). This work also highlights that most of the abundant 16S ASVs shared <95% similarity with the closest RefSeq match, emphasizing the taxonomic uniqueness of the deep sea.

One of the challenges facing natural product discovery is the re-discovery of known compounds. However, mining both metagenomic and genomic data has highlighted that only a small fraction of nature's polyketide biosynthetic potential has been realized. In this study, I leverage the webtool NaPDoS2 to deeply assess polyketide diversity from KS amplicon libraries. The 6,000 KS OBUs (80% similarity) identified from 68,000 abyssal sediment KS amplicons are

over twice the number of KS OBUs in the MIBiG database. Note that this is a conservative estimate, as across the abyssal KS amplicons more than 52,000 amplicon sequence variants (ASVs) were recovered. Further, >99% of the abyssal KS amplicons fell into OBUs that did not contain KSs from either the MIBiG or NCBI RefSeq databases, indicating extensive and poorly explored polyketide biosynthetic diversity in abyssal sediments. Additionally, five large clades were found within the modular *cis*-AT and hybrid *cis*-AT phylogenies, with four of the five lacking representatives from the MIBiG database. While it has long been suggested that the deep sea provides unique opportunities for microbial natural product discovery (Cong et al., 2022), our results provide empirical evidence in support of this hypothesis.

Over 90% of the abyssal KS amplicons were from the modular *cis*-AT, hybrid *cis*-AT, and *trans*-AT classes, which aligns with my previous work extracting type I KSs from global metagenomic datasets (Singh et al., 2023). Based on my prior work (chapter 2), I recommended alterations to published KS primers to maximize the recovery of KS amplicons from the hybrid *cis*-AT and *trans*-AT classes. Using these updated primers, 38.5% and 8.6% of the abyssal KS amplicons were assigned to these two classes, respectively. Additionally, the percentage of abyssal hybrid *cis*-AT KS amplicons was consistently high compared to that seen in nearshore continental margin sediments (20.4%). Thus, the modifications to this type I KS primer set were successful in recovering amplicons from a range of KS subclasses, while also providing specificity to resolve biosynthetic differences across these diverse marine sediment biomes.

Clade 1 within the *cis*-AT KS phylogeny lacked MIBiG representatives, with the closest neighbors being KS domains from the dutomycin and bisenarsan BGCs. Dutomycin is a *Streptomyces* antibacterial compound (Sun et al., 2016) encoded by a rare pathway that includes both an iterative-acting type I PKS and anthracycline type II PKS system (Sun et al., 2016). The

bisenarsan BGC contains a single KS that appears to act iteratively based on the proposed chemical structure and biosynthesis (Cruz-Morales et al., 2016). Unexpectedly, the clade 1 amplicons include a mixture of modular *cis*-AT (53.8%), olefin synthase (46.1%), and iterative aromatic (0.1%), suggesting that KSs from this clade are not well resolved. As such, unraveling how KSs from this clade are functioning within their corresponding PKS BGCs could lead to the addition of new KS classes for the NaPDoS2 classification scheme. Another large clade of *cis*-AT KS amplicons (clade 3) was exclusive of MIBiG sequences. The closest neighbor is a KS from the phenylnannolone BGC, which is found in Nannocystis and exhibits inhibitory activity against P-glycoprotein (Bouhired et al., 2014). The phenylnannolone BGC is a modular *cis*-AT pathway that contains four KS domains, with the one in our phylogeny belonging to the first module. This pathway is rare in that it contains a butyryl-CoA carboxylase (BCC) that gives an ethylmalonyl-CoA precursor that interacts with the first KS module to give the rare ethyl-substituted moiety in phenylnannolone (Bouhired et al., 2014).

The largest *cis*-AT amplicon clade seen (clade 2 - 5,295 KSs) cladded with KS domains from the BGCs for psymberin, chondrochloren, guadinominic acid, and cylindrospermopsin. Psymberin is a noted cytotoxic compound that was isolated from symbionts within the sponge *Psammocinia* (Bielitza et al., 2013). The psymberin BGC is extremely unusual as it contains 10 *trans*-AT KS domains and ends with a modular *cis*-AT KS domain (Bielitza et al., 2013). Chondrochloren A and B are also unusual *Chondromyces*-produced compounds in that they contain a chloro-hydroxy-styryl group (Rachid et al., 2009). While this PKS BGC appears to be a modular *cis*-AT pathway, the KS domain in our amplicon clade is in the cndE module, which contains an O-MT (O-methyltransferase) that is hypothesized to act iteratively, raising the possibility that the fifth and final KS domain also acts iteratively (Rachid et al., 2009). Also of

note, the cndE module installs an ethoxy moiety in chondrochloren B (compared to methoxy in chondrochloren A), and ethoxy moieties, while common in plant metabolites, are rare in bacteria (Rachid et al., 2009). Finally, while the guadinominic acid and cylindrospermopsin pathways are from different phyla (Actinobacteria and Cyanobacteria, respectively) they both are unique to other modular *cis*-AT PKS pathways in that they incorporate guanidinoacetate as a starter unit (Holmes et al., 2012).

The two large hybrid *cis*-AT KS amplicon clades both lacked representatives from the MIBiG database, with the first (clade 4) grouping closest to hybrid *cis*-AT KS domains within the anachelin and DKxanthene pathways. These compounds play different roles - DKxanthene is a myxobacterial yellow pigment that is hypothesized to play a role in cell development (Meiser et al., 2008), while anachelin is a cyanobacterial siderophore (Calteau et al., 2014). Notably, these two pathways share an oxazole moiety, which is common in many bioactive marine peptides (Mhlongo et al., 2020). As expected, the hybrid *cis*-AT KS domains from both the anachelin and DKxanthene pathways catalyze the first extension of the growing polyketide chain immediately after the NRPS-installed oxazole moiety (Meiser et al., 2008; Calteau et al., 2014), suggesting that the abyssal hybrid *cis*-AT KS amplicons in this clade might share the same functionality. The other large hybrid *cis*-AT clade was closest to a KS domain from the *Streptomyces* butyrolactol pathway, which is intriguing in that it has a tert-butyl group (which has only otherwise been seen in Cyanobacteria) and that it contains eight consecutive hydroxy groups, one of which forms the concluding γ-lactone ring (Harunari et al., 2017). This pathway includes 11 KS domains with the first eight being classified by NaPDoS2 as modular *cis*-AT and the last three being classified as hybrid *cis*-AT. However, this pathway is highly unusual in that it does not contain any NRPS machinery (Harunari et al., 2017), making it unclear why these last

three KS domains in the pathway are classified as hybrid *cis*-AT. Regardless, this highlights the potentially rare functionality that exists within this abundant abyssal sediment amplicon clade.

Metagenomic sequencing offers distinct advantages over KS amplicon sequencing in that it is not subject to primer bias, complete BGCs can be evaluated, and more reliable taxonomic assignments can be made. Notably, T1PKSs made up only 4.4% of the 922 BGCs recovered from the abyssal sediment metagenomes, with the majority annotated as terpenes (28.6%) and RiPPs (23.1%). In comparison, a previous study of BGCs from aquatic MAGs found that less than 5% were assigned to RiPPs. While my BGC sample size is smaller, this could indicate that RiPPs play important ecological roles in abyssal sediments. Of the T1PKSs recovered from the metagenomes, the majority were PUFAs, which our primer set is not designed to target. In contrast, only four modular *cis*-AT KS domains (all clustering into 1 OBU) were detected, and no KS domains from the hybrid *cis*-AT or *trans*-AT KS subclasses. This emphasizes the value of amplicon sequencing in accessing relatively low abundance sequences. Interestingly, the one modular *cis*-AT KS OBU recovered was seen in four separate Gemmatimonadetes MAGs, which is an understudied phylum that only recently has been shown to contain genes for polyketide biosynthesis (Crits-Cristoph et al., 2018). However, this has only been seen with culture-independent work, as to date there has not been a single experimentally characterized BGC from the Gemmatimonadetes phylum (Kautsar et al., 2020).

To date, marine polyketide natural products have been isolated from a variety of sources, including microbiomes from sponges, corals, and shallow continental margin sediments. However, almost all marine natural product discovery efforts have concentrated in shallow biomes, in part because they are more accessible to sample than the deep sea. As abyssal plains continue to gain traction as sites for deep-sea mining operations, it is crucial to understand the

biosynthetic diversity that could potentially be lost if they are exploited. Thus, our finding here of abyssal sediments carrying distinct and novel KS communities serves as a critical baseline and underscores the immense biosynthetic potential that awaits in abyssal sediments.

4.6 Conclusion

Abyssal sediments comprise 80% of the ocean floor and the associated microbiomes provide promising opportunities for natural product discovery. Here, I show that abyssal sediment polyketide biosynthetic potential reflects biogeochemical differences across three sites. Further, the vast majority of abyssal KS amplicon OBUs recovered were distinct from those seen in experimentally characterized pathways and nearshore continental margin sediments. This work highlights the significant biosynthetic potential of deep-sea sediment microbiomes.

4.7 Funding sources

This research was supported by the National Science Foundation Graduate Research Fellowship Program (grant no. DGE-2038238 to H.W.S.). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

4.8 Acknowledgements

Chapter 4, in full, is not a reprint of any materials that have been submitted for publication. The dissertation author was the main contributor to this work.

4.9 Supplementary Figures and Tables

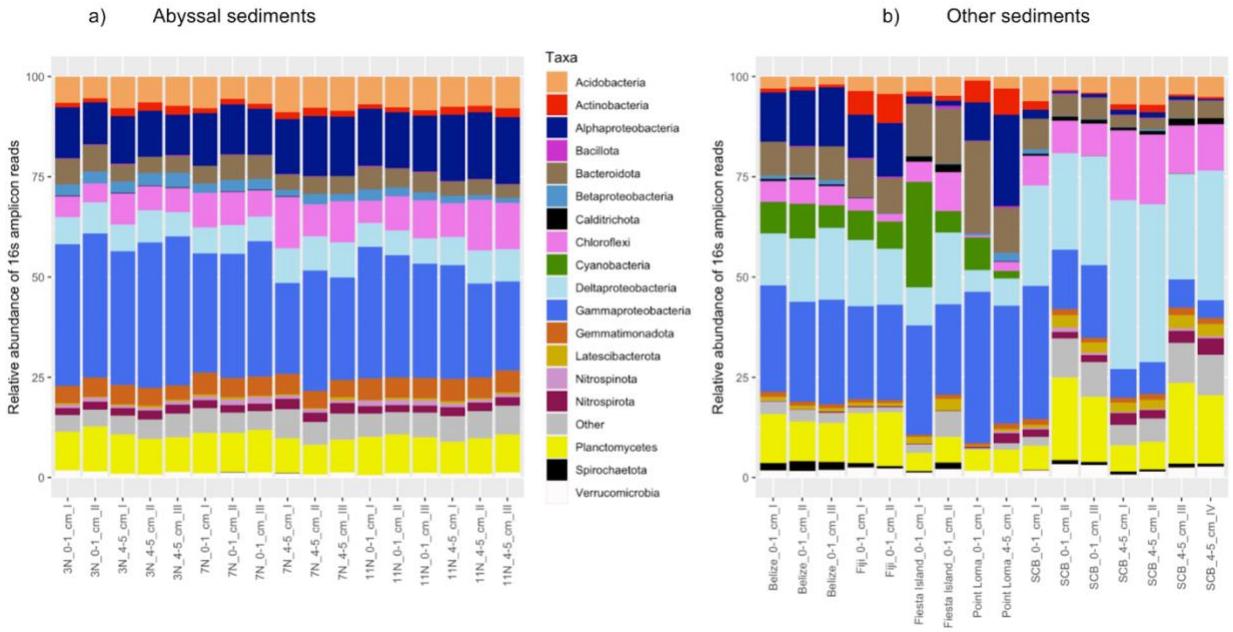

Figure 4.S1. Taxonomy (16S) of marine sediment amplicons. (a) abyssal (b) and continental margin sediments based on top SILVA database matches.

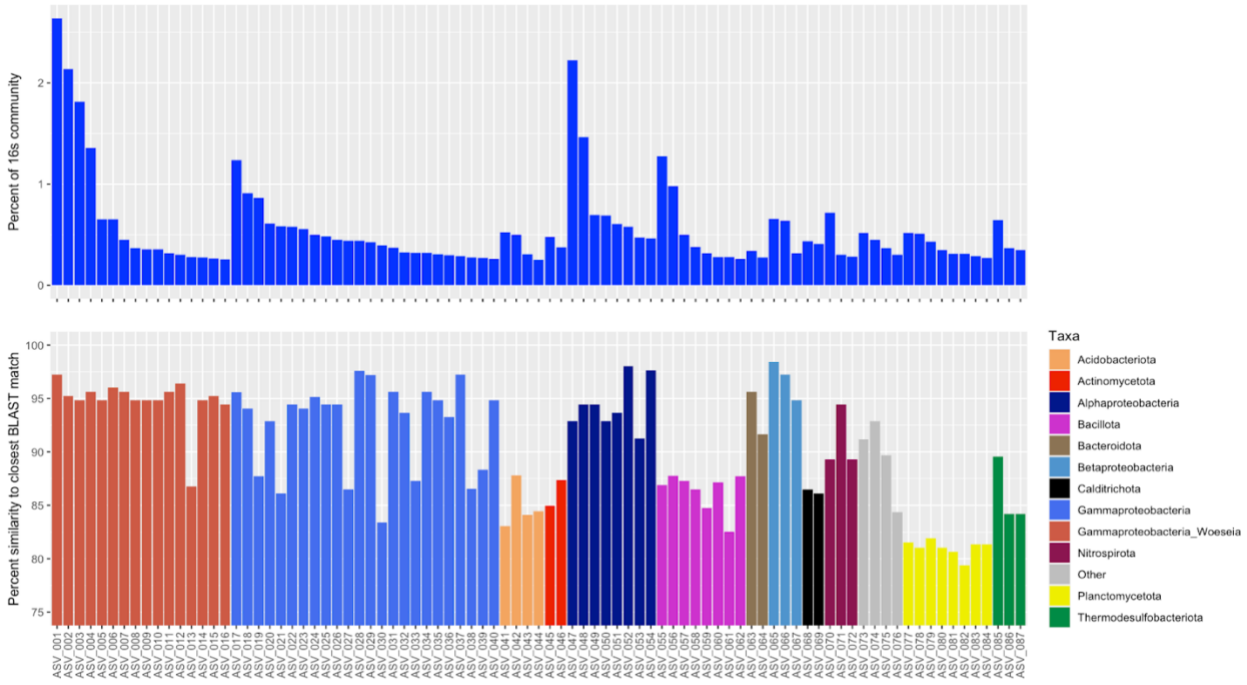

Figure 4.S2. Abundance and taxonomy of 16S rRNA ASVs. Most common (>0.25% relative abundance, upper) 16S ASVs across abyssal sediments ordered by taxa (lower). Percent identity of closest BLAST match (y-axis) to ASV in upper panel colored by phylum (lower).

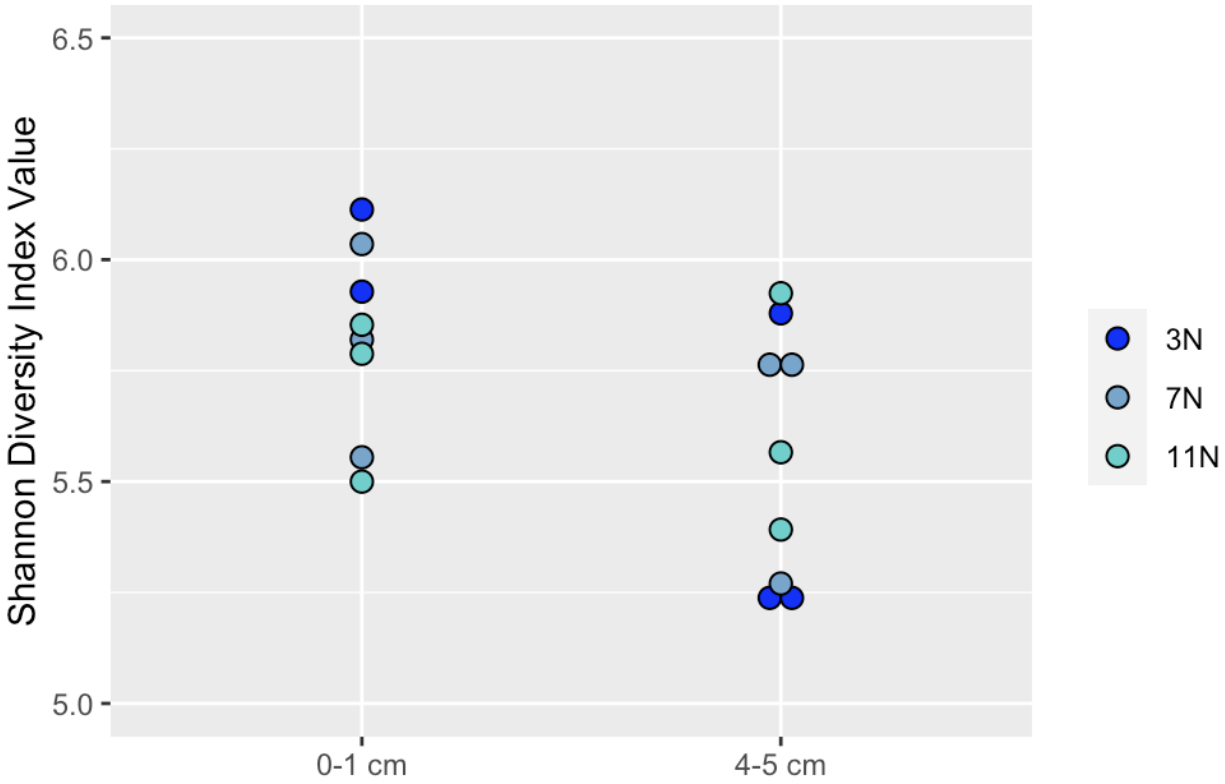

Figure 4.S3. Diversity of abyssal sediment 16S amplicons. Shannon Diversity Index values for 16S ASVs across the 0-1 cm and 4-5 cm abyssal sediment horizons. Values calculated for each sample colored by location.

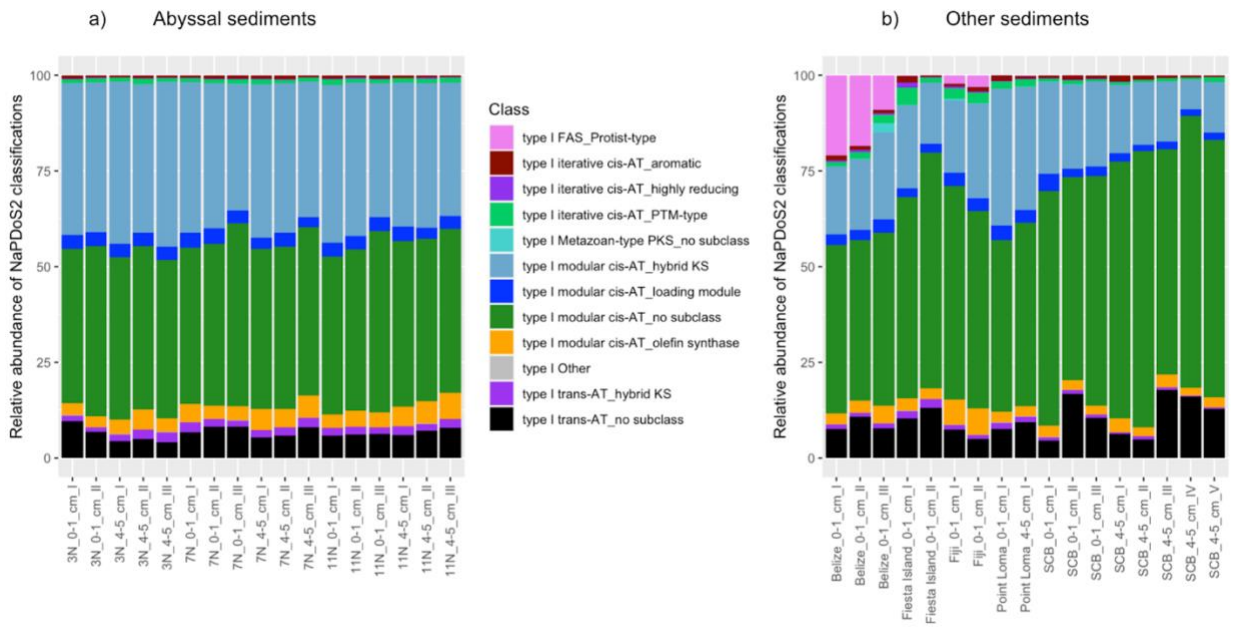

Figure 4.S4. NaPDoS2 classification of KS amplicons. (a) abyssal and (b) continental margin sediments.

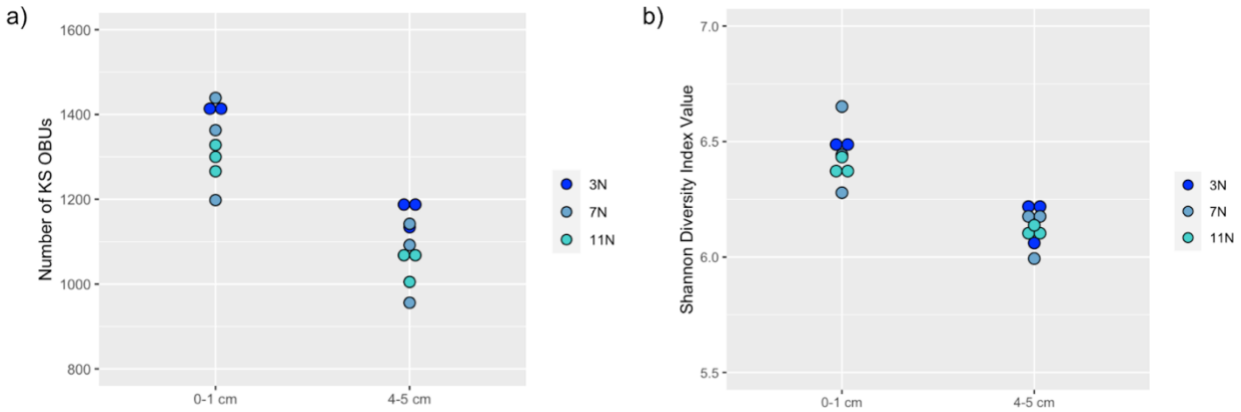

Figure 4.S5. Diversity of abyssal sediment KS amplicons (a) KS OBU totals (80% sequence identity) and (b) Shannon's diversity index for KSs per sediment core across the 0-1 cm and 4-5 cm abyssal sediment horizons, with datapoints colored by location.

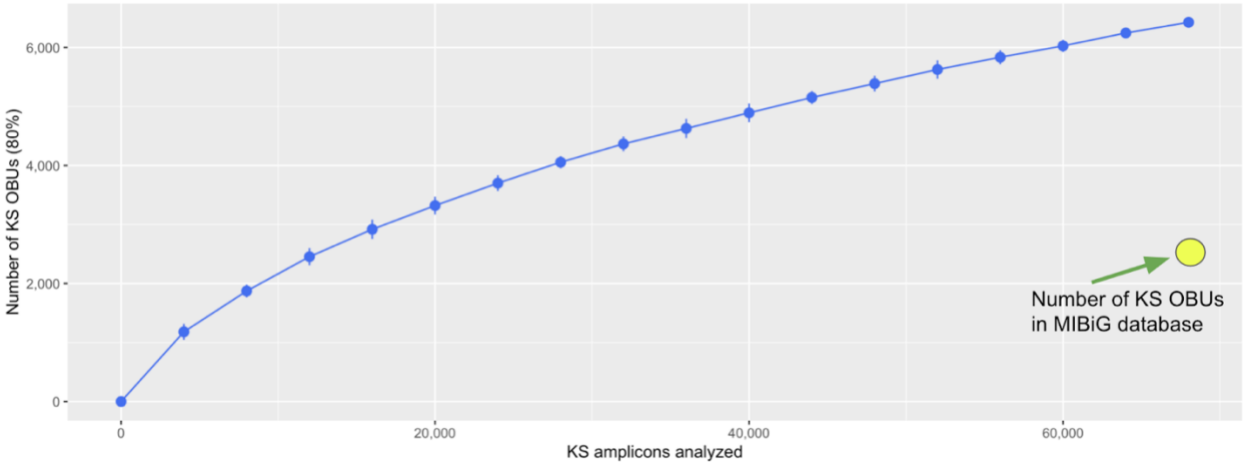

Figure 4.S6. Rarefaction curve for abyssal sediment KS OBUs (80% sequence identity). Yellow dot indicates the total number of similarly clustered KS OBUs in the MIBiG database.

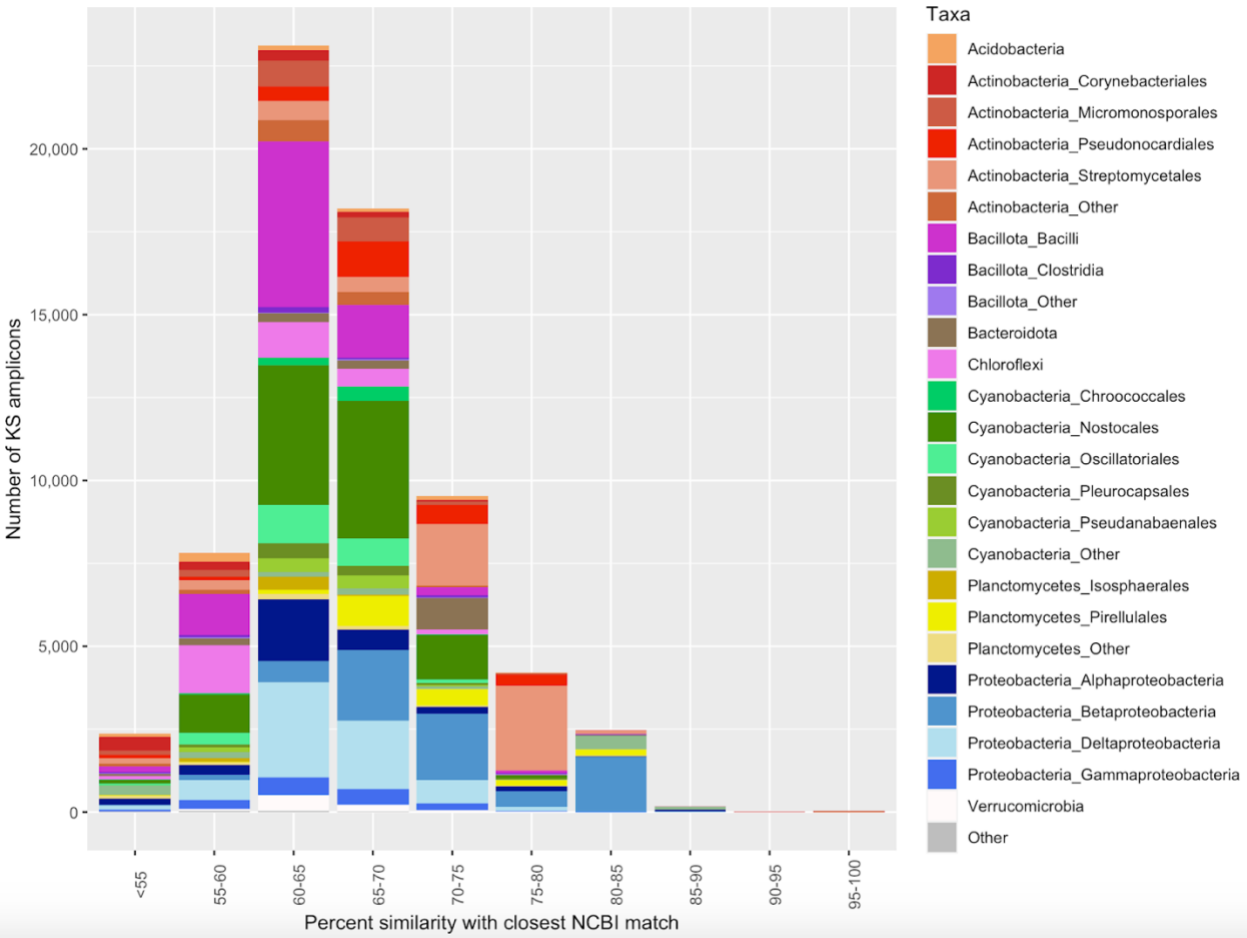

Figure 4.S7. NCBI RefSeq matches for abyssal KS amplicons. The percent similarity to the NCBI RefSeq match is shown on the x-axis. Colors indicate the taxonomy of the NCBI RefSeq match.

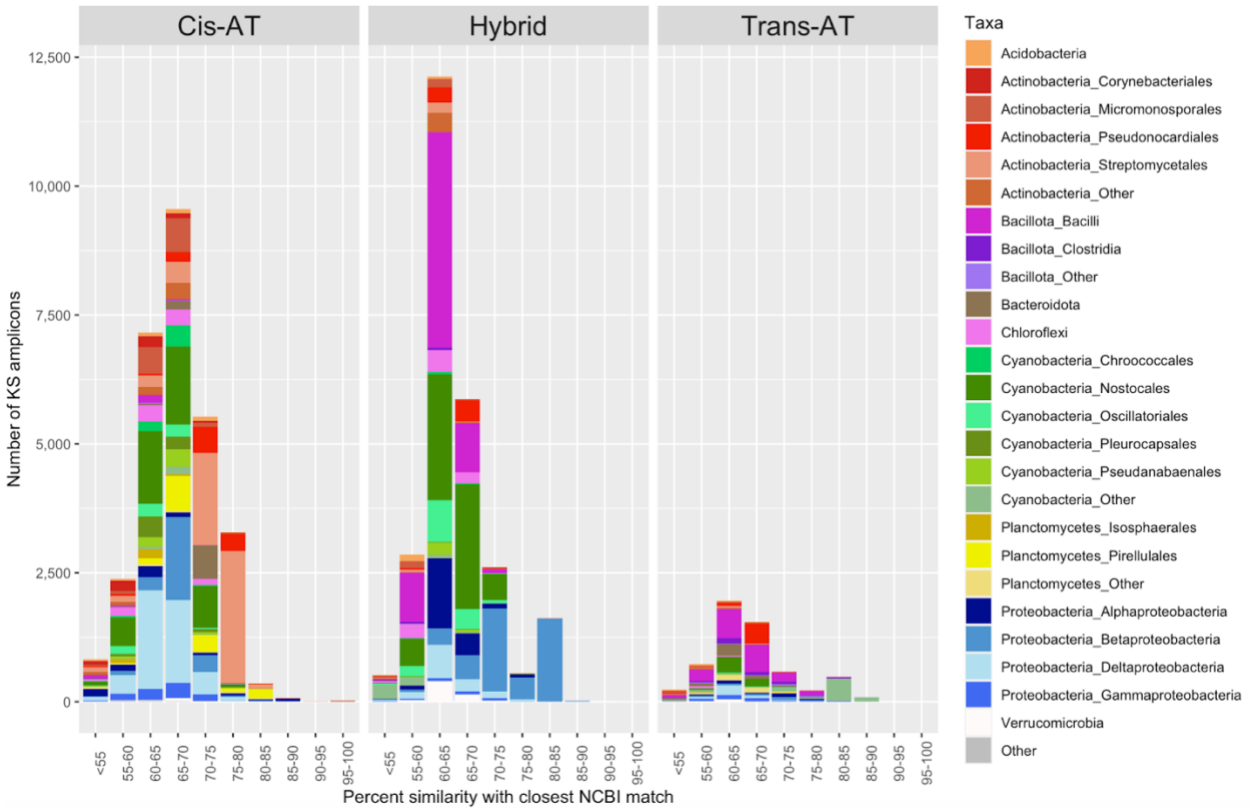

Figure 4.S8. NCBI RefSeq matches for abyssal KS amplicons split by KS subclass. Taxonomy of NCBI RefSeq matches to the three most common abyssal KS amplicon classes (modular *cis*-AT, hybrid *cis*-AT, *trans*-AT). Percent similarity to the RefSeq match is shown on the x-axis. Colors indicate the taxonomy of the closest NCBI RefSeq match.

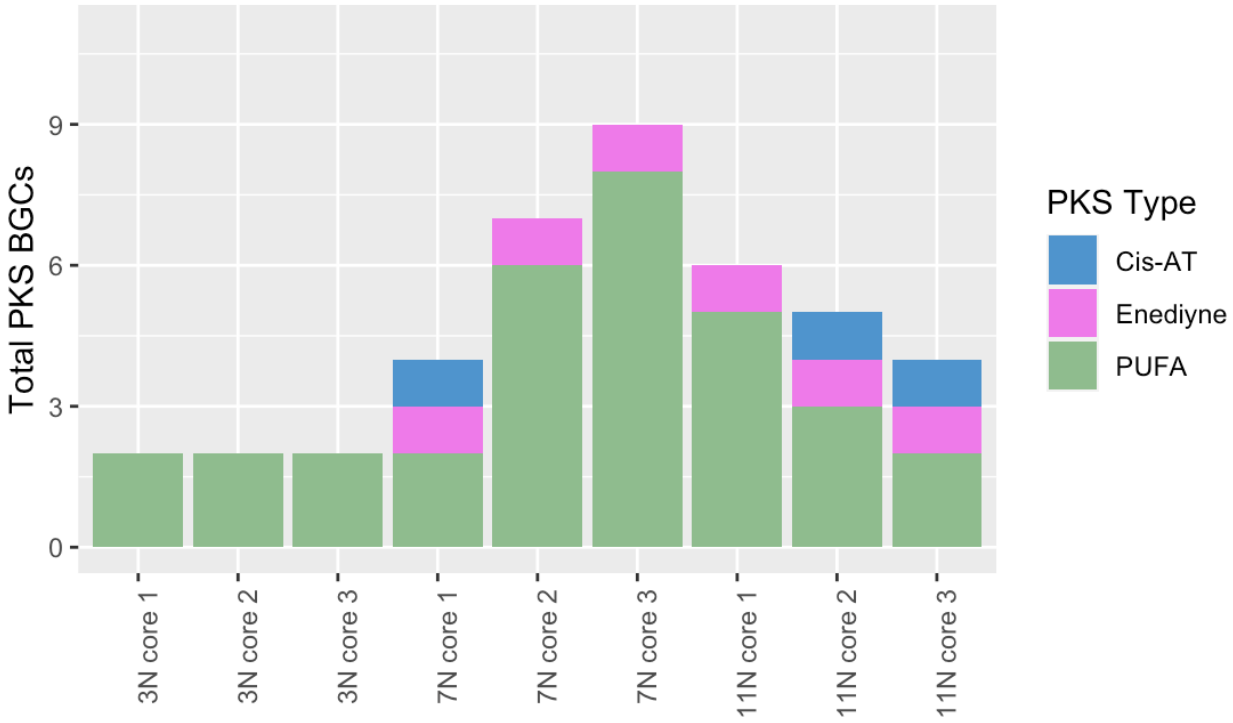

Figure 4.S9. T1PKS BGCs recovered from metagenomes across three abyssal sediment sites (3N, 7N, and 11N). Three metagenomes (three replicates) were generated from each of the three sites.

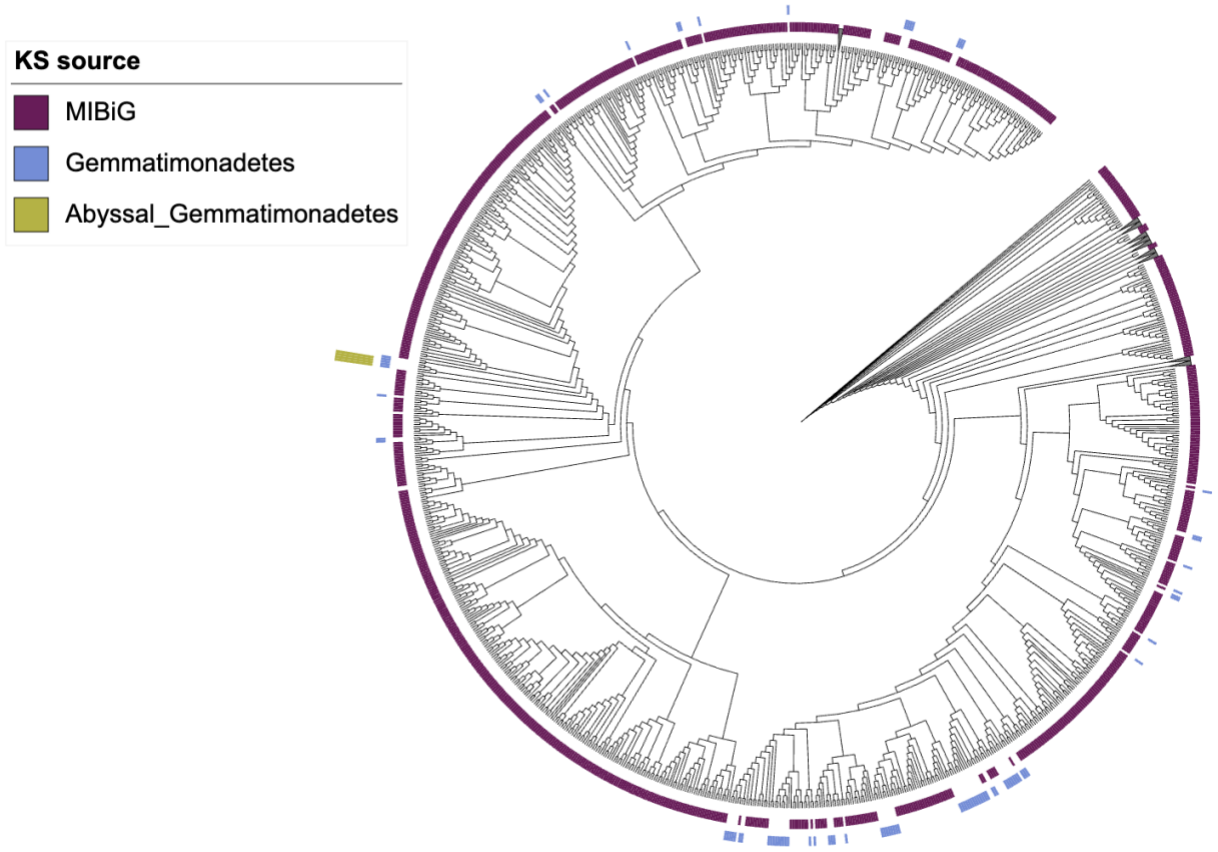

Figure 4.S10. Phylogeny of Gemmatimonadetes KS sequences. The FastME phylogeny includes abyssal metagenomic KSs within MAGs (n=4, yellow) and KS domain OBUs (80%) extracted from publicly available Gemmatimonadetes genomes (n=93, blue) and the MIBiG database (maroon).

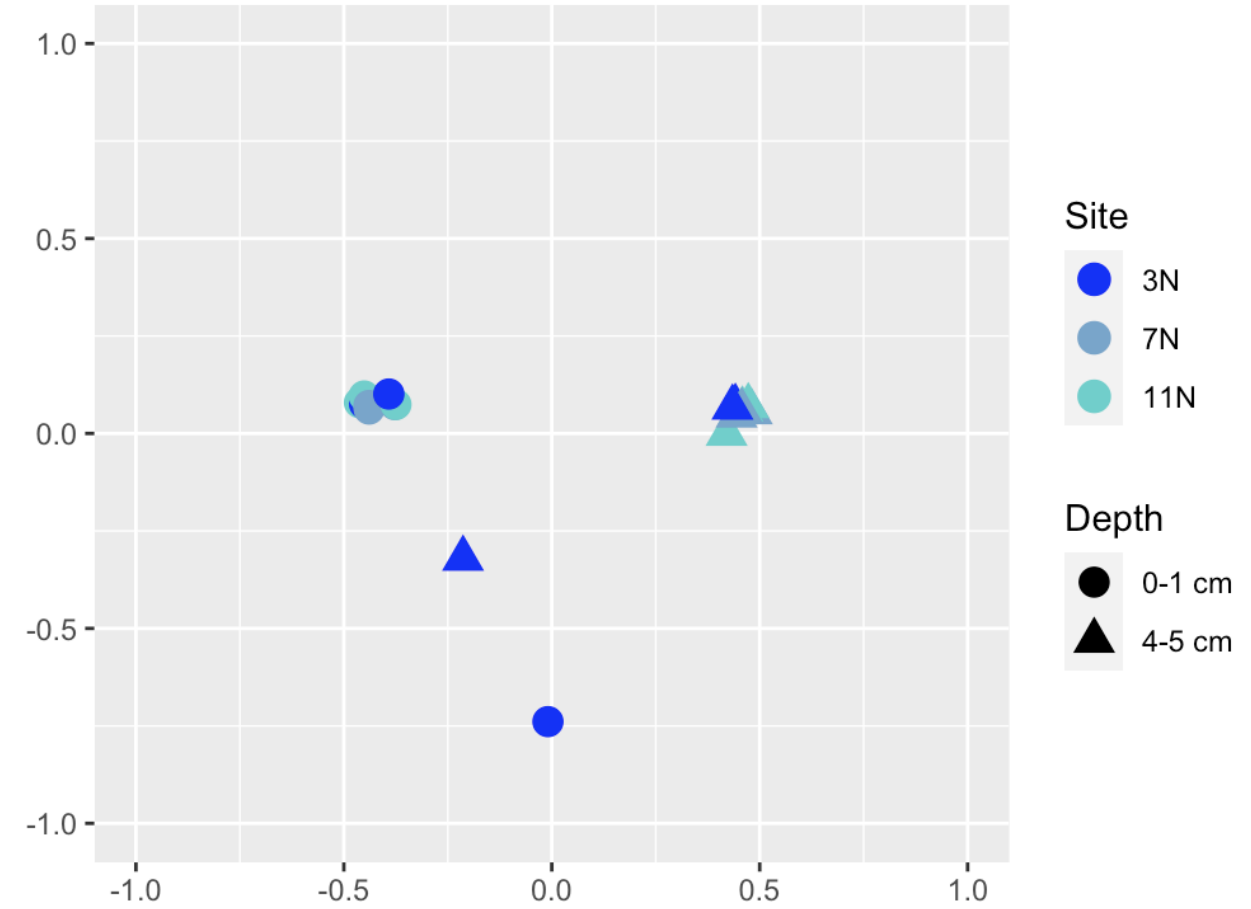

Figure 4.S11. PCoA using molecular features from abyssal sediments. Samples are delineated by site and the 0-1 cm or 4-5 cm sediment horizons.

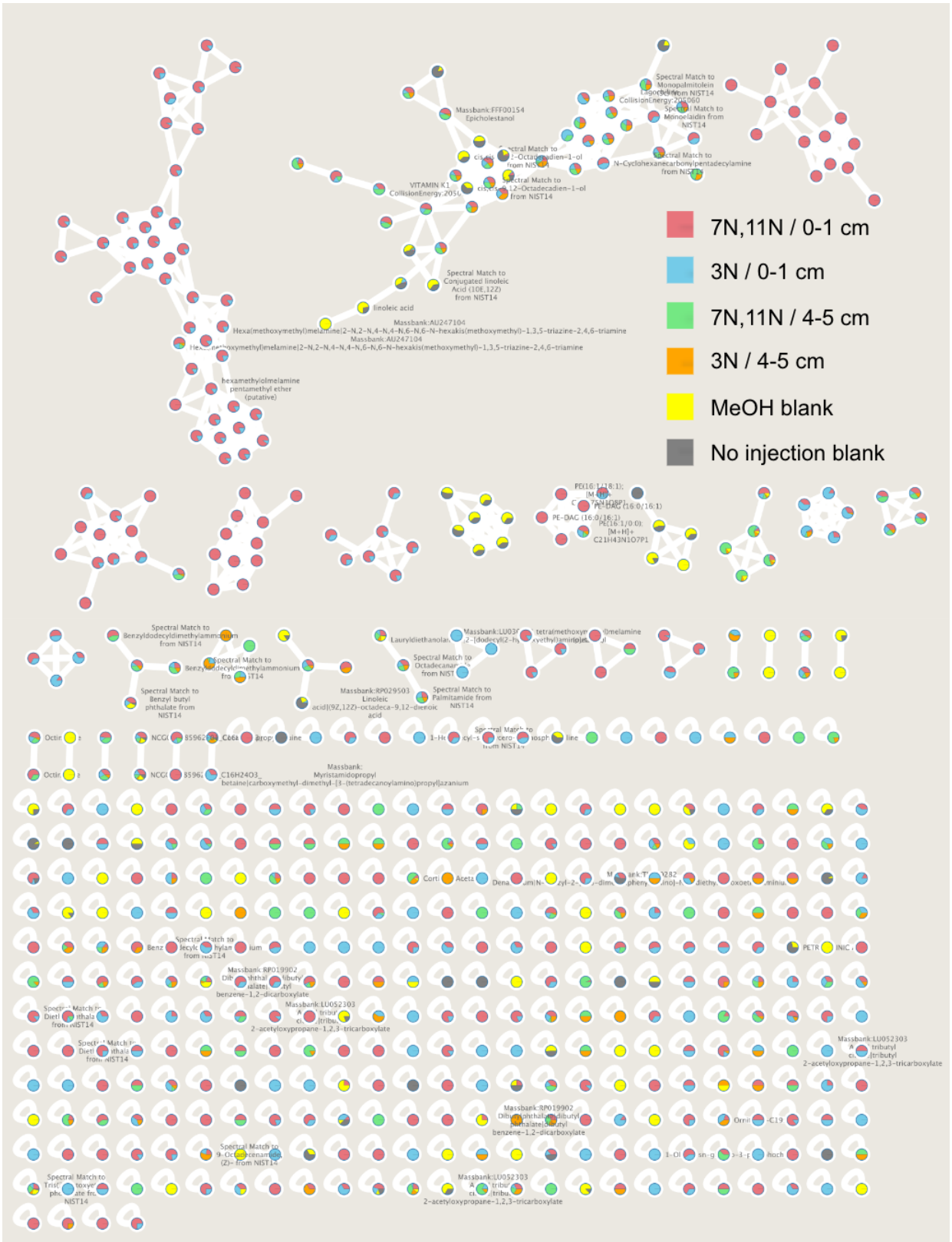

Figure 4.S12. Molecular network using features from abyssal sediments. Samples are delineated by site and the 0-1 cm or 4-5 cm sediment horizons. Text indicates molecular features annotated by GNPS.

4.10 References

Abdel-Razek AS, El-Naggar ME, Allam A, Morsy OM, Othman SI. 2020. Microbial Natural Products in Drug Discovery. Processes 8:470.

Amos GCA, Awakawa T, Tuttle RN, Letzel AC, Kim MC, Kudo Y, Fenical W, S. Moore B, Jensen PR. 2017. Comparative transcriptomics as a guide to natural product discovery and biosynthetic gene cluster functionality. Proc Natl Acad Sci USA 114.

Aron AT, Gentry EC, McPhail KL, Nothias L-F, Nothias-Esposito M, Bouslimani A, Petras D, Gauglitz JM, Sikora N, Vargas F, Van Der Hooft JJJ, Ernst M, Kang KB, Aceves CM, Caraballo-Rodríguez AM, Koester I, Weldon KC, Bertrand S, Roullier C, Sun K, Tehan RM, Boya P. CA, Christian MH, Gutiérrez M, Ulloa AM, Tejeda Mora JA, Mojica-Flores R, Lakey-Beitia J, Vásquez-Chaves V, Zhang Y, Calderón AI, Tayler N, Keyzers RA, Tugizimana F, Ndlovu N, Aksenov AA, Jarmusch AK, Schmid R, Truman AW, Bandeira N, Wang M, Dorrestein PC. 2020. Reproducible molecular networking of untargeted mass spectrometry data using GNPS. Nat Protoc 15:1954–1991.

Ayuso-Sacido A, Genilloud O. 2005. New PCR primers for the screening of NRPS and PKS-I systems in actinomycetes: detection and distribution of these biosynthetic gene sequences in major taxonomic groups. Microb Ecol 49:10–24.

Baker BJ, Appler KE, Gong X. 2021. New Microbial Biodiversity in Marine Sediments. Annu Rev Mar Sci 13:161–175.

Bech PK, Lysdal KL, Gram L, Bentzon-Tilia M, Strube ML. 2020. Marine Sediments Hold an Untapped Potential for Novel Taxonomic and Bioactive Bacterial Diversity. mSystems 5:e00782-20.

Bielitza M, Pietruszka J. 2013. The Psymberin Story—Biological Properties and Approaches towards Total and Analogue Syntheses. Angew Chem Int Ed 52:10960–10985.

Blin K, Shaw S, Kloosterman AM, Charlop-Powers Z, van Wezel GP, Medema MH, Weber T. 2021. AntiSMASH 6.0: improving cluster detection and comparison capabilities. Nucleic Acids Res 49:W29–W35.

Bogdanov A, Salib MN, Chase AB, Hammerlindl H, Muskat MN, Luedtke S, Da Silva EB, O'Donoghue AJ, Wu LF, Altschuler SJ, Molinski TF, Jensen PR. 2024. Small molecule in situ resin capture provides a compound first approach to natural product discovery. Nat Commun 15:5230.

Bouhired SM, Crüsemann M, Almeida C, Weber T, Piel J, Schäberle TF, König GM. 2014. Biosynthesis of Phenylnannolone A, a Multidrug Resistance Reversal Agent from the Halotolerant Myxobacterium Nannocystis pusilla B150. ChemBioChem 15:757–765.

Bushnell, B. 2014. BBMap: A Fast, Accurate, Splice-Aware Aligner. Lawrence Berkeley National Laboratory. LBNL Report #: LBNL-7065E. Retrieved from https://escholarship.org/uc/item/1h3515gn

Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. 2016. DADA2: High-resolution sample inference from Illumina amplicon data. Nat Methods 13:581–583.

Calteau A, Fewer DP, Latifi A, Coursin T, Laurent T, Jokela J, Kerfeld CA, Sivonen K, Piel J, Gugger M. 2014. Phylum-wide comparative genomics unravel the diversity of secondary metabolism in Cyanobacteria. BMC Genomics 15:977.

Caporaso J, Ackermann G, Apprill A, Bauer M, Berg-Lyons D, Betley J, Fierer N, Fraser L, A. Fuhrman J, A. Gilbert J, Gormley N, Humphrey G, Huntley J, K. Jansson J, Knight R, L. Lauber C, A. Lozupone C, McNally S, M. Needham D, M. Owens S, E. Parada A, Parsons R, Smith G, R. Thompson L, Thompson L, J. Turnbaugh P, A. Walters W, Weber L. 2018. EMP 16S Illumina Amplicon Protocol v1.

Cappello E, Nieri P. 2021. From Life in the Sea to the Clinic: The Marine Drugs Approved and under Clinical Trial. Life 11:1390.

Carlin DE, Demchak B, Pratt D, Sage E, Ideker T. 2017. Network propagation in the cytoscape cyberinfrastructure. PLoS Comput Biol 13:e1005598.

Charlop-Powers Z, Owen JG, Reddy BVB, Ternei MA, Brady SF. 2014. Chemical-biogeographic survey of secondary metabolism in soil. Proc Natl Acad Sci U S A 111:3757–3762.

Charlop-Powers Z, Owen JG, Reddy BVB, Ternei MA, Guimarães DO, de Frias UA, Pupo MT, Seepe P, Feng Z, Brady SF. 2015. Global biogeographic sampling of bacterial secondary metabolism. Elife 4:e05048.

Chase AB, Bogdanov A, Demko AM, Jensen PR. 2023. Biogeographic patterns of biosynthetic potential and specialized metabolites in marine sediments. The ISME Journal 17:976–983.

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

Cong M, Pang X, Zhao K, Song Y, Liu Y, Wang J. 2022. Deep-Sea Natural Products from Extreme Environments: Cold Seeps and Hydrothermal Vents. Marine Drugs 20:404.

Crits-Christoph A, Diamond S, Butterfield CN, Thomas BC, Banfield JF. 2018. Novel soil bacteria possess diverse genes for secondary metabolite biosynthesis. Nature 558:440–444.

Cruz-Morales P, Kopp JF, Martínez-Guerrero C, Yáñez-Guerra LA, Selem-Mojica N, Ramos-Aboites H, Feldmann J, Barona-Gómez F. 2016. Phylogenomic Analysis of Natural Products Biosynthetic Gene Clusters Allows Discovery of Arseno-Organic Metabolites in Model Streptomycetes. Genome Biol Evol 8:1906–1916.

Demko AM, Patin NV, Jensen PR. 2021. Microbial diversity in tropical marine sediments assessed using culture-dependent and culture-independent techniques. Environmental Microbiology 23:6859–6875.

D'Hondt S, Inagaki F, Zarikian CA, Abrams LJ, Dubois N, Engelhardt T, Evans H, Ferdelman T, Gribsholt B, Harris RN, Hoppie BW, Hyun J-H, Kallmeyer J, Kim J, Lynch JE, McKinley CC, Mitsunobu S, Morono Y, Murray RW, Pockalny R, Sauvage J, Shimono T, Shiraishi F, Smith DC, Smith-Duque CE, Spivack AJ, Steinsbu BO, Suzuki Y, Szpak M, Toffin L, Uramoto G, Yamaguchi YT, Zhang G, Zhang X-H, Ziebis W. 2015. Presence of oxygen and aerobic communities from sea floor to basement in deep-sea sediments. Nature Geosci 8:299–304.

Dai M, Luo Y, Achterberg EP, Browning TJ, Cai Y, Cao Z, Chai F, Chen B, Church MJ, Ci D, Du C, Gao K, Guo X, Hu Z, Kao S, Laws EA, Lee Z, Lin H, Liu Q, Liu X, Luo W, Meng F, Shang S, Shi D, Saito H, Song L, Wan XS, Wang Y, Wang W, Wen Z, Xiu P, Zhang J, Zhang R, Zhou K. 2023. Upper Ocean Biogeochemistry of the Oligotrophic North Pacific Subtropical Gyre: From Nutrient Sources to Carbon Export. Reviews of Geophysics 61:3.

Estaki M, Jiang L, Bokulich NA, McDonald D, González A, Kosciolek T, Martino C, Zhu Q, Birmingham A, Vázquez-Baeza Y, Dillon MR, Bolyen E, Caporaso JG, Knight R. 2020. QIIME 2 Enables Comprehensive End-to-End Analysis of Diverse Microbiome Data and Comparative Studies with Publicly Available Data. CP in Bioinformatics 70:e100.

Fischbach MA, Walsh CT. 2006. Assembly-line enzymology for polyketide and nonribosomal peptide antibiotics: logic, machinery, and mechanisms. Chem Rev 106:3468–3496.

Gao Y, Zhao Y, He X, Deng Z, Jiang M. 2021. Challenges of functional expression of complex polyketide biosynthetic gene clusters. Current Opinion in Biotechnology 69:103–111.

Guraieb M, Mendoza G, Mizell K, Rouse G, McCarthy RA, Pereira OS, Levin LA. 2024. Deep-ocean macrofaunal assemblages on ferromanganese and phosphorite-rich substrates in the Southern California Borderland. PeerJ 12:e18290.

Harunari E, Komaki H, Igarashi Y. 2017. Biosynthetic origin of butyrolactol A, an antifungal polyketide produced by a marine-derived Streptomyces. Beilstein J Org Chem 13:441–450.

Hoffmann K, Bienhold C, Buttigieg PL, Knittel K, Laso-Pérez R, Rapp JZ, Boetius A, Offre P. 2020. Diversity and metabolism of Woeseiales bacteria, global members of marine sediment communities. The ISME Journal 14:1042–1056.

Holmes TC, May AE, Zaleta-Rivera K, Ruby JG, Skewes-Cox P, Fischbach MA, DeRisi JL, Iwatsuki M, Ōmura S, Khosla C. 2012. Molecular Insights into the Biosynthesis of Guadinomine: A Type III Secretion System Inhibitor. J Am Chem Soc 134:17797–17806.

Hoshino T, Doi H, Uramoto G-I, Wörmer L, Adhikari RR, Xiao N, Morono Y, D'Hondt S, Hinrichs K-U, Inagaki F. 2020. Global diversity of microbial communities in marine sediment. Proc Natl Acad Sci USA 117:27587–27597.

Jensen PR, Moore BS, Fenical W. 2015. The marine actinomycete genus *Salinispora*: a model organism for secondary metabolite discovery. Nat Prod Rep 32:738–751.

Jousset A, Bienhold C, Chatzinotas A, Gallien L, Gobet A, Kurm V, Küsel K, Rillig MC, Rivett DW, Salles JF, Van Der Heijden MGA, Youssef NH, Zhang X, Wei Z, Hol WHG. 2017. Where less may be more: how the rare biosphere pulls ecosystems strings. The ISME Journal 11:853–862.

Kallmeyer J, Pockalny R, Adhikari RR, Smith DC, D'Hondt S. 2012. Global distribution of microbial abundance and biomass in subseafloor sediment. Proc Natl Acad Sci USA 109:16213–16216.

Kautsar SA, Blin K, Shaw S, Navarro-Muñoz JC, Terlouw BR, van der Hooft JJJ, van Santen JA, Tracanna V, Suarez Duran HG, Pascal Andreu V, Selem-Mojica N, Alanjary M, Robinson SL, Lund G, Epstein SC, Sisto AC, Charkoudian LK, Collemare J, Linington RG, Weber T, Medema MH. 2020. MIBiG 2.0: a repository for biosynthetic gene clusters of known function. Nucleic Acids Res 48:D454–D458.

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

Klau LJ, Podell S, Creamer KE, Demko AM, Singh HW, Allen EE, Moore BS, Ziemert N, Letzel AC, Jensen PR. 2022. The natural product domain seeker version 2 (NaPDoS2) webtool relates ketosynthase phylogeny to biosynthetic function. J Biol Chem 298:102480.

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

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

Lewis K. 2020. At the Crossroads of Bioenergetics and Antibiotic Discovery. Biochemistry Moscow 85:1469–1483. Ziemert N, Alanjary M, Weber T. 2016. The evolution of genome mining in microbes-a review. Nat Prod Rep 33:988–1005.

Libis V, Antonovsky N, Zhang M, Shang Z, Montiel D, Maniko J, Ternei MA, Calle PY, Lemetre C, Owen JG, Brady SF. 2019. Uncovering the biosynthetic potential of rare metagenomic DNA using co-occurrence network analysis of targeted sequences. Nat Commun 10:3848.

Lloyd KG, Steen AD, Ladau J, Yin J, Crosby L. 2018. Phylogenetically Novel Uncultured Microbial Cells Dominate Earth Microbiomes. mSystems 3:10.1128/msystems.00055-18.

Meiser P, Weissman KJ, Bode HB, Krug D, Dickschat JS, Sandmann A, Müller R. 2008. DKxanthene Biosynthesis—Understanding the Basis for Diversity-Oriented Synthesis in Myxobacterial Secondary Metabolism. Chemistry & Biology 15:771–781.

Mhlongo JT, Brasil E, De La Torre BG, Albericio F. 2020. Naturally Occurring Oxazole-Containing Peptides. Marine Drugs 18:203.

Miyanaga A, Kudo F, Eguchi T. 2018. Protein-protein interactions in polyketide synthase-nonribosomal peptide synthetase hybrid assembly lines. Nat Prod Rep 35:1185–1209.

Moffitt MC, Neilan BA. 2003. Evolutionary affiliations within the superfamily of ketosynthases reflect complex pathway associations. J Mol Evol 56:446–457.

Nivina A, Yuet KP, Hsu J, Khosla C. 2019. Evolution and Diversity of Assembly-Line Polyketide Synthases: Focus Review. Chem Rev 119:12524–12547.

Orcutt BN, Sylvan JB, Knab NJ, Edwards KJ. 2011. Microbial Ecology of the Dark Ocean above, at, and below the Seafloor. Microbiol Mol Biol Rev 75:361–422.

Orsi WD. 2018. Ecology and evolution of seafloor and subseafloor microbial communities. Nat Rev Microbiol 16:671–683.

Patin NV, Kunin V, Lidström U, Ashby MN. 2013. Effects of OTU Clustering and PCR Artifacts on Microbial Diversity Estimates. Microb Ecol 65:709–719.

Peng Y, Leung HCM, Yiu SM, Chin FYL. 2011. Meta-IDBA: a de Novo assembler for metagenomic data. Bioinformatics 27:i94–i101.

Piel J. 2010. Biosynthesis of polyketides by *trans*-at polyketide synthases. Nat Prod Rep 27:996–1047.

Price MN, Dehal PS, Arkin AP. 2010. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. PLoS ONE 5:e9490.

Pye CR, Bertin MJ, Lokey RS, Gerwick WH, Linington RG. 2017. Retrospective analysis of natural products provides insights for future discovery trends. Proc Natl Acad Sci USA 114:5601–5606.

Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO. 2012. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Research 41:D590–D596.

Rachid S, Scharfe M, Blöcker H, Weissman KJ, Müller R. 2009. Unusual Chemistry in the Biosynthesis of the Antibiotic Chondrochlorens. Chemistry & Biology 16:70–81.

Rascher A, Hu Z, Viswanathan N, Schirmer A, Reid R, Nierman WC, Lewis M, Hutchinson CR. 2003. Cloning and characterization of a gene cluster for geldanamycin production in Streptomyces hygroscopicus NRRL 3602. FEMS Microbiol Lett 218:223–230.

Seemann T. 2014. Prokka: rapid prokaryotic genome annotation. Bioinformatics 30:2068–2069.

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

Singh HW, Creamer KE, Chase AB, Klau LJ, Podell S, Jensen PR. 2023. Metagenomic data reveals type I polyketide synthase distributions across biomes. mSystems 8:e00012-23.

Stratmann T, Soetaert K, Kersken D, Van Oevelen D. 2021. Polymetallic nodules are essential for food-web integrity of a prospective deep-seabed mining area in Pacific abyssal plains. Sci Rep 11:12238.

Sun L, Wang S, Zhang S, Shao L, Zhang Q, Skidmore C, Chang C-WT, Yu D, Zhan J. 2016. Characterization of Three Tailoring Enzymes in Dutomycin Biosynthesis and Generation of a Potent Antibacterial Analogue. ACS Chem Biol 11:1992–2001.

Wang B, Guo F, Huang C, Zhao H. 2020. Unraveling the iterative type I polyketide synthases hidden in Streptomyces. Proc Natl Acad Sci U S A 117:8449–8454.

Wei Y, Zhang L, Zhou Z, Yan X. 2018. Diversity of Gene Clusters for Polyketide and Nonribosomal Peptide Biosynthesis Revealed by Metagenomic Analysis of the Yellow Sea Sediment. Front Microbiol 9:295.

Weissman KJ. 2004. Polyketide biosynthesis: understanding and exploiting modularity. Philos Trans A Math Phys Eng Sci 362:2671–2690.

Ziemert N, Podell S, Penn K, Badger JH, Allen E, Jensen PR. 2012. The natural product domain seeker NaPDos: a Phylogeny based Bioinformatic tool to classify secondary metabolite gene diversity. PLoS One 7:e34064. Crossref PubMed. ISI.

CHAPTER 5. Final Remarks

Advances in genomic sequencing within the last couple of decades have helped broaden our understanding of the vast repertoire of microbial processes that affect our planet. Ecologically, microbes regulate global biogeochemical cycles (Falkowski et al., 2008), and understanding these processes is more important than ever, as we are faced with the rapidly accelerating threat of climate change. Past that, microbes play crucial roles as the base of global food chains, as essential symbionts within humans, in creating healthy soil systems for global food production, and as biofuel sources (Smetacek et al., 2002; Thursby et al., 2017; Salwan et al., 2022). Perhaps one of the largest ways microbes impact the human world, however, is through the production of specialized metabolites, as many microbial natural products have become the basis for essential therapeutics (Rani et al., 2021). Polyketides are one of the largest and most structurally diverse classes of specialized metabolites, and many are pharmaceutically important as antibiotics, antifungals, anti-cancer agents, antivirals, immunosuppressants, and cholesterol-lowering drugs (Korman et al., 2010).

Traditionally, the discovery of new natural products has occurred via screening crude extracts of microbial strains for bioactivity (Atanasov et al., 2021). In recent years, the success rate of this process has tailed off, with high compound rediscovery common (Pye et al., 2017). This is largely due to the isolation and screening of similar microbial strains. It also has limitations as the vast majority of microbes have not yet been cultured. This deficit is likely explained by common media conditions being unable to target viable but non-culturable cells (Lloyd et al., 2018). To bridge this gap within the natural products world, genome mining for biosynthetic gene clusters has proven effective in assessing the specialized metabolite potential

within uncultured organisms (Medema et al., 2021). Recent genome mining studies have highlighted that enormous amounts of BGC diversity remain to be realized, with one recent study showing that only an astonishing 3% of biosynthetic gene clusters were similar to those that encode for the production of known metabolites (Gavriilidou et al., 2022).

The goal of this dissertation was to evaluate how polyketide biosynthetic diversity varies across Earth. To do so, I first utilized metagenomic datasets to understand differences in KS composition and diversity across biomes. Next, I analyzed genomes across the tree of life to assess how KS repertoires varied across taxonomic lineages, with a focus on investigating understudied polyketide-producing taxonomic lineages. Finally, I used a multi-omics approach combining KS amplicons, metagenomic sequencing, and metabolomics to assess the biosynthetic potential of abyssal marine sediments.

Central to all three of these projects has been the development of the webtool NaPDoS2 (Klau et al., 2022), which identifies and classifies KS domains from genomic queries. Contrary to other genome mining tools such as antiSMASH (Blin et al., 2023) and PRISM (Skinnider et al., 2020), which characterize full-length BGCs, the KS sequence tag approach of NaPDoS2 evaluates PKS diversity within fragmented or amplicon-length sequence queries. Additionally, in scenarios where biosynthetic genes may not fall in the same chromosome such as in eukaryotes (Nutzmann et al., 2020), NaPDoS2 can still detect and assess KS diversity. Further, the KS domains detected using NaPDoS2 are saved as sequence outputs, making downstream phylogenetic analyses rapidly accessible. The recent expansion of the NaPDoS2 database allows for KS domains to be sorted into more than 40 KS subclasses, providing detailed predictions as to the types of polyketide metabolites likely seen in a given biome or organism (Klau et al., 2022).

In chapter 2, I analyzed over 240 GBP of metagenomic sequence data to extract over 35,000 KS domains from eight biomes (Singh et al., 2023). From this dataset, I showed that biomes differed in KS composition, with soils enriched in modular *cis*-AT and hybrid *cis*-AT KS domains, and marine biomes enriched in polyunsaturated fatty acid (PUFA) and enediyne KS domains. Further, I was able to investigate the phylogenetic diversity of certain KS subclasses highlighting biome-specific trends. For example, enediyne compounds are extremely cytotoxic and encode for important antitumor agents, but all enediyne compounds discovered to date have been traced to an actinobacterial producer (Shen et al., 2015). However, I found large enediyne KS clades that grouped separately from known enediynes in marine sediments (Singh et al., 2023). This marine sediment clade grouped with enediyne KSs from non-actinobacterial sources, suggesting that underexplored biomes and taxa carry significant enediyne biosynthetic potential. The findings from this meta-analysis can also guide future studies, as these metagenomic type I KSs were used to improve the design of primers used in chapter 4 of this dissertation. Lastly, one perhaps easy to overlook finding from this study is that on average less than 3% of type I KS domains within a metagenomic dataset were located within MAGs (Singh et al., 2023). In many metagenomic studies, there is a tendency to group contigs into high-quality MAGs and only analyze this portion of the dataset for functional genes of interest due to the high taxonomic confidence. The work presented in this thesis serves as a warning on the possible downfalls of such an approach, given that over 97% of type I KS domains within the metagenomic datasets analyzed fall outside of MAGs.

The goal of chapter 3 was to assess polyketide diversity across the tree of life. To do so, I analyzed more than 600,000 genomes across a broad taxonomic range, finding more than 50,000 KS domains. It came as no surprise to see that historically fruitful specialized metabolite

producing phyla such as Actinobacteria, Myxococcota, Cyanobacteria, and Ascomycota contain large numbers of KS domains, although I showed they differ in the types of KSs they contain. However, more surprising was the over 1000 KS OBUs (80% similarity) that were recovered from the understudied phyla of Acidobacteria, Chloroflexi, Planctomycetes, and Verrucomicrobia. Only three polyketides have been experimentally linked to BGCs from these four phyla (aurantoside A, lasonolide A, and palmerolide), highlighting the extraordinary amount of polyketide diversity that remains to be discovered. Another interesting takeaway from this analysis was in the fungal genomes, where a clade of 26 KS OBUs was classified by NaPDoS2 as hybrid *cis*-AT. This finding was corroborated by running the associated fungal genomes through antiSMASH, which found the KS domains to occur within NRPS (and not PKS) BGCs. To date, no fungal metabolites have been linked to a BGC that contains a hybrid *cis*-AT KS domain, suggesting that these KS domains could lead to novel structural diversity. This illustrates the power of NaPDoS2 for analyzing large datasets to generate predictions, which can then be complemented by other genome mining tools to understand the full BGC context.

The goal of chapter 4 of my dissertation was to evaluate the polyketide biosynthetic diversity within abyssal sediments using a multi-omics approach. Abyssal plains make up over half the Earth's surface and are rich sources of polymetallic nodules, a mineral resource containing copper, nickel, cobalt, iron, and manganese (Stratmann et al., 2021). Many of these are important in the development of electronic devices, making deep-sea mining operations more and more appealing to national and corporate interests. Nonetheless, abyssal plains are home to novel species across the kingdoms of life (Scheckenbach et al., 2010; Schwabe et al., 2008), making it imperative to assess the biosynthetic potential that could be lost or altered with deep-sea mining operations. Here, I found abyssal sediments to contain an astonishing degree of

biosynthetic diversity (over 6,000 KS OBUs), almost all of which was novel (less than 1% clustered with KS domains from the MIBiG database). By using KS amplicon sequencing, I was able to recover more than three orders of magnitude more modular *cis*-AT, hybrid *cis*-AT, and *trans*-AT KS domains than were obtained from the abyssal sediment metagenomes. This highlights one major advantage of amplicon sequencing, which is aided by the ability of the NaPDoS2 webtool to process large datasets. While KS amplicons were recovered from the abyssal sediments that matched closely with those in the salinlactam, salinisporamide, rifamycin, and quinolidomycin BGCs (all reported in marine actinomycetes), I was unable to find any of these compounds using untargeted metabolomics. The majority of the molecular features recovered (over 90%) were distinct from those seen in the GNPS database, which suggests that abyssal sediments contain distinct metabolites compared to more frequently sampled systems.

In conclusion, as the traditional natural product discovery approach of cultivation, chemical extraction, and bioassay-guided fractionation has fallen out of favor (Pye et al., 2017), an explosion in genomic capabilities has made mining for BGCs a viable alternative approach. Globally, antibiotic resistance is rising (Rossiter et al., 2017), making it more imperative than ever to discover novel chemical diversity. In this thesis, I assess the composition and diversity of KS domains across biomes, the tree of life, and unprecedently, in abyssal sediments. In each of these studies, I have highlighted novel KS lineages, which I hope future research programs can utilize for the discovery of novel polyketides.

5.1 References

Atanasov AG, Zotchev SB, Dirsch VM, Supuran CT. 2021. Natural products in drug discovery: advances and opportunities. Nat Rev Drug Discov 20:200–216.

Blin K, Shaw S, Augustijn HE, Reitz ZL, Biermann F, Alanjary M, Fetter A, Terlouw BR, Metcalf WW, Helfrich EJN, van Wezel GP, Medema MH, Weber T. 2023. antiSMASH 7.0: new and improved predictions for detection, regulation, chemical structures and visualisation. Nucleic Acids Research 51:W46–W50.

Falkowski PG, Fenchel T, Delong EF. 2008. The Microbial Engines That Drive Earth's Biogeochemical Cycles. Science 320:1034–1039.

Gavriilidou A, Kautsar SA, Zaburannyi N, Krug D, Müller R, Medema MH, Ziemert N. 2022. Compendium of specialized metabolite biosynthetic diversity encoded in bacterial genomes. Nat Microbiol 7:1324–1324.

Klau LJ, Podell S, Creamer KE, Demko AM, Singh HW, Allen EE, Moore BS, Ziemert N, Letzel AC, Jensen PR. 2022. The Natural Product Domain Seeker version 2 (NaPDoS2) webtool relates ketosynthase phylogeny to biosynthetic function. Journal of Biological Chemistry 298:102480.

Korman TP, Ames B, (Sheryl) Tsai S-C. 2010. Structural Enzymology of Polyketide Synthase: The Structure–Sequence–Function Correlation, p. 305–345. In Comprehensive Natural Products II. Elsevier.

Lloyd KG, Steen AD, Ladau J, Yin J, Crosby L. 2018. Phylogenetically Novel Uncultured Microbial Cells Dominate Earth Microbiomes. mSystems 3:10.1128/msystems.00055-18.

Medema MH, De Rond T, Moore BS. 2021. Mining genomes to illuminate the specialized chemistry of life. Nat Rev Genet 22:553–571.

Nützmann HW, Doerr D, Ramírez-Colmenero A, Sotelo-Fonseca JE, Wegel E, Di Stefano M, Wingett SW, Fraser P, Hurst L, Fernandez-Valverde SL, Osbourn A. 2020. Active and repressed biosynthetic gene clusters have spatially distinct chromosome states. Proc Natl Acad Sci USA 117:13800–13809.

Pye CR, Bertin MJ, Lokey RS, Gerwick WH, Linington RG. 2017. Retrospective analysis of natural products provides insights for future discovery trends. Proc Natl Acad Sci USA 114:5601–5606.

Rossiter SE, Fletcher MH, Wuest WM. 2017. Natural Products as Platforms To Overcome Antibiotic Resistance. Chem Rev 117:12415–12474.

Salwan R, Sharma V. 2022. Plant beneficial microbes in mitigating the nutrient cycling for sustainable agriculture and food security, p. 483–512. In Plant Nutrition and Food Security in the Era of Climate Change. Elsevier

Scheckenbach F, Hausmann K, Wylezich C, Weitere M, Arndt H. 2010. Large-scale patterns in biodiversity of microbial eukaryotes from the abyssal sea floor. Proc Natl Acad Sci USA 107:115–120.

Schwabe E. 2008. A summary of reports of abyssal and hadal Monoplacophora and Polyplacophora (Mollusca). Zootaxa 1866.

Shen B, Yan X, Huang T, Ge H, Yang D, Teng Q, Rudolf JD, Lohman JR. 2015. Enediynes: exploration of microbial genomics to discover new anticancer drug leads. Bioorg Med Chem Lett 25:9–15.

Singh HW, Creamer KE, Chase AB, Klau LJ, Podell S, Jensen PR. 2023. Metagenomic data reveals type I polyketide synthase distributions across biomes. mSystems 8:e00012-23.

Skinnider MA, Johnston CW, Gunabalasingam M, Merwin NJ, Kieliszek AM, MacLellan RJ, Li H, Ranieri MRM, Webster ALH, Cao MPT, Pfeifle A, Spencer N, To QH, Wallace DP, Dejong CA, Magarvey NA. 2020. Comprehensive prediction of secondary metabolite structure and biological activity from microbial genome sequences. Nat Commun 11:6058.

Smetacek V. 2002. Microbial food webs: The ocean’s veil. Nature 419:565–565.

Stratmann T, Soetaert K, Kersken D, Van Oevelen D. 2021. Polymetallic nodules are essential for food-web integrity of a prospective deep-seabed mining area in Pacific abyssal plains. Sci Rep 11:12238.

Thursby E, Juge N. 2017. Introduction to the human gut microbiota. Biochemical Journal 474:1823–1836.