Improving Natural Product Discovery Efforts Through Analysis of Biosynthetic Gene Populations in Sediment

BY

MARYAM ELFEKI

B.S. University of Illinois at Chicago, 2012

THESIS

Submitted as partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Pharmacognosy
in the Graduate College of the
University of Illinois at Chicago, 2020

Chicago, Illinois

Defense Committee:

Dr. Brian T. Murphy, Advisor and Chair
Dr. Stefan J Green, Biological Sciences, Co-Advisor
Dr. Jimmy Orjala,
Dr. Alessandra S. Eustáquio
Dr. Nadine Ziemert, Dept. Microbiology and Biotechnology, University of Tübingen

DEDICATION

To my beloved mom, Safana Fikry, whose love for me knew no bounds and who taught me the value of life, love, hard work, perseverance, and resilience.

***

To my father for his endless support, love, encouragement, advice, and endless stream of fruits.

***

To my brother Ahmed Elfeki for his support, love, care, and for being my eternal buddy and personal lawyer since birth.

***

To my uncle Mostafa Fikry and my grandmother Karam Faisal (may her soul rest in peace), for their love, care, support, endless encouragement, and wisdom and for filling our home with joy and bliss and delicious food.

***

To my caring family and amazing friends.

***

To past and future generations of scientists for their hard work and their future endeavors, may your lives be filled with continued success and may we continue to strive for a thriving global community.

***

ACKNOWLEDGEMENTS

I have received a great deal of support and assistance throughout the writing of this dissertation.

I would first like to express my gratitude for my Ph.D. advisor Dr. Brian T. Murphy, whose expertise and advice were invaluable in the formulation of the research topic. I thank you for the invaluable mentorship and opportunities you have provided me with. You have constantly motivated and empowered me with utter enthusiasm and support, and for that, I am incredibly fortunate and grateful to have worked under your mentorship.

I thank the members of my dissertation and preliminary exam committees for their precious guidance. I thank Dr. Stefan J. Green for his mentorship and support. I especially thank him for helping me troubleshoot and solve every mishap that came along with my experimental techniques – including, and not limited to, teaching me the proper pipetting technique. His insight and expertise were invaluable for the completion of my thesis project.

I thank Dr. Nadine Ziemert for her continuous guidance throughout my graduate studies and for hosting me in her lab and providing me with the opportunity to collaborate with her amazing team with whom I have published (and currently working on) with. I especially thank Mohammed Alanjary for his help with coding and for teaching me invaluable hacks that have been extremely useful to me and Shrikant Mantri for giving me the opportunity to collaborate.

Dr. Alessandra S. Eustáquio for always keeping her door open for professional and personal advice, and for her lab members for allowing me to continuously borrow equipment and tools. Special thanks to Dr. Jana Braesel for her advice and help with cloning and for being an amazing trip partner.

ACKNOWLEDGEMENTS (continued)

I thank all the members of the Murphy Lab for such a happy, supportive, and collaborative lab community. I thank them for all the great times during collection trips, conferences, and daily in lab.

I thank the many collaborators with whom I had the opportunity to work with.

I thank the Office of Technology Management (OTM) for their financial support and for providing me with endless support, for their friendship and meaningful discussions. Special thanks to Dr. Melissa Maderia and Dr. Nelson Grihalde for being the best managers one can ask for and for their mentorship. Thank you to Dr. Suseelan Pookote and to all the OTM team for always having the graduate student team's best interests in mind and for continuing to advocate for our progress.

Thanks to all instructors who guided me in the process of building knowledge and the critical thinking skills. Special thanks to and Dr. Jonathan Bisson and Dr. Charlotte Simmler for their friendship and meaningful discussions.

Thanks to everyone in the Center for Center for Biomolecular Sciences (CSB) for making it a kinetic and welcoming community.

I would like to thank the College of Pharmacy (COP), Molecular Biology Research Building (MBRB), and CBS staff who were always so accommodating: Elizabeth Ryan, Elizabeth Woods, Arletta K. Harris, Mei Zhang, Dan Lu, Rachel Morrow, and Omar Alvarado.

Special thanks to all my amazing friends, Shurooq Abdeljaber, Nitin Samuel Jayakumar, Xue Jiang, Amira Kefi, Nelson Grihalde, Teresa Grihalde, Maria Sofia Ramos da Costa, Camila Manoel Crnkovic, Mita De, and so many others. Thanks to Subramaniam Palghat

Balasubramanian, Marcela Ardengue, and Guilherme M. Balbim for the great memories and adventures in Chicago.

I thank Ahmed, Mostafa, Ossama, and my whole family for their unconditional support and endless encouragement. Finally, I would like to thank Safana Fikry for her wonderful friendship, infinite support, I could not ask for a better mother, friend, and partner.

Financial support was provided by the Department of Medicinal Chemistry and Pharmacognosy and by the OTM.

ME

CONTRIBUTION OF AUTHORS

Chapter 1 is an original work by Maryam Elfeki, it comprises a literature review that gives background into the field of natural product drug discovery that is relevant to this dissertation. Chapter 2 consists of a published manuscript (Elfeki, M., et al. *ACS Chem. Bio.* **2018,** 13, 8, 2074-2081) for which Maryam Elfeki was the primary author. Dr. Mohammad Alanjary assisted with the computational analysis. Dr. Stefan J. Green assisted with the sequencing design and strategy and provided insight with sequence data processing. Dr. Nadine Ziemert provided advice for the experimental work, the data analysis, and the conceptual impact of the work. Dr. Brian T. Murphy revised the manuscript and was the corresponding author on the publication. Chapter 3 consists of a planned manuscript (Elfeki, M., et al) for which Maryam Elfeki was the primary author. Shrikant Mantri assisted with the computational. Dr. Stefan J. Green assisted with the sequencing design and strategy and provided insight with sequence data processing. Dr. Nadine Ziemert provided insight into data analysis and interpretation. Chapter 4 describes the contribution of Maryam Elfeki to work done by collaborators. Chapter 4.1 consists of a published manuscript (Shaikh, A.F., Elfeki M., et al. *Nat Prod Sci* **2015**). Isolation and structure elucidation work performed by Anam F. Shaikh, phylogenetic analysis was performed by Maryam Elfeki under Dr. Stefan J. Green's advice. This work was advised and edited by Dr. Brian T. Murphy. Chapter 4.2 consists of a published manuscript (Wang, M., et al. *Nat Biotecnol* **2016**) for which Dr. Mingxun Wang was the primary author. Mass spectrometry data was acquired by various members of the Murphy lab including Dr. Skylar Carlson, Dr. Michael Mullowney, Dr. Hiyoung Kim, Anam F.Shaikh. Data was compiled and labelled the data, converted it to the format needed for the data curation of the GNPS database (http://gnps.ucsd.edu) under the supervision of Dr. Brian T. Murphy. Dr. Nuno Bandeira was the corresponding author on the publication. Chapter 4.2 consists of a published

CONTRIBUTION OF AUTHORS (continued)

manuscript (Petrovich, M., et al. *FEMS Microbiology Ecology*, **2018**) for which Morgan Petrovich was the primary author. Maryam Elfeki's input in this study consisted of providing Dr. Morgan Petrovich a curated list antibiotic classes of interest, under the supervision of Dr. Brian T. Murphy. Dr. George Wells was the corresponding author on the publication. Chapter 5 is an original work by Maryam Elfeki that summarizes the main findings from the previous chapters and highlights their relevance for natural product discovery. It also discusses limitations and opportunities for future work.

TABLE OF CONTENTS

LIST OF TABLES

TABLE **PAGE**

LIST OF FIGURES

FIGURE **PAGE**

LIST OF FIGURES

FIGURE **PAGE**

LIST OF ABBREVIATIONS

A : Adenylation

ACP : Acyl Carrier Protein

AT: Acyl Transferase

BGC : Biosynthetic Gene Cluster

BLAST : Basic Local Alignment Search Tool

C : Condensation

ER : EnoylReductase

eSNaPD : environmental Surveyor of Natural Product Diversity

HMMs : Hidden Markov Models

KR : KetoReductase

KS : KetoSynthase

MIBiG : Minimum Information about a Biosynthetic Gene cluster

MT : MethylTransferase

NGS : Next-Generation Sequencing

NP : Natural Productss

NRPS : Nonribosomal Peptide Synthetase

NRPs : Nonribosomal Peptide(s)

OBU : Operational Biosynthetic Unit

OTU : Operational Taxonomic Unit

PCP : Peptide Carrier Protein

PCR : Polymerase Chain Reaction

PK(s) : Polyketide(s)

LIST OF ABBREVIATIONS (continued)

PKS : Polyketide Synthase

RiPP(s) : Ribosomally-synthesized and Post-translationally-modified Peptide(s)

rRNA : Ribosomal Ribonucleic Acid

TE : Thioesterase

TPS : Terpene Synthase

SUMMARY

In this study, we examined the biosynthetic gene diversity that exists in Lake Huron sediment. We employed high-throughput amplicon sequencing to characterize geographic patterns of natural product (NP) production genes from 56 surface sediment samples across a nearly 60,000 square kilometer geographic area. From these data, we were able to map the occurrence of production genes from antibiotics, siderophores, and other bioactive compounds across lake sediment. Our results provided evidence that some NP classes exhibit sparse occurrence, while others exhibit more cosmopolitan occurrence throughout the lake. The results provide some of the first preliminary evidence to support the commonly accepted notion that extensive sample collection efforts are required to more fully capture the NP capacity that exists in sediment.

To further improve our understanding of chemical space available for use in drug discovery, we needed to understand the extent to which common cultivation techniques have accessed existing chemical space. Metagenomic studies have shown that cultivable bacteria represent a fraction of those that exist in the environment, and that uncultivated populations in sediment have genes that encode for biosynthetic enzymes that are capable of producing a high diversity of novel NPs. Quantifying these genes in both sediment and cultivatable bacterial populations allows us to assess how much diversity is present on nutrient agar and is critical to guiding the trajectory of future NP discovery platforms. We thus employed next generation amplicon sequencing to assess the NP biosynthetic gene populations present in two Lake Huron sediment samples, and compared these with populations from their corresponding cultivatable bacteria. We highlight three findings from our study: 1) after cultivation, we recovered between

7.7% and 23% of three common types of NP biosynthetic genes from the original sediment population; 2) between 76.3% and 91.5% of measured NP biosynthetic genes from nutrient agar have yet to be characterized in known biosynthetic gene cluster databases, indicating that readily cultivatable bacteria harbor potential to produce new NPs; 3) even though the predominant taxa present on nutrient media represented some of the major producers of bacterial NPs, the sediment harbored a significantly greater pool of NP biosynthetic genes that could be mined for structural novelty, and these likely belong to taxa that typically have not been represented in microbial drug discovery libraries.

1. INTRODUCTION

1.1 Historical perspective on the field of natural products research

The use of natural products (NPs) in traditional medicine dates back to the time of the early Chinese and Egyptian civilizations.[1] Rock paintings and fossil records of the Neolithic and the Middle Paleolithic ages represent plants that were already used as a remedy for our ancestors about 60,000 years ago.[2] Several ancient civilizations show widespread use of NPs for medicinal purposes: Mesopotamian clay tables dating from 2600 B.C. depicted oils from *Cupressus sempervirens* (Cypress) and *Commiphora* species (myrrh) in cuneiform.[3] These oils are still currently in use as cough, cold and inflammation remedies.[4] Egyptian Papyrus records document over 700 plant-based remedies dating from 2900 B.C.[5] Furthermore, Chinese documents dating from 1100 B.C. to 100 A.D. recorded a collection of prescriptions and medicinal drugs from herbal sources.[4] Likewise, documents ranging from 300 B.C to 100 A.D. illustrate Greek physicians and natural scientists stored and recoded their collection of medicinal herbs. This widespread knowledge of NP medicines was also prevalent in the Assyrian, Babylonian, Sumerian, Indian, and Native American cultures, among others. Today, NP medicines remain key to human health.[6]

1.1.1 Introduction to the conventional discovery of new NP structures

NPs discovery and research evolved with developing technology. Modern microbial NP drug discovery can be traced to the discovery and development of pyocyanase, penicillin, and tyrothricin roughly between 1899 and 1944.[7–9] These discoveries not only denoted the start of the "Golden Age of Antibiotics" (1930's-1970's), they also established microorganisms as a source of novel bioactive compounds. Prior to the Golden Age, microbial infections posed a serious health risk. This era revolutionized medicine and extended life expectancy by contributing many

antibiotics that are in use to the current day. NPs continue to play a leading role in the development of many currently used medicines. In fact, 49% of the small molecules approved from the 1940s until the end of 2014 are a NP or are NP derived.[10] Today, many of these are microbial-derived.[11]

1.1.1.1 Microbial primary and secondary metabolites

NPs are also called "secondary metabolites." Unlike primary metabolism, which involves the biosynthesis and breakdown of essential molecules such as proteins and fats, secondary metabolites are compounds that are generally not essential for life-sustaining processes of an organism.[12] It is believed that these specialized compounds confer some sort of a selective advantage to the producer, whether it be a defense mechanism against predators, a signaling molecule, or even a survival factor for extreme environments.[13] For this reason, secondary metabolism represents eras of evolution and adaptation subject to natural selection during microbial evolution.[13] Driven by this principle, researchers set out to distant parts of the globe for pharmaceutical lead-compound discovery.

1.1.1.2 History of environmental sample collection efforts

Prior to the discovery of penicillin, pyocyanase, and tyrothricin in the early 1900's, the majority of NP derived drugs were acquired from terrestrial plants.[14] Examples of these include alkaloids such as morphine and quinine. The success of penicillin in particular led to the expansion of drug discovery to terrestrial microorganisms. Both the scientific community and pharmaceutical companies responded by developing screening programs to identify soil microorganisms that can produce structurally diverse NPs that can be used against infectious diseases. Indeed, important contributions have been made by microorganisms collected in the terrestrial environment to the discovery of antibacterial agents including beta-lactams, aminoglycosides, tetracyclines, among other antibiotic classes.[15] It has been generally accepted that sampling understudied, or previously

unvisited locations will provide access to novel microbial genera with potential to harbor novel biosynthetic pathways and consequently structurally novel secondary metabolites.[16] Sample collection expeditions were thus devised with consideration for microbial diversity from different geographic regions with different environmental conditions.[17] Indeed, notable unique secondary metabolites were discovered from samples collected in geographically distinct areas across the globe.[17,18] However, the hypothesis that going to distinct places will yield distinct metabolites is met with opposing arguments. One such argument is that this hypothesis is based on incomplete knowledge of biological diversity; that more comprehensive sampling in terms of geographical coverage and species coverage can disprove this hypothesis.[18,19] Furthermore, this hypothesis can be invalidated once the full range of metabolites that any given species is able to produce are attainable.[18,19] This opposing argument is built from metagenomics studies that argue that "everything can be found everywhere" and that the limitation lies in limited cultivation techniques.[20,21] Since little is known about the extent to which cultivation techniques access diverse microbial and NP populations, it is difficult to evaluate how extensive sample collection expeditions should be.

1.1.1.3 Cultivation techniques and their limitations

Pharmaceutical companies and academic laboratories began the creation of complex biological and chemical libraries produced by the cultivation and extraction of microorganisms. Many procedures for the isolation, cultivation, and identification of these microorganisms were developed.[22,23] The traditional workflow of NP isolation (i.e. biological assay-guided isolation) consisted of sample collection, microbial isolation and library generation, cultivation and chemical extraction, biological activity testing, and structural elucidation. Though, the success of the workflow relies on the diversity of samples entered into the pipeline, outlining the importance of

sample collection and cultivation techniques to access a large part of the diversity pool in those samples. Thus, researchers began developing improved cultivation techniques. For example, low nutrient media that simulated the originating environment enabled the cultivation of microbial single cells separated from environmental samples.[6] Furthermore, different nutrients to mimic organic nitrogen and carbon in the environment and/or the addition of antifungal and antibacterial compounds allowed for the growth of both slow- and fast-growing microorganisms.[24] However, many organisms were not amenable to culturing using standard microbiological techniques; culture-dependent techniques typically used to describe bacterial communities cannot account for the large bacterial diversity observed in metagenome studies. Therefore, despite some advances in cultivation techniques, it is estimated that greater than 99% of microorganisms remain uncultured.[25] Consequently, a large range of potentially novel chemical diversity remains inaccessible to traditional NP discovery programs.

1.1.2 The problem of rediscovery of known compounds

The Golden Age of antibiotics identified microorganisms as a rich source of unique and novel chemical classes, however, over the last few decades, a large decline in the rate of discovery of new NPs has occurred.[26] This is due to some extent to the re-solation of known NP structures from terrestrial microorganisms. To reinvigorate the discovery pipeline, researchers expanded their efforts to obtain new organisms by including difficult-to-cultivate microorganisms, and by searching for unique and understudied sources of bacteria.

Interestingly, in their quest for novel structures, researchers devoted relatively little to no attention to marine microorganisms until the late 1960's.[27,28] Researchers turned their attention to marine microorganisms after realizing a vast diversity of microorganisms exists in the ocean.[29] Spongouridine, spongothymidine, and pentabromopseudilin were among the first described

bioactive compounds from marine sources.[30–32] They were isolated from the Caribbean sponge *Cryptotheca crypta* in the early 1950s and in 1966.[29] These efforts instigated a growing interest in the study of marine bacteria, which has been identified as a rich source of unique and novel chemical classes.

Nonetheless, the rediscovery of known NP structures remains an increasing challenge. The high degree of taxonomic and NP redundancy in libraries was one of several reasons that caused pharmaceutical companies to divest in NP research by the 2000's. The last 30 years alone have seen a drop in the number of approved antibiotics with significant stagnation over the last decade.[33] This is also due, in part, to a lack of innovation, a costly development pipeline, coupled with lower financial returns for developers, especially for antibiotic discovery.[34]

Yet, NPs remain the most valuable source of novel structural classes, and an indispensable source of new drugs.[35,36] Thus, to ensure whether NPs will remain valuable as sources of drug leads, researchers started evaluating how much of the chemical space and clinically relevant NPs has been discovered.[37]

1.1.3 Quantitative examination of available NP chemical space

This constant rediscovery of known antibiotic scaffolds forced researchers to ponder the capacity of microorganisms to provide novel structures. Have we captured the entire chemical space available in nature? Or has our narrow focus on specific genera and culture techniques limited us to a small percentage of available NP space?

1.1.3.1 Chemical space covered by microbial NPs

The promise of an unlimited supply of structures shifted away from NPs as the rate of discovery of novel scaffolds decreased. This was believed to be caused by researchers having

described all of the chemical space covered by NPs.[37] Researchers thus began investigating the chemical space of structures to evaluate whether to continue NP research or rely more heavily on alternative synthetic/combinatorial approaches.

It is important to understand what percentage of NP space we accessed in decades of cultivating bacteria for discovery efforts, in order to determine the trajectory of future efforts. Experimental and computational approaches designed to map chemical space have been undertaken. These approaches comprised of computational analysis of NPs databases,[37] and statistical and systematic comparisons of NPs collections with synthetic compounds.[37,38] In their dataset consisting of 52,395 unique molecules, Pye C.R et al. examined both number of compounds and compound novelty with the aim of answering central questions including 1) how has the rate of discovery of new NPs changed over time, 2) how has the average NP structural novelty evolved with time, 3) if studying novel taxonomic space will lead to novel structures discovery, and 4) if it is viable to approximate "how close we are to having described all of the chemical space covered by NPs."[37] Pye et al explained that unique structures represent a declining percentage of compounds isolated from natural sources, however, the discovery rate of novel NPs has increased since the field's inception.[37] They concluded that for efficient access to structural novelty, significant innovation will be required for the field to retain the impressive historical rate of novel compound discovery.[37]

Indeed, the Pye et al. study confirms that substantial opportunity for the discovery of novel structures exists, and that in order to access these and reinvigorate the NP discovery pipeline, we must steer away from traditional discovery and screening methods and from historically familiar sources. Instead, we must utilize new technologies, new scientific approaches, and expand to include promising underexplored sources.

1.1.4 Genetic research unveils NPs biosynthetic diversity

Advances in sequencing techniques and the increasing volume of genome sequence data have allowed the emergence of genomics as a resource for NP discovery.[39] Tools and techniques such as molecular phylogenetics have emerged allowing not only the profiling of strains for prioritization and dereplication, but also enabling bioinformatics-guided NP discovery and characterization. Furthermore, metagenomics has – and still is – enriching our understanding of the remarkably diverse biosynthetic capacities of uncultured communities – a potential that has been hidden from traditional approaches.[40–42]

1.1.4.1 Environmental DNA sequencing efforts to understand chemical space

High-throughput sequencing has broadened our understanding of the scope of biodiversity in ecological systems in terms of taxa, but only lately has it revealed nature's remarkable biosynthetic potential. [43] Proof of this biosynthetic potential greatly supported by genomics and metagenomics efforts to study microbial NP discovery.[44] It has provided opportunities to assess the secondary metabolite biosynthetic potential on the individual strain or community level without the requirement for cultivation.[45] Individual strains "are predicted to encode far more NPs than what we have characterized to date."[46] Furthermore, high-throughput sequencing allowed researchers to evaluate biosynthetic gene richness in environmental sample surveys such as soil and sediment samples, which proved tremendous unknown biosynthetic diversity in the environment.[41,42] Additionally, it secured researchers a better understanding of the evolution of NPs drug discovery and the biosynthetic process, allowing them to start building databases and phylogenetic relationships between enzymes.[47,48]

1.2 The creation of NPs gene databases

1.2.1 Primer on NPs biosynthesis

Genomic tools that were developed helped illuminate mechanisms of biosynthesis and allowed the correlation of specific genes to their corresponding compounds. The genomic basis for these compounds is encoded in Biosynthetic Gene Clusters (BGCs). The latter act as the "genetic blueprint of NP biosynthesis:" they contain the core machinery for NP production and the accessory enzymes necessary for their tailoring and export.[44] Major examples include the multi-modular enzyme complexes termed polyketide synthase (PKS) and nonribosomal peptide synthetase (NRPS).[49] They are responsible for two major classes of NPs – polyketides (PKs) and nonribosomal peptides (NRPs), respectively.[49] Other examples discovered later include the ribosomally-synthesized and post-translationally-modified peptides (RiPPs), terpene synthases (TPS), and other primary metabolism derivatives such as the saccharides.[10,50]

These modular mega-enzymes are of particular interest to the NP research community due to their ability to build a wide complexity of NP scaffolds using enzymatic chain reactions catalyzed by each module in an assembly line approach.[49] This is accomplished via stepwise reactions where starting monomers circulate from a starter module to an/many elongation module(s) and a termination module to form a growing chain that is then released and is subject to post-translational modifications (Figure 1). Each of these modules can be further subdivided into multiple domains, some of which are highly conserved. Modules of the PKS consist of a set of core catalytic domains that are present in most PKS systems and that catalyzes and condense acyl monomers: acyl transferase (AT), ketosynthase (KS), acyl carrier protein (ACP), and thioesterase (TE). Other domains can include ketoreductase (KR), dehydratase (DH), enoylreductase (ER), methyltransferase (MT), and special tailoring modules such as cyclization modules (Alanjary).

Additional types of PKS systems exist such as iterative type I, trans- AT, type II, and type III systems.[49]

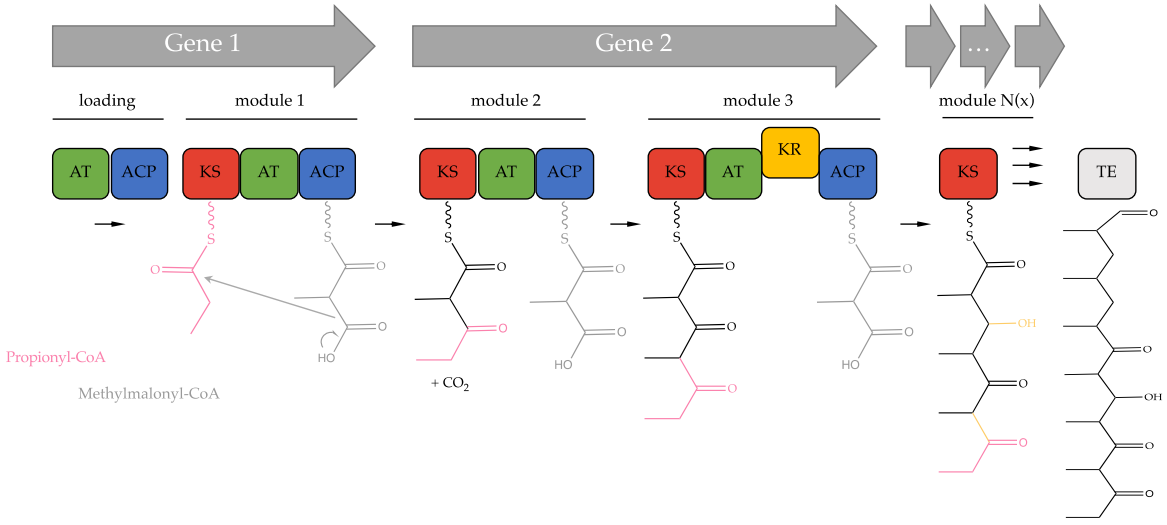


Figure 1. Conceptual example of modular PKS type I chain extension.
AT domain select for the starter and extender units. Additional domains not shown here. (Adapted from Alanjary M.M., 2018).

Like PKS systems, NRPS systems are also modular, with regular and non-proteinogenic amino acids used as starting and extender units instead of acyl monomers. The core domains for NRPS modules are the adenylation (A), peptide carrier protein (PCP), and condensation (C) domains. Hybrid systems have also been observed that fuse PKS and NRPS systems.

Due to the conserved nature of some fragments in these domains such the PKS KS domain and NRPS A and C domains, researchers were able to design probes to screen and identify genetic information responsible for a chemical structure.[40] These probes were used to capture specific areas of a genome from an environmental sample allowing us to systematically discover their chemical potential.

1.2.2 The development of NPs biosynthetic gene databases for gene annotation

Genome mining gained traction in drug discovery research thanks to the advancement of sequencing technologies and the increase in publicly available genomes. It allowed researchers to identify promising gene clusters via phylogenetic and functional metagenomic approaches, to predict the product of a specific pathway or even to probe genomes of uncultivable bacteria for their NP producing capacity directly from the environment using bioinformatic tools.[42,51–53] Databases were also created to annotate and store enzyme sequences in BGC pathways, in order to more rapidly identify and classify this NP biosynthetic potential.[41,47] A community-updated list of these cluster-mining tools can be found at the "The Secondary Metabolite Bioinformatics Portal" (http://secondarymetabolites.org). Many of these software algorithms use Hidden Markov Models (HMMs) to rapidly search for known signatures such as KS or A domains – a strategy we employed in our own research in order to filter our sequencing data.

These tools allowed us to assess how much BGC diversity is present in both sediment and cultivatable bacterial populations and whether a trend exists when it comes to the distribution of BGCs – a step which is critical to guiding the trajectory of future NP discovery platforms.

2. "ASSESSING THE EFFICIENCY OF CULTIVATION TECHNIQUES TO RECOVER NP BIOSYNTHETIC GENE POPULATIONS FROM SEDIMENT"

Adapted by permission from Springer Nature: ACS Chemical Biology Journal, Assessing the Efficiency of Cultivation Techniques to Recover NP Biosynthetic Gene Populations from Sediment. Elfeki, M., et al. Copyright 2018 American Chemical Society.

2.1 Abstract

Despite decades of cultivating microorganisms for use in drug discovery, few attempts have been made to measure the extent to which common cultivation techniques have accessed existing chemical space. Metagenomic studies have shown that cultivable bacteria represent a fraction of those that exist in the environment, and that uncultivated populations in sediment have genes that encode for a high diversity of novel NP biosynthetic enzymes. Quantifying these genes in both sediment and cultivatable bacterial populations allows us to assess how much diversity is present on nutrient agar and is critical to guiding the trajectory of future NP discovery platforms. Herein we employed next generation amplicon sequencing to assess the NP biosynthetic gene populations present in two Lake Huron sediment samples, and compared these with populations from their corresponding cultivatable bacteria. We highlight three findings from our study: 1) after cultivation, we recovered between 7.7% and 23% of three common types of NP biosynthetic genes from the original sediment population; 2) between 76.3% and 91.5% of measured NP biosynthetic genes from nutrient agar have yet to be characterized in known BGC databases, indicating that readily cultivatable bacteria harbor potential to produce new NPs; 3) even though the predominant taxa present on nutrient media represented some of the major producers of bacterial NPs, the sediment harbored a significantly greater pool of NP biosynthetic genes that could be mined for

structural novelty, and these likely belong to taxa that typically have not been represented in microbial drug discovery libraries.

2.2 Introduction

Since the discovery of penicillin in 1928, microorganisms have served as both a significant source of biologically active NPs and an inspiration for NP-derived small molecule scaffolds.[35] Their remarkable biosynthetic capacity allows them to produce NPs with high structural diversity and a wide range of biological activities. These small molecules have been the cornerstone for the treatment of an array of diseases, as 34% of FDA approved drugs between 2000 and 2014 were NPs or NP-derived, encompassing anticancer drugs, antibiotics, immunosuppressants, among others.[10]

Motivated by penicillin's discovery, massive sampling expeditions aimed at collecting cultivatable microorganisms from the environment continued for the next several decades. Though these successfully supplied us with an arsenal of drugs to treat a variety of diseases, as time passed, researchers began to re-isolate known molecules from these existing microbial libraries. This led to a divestment in microbial-based NP drug discovery,[28] in part due to the fact that the libraries contained a high degree of taxonomic and chemical redundancy. More recently, with the advent of advances in gene sequencing and bioinformatics, it has become possible to assess the microbial enzymatic machinery that builds NPs, and consequently the potential for a given community of microorganisms to harbor NP structural diversity; this has begun to re-shape the process of microbial drug discovery.

Two major classes of NPs – PKs and NRPs – are produced via multi-modular enzyme complexes termed PKS and NRPS, respectively.[49] These modular mega-enzymes build a wide range of NP

scaffolds using enzymatic reactions catalyzed by each module in an assembly line approach. Each module can be further subdivided into multiple domains, some of which have been amenable to probe and primer design due to high nucleotide conservation, thereby enabling molecular screening and discovery of new chemical structures.[40]

Prior studies have shown that immense PKS and NRPS biosynthetic diversity exists in sediment.[54–57] However, few attempts have been made to measure the extent to which common cultivation techniques are accessing this chemical space, and to assess how much genetic diversity remains to be recovered. For nearly a century, the majority of microbial drug discovery efforts have relied on cultivatable bacteria to supply the therapeutic lead pipeline. Since a large percentage of these prior cultivation efforts have focused on spore-forming microorganisms (which include bacterial genera within the phylum Actinobacteria such as *Streptomyces* and *Actinomyces*, and within the phylum Firmicutes such as *Bacillus*),[58] we employed six commonly used nutrient media to enrich for these bacteria from two Lake Huron sediment samples. High-throughput amplicon sequencing was used to characterize taxonomic (16S Ribosomal ribonucleic acid (rRNA) gene) diversity and the biosynthetic sequence diversity from sediment samples and their corresponding cultivated bacteria. From these data, recovery and novelty of chemical space from readily cultivatable bacteria were estimated. This study provides insight into past discovery efforts, allowing an estimation of the NP biosynthetic diversity that is available to researchers on commonly employed nutrient media. It also provides evidence that targeted innovations to current cultivation and genome mining techniques will afford a high probability of discovering novel NPs.

2.3 Results and discussion

2.3.1 Recovery of microbial 16S rRNA gene amplicon sequences from sediment

Microbial diversity was assessed in two distinct sediment samples collected in Lake Huron, and from corresponding heat-treated sediment plated on nutrient agar, using high throughput next-generation sequencing (NGS) of microbial 16S rRNA gene amplicons. A total of 664,384 sequence reads (21,990 from 2 sediment samples and 642,394 from 24 nutrient agar plates) were obtained; each of 6 nutrient agar samples from each sediment was analyzed in duplicate (Figure 2).

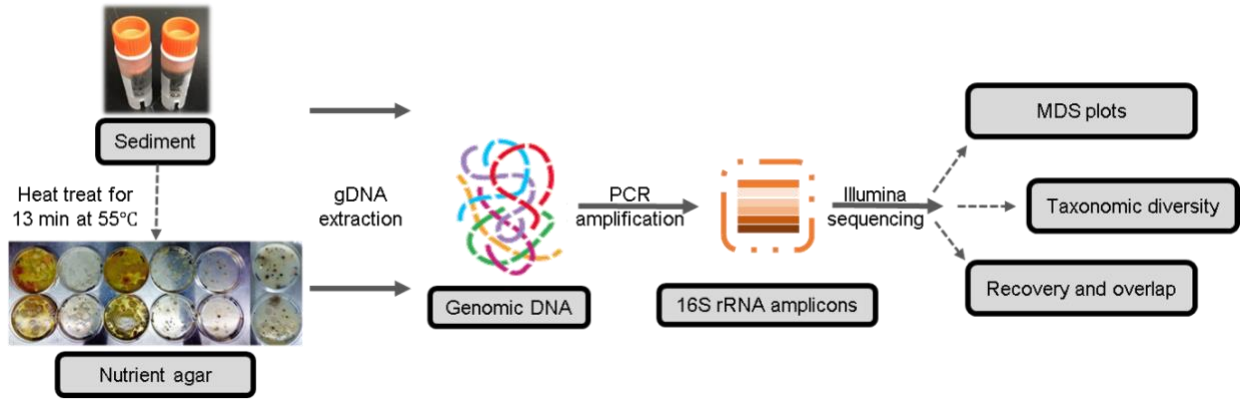


Figure 2. Genomic DNA extraction and 16S rRNA gene sequencing of samples.
Each sediment sample was heat-treated and plated on six different nutrient media, in duplicate. Genomic DNA was extracted from the sediment sample and the corresponding nutrient agar.

Combined sequences were clustered at 97% similarity, which resulted in 31,076 operational taxonomic units (OTUs; a common measure for bacterial diversity) before rarefaction analysis. 16S rRNA gene amplicon sequence data were rarefied prior to alpha diversity comparisons between samples (Table V-A).[59]

In total, we recovered an average of 8.13% of OTUs on nutrient agar from the original sediment samples, and these were predominantly from the phyla Firmicutes and Actinobacteria (Figure 3) shows the breakdown of the genera from Firmicutes sequence reads (Tables IV-B and IV-C show the percentage of sequences and OTUs and the number of OTUs belonging to different taxonomic

groups, respectively). We intended to enrich spore-forming bacteria on nutrient agar. Historically, these taxa have been extensively studied for NP drug discovery due to their ease of isolation and NP production capacity.[60] Sequences from spore-forming bacteria of the *Streptomycetaceae* and *Micromonosporaceae* families constituted approximately 87% of Actinobacterial sequences recovered from nutrient agar plates (Figure 3), while those families represented approximately 16% of Actinobacteria reads in sediment, and less than 2.4% of total reads in sediment. A major component of the sediment Actinobacterial community, as assessed by amplicon sequencing, was the *Gaiellaceae* family. Members of this family were not cultivated in this study, and elsewhere have been rarely cultivated under laboratory conditions.[61]

Similarly, the *Bacillaceae* and the *Paenibacillaceae* families constituted nearly 96% of Firmicutes sequence reads on nutrient agar, while they accounted for approximately 48% of Firmicutes sequence reads in sediment (Figure 14). The disproportionate abundance of sequence reads of Firmicutes on nutrient agar compared to Actinobacteria are likely due to the speed and extent to which *Bacillus* grows on our selected nutrient agar.

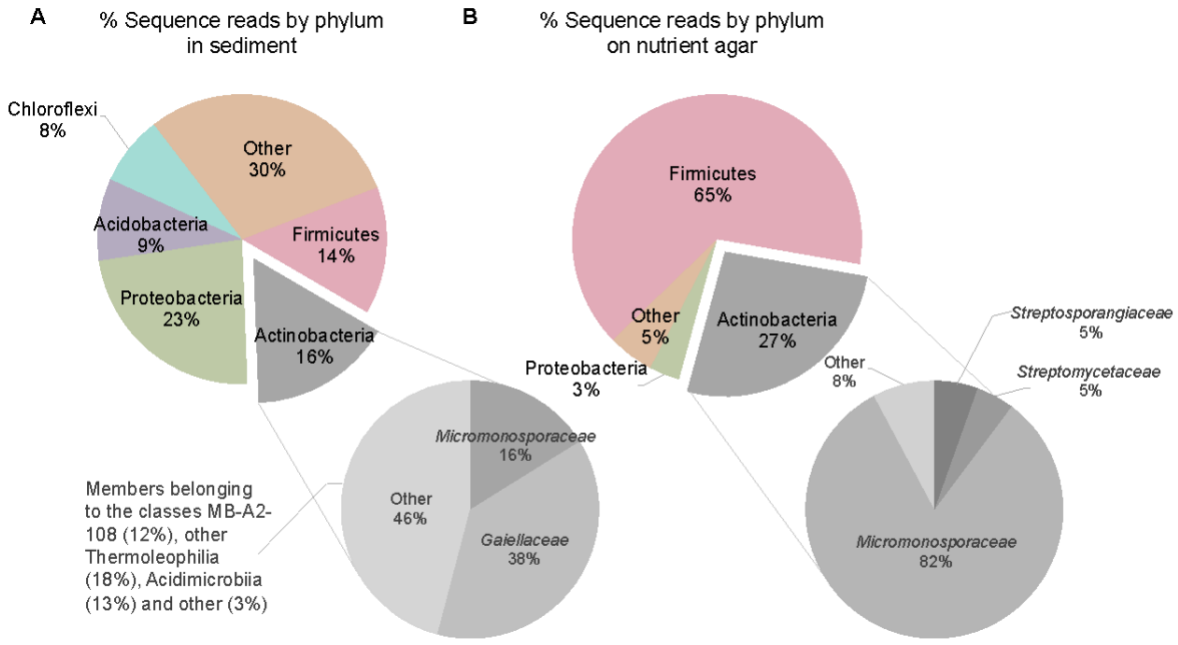


Figure 3. 16S rRNA gene amplicon sequence reads detected in sediment compared with those detected on nutrient agar at the level of phylum and family.
A more detailed breakdown of microbial composition is in Figure 14 and Tables IV-B and V. (a) Sequence read abundance of bacteria in sediment at the phylum level. (b) Sequence read abundance of bacteria at the phylum level after heat treatment of samples and cultivation on nutrient agar. Spore forming bacteria were successfully enriched, as Firmicutes and Actinobacteria constituted approximately 92% of sequences on nutrient agar.

2.3.2 Validation of percentage used to cluster BGC sequences

To assess the percent recovery of NP structural classes from sediment using common cultivation methods, we used previously designed degenerate primers to amplify the KS domain for PKS I, KSα domain for PKS II, and A domain for NRPS genes.[62–64] The KS, KSα, and A domains were selected because they are the most conserved catalytic domains of the PKS type I, type II, and NRPS gene clusters,[65,66] respectively, allowing the design of degenerate primers to amplify them. Furthermore, bioinformatic tools and databases have been developed to facilitate the prediction of NPs from these pathways.[67]

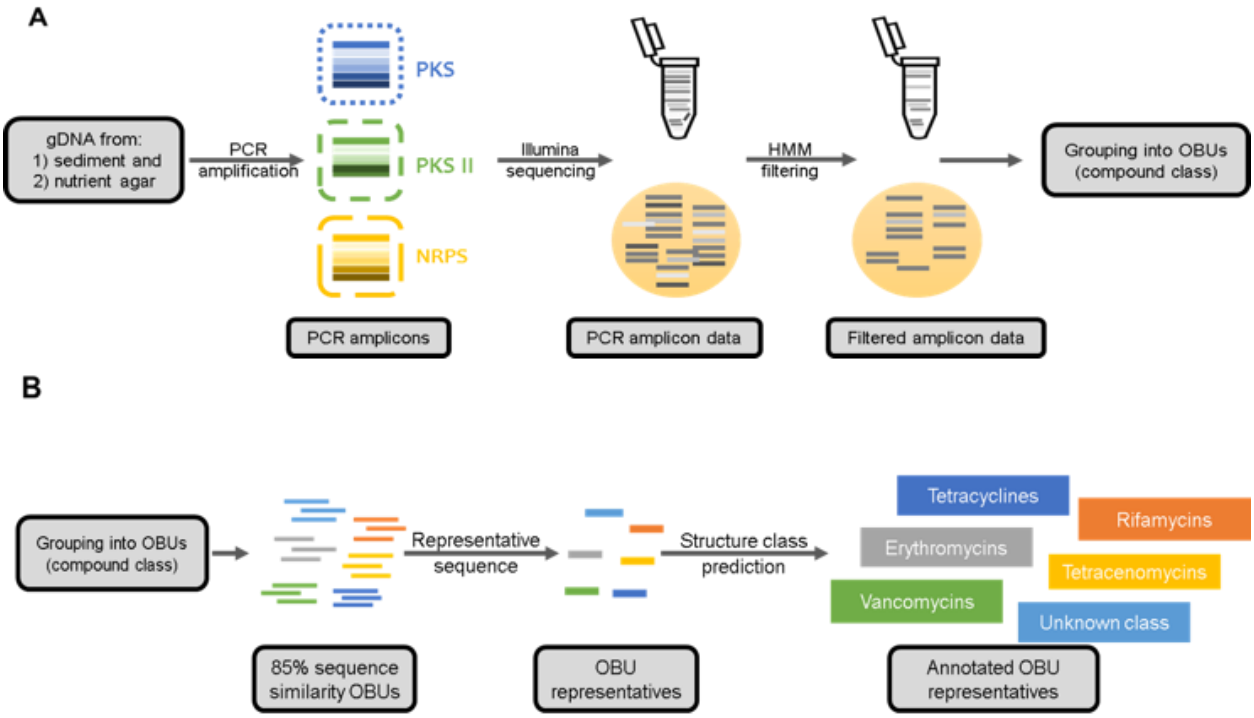


Figure 4. Experimental process to assess recovery of biosynthetic gene clusters from sediment. (a) Genomic DNA was extracted from sediment samples and nutrient agar. Conserved regions from PKS types I and II, and NRPS biosynthetic gene clusters were amplified and deeply sequenced. The sequences were then filtered using HMM models downloaded from the software package antiSMASH. (b) Filtered sequences were clustered into operational biosynthetic units (OBUs) at a threshold of 85% sequence similarity. A representative sequence was selected for each cluster and then subjected to Basic Local Alignment Search Tool (BLAST) analysis against the Minimum Information about a Biosynthetic Gene cluster (MIBiG) database.[68]

Amplified KS, KSα, and A domains were clustered by a furthest neighbor algorithm into OBUs according to percent similarity in an attempt to create compound class groupings, as described previously.[52,69] We expanded on this process of clustering in order to ensure that we did not over- or underestimate OBU diversity from our datasets. First, we extracted KS, KSα and A domain sequences from the manually curated and annotated BGC database MIBiG.[68] We then subjected

these sequences to the same bioinformatic analyses as our samples. In order to select a proper clustering percentage, we evaluated the accuracy of different thresholds for their ability to group entries according to similar biosynthetic origins. We found that the optimal clustering threshold fluctuates and is dependent on the specific compound class. However, since the optimal thresholds ranged from 80% to 90%, we selected 85% as the most suitable for our purposes (Figure 5 and Methods section 2.4.7).

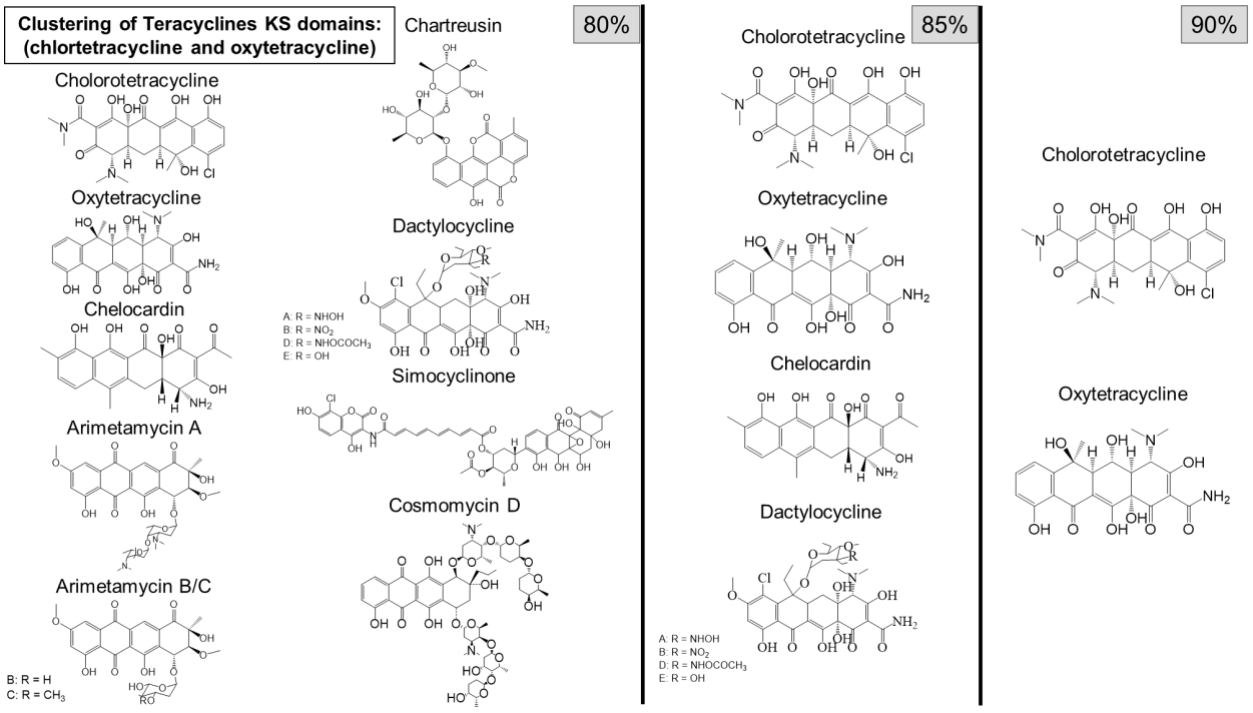


Figure 5. Clustering of KS domains extracted from the genomes of organisms producing tetracycline at different percentages.
KS domains were extracted from MIBiG entries. We clustered sequences associated with the tetracycline class of antibiotics at 80%, 85%, and 90% similarity. 85% was selected as the most suitable for our purposes.

2.3.3 Calculation of percent recovery of KS, KSα, and A domain OBUs from sediment samples

Next, the extent to which cultivation approaches recovered BGCs from sediment was measured. To calculate the recovery of OBUs using our cultivation techniques, we first rarefied sequence data, then divided the number of observed OBUs on nutrient agar by the combined total of observed OBUs in sediment and OBUs on nutrient agar (an explanation of the OBU recovery calculation appears in Figure 13). In the rarefied data, approximately 3.5-, 12-, and 5.4-fold greater KS, KSα, and A domain OBUs, respectively, were observed in sediment compared to nutrient agar (Tables IIA-B) at an 85% similarity threshold. The percent recovery of OBUs from sediment was calculated to be 23.3% for KS, 7.7% for KSα, and 15.8% for A domains at this similarity threshold. This estimate, however, represents the upper limit of percent recovery, since calculations were based on a conservative estimate of OBUs in sediment derived from rarified data (sequencing depth for OBU populations on nutrient agar was closer to saturation compared with those in sediment). Moreover, the Shannon diversity index highlights a large disparity in NP biosynthetic richness between sediment and nutrient agar populations (Table IA-B).

Table I. Diversity analysis and OBU recovery of filtered sequences from (a) sediment and (b) nutrient agar samples.

The Shannon diversity index for sediment and nutrient agar was calculated to highlight the disparity in OBU diversity. The number of observed KS, KSα, and A domain OBUs in sediment was approximately 3.5-, 12-, and 5.4-fold greater compared to those on nutrient agar.

(A) Diversity indices and number of observed sequences and OBUs from sediment after rarefaction.

Average # of sequences per sample (total)	*Average # of observed OBUs per sample (total)*	*Average Shannon index*

KS	1,354 (2,708)	738.5 (1,448)	7.86
KSα	1,058 (2,116)	862.5 (1,719)	8.64
A	2,269 (4,538)	1,014 (1,879)	8.62

(B) Diversity indices and number of observed sequences and OBUs from nutrient agar after rarefaction.

	Average # of sequences per sample (total)	*Average # of observed OBUs per sample (total)*	*Average Shannon index*	*% recovery of OBUs**
KS	916 (19,227)	45.0 (413)	3.65	23.3
KSα	518 (11,935)	14.3 (142)	1.98	7.7
A	1,010 (23,234)	24.9 (350)	3.26	15.8

*For a detailed explanation of calculation of the % recovery of OBUs, see Figure 13.

We cultivated a minority of the sediment microbial population, which consequently yielded a relatively small percent of the NP biosynthetic capacity that existed in the sediment samples. Even though the predominant cultivated taxa represent some of the major producers of bacterial NPs (bacteria from the genus *Streptomyces* are responsible for more than half of medically important antimicrobial and antitumor agents,[70] while those from the genera *Bacillus* and *Micromonospora* have been well studied and produce a diverse array of biologically active compounds),[71,72] the estimated BGC recovery indicates that these taxa harbor a minority of the chemical space present in the sediment. This is supported by 16S rRNA gene analyses (Tables IVB-C) that indicate *Streptomyces*, *Micromonospora*, and *Bacillus* together account for a small percent of OTUs in the

sediment samples (average of 5.2%), while being relatively abundant on nutrient media. When averaging the total contribution of these genera on each nutrient plate, they accounted for 56.1% of OTUs, suggesting that they are responsible for a large share of the OBUs that were detected from the cultivated bacterial population. This was supported by an alignment of representative OBU sequences against the NCBI-NT nucleotide database to infer assignment of BGC sequence reads to different bacterial taxa (Figure 19). *Streptomyces* and *Micromonospora* accounted for the majority of Actinobacteria OTUs on nutrient media (average of 85.0%), while representing an average of 6.3% of Actinobacteria OTUs in sediment. One explanation for the observed number of OTUs belonging to *Streptomyces* and *Micromonospora* can be attributed to their spore-forming ability, allowing them to withstand the heat shock treatment prior to their cultivation. These results also suggest that understudied actinomycetes not cultivated in our study may contribute to the large OBU population found in sediment.

In these Lake Huron samples, the sediment harbored a significantly greater pool of OBUs that could be mined for NP novelty. The uncultivated organisms from which these OBUs are derived may belong to taxa that typically have not been represented in drug discovery libraries, such as Proteobacteria and phyla that make up the "30% other" category (*e.g.*, Nitrospirae, Acidobacteria, Chloroflexi and Planctomycetes; Figure 3 and Table VI). For example, we detected 413 KS OBUs on nutrient media, but a minimally overlapping 1,448 OBUs in corresponding sediment (approximately 3.5-fold increase). This effect was more pronounced for KSα and A domains, at 12- and 5.4-fold increases, respectively. These results further validate the hypothesis that growing and/or mining the "uncultivated" majority will increase probability for discovering novel NP scaffolds, as exemplified in recent studies.[73–75] One way to explain this disparity is that several NP producers were not present/abundant on cultivation media due to our selection for spore-forming

bacteria. Those selected against included Cyanobacteria, Betaproteobacteria (*e.g.*, *Burkholderia* spp.), Gammaproteobacteria (*e.g.*, *Vibrio*, *Alteromonas*, *Pseudomonas* spp.), Deltaproteobacteria (*e.g.*, myxobacteria), among others. Although NP production capacity varies widely between taxa, these phyla and classes all contain strains with high levels of NP biosynthetic genes (up to nearly 20% of their coding capacity). Additionally, several studies document this phenomenon, and emphasize that increased NP production capacity correlates well with increased genome size.[76,77]

2.3.4 Approximation of undescribed BGC diversity from the cultivatable bacterial population

In order to assess whether the BGC sequences detected from the cultivatable bacterial population could be assigned to putatively novel chemotypes, a representative sequence from each OBU group was aligned with domain-specific reference sequences from the MIBiG database. Sequences derived from plates were assigned to known compound classes if they had a minimum similarity of 85% to known sequences over a minimum length of 84 amino acids. The remaining sequences were classified either as known but unannotated, or putatively novel. As shown in Figure 6, 81.8% KS, 76.3% KSα, and 91.5% A domain OBUs from nutrient agar and 98.1% KS, 99.8% KSα, and 99.6% A domain OBUs from sediment have yet to be characterized in peer-reviewed literature or deposited into the MIBiG database. Our results corroborate previous efforts to document BGC presence and diversity in environmental samples. [42,78]

Charlop-Powers Z. et al. (2014) compared biosynthetic gene richness and diversity from 96 different sediment samples located throughout the southwestern and northeastern regions of the United States.[79] The Chao1 diversity metric estimated the presence of 1,000 to greater than 7,000 OBUs clustered at 95% sequence identity per soil microbiome. Moreover, based on our

calculations from their published data, only 30% of these OBUs identified to KS and A domain fragments found in functionally characterized gene clusters. In a separate study, Charlop-Powers Z. et al. (2015) compared NP biosynthetic potential of soil samples from a diverse array of environmental microbiomes.[42] They showed that 185 biomes predicted greater than 350,000 OBUs for each of the two studied domains, KS and A, with rarefaction analysis suggesting that the sequence space had not yet been saturated. In addition, the authors found that only 5-10% of the total KS and A domain sequences originating from all 185 biomes were confidently assigned to known gene clusters using the environmental Surveyor of Natural Product Diversity (eSNaPD) algorithm. Our results corroborate these efforts. We observed approximately 3.5-, 12-, and 5.4-fold greater KS, KSα, and A domain OBUs in sediment compared to those on nutrient agar. Moreover, the Shannon diversity index was found to be significantly greater in sediment compared with nutrient agar.[80] These results highlight the disparity in NP biosynthetic diversity between the sediment and the nutrient agar populations.

Importantly, in addition to the high number of uncharacterized OBUs in these sediment samples, there is still a large pool of OBUs in bacteria that are readily cultivatable. It is possible that some of these sequences encode for known compounds and are simply unannotated in MIBiG. However, it is also possible that these sequences encode for the production of novel NPs (particularly since freshwater environments are relatively under-explored for their NP biosynthetic capacity).[81] It is unclear which of these possibilities is more prevalent, but this highlights that readily cultivatable bacteria may still represent a source of undescribed NPs. Since many cultivation practices have not changed substantially since the 1930s, innovative approaches are needed exploit the full NP-producing capacity of nature's cultivatable microbiome; we recently reported a technique and bioinformatics pipeline to address this.[82] Additional strategies include employing less biased

techniques to select colonies from nutrient media for addition to strain libraries, decreasing reliance on colony morphology and taxonomy to estimate NP production capacities, and improving methods to assess NP production directly from colonies under multiple nutrient media in high-throughput in the front-end of drug discovery efforts. Ultimately, improving recovery of novel NPs will require a combination of novel cultivation strategies, isolation from novel environmental samples, improved screening of current cultivation efforts, corroboration of cultivation data with cultivation-independent shotgun metagenomic data, and functional metagenomic strategies.

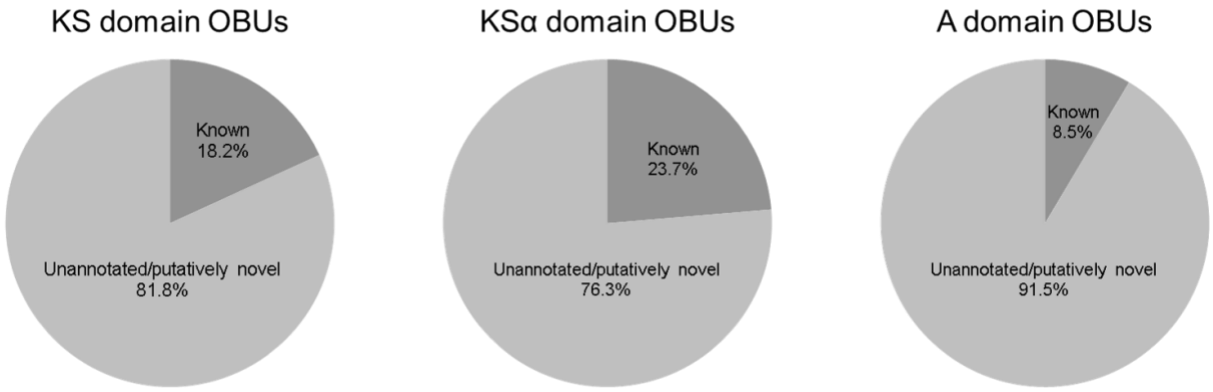


Figure 6. Percent of known and unannotated/putatively novel OBUs on nutrient agar.
A representative sequence from each OBU was blasted against the MIBiG database. Nutrient agar plate sequences with an identity of 85% over a minimum of 84 amino acids were assigned to known compound classes. The remaining sequences were classified as either known but unannotated in the MIBiG database, or putatively novel.

There are a few limitations to the current study. The low degree of overlap between OBUs from sediment and nutrient agar affects the accuracy of our recovery estimates. We accounted for this phenomenon by assuming sequences observed on nutrient agar also existed in sediment, and added the former to the latter when calculating percent recovery (Figure 13). Furthermore, these environmental diversity assessments performed in this study most likely represent substantial

underestimates of the true diversity of the sediment. First, these sediment samples were not sequenced to saturation, and thus, additional diversity likely remains to be sequenced with the methods employed. Second, polymerase chain reaction (PCR) amplification – particularly with degenerate primers – can introduce substantial bias, thereby decreasing or eliminating the observed frequency of some sediment sequences in the data set.[83] In addition, the target range of some of the primer sets is taxonomically limited. For example, the primers employed were designed specifically for Actinobacteria sequences (A domain primers) or a small subset of Actinobacteria, such as *Streptomyces* spp. (KSα domain primers). In addition, by clustering at 85% similarity, distinct sequences were clustered together, increasing the appearance of similarity between the sequences derived from cultivated organisms and those found in the sediment.

Finally, we relied on the MIBiG database to assess sequence novelty. The number of existing NP s far outnumbers the number of entries in MIBiG, underlining the limitation of existing databases. The expansion of entries in NP databases such as MIBiG,[68] antiSMASH[76] and NaPDoS[47] to more fully document the library of BGCs available along with their associated molecular products, is critical toward calculating more accurate assessments of existing NP structural diversity. With a more expansive NP BGC database, we can better prioritize and translate existing chemical space in a set of cultivatable sediment microbiomes into a higher potential to produce structurally novel therapeutic leads.

2.4 Conclusion

Despite decades of cultivating microorganisms for use in drug discovery, few attempts have been made to measure the extent to which common cultivation techniques have accessed existing chemical space. In our study of Lake Huron sediment samples, we highlight three findings: 1) after

cultivation, we recovered between 7.7% and 23% of three common types of NP biosynthetic genes from the original sediment population, and these likely represent upper limits, as they are based on a conservative estimate of OBUs in sediment; 2) between 76.3% and 91.5% of measured NP biosynthetic genes from nutrient agar have yet to be characterized in known BGC databases, indicating that readily cultivatable bacteria harbor potential to produce new NPs; 3) even though the predominant taxa present on nutrient media represented some of the major producers of bacterial NPs, the sediment harbored a significantly greater pool of NP biosynthetic genes that could be mined for structural novelty, and these likely belong to taxa that typically do not constitute modern microbial drug discovery libraries. In a recent study, R. Pye et al. analyzed 52,395 microbial and marine-based NPs discovered between 1941 and 2015 to estimate how much of the NP chemical space remains unknown.[37] Despite some trends that highlighted the continuous discovery of known compound classes, they predicted that innovative discovery methods will continue to yield unique structures. Our study supports this recommendation and suggests that there is a large untapped chemical diversity in *both* readily cultivatable and total sediment bacterial populations. In addition to rapidly developing functional genomics techniques designed to access BGCs that are silent or part of "uncultivated" bacteria, we predict that improved microbial cultivation, unbiased colony selection, and more thorough structural and functional characterizations of NP BGC pathways are needed to access the large portion of microbial and chemical space revealed in our study.

2.5 Methods

2.5.1 Collection of sediment samples.

Sediment samples H054 and NC68 were collected in Lake Huron at a surface depth of 134.9 m and 17.3 m, respectively during a research expedition aboard the Environmental Protection Agency's (EPA's) Lake Guardian Research Vessel. The samples were collected using a PONAR grab in summer 2012 from Georgian Bay and the Northern Channel (Figure 18). The top layer of sediment was homogenized, and two aliquots were placed into two sterile 50 mL conical tubes containing 20% glycerol. They were stored in cryogenic vials in a Dewar.

2.5.2 Cultivating sediment bacteria on nutrient agar.

Conical tubes were thawed and homogenized, and aliquots of two sediment samples were individually collected and placed in 4 mL vials for duplicates processing (sample A and sample B). Samples were diluted with filter-sterilized deionized (DI) water to a 1/10th concentration and incubated in a 57 °C water bath for 15 minutes. A 50 µL aliquot of the sediment dilution was spread onto the surface of an agar plate. Six different media types were used to make nutrient agar diversity plates: A1, 1/10th dilution of A1 (M1), ISP2, 1/10th dilution of ISP2 (DISP2), minimal agar media (LWA), and chitin (Table X).

2.5.3 Genomic DNA isolation from sediment and nutrient agar.

Each nutrient agar plate was split into two halves; each half was placed in a 50 mL conical tube with added filter sterilized TRIS buffer to avoid re-solidification after melting. The conical tubes containing nutrient agar were microwaved for 2 min, with vortexing every 30 sec at speed 7 (1890 RPM). Two 250 µL aliquots were pipetted from each conical tube and placed into a sterile Eppendorf® tube for DNA isolation. DNA was extracted using the PowerSoil® DNA Isolation Kit according to manufacturer's protocol (Mo Bio Laboratories, Inc). gDNA was extracted from all sediment samples, using the PowerSoil® DNA Isolation Kit according to manufacturer's

instructions.

2.5.4 16S rRNA gene amplification and sequencing.

The V4 region of the small subunit ribosomal RNA genes (16S rRNA) was PCR amplified from genomic DNA using a two-stage PCR protocol, as described previously.[84] Domain-level primers 515F (5′- GTGCCAGCMGCCGCGGTAA-3′) and 806R (5′-GGACTACHVGGGTWTCTAAT-3′) were synthesized with 5′ linker sequences CS1 (forward primer; ACACTGACGACATGGTTCTACA) and CS2 (reverse primer; TACGGTAGCAGAGACTTGGTCT).[85] Each 25 µL PCR reaction mixture consisted of 0.5 µL of DNA, 0.8 µL of 10 µM of 515F, 10 µM of 806R, 12.5 µL KAPA Taq 2X ReadyMix (Kapa Biosystems), and 10.4 µL of D.I. water. The thermal cycler conditions were set to a denaturation step at 95°C for 5 min, 28 cycles of 95°C for 30 secs, 45°C for 60 secs, and 68°C for 90 secs, and a final elongation step at 68°C for 7 min. Amplification products were observed by agarose gel electrophoresis and purified using Qiagen's QIAquick PCR cleanup kit according to the manufacturer's protocol (Qiagen Inc.). Subsequently, a second PCR amplification was performed to incorporate Illumina sequencing adapters and a sample-specific barcode into the amplicons. Each reaction received a separate primer pair with a unique 10-base barcode, obtained from the AccessArray Barcode Library for Illumina (Fluidigm, South San Francisco, CA). In addition to Illumina adapter sequences and sample-specific barcodes, these AccessArray primers contained the CS1 and CS2 linkers at the 3′ ends of the oligonucleotides. Cycling conditions were as follows: 95°C for 5 min, followed by 8 cycles of 95°C for 30", 60°C for 30" and 72°C for 60". The pooled libraries, with a 20% phiX spike-in, were loaded onto MiSeq V2 flow cells, and sequenced. Fluidigm sequencing primers, targeting the CS1 and CS2 linker regions, were used to initiate

paired-end 2x250 base read sequencing. Library preparation, pooling, and sequencing were performed at the University of Illinois at Chicago Sequencing Core (UICSQC).

2.5.5 Bioinformatic analyses of 16S rRNA sequence data.

Approximately 1.3 million 16S rRNA sequencing reads were obtained for 2 sediment samples and 6 nutrient agar plates in duplicate (a total of 26 samples). All sequence data generated from the Illumina MiSeq sequencer were first processed using the QIIME-1.9.1 pipeline[86] at the UIC RRC facility. Bar-coded 16S rRNA gene sequences were demultiplexed, primers and chimeras were removed, and the reads were filtered according to Phred quality scores. Forward and reverse reads were merged, labelled according to sample source, and concatenated for OTU clustering. OTU clustering was performed at 97% identity using uclust implemented in QIIME, resulting in 31,076 OTUs. A sequence representative was extracted from each OTU and was classified using the Silva_128 database.[87] A taxon-by-sample abundance matrix (biological observation matrix, BIOM)[88] file was then created. The BIOM file was used to generate Figure 3, Figure 14, and was used for statistical analysis. Diversity at the OTU level was measured by the Shannon index for each sample. The formula used, and the resulting indices are reported in Table V. The BIOM was rarefied to 5,834 sequences per sample within QIIME to avoid analytical issues associated with variable sequence number between samples.

2.5.6 KS, KSα, and A domain amplification and sequencing.

KS, KSα, and A domain amplicon sequencing was performed using the same two-step PCR strategy described above. A 700-bp fragment of the KS domain was PCR amplified from gDNA using degenerate oligonucleotides KSLF (5′-CCSCAGSAGCGCSTSYTSCTSGA-3′) and KSLR (5′-GTSCCSGTSCCGTGSGYSTCSA-3′).[64] A 613 bp fragment of the KSα (α-ketoacyl synthase)

was amplified using degenerate oligonucleotides (5′-TSGCSTGCTTCGAYGCSATC-3′) and (5′-TGGAANCCGCCGAABCCGCT-3′).[63] Degenerate oligonucleotides A3F (5′GCSTACSYSATSTACACSTCSGG3′) and A7R (5′SASGTCVCCSGTSCGGTAS3′)[64] annealed in conserved motifs in the NRPS A domain and amplified a 700-bp fragment. All primers contained a locus-specific sequence as well as a universal 5′ tail *(i.e.*, CS1 and CS2 linkers). 20 µL PCR reaction mixture consisted of 1 µL of DNA, 1 µL of a 10 µM solution of each primer, 10 µL KAPA Taq 2X ReadyMix (Kapa Biosystems), 0.8 µL of DMSO, 3.2 µL of 100 mg/mL Bovine Albumin Serum, and 3 µL of DI water. The thermal cycler conditions were set to an initial denaturation step at 95°C for 5 min, 7 cycles of 1 min at 95°C, 1 min at 65°C (annealing temperature was lowered 1°C per cycle), and 1 min at 72°C, and 40 cycles of 1 min at 95°C, 90 sec at 58°C, and 1 min at 72°C, and a final elongation step at 72°C for 5 min. Amplification products were observed by agarose gel electrophoresis and purified using Qiagen's QIAquick PCR cleanup kit according to the manufacturer's protocol (Qiagen Inc). The resulting PCR amplicons were used as templates for the second PCR step, as described above, to incorporate sequencing adapters and sample-specific barcodes. Pooled and purified amplicon libraries, with a 20% phiX spike-in, were loaded onto a MiSeq V3 flow cell, and sequenced using paired-end 2x300 reads.

2.5.7 Bioinformatic analyses of BGC data.

All sequences generated from the Illumina MiSeq sequencer were 6-frame translated into amino acid sequences using TranslatorX.[89] Only frames with no internal stop codons were kept using TranslatorX's "guess most likely reading frame" option. Amino acid sequences were then filtered via HMMER[90] using HMM pre-build generic detection models downloaded from antiSMASH v3.0.5.[76]

The models used were: PKS_KS.hmm for PKS type I, AMP-binding and A-OX for A domain, and t2ks and t2pks2 for PKS type II. Only sequences that passed the default e-value thresholds were kept; the corresponding nucleotide sequences from each sample were grouped and clustered at 85% using USEARCH v10's UCLUST cluster_fast greedy algorithm via the cluster_fast command.[91] A representative sequence from each cluster – labelled an OBU – was extracted to a separate file using USEARCH v10's -makeudb_usearch command and the file was aligned against the MIBiG database using DIAMOND.[92] The molecular identity of the sequence was appended to the sequence only if sequence identity was equal to or higher than 85% for OBUs clustered at 85%, 90% for OBUs clustered at 90%, and so on. The 85% similarity threshold was selected for subsequent analyses and the OBU representative sequence was annotated with its BLAST identity only if the pairwise identity was at least 85% and coverage over at least 84 amino acids. An OBU-by-sample BIOM[88] file was then created and rarefied to 5,834 sequences per sample within QIIME. The BIOM file was used to generate Figure 3 and Figure 14.

2.5.8 Bioinformatic method validation using reference sequences from the MIBiG database. Gene entries from the MIBiG database were subjected to the same bioinformatics procedures as the BGC sequence data. We selected 12 common antibiotics representing 8 antibiotic classes to test the accuracy of our clustering threshold: ansamysin (rifamycin and geldanamycin), macrolide (erythromycin), and tetracycline antibiotics (chlortetracycline and oxytetracyline) for type I KS domains; aromatic polyketides such as the benzoisochromanequinone compounds (actinorhodin) and type II tetracycline antibiotics (tetracenomycin) for KSα domains; and streptogramins (pristinamycin and virginiamycin), lipopeptide (daptomycin), non-ribosomal cyclic peptide (bacillibactin), and glycopeptide (vancomycin) antibiotics for NRPS A domains. KS, KSα, and A domains were extracted from MIBiG.

Since PKS and NRPS clusters usually contain more than one KS/KSα and A domain, respectively, in most cases multiple KS, KSα and A domains were extracted from each MIBiG entry. For example, we clustered sequences associated with the ansamycin antibiotic rifamycin at 80%, and obtained a total of 30 KS domains, which clustered into 10 OBUs representing 9 compounds. Only 2 of these candidate sequences belonged to the ansamycin class of antibiotics (rifamycin and rubradirin). Additionally, we clustered the sequences at 90% similarity and obtained a total of 26 KS domains, which grouped into 10 OBUs and represented only 3 compounds: rifamycin, naphthomycin, and chaxamycin analogues A/B/C/D (each produced by the same cluster and modified post-translationally) – all of which belong to the ansamycin class. Finally, we clustered the sequences at 85% similarity, and obtained a total of 40 KS domains, which clustered into 4 OBUs that represented 4 molecules: rifamycin, rubradirin, naphthomycin, and chaxamycin analogues A/B/C/D. This analysis was repeated for all the aforementioned antibiotic classes. We found that the optimal clustering threshold fluctuates and is dependent on the specific compound class. However, since the optimal thresholds ranged from 80 to 90%, we selected 85% as the most suitable for our purposes.

2.5.9 Accession Codes.

Project SRA accession: SRP145045. For accession numbers of individual samples within the project, see Table XI.

Supporting Information Available: This material, which includes additional experimental procedures, tables, and figures, is available free of charge via the internet at http://pubs.acs.org.

2.6 Acknowledgements

The authors wish to acknowledge the following contributors: A. Li (UIC), K. Rockne (UIC), S. Carlson (formerly UIC), and crew of EPA RV Lake Guardian for assistance with sediment collection; G. Chlipala of UIC's Core for Research Informatics and A. Naqib of UIC's DNA Services Facility for assistance processing data. This publication was supported by Vahlteich Scholar research funds, the Illinois-Indiana Sea Grant, the Office of Technology Management at UIC, and UIC startup funds.

3. EVALUATING DISTRIBUTION OF BACTERIAL NATURAL PRODUCT BIOSYNTHETIC GENE CLUSTERS IN LAKE HURON SEDIMENT

3.1 Abstract

Environmental microorganisms continue to serve as a major source of bioactive NPs (NPs) and as an inspiration for many other scaffolds in the toolbox of modern medicine. However, improving the discovery rate of novel NPs will necessitate an understanding of their distribution in nature. In this study, the NP biosynthetic gene cluster (BGC) diversity in Lake Huron sediment was examined through biogeographic analysis of BGC domain architecture. High-throughput amplicon sequencing was employed to document geographic occurrences of NP biosynthetic domains from 59 surface sediment samples across a nearly 60,000 square kilometer geographic area. From these data, the occurrence of several classes of NPs were mapped, including antibiotics, siderophores, and other bioactive compounds across lake sediment. These maps provided evidence that some NP classes exhibit sparse occurrence, while others exhibit more cosmopolitan distribution throughout the lake. These results present some of the first preliminary evidence to support the commonly accepted notion that extensive sample collection efforts are required to more fully capture the NP capacity that exists in sediment.

3.2 Introduction

The preparation of Pyocyanase in 1899 and the discovery of bioactive secondary metabolites such as gramicidin and penicillin in 1939 and 1928, respectively, marked the beginning of modern microbial drug discovery efforts.[7,9,93,94] Since then, environmental microorganisms have served as a major source of bioactive NPs and as an inspiration for a plethora

of therapeutic scaffolds. These small molecules have been foundational for human health as therapies for an array of diseases such as cancer, bacterial infections, immune disorders, among others, as 34% of FDA approved drugs from 2000 to 2014 were NPs or NP-derived.[10] Importantly, nearly all of these microbial NP-inspired therapies resulted from field expeditions to collect samples from the environment, arguably one the most important steps in NP drug discovery. Generally speaking, these field expeditions have been guided by the hypothesis that environments in diverse geographic locations contain different evolutionary pressures, and as a result harbor minimally-overlapping populations of NP biosynthetic pathways.[95–97] For over a century of drug discovery, expeditions have seldom expanded on this philosophy, which is reliant on a high degree of serendipity.

Despite the importance of sample collection expeditions toward yielding new drugs, few studies have attempted to document the extent to which NPs or their corresponding production genes are distributed in an environment. Charlop-Powers et al. (2015) compared NP biosynthetic potential of soil samples from a diverse array of environmental microbiomes.[42] Their analyses of 185 soil microbiomes collected from five continents suggested that geographic distance and local environment play important roles in biosynthetic diversity differences observed between samples.[42] Additionally, Lemetre et al. found that latitude and biosynthetic domain composition modifications correlated on a continent-wide scale.[98] Borsetto et al. implicated the observed differences in biosynthetic gene cluster (BGC) diversity in a range of soils within metagenome data, with the microbial community present at each site and with geographic location, and suggested that environmental variables can influence the biosynthetic potential at a given site.[99] This body of studies suggests that biosynthetic domain composition differs, at least in part, with changing geography. Thus, understanding the distribution of BGCs responsible for the production

of NPs across a particular geographic area will inform front-end discovery practices such as sample collection and microbial library generation, which have traditionally been based off of a high degree of uncertainty.

Due to decreasing sequencing costs and availability of online tools, probing the microbial-based chemical diversity in nature has become attainable without relying on cultivation techniques to supply NPs. In this study, KS domains from PKs and A domains from NRPs were examined as they represent two of the most conserved domains present in two of the most abundant classes of NPs.

To gain insight into how certain NP classes are distributed in an environment, the geographic occurrence of NP domains was documented. These NPs are encoded to produce antibiotics, siderophores, and other bioactive compounds from 59 Lake Huron surface sediment samples covering a 59,590 square kilometer region. Investigating BGC distribution patterns and dynamics in Lake Huron represents an essential initial step toward the design of a more methodical environmental sample collection approach, a critical front-end process that has been largely unchanged since the dawn of modern antibiotic discovery efforts in the early 20th century.

3.3 Results and Discussion

3.3.1 Characterization of BGC Domain Sequence Diversity in Sediment

In August and September of 2014, 59 samples were collected from Lake Huron – a geographic region that spans 59,590 square kilometers (Table XII). To assess the occurrence of NPs at each collection site, previously designed degenerate primers were used to amplify the KSα domain for PKS II,[63] and the A domain for NRPS genes from environmental DNA (eDNA) extracted from sediment.[64] The KSα and A domains were selected because they contain some of the most

conserved catalytic domains of the PKS type II and NRPS gene clusters respectively, allowing the design of degenerate primers to amplify them.[62–64] Furthermore, bioinformatic tools and databases have been developed to facilitate the annotation and prediction of NPs from these pathways.[67,100] Amplicons were sequenced using an Illumina MiSeq. Sequence data was then filtered using pHMMs downloaded from antiSMASH's hmm detection modules.[101] These models are based on known and predicted KSα and A domains, and therefore only identify BGCs with known biosynthesis mechanisms.[102]

Sequences were clustered at 85% similarity to indicate approximate molecular class designations, and to avoid the over-estimation of chemical diversity in sediment. This process is described in further detail in Chapter 2.[103] In summary, sequences were extracted from the manually curated and annotated BGC database MIBiG[100], were subjected to different clustering thresholds, and were evaluated for their ability to group according to similar biosynthetic origins/molecular products. The optimal clustering threshold fluctuated and was dependent on the specific compound class (ranged from 80% to 90%). Therefore, analysis proceeded using the 85% similarity threshold. At 85% similarity, the sequence groupings – or OBUs – represent an estimation of molecular classes. To further scrutinize this clustering method, amplicons from a fully sequenced, in-house *Micromonospora* strain B006 were subjected to this process.[104] B006 produces one KSα domain-containing compound and seven A domain-containing compounds.[104] Analysis of B006 amplicons at 85% similarity afforded and two KSα domain OBUs and seven A domain OBUs, and confirmed this as a suitable threshold to organize 300 bp fragments into groups that represent molecular class.

Three samples for KSα and one sample for A domain didn't return sufficient quality data to be included in the analysis. In total, 1,818 KSα OBUs throughout 53 sediment samples, and

171,527 A domain OBUs throughout 58 sediment samples were observed. On average, this translates to approximately 34 KSα and 2,957 A domain OBUs per sediment sample (Table II). These original numbers were eventually corrected to account for suspected overestimation of chemical diversity, as described in the following section. The disparity in KSα and A domain OBU counts may be attributed to the size of the family to which these domains belong. A domains belong to a large superfamily of adenylate-forming enzymes,[105] in contrast to the smaller KSα (α-ketoacyl synthases) domain family, which are only known to produce aromatic-polyketides and polyenes.[106,107] The number of sequences and OBUs for each compound class are listed in Tables XIIIA-B.

3.3.2 Analysis of NP BGC Distribution in Lake Sediment

In order to assess the occurrence of specific NP BGCs classes across Lake Huron sediment, the identity of each OBU must be verified. To accomplish this, a sequence representative from each OBU was aligned against the MIBiG database and a list of hits was generated.[100] MIBiG associates BGCs with known NP structures, allowing us to predict the product of each OBU and as a result, estimate the chemical diversity at each sample site. To ensure that a 300 bp amplicon is sufficient for structural annotation, sequences from a control strain, *Streptomyces coelicolor* A3(2) were amplified, sequenced, and aligned.[108] Amplified KSα and A domain sequences from *S. coelicolor* A3(2) aligned appropriately against coelichelin, coelibactin, and select calcium-dependent antibiotic (CDA) sequences from *S. coelicolor* in MIBiG at a maximum e-value of 3.90 e^{-43}. In general, an e-value smaller than 0.01 is considered a reliable hit for homology matches, while an e-value in the range of 1e^{-50} is considered a match of high reliability.[109] These results were used as a guide to select a list of annotated OBUs to map across lake sediment. A maximum

e-value threshold of 1.2 e^{-15} was selected for KSα domain OBUs and 1.3 e^{-11} for A domain OBUs. These stringent cutoffs allowed only high-confidence OBU assignments to be used in the study.

In total, of the 1,818 KSα domain OBUs that were observed across 53 samples, 32 were assigned to known compound classes. Similarly, of 171,527 total A domain OBUs, 108 were assigned to known compound classes. Once OBU sequence representatives were aligned against MIBiG, the majority of these (98% and 93%, respectively) could not be assigned to known compound classes. Of particular note is that some OBU sequence representatives were assigned to the same molecule (for example, five separate OBU sequence representatives aligned to rifamycin), which resulted in an overestimation of molecular classes present in sediment. To correct for this, the average number of times a molecular class was split into separate OBUs was estimated in the dataset (a split correction factor), and the total number of observed OBUs was divided by that factor for each domain type. The resulting corrected numbers are a total of 1,193 KSα domain OBUs, of which 21 were known compound classes, and a total of 90,528 A domain OBUs, of which 57 were known compound classes. Further details are listed in Tables XIIA-B.

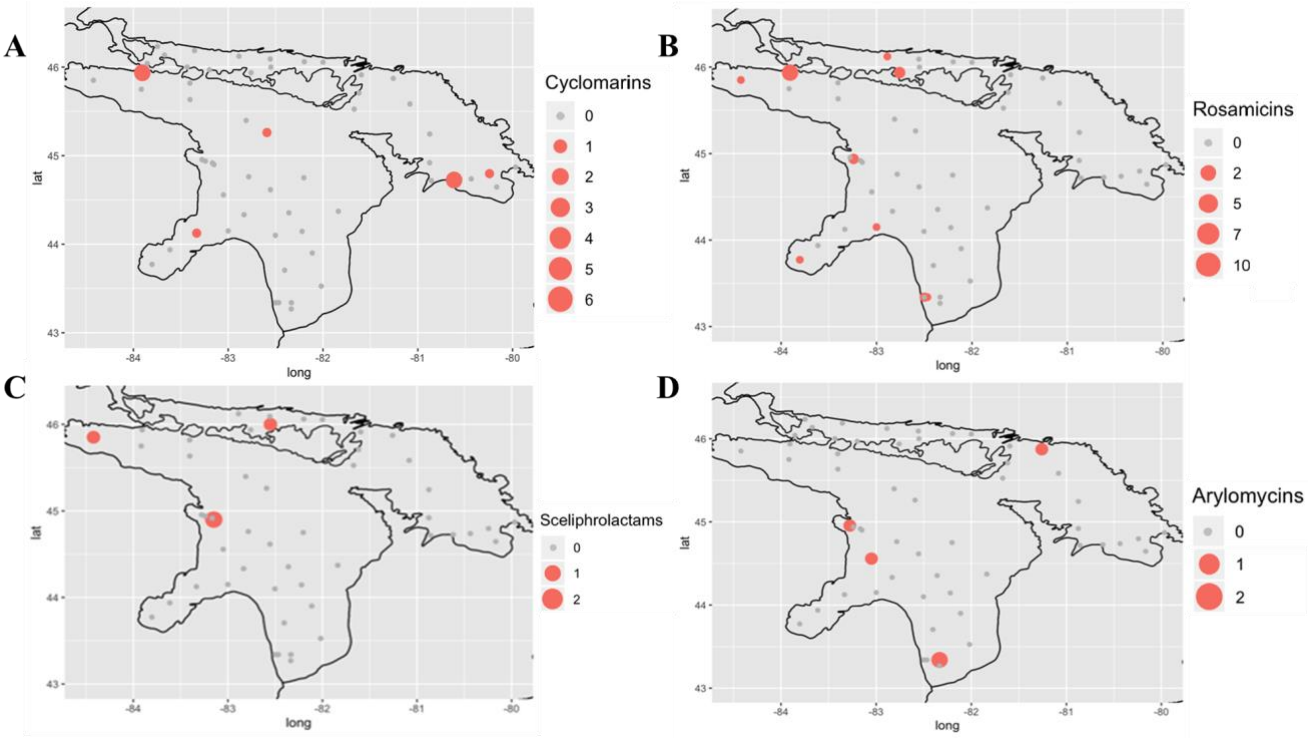
Table II. A and KSα domain abundances in sediment.

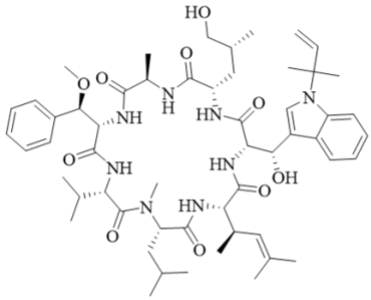
	KSα	**A**
Total # of OBUs	1,818	171,527
Total # of OBUs after adjustment by the split correction factor	1,193	90,528
Average # of OBUs per sample after adjustment by the split correction factor	23	1,561

Of the 78 known classes of PKS (21) and NRPS (57) NPs, distribution maps of compounds that occurred in at least two distinct locations at an abundance of at least two sequences per sample were generated. A total of 30 OBUs met these criteria. From this list of OBUs, eleven antibiotics were selected, and their patterns of occurrence were assessed across lake sediment. The sequence

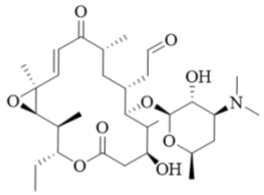
read abundance at each collection site was mapped and represented as different sized circles (four are shown in Figure 7). Figures 7A-D show the distribution of cyclomarins, rosamicins, sceliphrolactams, and arylomycins. Cyclomarins are potent anti-inflammatory cyclic heptapeptides containing four unusual amino acids first described from an estuarine streptomycete, strain CNB-982.[110] Rosamicins are macrolide antibiotics with broad-spectrum activity.[111] Sceliphrolactams are polyene macrocyclic lactams that exhibit antifungal activity.[112] Finally, the arylomycins are lipopeptide antibiotics being investigated for their potent activities against gram-negative bacteria.[113] For example, sequence reads for cyclomarin-type antibiotics were detected in five distinct geographic locations across the lake, meanwhile those for rosamicin-type antibiotics were detected in nine, with only one overlapping location. All eleven antibiotics exhibited distinct distribution profiles across lake sediment. (Figure 21).

Figure 7. Domain sequence distribution of select antibiotic classes across Lake Huron sediment and representative structures from each of the four antibiotic classes.

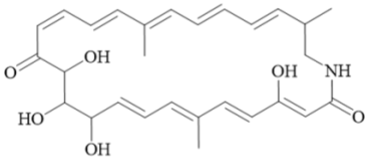




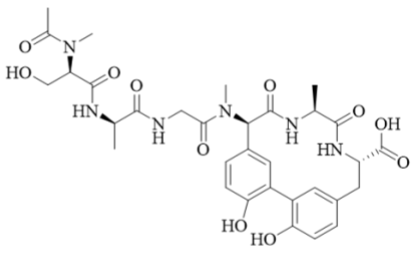
Cyclomarins: Cyclomarin D



Rosamicins: Rosamicin



Sceliphrolactams: Sceliphrolactam

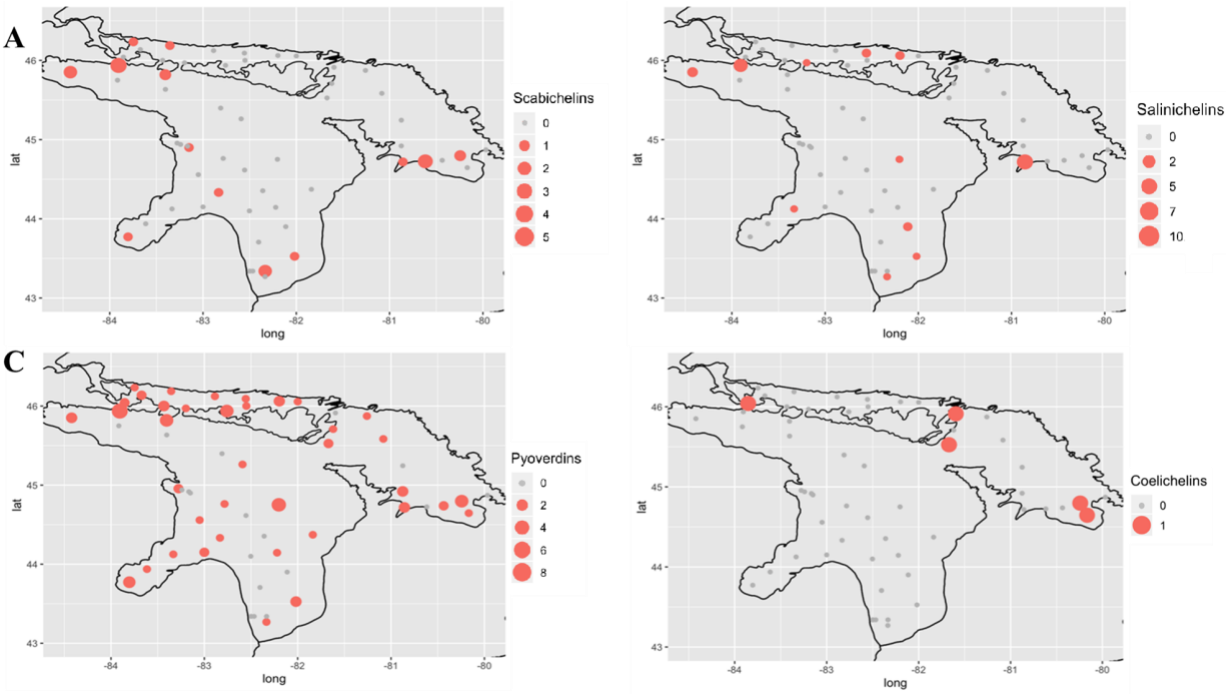


Arylomycins: Arylomycin A2

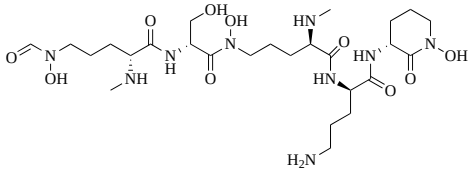
Figure 7. A-D show the domain sequence distribution of select antibiotic classes across Lake Huron sediment: cyclomarins, rosamicins, sceliphrolactams, and arylomycins, respectively. Different sized circles represent thresholds of sequence read abundance at each collection site in Lake Huron. E-H show representative structures from each of the four antibiotic classes. No discernable distribution pattern was observed among eleven antibiotics analyzed (Figure 21).

In order to assess whether similar results could be observed across other NP classes, a group of 8 siderophores were selected from the 30 known and mapped OBUs, and the same analysis was performed. Figures 8A-D show the domain sequence distribution of select types of siderophores across Lake Huron sediment: scabichelins, salinichelins, pyoverdins, and coelichelins, respectively. Scabichelins are tris-hydroxamate siderophores produced by the plant pathogen *Streptomyces scabies* 87.22.[114] Salinichelins were first reported from *Salinispora* strains that lack the desferrioxamine biosynthesis genes.[115] Pyoverdins, formerly called fluorescein due to its with yellowish green fluorescence, are known to be produced by *P. aeruginosa* and function as siderophores.[116] Finally, coelichelins are tripeptide siderophores first reported from *Streptomyces coelicolor*.[117] Sequence reads for scabichelin-type siderophores (Figure 8A) were detected in 13 distinct geographic locations across the lake, meanwhile those for pyoverdin-type siderophores (Figure 8C) were detected in 38, with 10 overlapping locations. In general, the profiles among the eight siderophores tested exhibited no obvious distribution patterns in lake sediment, however, generally speaking, siderophores appeared more frequent in lake sediment than antibiotics (Figure 22 and Supplementary discussion in Appendix B).

Figure 8. Domain sequence distribution of select siderophore classes across Lake Huron sediment and representative structures from each of the four siderophore classes.

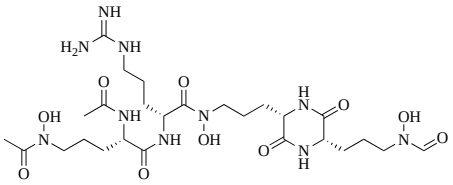


E



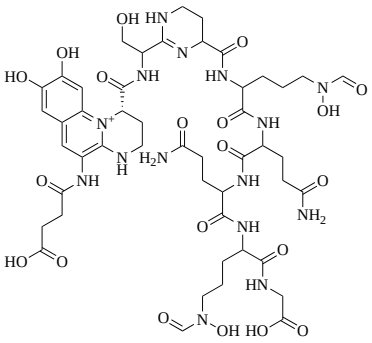
Scabichelins: Scabichelin

F



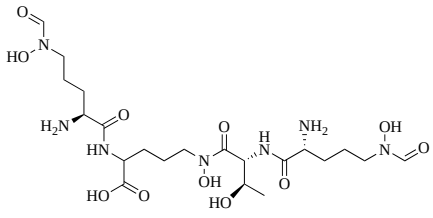
Salinichelins: Salinichelin A

G



Pyoverdins: Pyoverdin Pa6

H

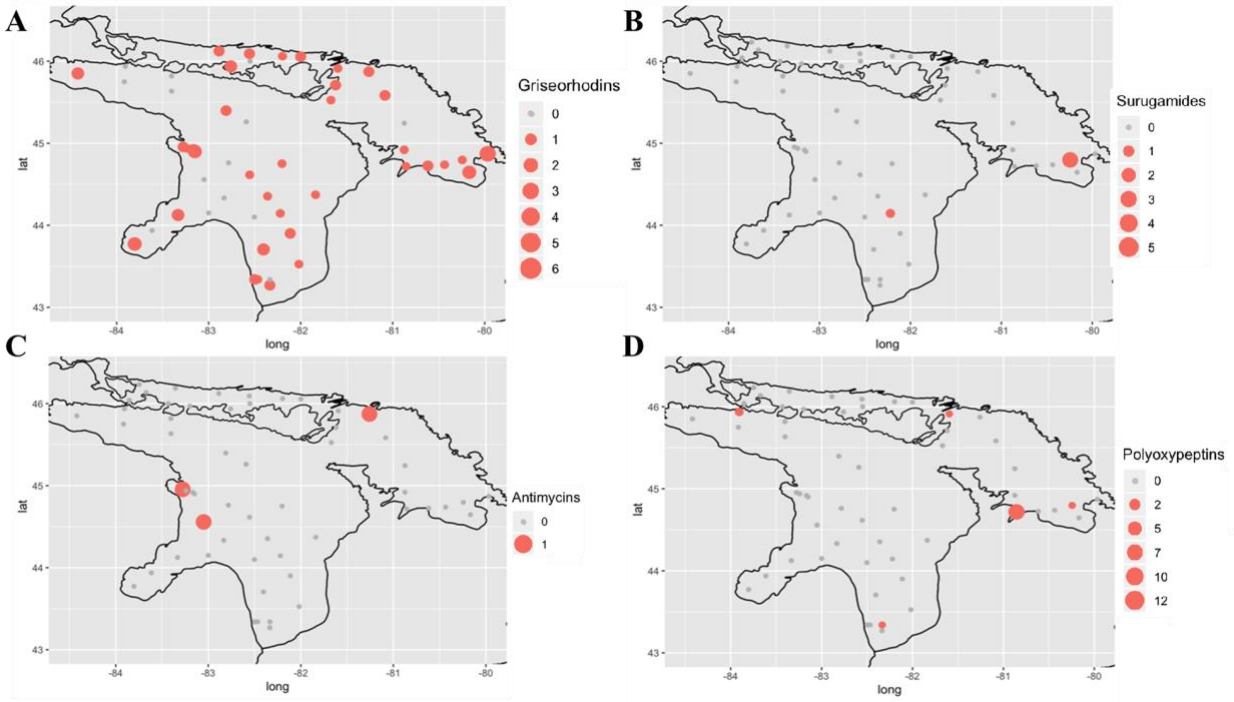


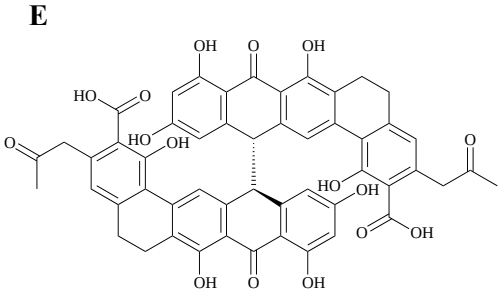
Coelichelins: Coelichelin

Figure 8. A-D show the domain sequence distribution of select siderophores across Lake Huron sediment: scabichelins, salinichelins, pyoverdins, and, coelichelins, respectively. Different sized circles represent thresholds of sequence read abundance at each collection site in Lake Huron. E-H show representative structures from each of the four siderophore classes. No discernable distribution pattern was observed among eight siderophores analyzed (Figure 22).

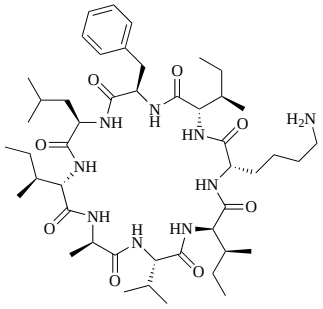
From the 30 known and mapped OBUs, a group of 11 bioactive NPs classes (such as anticancer and antiviral compounds) were also subjected to the same analysis. Figure 9A-D shows the domain sequence distribution of select types of bioactive NPs across Lake Huron sediment: griseorhodins, surugamides, antimycins, and, polyoxypeptins, respectively. Griseorhodins are members of the rubromycin family that inhibit HIV reverse transcriptase and human telomerase.[118] Surugamides are cyclic octapeptides that inhibit cathepsin B.[119] Antimycins are depsipeptides, consisting of a macrocyclic ring that has been reported to have generally cytotoxic bioactivities, including antifungal, insecticidal, and nematocidal properties.[120,121] Finally, polyoxypeptins are also bioactive cyclic depsipeptides that were shown to induce apoptosis in human pancreatic carcinoma.[122] Sequence reads for griseorhodins-type NPs (Figure 9A) were detected in 39 geographic locations across the lake. In contrast, the cyclic depsipeptides such as surugamides (Figure 9B), antimycins (Figure 9C) and polyoxypeptins (Figure 9D) appear less frequently. In general, the profiles among the eleven bioactive NPs tested exhibited no obvious distribution patterns in lake sediment (Figure 23).

Figure 9. Domain sequence distribution of select bioactive NP classes across Lake Huron sediment and representative structures from each of the four siderophore classes.

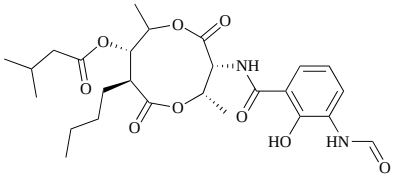




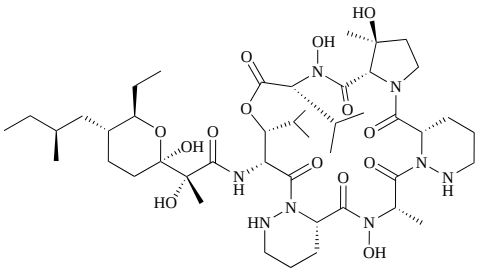
Griseorhodins: Griseorhodin D1/D2



Surugamides: Surugamide D



Antimycins: Antimycin A3



Polyoxypeptins: Polyoxypeptin B

Figure 9. A-D show the domain sequence distribution of select bioactive NP classes across Lake Huron sediment: griseorhodins, surugamides, antimycins, and, polyoxypeptins, respectively. Different sized circles represent thresholds of sequence read abundance at each collection site in Lake Huron. E-H show representative structures from each of the four bioactive NP classes. No discernable distribution pattern was observed among eleven bioactive NP classes analyzed (Figure 23).

3.3.3 Interpretation of NP distribution Profiles Across Lake Huron Surface Sediment.

This study aimed to generate a preliminary assessment of how NPs are distributed in the environment. As partly shown in Figures 7-9, among the select 30 characterized OBUs that were analyzed, no discernable patterns of distribution in Lake Huron surface sediment were observed. Some NP OBUs exhibited frequent occurrences in sediment across the geographic locations sampled (*e.g.* the pyoverdins in Figure 8D and the griseorhodins in Figure 9A), while others were confined to select sample sites (*e.g.* the antibiotics in Figures 7A-D, among others).

In context of applying this knowledge toward designing sample collection expeditions in Lake Huron, these results are quite preliminary, as they are among few attempts to document distribution of specific classes of NPs across an environment representative of a potential collection expedition. Given the limitations of this study (discussed below), the observed NP distribution profiles lend experimental evidence to a few predictable phenomena, that to the best of our knowledge have yet to be demonstrated on a large scale. First, some profiles, particularly those that represent bioactive NPs (antibiotics, anticancer, etc), occur selectively across Lake Huron sediment. To speculate, this suggests that the NPs with scarce distribution profiles either occur coincidentally within the given microbial species in a manner that is unrelated to their

ecological survival, or that the NPs serve a particular function that is required in that specific geographic location. Second, some profiles occur more frequently across the 58 and 53 collection sites (A and KSα, respectively), such as the pyoverdins and griseorhodins. This may suggest that either the NP contains a common ecological function, is located on a mobile genetic element that is commonly transferred between species, or both. Regardless, of the 30 NP profiles analyzed, none provide evidence for discernable patterns of NP occurrence across Lake Huron sediment, and suggest that sampling an environment of this magnitude should be as frequent as possible and limited by budgetary constraints and sample processing capacity, and not by a fear of collecting too many samples that contain redundant NP capacity.

Only 30 characterized OBUs were discussed in this study. However, the vast majority of OBUs observed (98% or 1,172 KSα domain OBUs and 93% or 90,471 A domain OBUs) could not be assigned to known compound classes. When calculating correlation coefficients (Table XIV), some of these uncharacterized OBUs display a strong positive correlation (a correlation score of 0.999987 reported as 1 in Table XIVB) in their distribution patterns (Table XIV). This suggests that some NPs may co-occur in the environment, and may provide evidence of either phylogenetic or ecological forces that drive regional NP distribution.

To assess whether one particular site was a frequent hotspot for antibiotics (or other NPs), the Shannon diversity index for each sample site and for each domain was computed before and after rarefaction to the sample number of sequences per sample (Table XV). The Shannon index is a common measure of diversity used in ecological studies that takes into account species richness and evenness.[80] In general, samples with a relatively high number of OBUs also had a relatively high diversity index, indicating a greater OBU diversity and abundance. This suggests that hotspots for NP diversity might exist. However, since not all samples exhibiting a high number of OBUs

displayed high Shannon indexes, this suggests that OBU frequency does not necessarily correlate with OBU diversity.

Nevertheless, in order to claim that distinct OBU patterns occur in nature, many more comprehensive analyses of OBU distribution, in addition to an enhanced understanding of the frequency and impact of horizontal and vertical gene transfer of NP BGCs, is needed. Our study provides preliminary evidence to suggest that NPs are not all ubiquitously occurring, nor do they follow obvious occurrence patterns. However, surface-level analysis of the majority of OBUs detected in Lake Huron sediment, which belong to unknown/uncharacterized NPs, indicate that co-occurrence patterns may indeed exist in the environment.

It is also important to note that the maps generated in Figures 7-9 represent less than 0.06% of sediment collected from a 216 cubic inch PONAR grab. Despite homogenization of surface sediment upon collection and processing, microbial diversity, and thus NP diversity, may very well differ even within one block of sediment, particularly as the environment becomes more microaerophilic/anoxic with increasing depth. A similar study performed on one PONAR of sediment would provide valuable information on the available NP chemical space within. Therefore, results presented herein should be viewed as a seasonal snapshot of one collection trip that occurred in the summer of 2012, where 59 surface sediment samples were collected across diverse locations in Lake Huron (58 were analyzed for A domain diversity, and 53 were analyzed for KSα domain diversity). Events (i.e., algal blooms or other localized environmental phenomena at the time of collection) could have influenced results from any of the locations. A vast expanse of studies would be required to make broad claims of NP distribution patterns, however these studies become more likely as the repertoire of available BGCs in databases is expanded and as

new and improved primers are designed, which capture a more comprehensive array of NP diversity.

There are a few experimental limitations to the current study. First, the low abundance of sequence reads belonging to NPs can be attributed to limited eDNA extracted from sediment and biases generated from PCR amplification using highly degenerate primers. In addition, the resulting amplicons are only partially representative of the BGC population present in sediment: (1) the eDNA extraction step is biased towards non-spore forming bacteria, (2) the primers used in this study target a limited range of bacterial taxa, since they were designed specifically for Actinobacteria sequences (A domain primers) or a small subset of Actinobacteria, such as Streptomyces spp. (KSα domain primers), and (3) PCR amplification itself yields a distorted representation of the true distribution of gene targets. Yet, these primers and PCR conditions are currently commonly used to evaluate BGC diversity in eDNA from various environments. The design of new, more inclusive primers will be vital for the discovery of new non-traditional BGCs. Similarly, alternative, non-PCR-based approaches may also be necessary. Such approaches include deep shotgun metagenome sequencing coupled with long-read sequence data (e.g. Oxford Nanopore, PacBio, Loop Genomics), or enrichment sequencing (e.g., Oxford Nanopore selective sequencing, hybridization capture+shotgun metagenome sequencing). Finally, the MIBiG database was used to assess molecular classes.16 The number of existing NPs greatly outnumbers the number of entries in MIBiG, underlining the limitation of existing databases to identify NPs. The expansion of entries in NP databases to more fully document BGCs in the environment, along with their associated molecular products is critical toward evaluating existing NP structural diversity.

3.4 Conclusion

Despite decades of collecting soil microorganisms for use in drug discovery, few attempts have been made to measure the extent to which NP production genes are distributed in the environment. In this study, domain amplicon sequencing was used to document distribution profiles of 30 NPs across Lake Huron surface sediment collected from 59 locations. Overall, no discernable patterns in NP distribution was observed when comparing OBUs from 30 NP classes. In some instances, NP BGC domains appeared more frequently across the lake (e.g., griseorhodins, pyoverdins, among others), while other instances NP BGC domains were more scarcely detected (e.g. cyclomarins, rosamicins, among others). These results suggest that some NPs may be endemic to select geographic locations (perhaps due to unique environmental pressures), while others may be of cosmopolitan distribution, supporting the hypothesis that "some antibiotic gene clusters are cosmopolitan, while others have cameo roles."[95] Investigating BGC distribution patterns and dynamics in Lake Huron represents an essential first step towards a more methodical environmental sample collection approach and contributes to the design of future environmental samples collection expeditions, a great unmet need in NP drug discovery.

3.5 Methods

3.5.1 Collection of Sediment Samples, Cultivation of Sediment Bacteria on Nutrient Agar

Sediment samples were collected using a PONAR grab in the summer of 2012 from Lake Huron, the Georgian Bay, and the Northern Channel during a research expedition aboard the EPA's Lake Guardian Research Vessel. Surface depths of sediment are listed in Table XII. Approximately 1 cm^3 of sediment was homogenized, and an aliquot was placed into a 2 mL cryovial containing

20% glycerol. They were stored in cryogenic vials in a Dewar until transported back to the laboratory where they were stored in a -20°C freezer.

3.5.2 Genomic DNA Isolation from Sediment and Nutrient Agar

Cryogenic vials were thawed at room temperature, and eDNA was extracted from approximately 0.25 g of sediment, using the PowerSoil DNA Isolation Kit (now called DNeasy PowerSoil Kit, Qiagen, Netherlands) according to the manufacturer's instructions.

3.5.3 KSα and A Domain Amplification and Sequencing

KSα and A domain amplicon sequencing was performed using the same two-step PCR strategy described above. A 613 bp fragment of the KSα (β-ketoacyl synthase) was amplified using degenerate oligonucleotides (5′-TSGCSTGCTTCGAYGCSATC-3′) and (5′-TGGAANCCGCCGAABCCGCT-3′).[63] 700-bp NRPS A domain gene fragments were amplified using degenerate oligonucleotides A3F (5′-GCSTACSYSATSTACACSTCSGG-3′) and A7R (5′SASGTCVCCSGTSCGGTAS-3′).[64] All primers contained a locus-specific sequence as well as a universal 5′ tail (i.e., CS1 and CS2 linkers). 20 μL of PCR reaction mixture consisted of 1 μL of DNA, 1 μL of a 10 μM solution of each primer, 10 μL KAPA Taq 2X ReadyMix (Kapa Biosystems), 0.8 μL of DMSO, 3.2 μL of 100 mg mL^{-1} Bovine Albumin Serum, and 3 μL of DI water. The thermal cycling conditions were set to an initial denaturation step at 95 °C for 5 min; 7 cycles of 1 min at 95 °C, 1 min at 65 °C (annealing temperature was lowered 1 °C per cycle), and 1 min at 72 °C; and 40 cycles of 1 min at 95 °C, 90 s at 58 °C and 1 min at 72 °C; and a final elongation step at 72 °C for 5 min. Amplification products were detected on agarose gel electrophoresis and purified using Qiagen's QIAquick PCR cleanup kit according to the manufacturer's protocol (Qiagen, Inc.). The resulting PCR amplicons were used as templates for

the second PCR step, as described above, to incorporate sequencing adapters and sample-specific barcodes. Pooled and purified amplicon libraries, with a 20% phiX spike-in, were loaded onto a MiSeq V3 flow cell, and sequenced using paired-end 2 × 300 reads.

3.5.4 Bioinformatic Analyses of BGC Data

Only forward reads were used in further analysis due to the low quality of reverse reads. All sequences generated from the Illumina MiSeq sequencer were 6-frame translated into amino acid sequences using TranslatorX.[89] Only frames with no internal stop codons were kept using TranslatorX's "guess most likely reading frame" option. Amino acid sequences were then filtered via HMMER[90] using HMM prebuild generic detection models downloaded from antiSMASH v5.0.0.[101] The following models were used: AMP-binding and A-OX for A domain, and t2ks and t2pks2 for PKS type II. Only sequences that passed the default e-value thresholds were kept. Sequences were then clustered at 80%, 85%, 90%, and 95%. A taxon-by-sample abundance matrix (a feature table or biological observation matrix, BIOM)[88] file was then created. The 85% similarity threshold was selected for subsequent analyses and the OBU representative sequence was annotated with its BLAST identity only if the pairwise identity was at least 85% and coverage over at least 84 amino acids. An OBU-by-sample BIOM file was then created and rarefied to the minimum number of sequences within samples.

3.5.5 Bioinformatic Method Validation Using Reference Strains

Two strains were included in wet lab and bioinformatics analysis to ensure clustering methods and molecular identities were valid: *Micromonospora* strain B006 and *Streptomyces Coelicolor* A3(2). These strains were subjected to the same amplification procedure using the generic primers.

Amplicons were sequenced using the same sequencing strategy used previously, and sequence data was subjected to the same bioinformatics procedures used.

3.5.6 Acknowledgments

The authors wish to acknowledge the following contributors: A. Li (UIC), K. Rockne (UIC), S. Carlson (formerly UIC), and crew of EPA RV Lake Guardian for assistance with sediment collection; G. Chlipala of UIC's Core for Research Informatics and A. Naqib of UIC's DNA Sequencing Core for assistance processing data; C. Clark for assistance with data mapping. This publication was supported by Vahlteich Scholar research funds, the Illinois–Indiana Sea Grant, the Office of Technology Management at UIC, and UIC startup funds.

4. CONTRIBUTION TO OTHER NP RESEARCH PROJECTS

4.1 Deuteromethylactin B from a Freshwater-derived *Streptomyces* sp.

Adapted with permission from Shaikh, A.F., Elfeki, M., Landolfa, S.; Tanouye, U.; J. Green, S.J.; and Murphy, B.T. Deuteromethylactin B from a Freshwater-derived *Streptomyces* sp. *Nat Prod Sci.* **2015.**

Terrestrial actinomycete bacteria are among the most prolific sources of NPs, accounting for approximately 7000 of the compounds reported in the Dictionary of NPs (insert source). Many of the well-studied genera in the terrestrial sphere, such as the *Streptomyces* genus, have been explored for their bioactive molecules. However, the focus on terrestrial actinomycetes began to decline as researchers continuously isolated known bioactive chemical scaffolds. For this reason, discovery efforts have focused on expanding the access to a wide diversity of unique and underexploited habitats in order to ensure the isolation of novel Actinobacterial species that potentially produce novel compounds.

Among these novel habitats, freshwater environments have been underexplored. Since freshwater systems harbor distinct environmental selection pressures and growth conditions, it was suspected that globally endemic freshwater microbial populations existed. Indeed, several studies have shown that some freshwater-derived actinomycetes were taxonomically distinct in comparison to their terrestrial and marine counterparts (e.g., the acI – acIV clades), which are well summarized by Newton R. J. et al.[123] In this study, we explored the capacity of a freshwater-derived sediment actinomycete bacterium to produce novel secondary metabolites. We used spectroscopic and chemical derivitization techniques to characterize a class of cytotoxic lactones and its corresponding novel, unnatural degradation product. Furthermore, we presented a brief

analysis of the bacterial community that existed in four locations of Lake Michigan sediment, and assessed the corresponding cultivatable actinomycete populations.

4.1.1 Cultivation-independent analysis of Actinobacteria in Lake Michigan sediment

In order to assess the population of actinomycetes in Lake Michigan sediment, we collected four sediment samples off the coast of Milwaukee, Wisconsin, from depths ranging between 56 and 145 m. We were interested in assessing the actinobacterial population present in sediment and the corresponding cultivatable actinomycete population. After extracting genomic DNA from sediment samples, we PCR amplified the DNA with primers targeting the V4 region of the small subunit 16S rRNA gene of bacteria. After next generation sequencing and subsequent bioinformatic analysis of sequences from all four locations, we assessed the distribution of bacterial communities in sediment. Sequences derived from members of the phylum Proteobacteria accounted for the majority (54%) of the reads, while sequences from Actinobacteria were approximately 11% of the total sequence library. Interestingly, according to cultivation-independent analysis within the phylum Actinobacteria, two of the most common families that account for the majority of actinomycete secondary metabolites were scarcely represented (*Streptomycetaceae* and *Micromonosporaceae*, 0.04 and 0.11% of total Actinobacteria reads, respectively) while 31% of reads were attributed to families that remain uncharacterized.

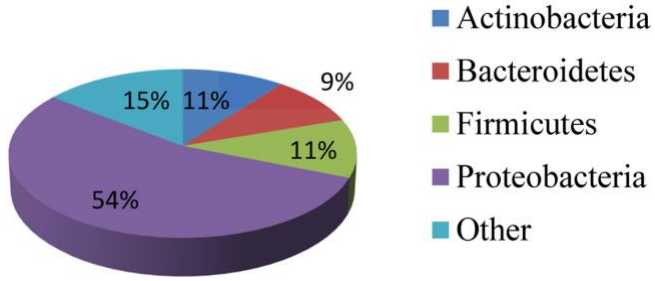


Figure 10. Composition of bacterial community in collected Lake Michigan sediment
Proteobacteria is the most abundant phylum in Lake Michigan surfact sediment (54% of all sequence reads), followed by Actinobacteria (15% of all sequence reads).

4.1.2 Phylogenetic analysis of cultivatable Lake Michigan actinomycetes

The majority of isolates obtained through cultivation of strains from the four Lake Michigan sediment samples were members of the families *Streptomycetaceae* and *Micromonosporaceae* (Figure 11).

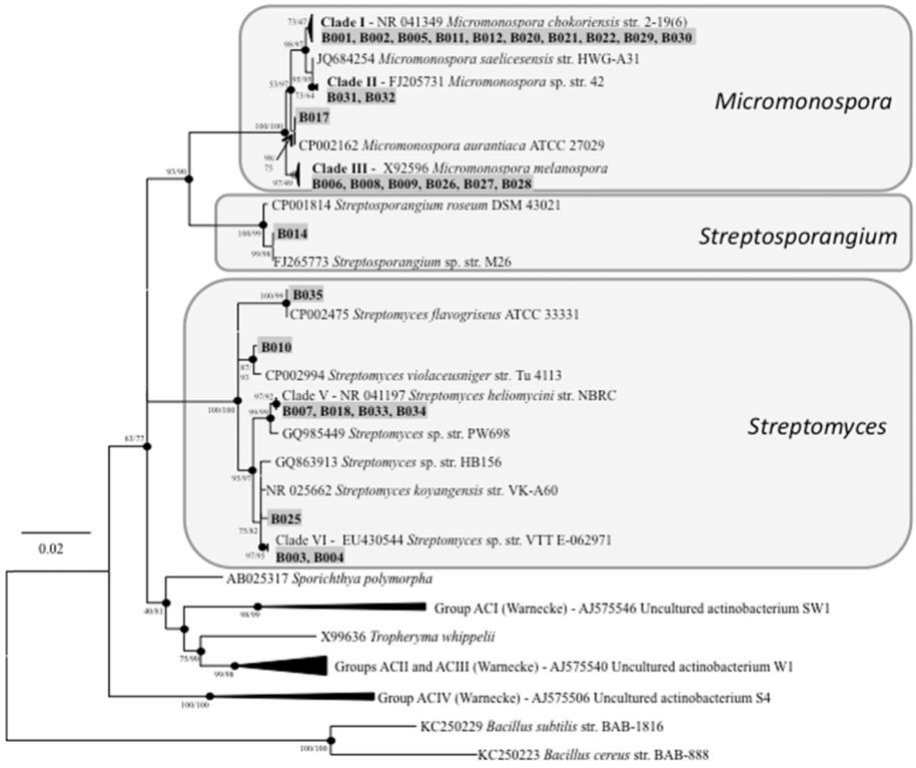


Figure 11. Phylogenetic analysis of cultivatable actinomycete isolates from Lake Michigan

Phylogenetic analysis of Actinobacteria isolates from Lake Michigan revealed that the organisms belonged to three different genera: *Micromonospora*, *Streptosporangium spp.* and *Streptomyces*. Most of the isolates (19 of 29) belonged to the genus *Micromonospora*. No organisms with rRNA

genes similar to so-called “lake Actinobacteria” (i.e. clades acI-acIV) were detected (Warnecke, F.; Amann, R., 2004). Actinobacterial lineages that so far are known as exclusive to freshwater environments have primarily been detected through cultivation-independent molecular assays, and further efforts will be needed to isolate representatives of these more divergent lineages. However, it is important to note that traditionally, the process of isolating actinomycetes from petri dishes for drug-lead discovery has been heavily biased toward larger, spore-forming colonies with specific morphological features, and in our case this collection of strains does not represent the true population that exists on the plates. Other studies from our lab have sought to address this obstacle.[124]

Although many isolates had identical or nearly identical SSU rRNA gene sequences, they did not necessarily have similar profiles of secondary metabolite production. For example, Strain B006 is a producer of the diazaquinomycin antibiotic class.[125] Five strains clustered together with strain B006 (B026, B009, B008, B027, and B028), but via LCMS analysis of fermentation extracts only B006 and B026 produced that antibiotic class. This observation highlights the complexity of using taxonomic uniqueness to guide the discovery of novel chemistry.

Even though the isolates recovered in this study were highly similar by 16S rRNA gene sequence to previously isolated organisms, we were still able to identify strains with novel secondary metabolites. Of the strains classified under *Streptomycetaceae*, we identified B025 as being distinct from its *Streptomyces* counterparts. Upon fermentation it was found to produce a rare class of eight-membered lactone secondary metabolites, a member of which has been investigated for its antitumor properties. Anam F. Shaikh completed the characterization of octalactin B in addition to its corresponding novel, unnatural degradation product using spectroscopic and chemical derivitization techniques.[126]

4.1.3 Methods

Note: methods highlighted here reflect contributions of the author of the dissertation

4.1.3.1. DNA extraction and cultivation-independent analysis of bacterial communities from Lake Michigan sediment samples

Each of the four sediment samples used in these analyses was collected using PONAR, from the RV Neeksay (Dr. Russell Cuhel, University of Wisconsin-Milwaukee). The location of each sediment sample is presented in Table III.

Table III. Coordinates of sediment samples

Sample	Longitude	Latitude	Depth
1	43°21'3.16" N	87°34'11.63" W	145 m
2	43°16'48.00" N	87°34'12.55" W	80 m
3	43°16'12.76" N	87°34'12.22" W	56 m
4	43°13'27.63" N	87°34'10.62" W	56 m

The top 1 cm of sediment was used for analysis. Genomic DNA was extracted from all sediment samples in duplicate using the PowerSoil® DNA Isolation Kit. Fragments of microbial small subunit ribosomal RNA (SSU or 16S rRNA) genes were PCR amplified from genomic DNA using primers 515F and 806R, as described previously.[88] Following pooling and cleanup of samples, the final pool was loaded onto an Illumina MiSeq sequencer, employing V2, 2x250 read chemistry.[86,88]

4.1.3.2 Analysis of microbial community composition

Sequence data were initially processed using the software package CLC Genomics Workbench (CLC Bio, Qiagen) to merge forward and reverse reads, trim poor quality data, and remove primer sequences. Sequences were then processed using the software package QIIME to remove chimeric sequences, perform clustering and annotation. Briefly, sequences were screened for chimeras using

the Usearch61 algorithm and putative chimeric sequences were removed from the dataset.[86,91] Subsequently, each sample sequence set was sub-sampled to 9,000 sequences per sample to avoid analytical issues associated with variable library size.[59] Sub-sampled data were pooled, renamed and clustered into operational taxonomic units (OTU) at 97% similarity. Representative sequences from each OTU were extracted, and these sequences were classified using the "assign_taxonomy" algorithm implementing the uclust consensus taxonomy assigner, with the Greengenes reference OTU build.[88,127] A biological observation matrix (BIOM) was generated at taxonomic levels from phylum to genus using the "make_OTU_table" algorithm. The BIOMs were imported into the software package Primer6 for analysis and visualization.[128] Figures were generated using the software package Origin Pro8.5 (OriginLab Corporation, Northampton, MA).

4.1.3.3 Isolation of actinomycete strains from Lake Michigan sediment

Sediment samples were placed in glass vials and treated with heat (55ºC for 6 minutes), vortexed, and inoculated onto agar plates (50 µL) on five different types of solid media (A1, M1, ISP1, 1/10th ISP2, and LWA – filtered Lake Michigan water and agar; each containing 28 µM of the antifungal agent cycloheximide; Table X). Actinomycete colonies appeared between two and four weeks and upon observation of branched hyphae or sporulation, individual strains were isolated with sterile toothpicks and re-plated on A1 media to assess their purity.

4.1.3.4 Phylogenetic analysis of cultivatable Lake Michigan strains

Genomic DNA was extracted from individual colonies using the MasterPure Gram Positive DNA Purification (Epicentre) kit. Near-full length 16S rRNA genes were PCR amplified from gDNA extracts using the primers 27F-1492R, as described previously.[59] Sequencing was performed on an ABI 3730XL DNA Analyzer Sequencer at the Sequencing Core at the University of Illinois at Chicago. SSU rRNA gene sequences of isolates recovered in this study, and those of

the most similar sequences were aligned using the software package Greengenes.[129] This alignment was imported into the software package ARB, and filtered using the Actinobacterial conservation filter, removing from the analysis positions where fewer than 50% of the sequences shared the same base.[130] This filtered alignment was imported into the phylogenetic software package MEGA5 and into the software package MrBayes v3.1.2 for phylogenetic tree construction.[131,132] Neighbor-joining phylogenetic trees were constructed with aligned sequences using the maximum composite likelihood substitution model with complete deletion of gapped positions. The robustness of inferred tree topologies was evaluated by 1000 bootstrap resamplings of the data. For maximum likelihood trees, the general time reversible substitution model was employed, with complete deletion of gapped positions, and 1000 bootstrap re-samplings of the data. Additionally, Bayesian analyses were performed on the aligned sequence data by running five simultaneous chains (four heated, one cold) for ten million generations, sampling every 1000 generations. The selected model was the general time reversible (GTR) using empirical base frequencies and estimating the shape of the gamma distribution and proportion of invariant sites from the data. A resulting 50% majority-rule consensus tree (after discarding the burn-in of 25% of the generations) was determined to calculate the posterior probabilities for each node. The split-differential at ten million generations was below 0.01.

Figure 11 depicts a phylogenetic tree reflecting the relationships of SSU rRNA gene sequences from select isolates. The tree topology was obtained from a bootstrapped neighbor-joining analysis, as described in the text. Nodes for which bootstrap values equaled or exceeded 70% are indicated by a numerical value. The bootstrap value derived from maximum likelihood analysis is also indicated (NJ/ML). Nodes supported by Bayesian analysis, with posterior probability values greater than 95%, are indicated with black circles. Nodes with posterior probability values greater

than 70% are indicated with gray circles. Polytomies indicate branching points that were not consistently supported by bootstrap or Bayesian analyses. The scale bar indicates 0.02 substitutions per nucleotide position. Isolates are highlighted in gray.

4.1.3.5 Phylogenetic analysis of strain B025

Strain B025 (GenBank Accession number KM678242) was isolated from a sediment sample collected from Lake Michigan. It shared 99% 16S rRNA gene sequence identity with the most closely related type strain, *Streptomyces koyangensis* (GenBank accession number NR025662).[133]

4.2 Sharing and community curation of mass spectrometry data with Global NPs Social Molecular Networking

Adapted with permission from Wang, M., et al. Sharing and community curation of mass spectrometry data with Global NPs Social Molecular Networking. *Nat Biotechnol* **2016.**

Mass spectrometry is heavily utilized in NP research laboratories for structure elucidation and sample analysis. However, mass spectrometry datasets can be too large for manual analysis. Furthermore, comprehensive software and proper computational infrastructure are not readily available and only low throughput sharing of either raw or annotated spectra is feasible, even among members of the same lab. The potentially useful information in tandem (MS/MS) datasets can thus remain buried in papers, laboratory notebooks, and private databases, hindering retrieval, mining, and sharing of data and knowledge. As a response, Dr. Mingxun Wang et al. created Global NPs Social Molecular Networking (GNPS, available at gnps.ucsd.edu). GNPS is a data-driven platform for the storage, analysis, and knowledge dissemination of MS/MS spectra that enables community sharing of raw spectra, continuous annotation of deposited data, and collaborative curation of reference spectra (referred to as spectral libraries) and experimental data (organized as datasets).

GNPS provides the ability to analyze an MS/MS dataset and to compare it to all publicly available data. By building on the computational infrastructure of the University of California San Diego (UCSD) Center for Computational Mass Spectrometry (CCMS), GNPS provides public dataset deposition/retrieval through the Mass Spectrometry Interactive Virtual Environment (MassIVE) data repository. The GNPS analysis infrastructure further enables online dereplication, automated molecular networking analysis, and crowdsourced MS/MS spectrum curation.

Through an enormous effort by Wang et al, researchers across the globe contributed to the >93 million spectra in >250,000 private LC/MS runs available on GNPS. The data provided at the time of publication are provided in Table IV.

Table IV. Sample experimental data

Compound Name	Ion Source	Instrument	Scan number	Molecular weight	Precursor MZ	Ion Mode	Adduct	Compound Source
Antimycin A6	APCI	IT-TOF	1900	478	477.0501	Negative	M-H	Isolate
Kitamycin A	APCI	IT-TOF	1320	464	463.8819	Negative	M-H	Isolate
N-formylantimycic acid methyl ester	LC-ESI	IT-TOF	520	296	295.351	Negative	M-H	Isolate
Antimycin A2b	LC-ESI	IT-TOF	384	534	533.7116	Negative	M-H	Isolate
Diazaquinomycin A	LC-ESI	IT-TOF	1264	354	355.4607	Positive	M+H	Synthetic
Diazaquinomycin E	LC-ESI	IT-TOF	1169	396	397.1839	Positive	M+H	Isolate
Diazaquinomycin H	LC-ESI	IT-TOF	781	382	369.7575	Positive	M+H	Isolate
Diazaquinomycin I	LC-ESI	IT-TOF	916	396	397.8423	Positive	M+H	Isolate
Diazaquinomycin C	LC-ESI	IT-TOF	1059	382	383.1668	Positive	M+H	Isolate
Diazaquinomycin D	LC-ESI	IT-TOF	891	410	411.1632	Positive	M+H	Isolate
Diazaquinomycin F&G	LC-ESI	IT-TOF	975	368	369.5294	Positive	M+H	Isolate
Resistomycin	LC-ESI	IT-TOF	1437	376	375.6834	Negative	M-H	Crude
Resistomycin	LC-ESI	IT-TOF	1440	376	375.4659	Negative	M-H	Crude
Piericidin A	LC-ESI	IT-TOF	1146	414	414.9027	Negative	M-H	Isolate
Indolmycin	LC-ESI	IT-TOF	689	257	258.6271	Positive	M+H	Isolate
HMP-M2	LC-ESI	IT-TOF	772	366	365.2758	Negative	M-H	Isolate
Geldanamycin	LC-ESI	IT-TOF	931	560	559.0659	Negative	M-H	Isolate

GNPS provides a community-led knowledge space in which NP data can be shared, analyzed and annotated by researchers worldwide. It enables a cycle of annotation, in which users curate data, continuous dereplication for product identification, and houses a knowledge base of reference spectral libraries and public datasets.

4.2.1 Methods

4.2.1.1 General Experimental Procedures

All samples were dissolved into methanol. HRESIMS data were obtained on a Shimadzu ion trap-time of flight (IT-TOF) spectrometer at the University of Illinois at Chicago Research Resources Center (UIC RRC). All samples were analyzed using reversed-phase C18 (RP-C18) HPLC, equipped with a photodiode array detector (PDA). The first 10 min of the run was a gradient from

10% aqueous MeOH to 100% MeOH, followed by an isocratic flow of 100% MeOH for 10 min. A flow rate of 0.5 mL/min was employed. A Phenomenex Luna C18 (2), 100 × 4.6 mm, 5 μm pore size column was used. Spectra were exported as an mzXML file formal before submission into GNPS.

4.3 Antibiotic resistance genes show enhanced mobilization through suspended growth and biofilm-based wastewater treatment processes

Adapted with permission from Petrovich, M. et al. *FEMS Microbiol Ecol.* **2018.**

Antibiotic resistance is an urgent challenge in modern public health, and is exacerbated by the prevalent use of antibiotics in livestock operations and medicine.[134–140] Wastewater treatment plants (WWTPs) are known to harbor antibiotic resistance genes (ARGs) and can be significant point sources of ARG input into receiving water bodies.

In this study led by previous graduate student Morgan Petrovich, we employed a shotgun metagenomics approach to assess dynamics of ARGs, antibiotic production genes (APGs) and mobile genetic elements (MGEs) in two full-scale WWTPs that employ two different secondary treatment process variations. Furthermore, co-occurrences between ARGs and corresponding classes of APGs in different types of WWTP bioreactors were investigated to test whether positive correlations exist between ARGs and APGs.

The author of the dissertation's input in this study consisted of providing Dr. Morgan Pertrovich a curated list of amino acid sequences containing 11 different antibiotic classes of interest from the MIBiG database.[100] This list was used to detect APGs in the samples using Blastp. Of the 11 APG classes studied, 9 were detected in the suspended growth system, and 6 were found in the biofilm system (Figure 12). Throughout both WWTPs, aminoglycoside and rifamycin-like APGs were

most abundant. It should be noted that aminoglycoside-like APGs are the most highly represented of all APG classes in the MIBiG database, which likely contributes to their high abundances detected in the WWTP samples. However, rifamycin-like APGs are only the fourth most highly represented, after aminoglycoside, beta-lactam and streptogramin-like APGs (in that order).

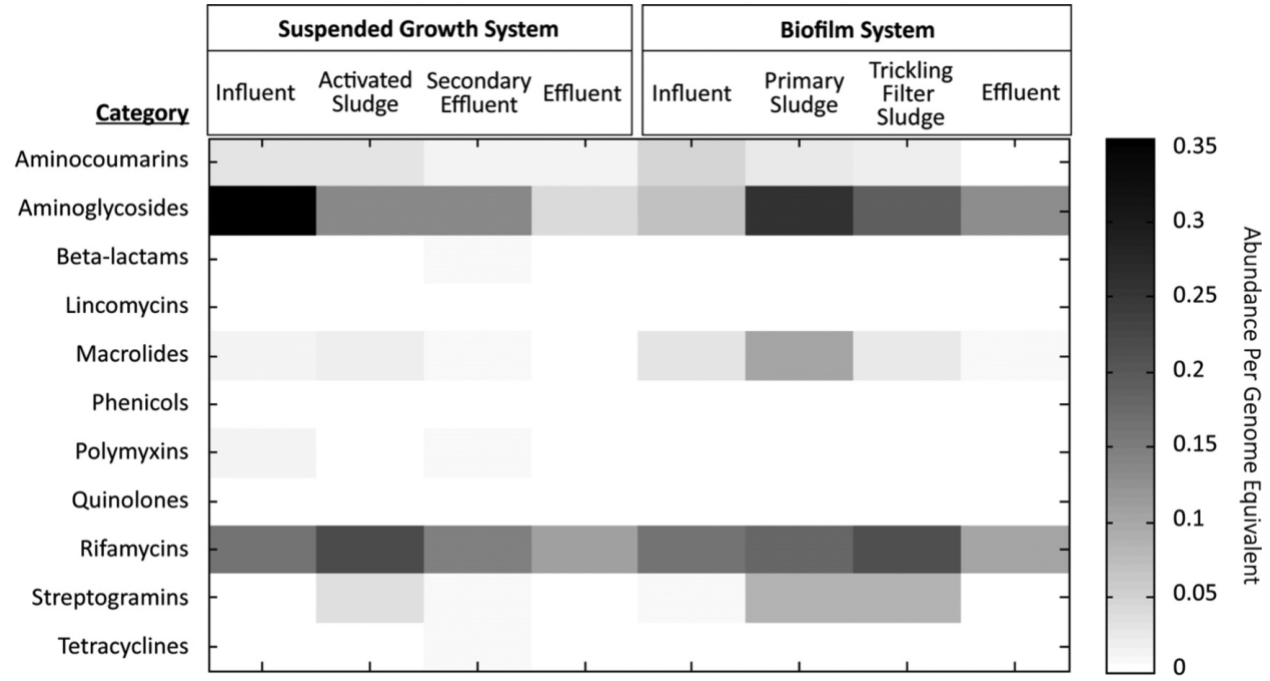


Figure 12. Heatmap of antibiotic production gene (APG) abundances for Suspended Growth and Biofilm Growth WWTPs.
Abundances are based on coverage of gene alignments to genes using the MiBIG database normalized to average coverage of single copy genes.

Positive associations between presence of ARGs and APGs were identified for the aminoglycoside antibiotic class in the suspended growth system and for the streptogramin antibiotic class in the biofilm system, suggesting that *in situ* production of some types of antibiotics may lead to selection for ARGs conferring resistance to those compounds.

4.3.1 Methods

To detect APGs in the samples, genes were compared with Blastp to the MIBiG[68] amino acid database containing 11 different antibiotic classes of interest. Cutoffs of e-value $\leq 10^{-10}$, % identity ≥70% and bit score ≥50 were used.

5. CONCLUSION AND PERSPECTIVES

5.1 Conclusion

Environmental microorganisms continue to serve as a major source of bioactive NPs and as an inspiration for many other scaffolds in the toolbox of modern medicine. Importantly, nearly all of these microbial NP-inspired therapies resulted from field expeditions to collect samples from the environment, arguably one the most important steps in NP drug discovery. Yet, the field of NP drug discovery faces major challenges due to the serendipitous nature of collection expeditions and the limited knowledge of NP genes present in both sediment and cultivatable bacterial populations. Furthermore, few attempts have been made to measure the extent to which common cultivation techniques have accessed existing chemical space. We addressed these challenges by exploring uncultivated and cultivated sediment biosynthetic potential for NPs using amplicon sequencing as a strategy to identify chemical space and to discover NP production genes in samples collected in Lake Huron.

In Chapter 2, we cultivated a minority of the sediment microbial population, which consequently yielded a relatively small percent of the NP biosynthetic capacity that existed in the sediment samples. This explains in part why traditional discovery programs face increasing rediscovery rates. Additionally, even though the predominant taxa on nutrient agar represent some of the major producers of bacterial NPs, the estimated BGC recovery percentages reported in Chapter 2 (23.3% for KS, 7.7% for KSα, and 15.8% for A domains) indicate that these taxa harbor a minority of the chemical space present in the sediment. Furthermore, most of these domains in both cultivatable and in sediment bacteria have yet to be characterized in peer-reviewed literature. It is possible that some of these sequences encode for novel compounds, or that the gene clusters

produce known compounds that have yet to be characterized in peer-reviewed literature or deposited into the MIBiG database.[100,103] . Taken together, these results suggest that there remains a large untapped chemical diversity in *both* readily cultivatable and total sediment bacterial populations. Finally, we were not able to comment on BGC recovery percentages in context of comparing them to others in the field since we were not able to find a precedent in the literature of a study specifically quantifying NPs from nutrient agar plates.

In Chapter 3, we assessed the NP BGC domain diversity from 59 sediment samples collected across a nearly 60,000 square kilometer geographic area of Lake Huron. We identified 30 known compound classes, none of which provide evidence for discernable patterns of NP occurrence across Lake Huron sediment. Some exhibited frequent distribution profiles across the geographic locations sampled, while others were confined to select sample sites. To speculate, a selective occurrence suggests the NPs either occur coincidentally within the given microbial species in a manner that is unrelated to their ecological survival, or that the NPs serve a particular function that is required in that specific geographic location. In contrast, frequently occurring NPs suggest that either the NP contains a common ecological function, is located on a mobile genetic element that is commonly transferred between species, or both. However, when we examined OBUs that were assigned to known compound classes – which constitute the majority of OBUs – we found correlations between different OBUs. This suggests that some NPs may co-occur in the environment, providing evidence of either phylogenetic or ecological forces that drive regional NP distribution.

Nevertheless, these results are quite preliminary, as they are among few attempts to document NP distribution of specific classes of NPs across an environment representative of a potential collection expedition. In order to claim that distinct OBU patterns occur in nature, many

more comprehensive analyses of OBU distribution (and of OBU pattern detection), in addition to an enhanced understanding of the frequency and impact of horizontal and vertical gene transfer of NP BGCs, is needed. We plan on including some these analysis for our future publication that will include data from Chapter 3.

5.2 Perspectives

5.2.1 Designing better PCR primers to increase coverage of amplified PKS and NRPS genes.

In NP drug discovery, culture-independent assessment of biosynthetic diversity present in samples relies on a PCR-based approach that relies on decade-old primers. Although PCR-based studies provide a rapid means of surveying a sample to screen soil directly without the need for cultivation, it has a few limitations. Thanks to their prolific abilities to produce NPs, members of the *Streptomycetaceae* and *Micromonosporaceae* families have been over-represented in NP drug discovery. PCR primers designed to amplify PKS and NRPS domains are thus biased towards members of these two families, or towards genera that have a high GC nucleotide content. For instance, the traditional KS primers used in this study and most other similar studies were designed using an alignment of known DNA sequences of eleven PKS gene clusters from Actinobacteria for the purpose of the rapid detection of PKS and NRPS genes specifically from major actinomycete lineages.[64] Moreover, the KSα primers were specifically designed to amplify genes from Actinobacteria. Their design was based on an alignment of sequences from *Streptomyces* spp. producing polyketide antibiotics, as these are species that are responsible for the production of many of the antibiotics currently employed in clinic.[63,141,142] Hence, the use of these primers was not indented for use with genera other than Actinobacteria. However, when testing these on a

selection of soil isolates, Metsä-Ketelä et al. appeared to have also amplified KSα sequences that do not belong to any of GenBank strains selected at the time of the study,[63] justifying the use of these primers for general screens like ours from sediment and cultivatable bacteria.

Yet, these primers are widely employed to this day, with Ayuso-Sacido's study cited 347 times,[64] of which 34 were in 2019 alone; Metsä-Ketelä's study was cited 176 times,[63] of which 13 were in 2019 alone, indicating the field's heavy use of these primers to study biosynthetic diversity. New and more inclusive BGC primers will enable a less biased screening of soil and sediment microorganism for their biosynthetic potential. One could argue that it is not possible to design one set of primers to encompass all KS domains, and that finding new primer will necessitate a different approach other than an alignment of known sequences. Rather, the biochemistry and the boundaries of these domains within their modules must be elucidated and understood for each phylum. Commonalities between KS architecture belonging to different phyla might allow the design of more inclusive primers, or a comprehensive primer library. These efforts will require more structural and evolutionary studies of secondary metabolism at the population level.[47,143]

The degenerate primers biases cannot fully account for the discrepancy between the number of KS/KSα domains amplified and the numbers of A domains amplified in Chapter 3. One explanation is that NRPs are a much bigger family of enzymes that employ a uniform mode for biosynthesis of peptides with an ability to activate more than 300 nonproteinogenic residues.[65] A domains are thought to be functionally independent of the enzyme, and the region identified in the design of A domain primers includes five conserved motifs, some of which are responsible for the domain's activity.[64,65,144,145] This ability to activate so many substrates, while maintaining core sequences responsible for the domain's activity, might explain the improved ability of the A

domain primers to amplify a large diversity of A domains. A similar depth of understanding of the various KS domains will allow a more effective design of novel KS primers.

5.2.2 Improving BGC annotation tools.

Even with their biases, most domains amplified using the primers in Chapters 2 and 3 were uncharacterized in the literature. Several studies also reported a large proportion of uncharacterized OBUs.[99,146–148] It is vital for the advancement of the field that a larger effort is spent on identifying these OBUs. This is because many research groups prioritize samples using genome mining strategies in which characterization of BGC products is predicted computationally.[149–152] Indeed, many remarkable tools have been designed to address these shortcomings. The MIBiG database, for example, was created due to a need for a centralized database that associates BGCs with their structural products, as this information is "usually buried inside the text of scientific articles".[100] Several other algorithms were developed for the annotation and prediction of products from these pathways.[47,101,149,153–156] However, some of these are designed for specific uses such as the detection of characterized classes, while others that identify clusters of unknown classes require a minimum sequence length of 1000 bps, a longer sequence than that amplified by the primers used in Chapters 2 and 3. Not many tools are designed to annotate metagenomic or mixed amplicon sequence reads with their structural products. To do so, most studies rely on using HMMs or Basic Local Alignment Search Tool (BLAST) against the MIBiG database, which only contains 2,036 entries.[100,157] Annotation of metagenomic amplicons generated from eDNA can be enhanced with expanded databases. Alternatively, a machine learning algorithm similar to HMMs can be developed using sequences from the MIBiG database for the purpose of associating shorter amplicon sequences with a molecular structure.

5.2.3 Exploring novel taxonomic space to accelerate the rate of novel compound discovery.

It has been estimated that only 0.1% of all bacterial strains have been cultured.[25] Developing innovative cultivation methods can open up new avenues to discover novel taxa and, by extension, discover new compounds. Indeed, several studies confirm this hypothesis. Technologies such as the isolation chip (Ichip) were developed to tap into not-yet cultured bacteria.[158] The Ichip contains permeable membranes that permit a sample to be cultured in situ, allowing colonies to be grown directly where they were found.[158] This technique had led to the discovery of teixobactin, the first of a new class of antibiotics, that was produced by a new species of beta-proteobacteria, named *Eleftheria terrae*.[74] Other innovative approaches were designed to overcome bacterial cultivation hurdles and screen for new compounds were also developed. One such example is the droplet-based microfluidic platform couple with mass spectrometry which allowed for the high throughput detection of novel antimicrobials.[159] Another example is the microfluidic picolitre droplets cultivation and sorting of more than 600,000 soil-derived Actinobacteria cultures per hour. Such powerful tools may enable the detection of novel promising compounds from soil microbes.[160]

Efforts to search for new sources in different environments such as specific niches (extreme environments or underexplored microbiomes), represent another promising expansion of NP discovery endeavors. These different environments can harbor useful compounds; a few examples include mangroves,[161] deep-sea vent microorganisms,[162] and iron-rich environments.[163] However, results from Chapter 2 indicate that novel cultivatable taxa readily exist on nutrient plates even from 'standard environments'. They simply haven't been picked due to biased morphology-based colony picking practices.[124]

Other groups focused on isolating less common species among chemically talented genera such as strains from the order *Myxococcales* within the myxobacteria[164] and rare marine actinobacteria[165,166] and Cyanobacteria whose investigation only began in the late 1970's.[16,167] Previously understudied and uncultivated phyla such as members of the Acidobacteria, Verrucomicobia and Gemmatimonadetes, and the candidate phylum Rokubacteria[75] also offer opportunities for natural product discovery. Indeed, Chapter 3 supports these claims, as members of the *Streptomycetaceae* and *Micromonosporaceae* families only accounted for about half of the OBU detected on nutrient plates. Focusing on strains that are not typically studied in drug discovery programs can be a successful strategy for the design of a novel drug discovery program, and it might result in decreasing the rediscovery rates of known compounds.

5.2.4 Computational and bioinformatics training for future NP scientists

NP drug discovery is a rapidly changing field. It evolved from requiring only minor microbiology skills (for strain isolation and growth) and chromatographic/spectroscopic expertise (for the isolation and structure elucidation of NPs), to requiring various skills, including but not limited to molecular biology for cloning and heterologous expression of biosynthetic gene clusters, evolutionary biology for phylogeny-based studies, and most importantly, bioinformatics skills for 'big' data analysis. Tools such as MALDI-TOF MS and sequencing data analysis require a minimum working knowledge of programming software such as R and python. However, current NP graduate programs do not require computational skills, yet, incoming graduate students with a computational background or experience are tremendously valued. Graduate programs in NP drug discovery would greatly benefit from the incorporation of computer science and data science into their curriculum to allow for this field to adapt to a rapidly evolving NP landscape. For example, new algorithms can be designed to predict the structural output from shorter amplicon sequences,

or implementation of statistical components into student projects can render stronger research outcomes, and can enable a student to avoid quantitative shortages in statistical analyses. Students can design pipelines or wrappers to implement pre-existing tools into their own analysis, or work with university IT systems administrators to incorporate data storage and sharing between different research groups. Incorporating data science topics that have not traditionally been discussed in NP drug discovery programs can have tremendous outcomes for both students and research groups.

CITED LITERATURE

(1) VARTIA, K. O. Chapter 17 – ANTIBIOTICS IN LICHENS. In *The Lichens*; 1973. https://doi.org/10.1016/B978-0-12-044950-7.50022-2.

(2) Solecki, R. S. Shanidar IV, a Neanderthal Flower Burial in Northern Iraq. *Science (80-.).* **1975**. https://doi.org/10.1126/science.190.4217.880.

(3) Dias, D. A.; Urban, S.; Roessner, U. A Historical Overview of NPs in Drug Discovery. *Metabolites* **2012**, *2* (2), 303–336. https://doi.org/10.3390/metabo2020303.

(4) Cragg, G. M.; Newman, D. J. Plants as a Source of Anti-Cancer Agents. *Journal of Ethnopharmacology*. Elsevier Ireland Ltd August 22, 2005, pp 72–79. https://doi.org/10.1016/j.jep.2005.05.011.

(5) Cragg, G. M.; Boyd, M. R.; Khanna, R.; Kneller, R.; Mays, T. D.; Mazan, K. D.; Newman, D. J.; Sausville, E. A. International Collaboration in Drug Discovery and Development: The NCI Experience. *Pure Appl. Chem.* **1999**. https://doi.org/10.1351/pac199971091619.

(6) Newman, D. J.; Cragg, G. M. NPs as Sources of New Drugs over the 30 Years from 1981 to 2010. *Journal of NPs*. March 23, 2012, pp 311–335. https://doi.org/10.1021/np200906s.

(7) Fleming, A. On the Antibacterial Action of Cultures of a Penicillium, with Special Reference to Their Use in the Isolation of B. Influenzae. 1929. *Bull. World Health Organ.* **2001**, *79* (8), 780–790. https://doi.org/10.1093/clinids/2.1.129.

(8) Gause, G. F.; Brazhnikova, M. G. Gramicidin S and Its Use in the Treatment of Infected Wounds [3]. *Nature*. 1944, p 703. https://doi.org/10.1038/154703a0.

(9) Caltrider, P. G. Pyocyanine. In *Antibiotics*; Springer Berlin Heidelberg, 1967; pp 117–121. https://doi.org/10.1007/978-3-662-38439-8_7.

(10) Newman, D. J.; Cragg, G. M. NPs as Sources of New Drugs from 1981 to 2014. *J. Nat.*

Prod. **2016**, *79* (3), 629–661. https://doi.org/10.1021/acs.jnatprod.5b01055.

(11) Bérdy, J. Bioactive Microbial Metabolites: A Personal View. *Journal of Antibiotics*. Japan Antibiotics Research Association 2005, pp 1–26. https://doi.org/10.1038/ja.2005.1.

(12) Hadacek, F.; Bachmann, G.; Engelmeier, D.; Chobot, V. Hormesis and a Chemical Raison d'être for Secondary Plant Metabolites. *Dose-Response* **2011**. https://doi.org/10.2203/dose-response.09-028.Hadacek.

(13) Firn, R. D.; Jones, C. G. The Evolution of Secondary Metabolism - a Unifying Model. *Mol. Microbiol.* **2000**, *37* (5), 989–994. https://doi.org/10.1046/j.1365-2958.2000.02098.x.

(14) Chin, Y. W.; Balunas, M. J.; Chai, H. B.; Kinghorn, A. D. Drug Discovery from Natural Sources. In *Drug Addiction: From Basic Research to Therapy*; Springer New York, 2008; pp 17–39. https://doi.org/10.1007/978-0-387-76678-2_2.

(15) Dewick, P. M. *Medicinal NPs: A Biosynthetic Approach: Third Edition*; 2009. https://doi.org/10.1002/9780470742761.

(16) Welker, M.; Dittmann, E.; Von Döhren, H. Cyanobacteria as a Source of NPs. In *Methods in Enzymology*; 2012. https://doi.org/10.1016/B978-0-12-404634-4.00002-4.

(17) Genilloud, O.; González, I.; Salazar, O.; Martín, J.; Tormo, J. R.; Vicente, F. Current Approaches to Exploit Actinomycetes as a Source of Novel NPs. *Journal of Industrial Microbiology and Biotechnology*. March 2011, pp 375–389. https://doi.org/10.1007/s10295-010-0882-7.

(18) Bull, A. T.; Ward, A. C.; Goodfellow, M. Search and Discovery Strategies for Biotechnology: The Paradigm Shift. *Microbiol. Mol. Biol. Rev.* **2000**, *64* (3), 573–606. https://doi.org/10.1128/mmbr.64.3.573-606.2000.

(19) Tiedje, J. M.; Asuming-Brempong, S.; Nüsslein, K.; Marsh, T. L.; Flynn, S. J. Opening the

Black Box of Soil Microbial Diversity. In *Applied Soil Ecology*; 1999; Vol. 13, pp 109–122. https://doi.org/10.1016/S0929-1393(99)00026-8.

(20) Reen, F.; Romano, S.; Dobson, A.; O'Gara, F. The Sound of Silence: Activating Silent Biosynthetic Gene Clusters in Marine Microorganisms. *Mar. Drugs* **2015**, *13* (8), 4754–4783. https://doi.org/10.3390/md13084754.

(21) O'Malley, M. A. "Everything Is Everywhere: But the Environment Selects": Ubiquitous Distribution and Ecological Determinism in Microbial Biogeography. *Stud. Hist. Philos. Sci. Part C Stud. Hist. Philos. Biol. Biomed. Sci.* **2008**, *39* (3), 314–325. https://doi.org/10.1016/j.shpsc.2008.06.005.

(22) Leeds, J. A.; Schmitt, E. K.; Krastel, P. Recent Developments in Antibacterial Drug Discovery: Microbe-Derived NPs - From Collection to the Clinic. *Expert Opinion on Investigational Drugs*. March 2006, pp 211–226. https://doi.org/10.1517/13543784.15.3.211.

(23) Fujimori, F.; Okuda, T. Application of the Random Amplified Polymorphic DNA Using the Polymerase Chain Reaction for Efficient Elimination of Duplicate Strains in Microbial Screening. I. Fungi. *J. Antibiot. (Tokyo).* **1994**, *47* (2), 173–182. https://doi.org/10.7164/antibiotics.47.173.

(24) Barka, E. A.; Vatsa, P.; Sanchez, L.; Gaveau-Vaillant, N.; Jacquard, C.; Klenk, H.-P.; Clément, C.; Ouhdouch, Y.; van Wezel, G. P. Taxonomy, Physiology, and NPs of Actinobacteria. *Microbiol. Mol. Biol. Rev.* **2016**, *80* (1), 1–43. https://doi.org/10.1128/mmbr.00019-15.

(25) Staley, J. T.; Konopka, A. Measurement of in Situ Activities of Nonphotosynthetic Microorganisms in Aquatic and Terrestrial Habitats. *Annu. Rev. Microbiol.* **1985**, *39* (1),

321–346. https://doi.org/10.1146/annurev.mi.39.100185.001541.

(26) Davies, J. Origins and Evolution of Antibiotic Resistance. *Microbiología (Madrid, Spain)*. 1996, pp 9–16. https://doi.org/10.1128/mmbr.00016-10.

(27) Jensen, P. R.; Fenical, W. Strategies for the Discovery of Secondary Metabolites from Marine Bacteria: Ecological Perspectives. *Annu. Rev. Microbiol.* **1994**, *48* (1), 559–584. https://doi.org/10.1146/annurev.mi.48.100194.003015.

(28) Fenical, W.; Jensen, P. R. Developing a New Resource for Drug Discovery: Marine Actinomycete Bacteria. *Nature Chemical Biology*. Nature Publishing Group 2006, pp 666–673. https://doi.org/10.1038/nchembio841.

(29) Chin, Y.-W.; Balunas, M. J.; Chai, H. B.; Kinghorn, A. D. Drug Discovery from Natural Sources. *AAPS J.* **2006**, *8* (2), E239--E253. https://doi.org/10.1007/bf02854894.

(30) Bergmann, W.; Burke, D. C. Contributions to the Study of Marine Products. XXXIX. the Nucleosides of Sponges. III.1 Spongothymidine and Spongouridine. *J. Org. Chem.* **1955**. https://doi.org/10.1021/jo01128a007.

(31) Bergmann, W.; Feeney, R. J. Contributions to the Study of Marine Products. XXXII. the Nucleosides of Sponges. I. *J. Org. Chem.* **1951**. https://doi.org/10.1021/jo01146a023.

(32) Burkholder, P. R.; Pfister, R. M.; Leitz, F. H. Production of a Pyrrole Antibiotic by a Marine Bacterium. *Appl. Microbiol.* **1966**. https://doi.org/10.1128/aem.14.4.649-653.1966.

(33) Zhu, C.; Schneider, E. K.; Wang, J.; Kempe, K.; Wilson, P.; Velkov, T.; Li, J.; Davis, T. P.; Whittaker, M. R.; Haddleton, D. M. A Traceless Reversible Polymeric Colistin Prodrug to Combat Multidrug-Resistant (MDR) Gram-Negative Bacteria. *J. Control. Release* **2017**. https://doi.org/10.1016/j.jconrel.2017.02.005.

(34) Fernandes, P.; Martens, E. Antibiotics in Late Clinical Development. *Biochemical*

Pharmacology. Elsevier Inc. June 2017, pp 152–163. https://doi.org/10.1016/j.bcp.2016.09.025.

(35) Newman, D. J.; Cragg, G. M.; Snader, K. M. NPs as Sources of New Drugs over the Period 1981-2002. *J. Nat. Prod.* **2003**, *66* (7), 1022–1037. https://doi.org/10.1021/np030096l.

(36) Harvey, A. L. NPs in Drug Discovery. *Drug Discovery Today*. October 2008, pp 894–901. https://doi.org/10.1016/j.drudis.2008.07.004.

(37) Pye, C. R.; Bertin, M. J.; Lokey, R. S.; Gerwick, W. H.; Linington, R. G. Retrospective Analysis of NPs Provides Insights for Future Discovery Trends. *Proc. Natl. Acad. Sci.* **2017**, *114* (22), 5601–5606. https://doi.org/10.1073/pnas.1614680114.

(38) Feher, M.; Schmidt, J. M. Property Distributions: Differences between Drugs, NPs, and Molecules from Combinatorial Chemistry. *J. Chem. Inf. Comput. Sci.* **2003**, *43* (1), 218–227. https://doi.org/10.1021/ci0200467.

(39) Winter, J. M.; Behnken, S.; Hertweck, C. Genomics-Inspired Discovery of NPs. *Current Opinion in Chemical Biology*. February 2011, pp 22–31. https://doi.org/10.1016/j.cbpa.2010.10.020.

(40) Ziemert, N.; Alanjary, M.; Weber, T. The Evolution of Genome Mining in Microbes – a Review. *Nat. Prod. Rep.* **2016**, *33* (8), 988–1005. https://doi.org/10.1039/C6NP00025H.

(41) Charlop-Powers, Z.; Owen, J. G.; Reddy, B. V. B.; Ternei, M. A.; Brady, S. F. Chemical-Biogeographic Survey of Secondary Metabolism in Soil. *Proc. Natl. Acad. Sci. U. S. A.* **2014**, *111* (10), 3757–3762. https://doi.org/10.1073/pnas.1318021111.

(42) Charlop-Powers, Z.; Owen, J. G.; Reddy, B. V. B.; Ternei, M.; Guimaraes, D. O.; De Frias, U. A.; Pupo, M. T.; Seepe, P.; Feng, Z.; Brady, S. F. Global Biogeographic Sampling of Bacterial Secondary Metabolism. *Elife* **2015**, *2015* (4). https://doi.org/10.7554/eLife.05048.

(43) Dunbar, K. L.; Mitchell, D. A. Revealing Nature's Synthetic Potential through the Study of Ribosomal Natural Product Biosynthesis. *ACS Chemical Biology*. American Chemical Society March 15, 2013, pp 473–487. https://doi.org/10.1021/cb3005325.

(44) Gerwick, W. H.; Moore, B. S. Lessons from the Past and Charting the Future of Marine NPs Drug Discovery and Chemical Biology. *Chemistry and Biology*. January 2012, pp 85–98. https://doi.org/10.1016/j.chembiol.2011.12.014.

(45) Letzel, A. C.; Li, J.; Amos, G. C. A.; Millán-Aguiñaga, N.; Ginigini, J.; Abdelmohsen, U. R.; Gaudêncio, S. P.; Ziemert, N.; Moore, B. S.; Jensen, P. R. Genomic Insights into Specialized Metabolism in the Marine Actinomycete Salinispora. *Environ. Microbiol.* **2017**, *19* (9), 3660–3673. https://doi.org/10.1111/1462-2920.13867.

(46) Rutledge, P. J.; Challis, G. L. Discovery of Microbial NPs by Activation of Silent Biosynthetic Gene Clusters. *Nature Reviews Microbiology*. Nature Publishing Group July 16, 2015, pp 509–523. https://doi.org/10.1038/nrmicro3496.

(47) Ziemert, N.; Podell, S.; Penn, K.; Badger, J. H.; Allen, E.; Jensen, P. R. The Natural Product Domain Seeker NaPDoS: A Phylogeny Based Bioinformatic Tool to Classify Secondary Metabolite Gene Diversity. *PLoS One* **2012**, *7* (3), 1–9. https://doi.org/10.1371/journal.pone.0034064.

(48) Vobruba, S.; Kadlcik, S.; Gazak, R.; Janata, J. Evolution-Guided Adaptation of an Adenylation Domain Substrate Specificity to an Unusual Amino Acid. *PLoS One* **2017**, *12* (12). https://doi.org/10.1371/journal.pone.0189684.

(49) Fischbach, M. A.; Walsh, C. T. Assembly-Line Enzymology for Polyketide and Nonribosomal Peptide Antibiotics: Logic Machinery, and Mechanisms. *Chem. Rev.* **2006**, *106* (8), 3468–3496. https://doi.org/10.1021/cr0503097.

(50) Medema, M. H.; Cimermancic, P.; Sali, A.; Takano, E.; Fischbach, M. A. A Systematic Computational Analysis of Biosynthetic Gene Cluster Evolution: Lessons for Engineering Biosynthesis. *PLoS Comput. Biol.* **2014**, *10* (12), e1004016. https://doi.org/10.1371/journal.pcbi.1004016.

(51) Bachmann, B. O.; Van Lanen, S. G.; Baltz, R. H. Microbial Genome Mining for Accelerated NPs Discovery: Is a Renaissance in the Making? *Journal of Industrial Microbiology and Biotechnology*. February 2014, pp 175–184. https://doi.org/10.1007/s10295-013-1389-9.

(52) Woodhouse, J. N.; Fan, L.; Brown, M. V.; Thomas, T.; Neilan, B. A. Deep Sequencing of Non-Ribosomal Peptide Synthetases and Polyketide Synthases from the Microbiomes of Australian Marine Sponges. *ISME J.* **2013**, *7* (9), 1842–1851. https://doi.org/10.1038/ismej.2013.65.

(53) Daniel, R. The Soil Metagenome - A Rich Resource for the Discovery of Novel NPs. *Current Opinion in Biotechnology*. Elsevier Current Trends June 1, 2004, pp 199–204. https://doi.org/10.1016/j.copbio.2004.04.005.

(54) Gontang, E. A.; Gaudêncio, S. P.; Fenical, W.; Jensen, P. R. Sequence-Based Analysis of Secondary-Metabolite Biosynthesis in Marine Actinobacteria. *Appl. Environ. Microbiol.* **2010**, *76* (8), 2487–2499. https://doi.org/10.1128/AEM.02852-09.

(55) Prieto-Davó, A.; Villarreal-Gómez, L. J.; Forschner-Dancause, S.; Bull, A. T.; Stach, J. E. M.; Smith, D. C.; Rowley, D. C.; Jensen, P. R. Targeted Search for Actinomycetes from Nearshore and Deep-Sea Marine Sediments. *FEMS Microbiol. Ecol.* **2013**, *84* (3), 510–518. https://doi.org/10.1111/1574-6941.12082.

(56) Reddy, B. V. ija. B.; Milshteyn, A.; Charlop-Powers, Z.; Brady, S. F. ESNaPD: A Versatile, Web-Based Bioinformatics Platform for Surveying and Mining Natural Product

Biosynthetic Diversity from Metagenomes. *Chem. Biol.* **2014**, *21* (8), 1023–1033. https://doi.org/10.1016/j.chembiol.2014.06.007.

(57) Owen, J. G.; Reddy, B. V. B.; Ternei, M. A.; Charlop-Powers, Z.; Calle, P. Y.; Kim, J. H.; Brady, S. F. Mapping Gene Clusters within Arrayed Metagenomic Libraries to Expand the Structural Diversity of Biomedically Relevant NPs. *Proc. Natl. Acad. Sci.* **2013**, *110* (29), 11797–11802. https://doi.org/10.1073/pnas.1222159110.

(58) Hadjithomas, M.; Chen, I. M. A.; Chu, K.; Ratner, A.; Palaniappan, K.; Szeto, E.; Huang, J.; Reddy, T. B. K.; Cimermančič, P.; Fischbach, M. A.; et al. IMG-ABC: A Knowledge Base to Fuel Discovery of Biosynthetic Gene Clusters and Novel Secondary Metabolites. *MBio* **2015**, *6* (4), 1–10. https://doi.org/10.1128/mBio.00932-15.

(59) Gihring, T. M.; Green, S. J.; Schadt, C. W. Massively Parallel RRNA Gene Sequencing Exacerbates the Potential for Biased Community Diversity Comparisons Due to Variable Library Sizes. *Environmental Microbiology*. 2012, pp 285–290. https://doi.org/10.1111/j.1462-2920.2011.02550.x.

(60) Demain, A.; Fang, A. The Natural Functions of Secondary Metabolites History of Modern Biotechnology I. *Adv. Biochem. Eng. Biotechnol.* **2000**, *69*, 1–39. https://doi.org/10.1007/3-540-44964-7.

(61) Albuquerque, L.; da Costa, M. S. The Family *Gaiellaceae*. In *The Prokaryotes*; Springer Berlin Heidelberg: Berlin, Heidelberg, 2014; pp 357–360. https://doi.org/10.1007/978-3-642-30138-4_394.

(62) Ginolhac, A.; Jarrin, C.; Gillet, B.; Robe, P.; Pujic, P.; Tuphile, K.; Bertrand, H.; Vogel, T. M.; Perrière, G.; Simonet, P.; et al. Phylogenetic Analysis of Polyketide Synthase I Domains from Soil Metagenomic Libraries Allows Selection of Promising Clones. *Appl. Environ.*

Microbiol. **2004**, *70* (9), 5522–5527. https://doi.org/10.1128/AEM.70.9.5522-5527.2004.

(63) Metsä-Ketelä, M.; Salo, V.; Halo, L.; Hautala, A.; Hakala, J.; Mäntsälä, P.; Ylihonko, K. An Efficient Approach for Screening Minimal PKS Genes from *Streptomyces*. *FEMS Microbiol. Lett.* **1999**, *180* (1), 1–6. https://doi.org/10.1016/S0378-1097(99)00453-X.

(64) Ayuso-Sacido, A.; Genilloud, O. 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.* **2005**, *49* (1), 10–24. https://doi.org/10.1007/s00248-004-0249-6.

(65) Konz, D.; Marahiel, M. A. How Do Peptide Synthetases Generate Structural Diversity? **1999**.

(66) Nguyen, T.; Ishida, K.; Jenke-Kodama, H.; Dittmann, E.; Gurgui, C.; Hochmuth, T.; Taudien, S.; Platzer, M.; Hertweck, C.; Piel, J. Exploiting the Mosaic Structure of Trans-Acyltransferase Polyketide Synthases for Natural Product Discovery and Pathway Dissection. *Nat. Biotechnol.* **2008**, *26* (2), 225–233. https://doi.org/10.1038/nbt1379.

(67) Weber, T.; Kim, H. U. The Secondary Metabolite Bioinformatics Portal: Computational Tools to Facilitate Synthetic Biology of Secondary Metabolite Production. *Synth. Syst. Biotechnol.* **2016**, *1* (2), 69–79. https://doi.org/10.1016/j.synbio.2015.12.002.

(68) Medema, M. H.; Kottmann, R.; Yilmaz, P.; Cummings, M.; Biggins, J. B.; Blin, K.; De Bruijn, I.; Chooi, Y. H.; Claesen, J.; Coates, R. C.; et al. Minimum Information about a Biosynthetic Gene Cluster. *Nature Chemical Biology*. 2015. https://doi.org/10.1038/nchembio.1890.

(69) Ziemert, N.; Lechner, A.; Wietz, M.; Millan-Aguinaga, N.; Chavarria, K. L.; Jensen, P. R. Diversity and Evolution of Secondary Metabolism in the Marine Actinomycete Genus

Salinispora. *Proc. Natl. Acad. Sci.* **2014**, *111* (12), E1130–E1139. https://doi.org/10.1073/pnas.1324161111.

(70) Liu, G.; Chater, K. F.; Chandra, G.; Niu, G.; Tan, H. Molecular Regulation of Antibiotic Biosynthesis in *Streptomyces*. *Microbiol. Mol. Biol. Rev.* **2013**, *77* (1), 112–143. https://doi.org/10.1128/MMBR.00054-12.

(71) Zhao, H.; Shao, D.; Jiang, C.; Shi, J.; Li, Q.; Huang, Q.; Rajoka, M. S. R.; Yang, H.; Jin, M. Biological Activity of Lipopeptides from *Bacillus*. *Appl. Microbiol. Biotechnol.* **2017**, *101* (15), 5951–5960. https://doi.org/10.1007/s00253-017-8396-0.

(72) Boumehira, A. Z.; El-Enshasy, H. A.; Hacène, H.; Elsayed, E. A.; Aziz, R.; Park, E. Y. Recent Progress on the Development of Antibiotics from the Genus *Micromonospora*. *Biotechnol. Bioprocess Eng.* **2016**, *21* (2), 199–223. https://doi.org/10.1007/s12257-015-0574-2.

(73) Hover, B. M.; Kim, S.-H.; Katz, M.; Charlop-Powers, Z.; Owen, J. G.; Ternei, M. A.; Maniko, J.; Estrela, A. B.; Molina, H.; Park, S.; et al. Culture-Independent Discovery of the Malacidins as Calcium-Dependent Antibiotics with Activity against Multidrug-Resistant Gram-Positive Pathogens. *Nat. Microbiol.* **2018**, 1. https://doi.org/10.1038/s41564-018-0110-1.

(74) Ling, L. L.; Schneider, T.; Peoples, A. J.; Spoering, A. L.; Engels, I.; Conlon, B. P.; Mueller, A.; Schäberle, T. F.; Hughes, D. E.; Epstein, S.; et al. A New Antibiotic Kills Pathogens without Detectable Resistance. *Nature* **2015**, *517* (7535), 455–459. https://doi.org/10.1038/nature14098.

(75) Crits-Christoph, A.; Diamond, S.; Butterfield, C. N.; Thomas, B. C.; Banfield, J. F. Novel Soil Bacteria Possess Diverse Genes for Secondary Metabolite Biosynthesis. *Nature* **2018**,

558 (7710), 440–444. https://doi.org/10.1038/s41586-018-0207-y.

(76) Weber, T.; Blin, K.; Duddela, S.; Krug, D.; Kim, H. U.; Bruccoleri, R.; Lee, S. Y.; Fischbach, M. A.; Müller, R.; Wohlleben, W.; et al. AntiSMASH 3.0-A Comprehensive Resource for the Genome Mining of Biosynthetic Gene Clusters. *Nucleic Acids Res.* **2015**, *43* (W1), W237–W243. https://doi.org/10.1093/nar/gkv437.

(77) Baltz, R. H. Gifted Microbes for Genome Mining and Natural Product Discovery. *J. Ind. Microbiol. Biotechnol.* **2017**, *44* (4–5), 573–588. https://doi.org/10.1007/s10295-016-1815-x.

(78) Charlop-Powers, Z.; Owen, J. G.; Reddy, B. V. B.; Ternei, M. A.; Brady, S. F. Chemical-Biogeographic Survey of Secondary Metabolism in Soil. *Proc. Natl. Acad. Sci.* **2014**, *111* (10), 3757–3762. https://doi.org/10.1073/pnas.1318021111.

(79) Charlop-Powers, Z.; Milshteyn, A.; Brady, S. F. Metagenomic Small Molecule Discovery Methods. *Current Opinion in Microbiology*. Elsevier Ltd 2014, pp 70–75. https://doi.org/10.1016/j.mib.2014.05.021.

(80) Shannon, C. E. A Mathematical Theory of Communication. *Bell Syst. Tech. J.* **1948**, *27* (July 1928), 379–423. https://doi.org/10.1145/584091.584093.

(81) Mullowney, M. W.; Hwang, C. H.; Newsome, A. G.; Wei, X.; Tanouye, U.; Wan, B.; Carlson, S.; Barranis, N. J.; Ó hAinmhire, E.; Chen, W. L.; et al. Diaza-Anthracene Antibiotics from a Freshwater-Derived Actinomycete with Selective Antibacterial Activity toward *Mycobacterium Tuberculosis*. *ACS Infect. Dis.* **2016**, *1* (4), 168–174. https://doi.org/10.1021/acsinfecdis.5b00005.

(82) Clark, C. M.; Costa, M. S.; Sanchez, L. M.; Murphy, B. T. Coupling MALDI-TOF Mass Spectrometry Protein and Specialized Metabolite Analyses to Rapidly Discriminate

Bacterial Function. *Proc. Natl. Acad. Sci.* **2018**. https://doi.org/10.1073/pnas.1801247115.

(83) Green, S. J.; Venkatramanan, R.; Naqib, A. Deconstructing the Polymerase Chain Reaction: Understanding and Correcting Bias Associated with Primer Degeneracies and Primer-Template Mismatches. *PLoS One* **2015**, *10* (5), e0128122. https://doi.org/10.1371/journal.pone.0128122.

(84) Naqib, A.; Poggi, S.; Wang, W.; Hyde, M.; Kunstman, K.; Green, S. J. Making and Sequencing Heavily Multiplexed, High-Throughput 16S Ribosomal RNA Gene Amplicon Libraries Using a Flexible, Two-Stage PCR Protocol. In *Methods in molecular biology (Clifton, N.J.)*; 2018; Vol. 1783, pp 149–169. https://doi.org/10.1007/978-1-4939-7834-2_7.

(85) Caporaso, J. G.; Lauber, C. L.; Walters, W. A.; Berg-Lyons, D.; Lozupone, C. A.; Turnbaugh, P. J.; Fierer, N.; Knight, R. Global Patterns of 16S RRNA Diversity at a Depth of Millions of Sequences per Sample. *Proc. Natl. Acad. Sci. U. S. A.* **2011**, *108 Suppl 1* (Suppl 1), 4516–4522. https://doi.org/10.1073/pnas.1000080107.

(86) Caporaso, J. G.; Kuczynski, J.; Stombaugh, J.; Bittinger, K.; Bushman, F. D.; Costello, E. K.; Fierer, N.; Peña, A. G.; Goodrich, J. K.; Gordon, J. I.; et al. QIIME Allows Analysis of High-Throughput Community Sequencing Data. *Nat. Methods* **2010**, *7* (5), 335–336. https://doi.org/10.1038/nmeth.f.303.

(87) Quast, C.; Pruesse, E.; Yilmaz, P.; Gerken, J.; Schweer, T.; Yarza, P.; Peplies, J.; Glöckner, F. O. The SILVA Ribosomal RNA Gene Database Project: Improved Data Processing and Web-Based Tools. *Nucleic Acids Res.* **2013**, *41* (D1), 590–596. https://doi.org/10.1093/nar/gks1219.

(88) McDonald, D.; Clemente, J. C.; Kuczynski, J.; Rideout, J. R.; Stombaugh, J.; Wendel, D.;

Wilke, A.; Huse, S.; Hufnagle, J.; Meyer, F.; et al. The Biological Observation Matrix (BIOM) Format or: How I Learned to Stop Worrying and Love the Ome-Ome. *Gigascience* **2012**, *464* (1), 1–6. https://doi.org/10.1186/2047-217X-1-7.

(89) Abascal, F.; Zardoya, R.; Telford, M. J. TranslatorX: Multiple Alignment of Nucleotide Sequences Guided by Amino Acid Translations. *Nucleic Acids Res.* **2010**, *38* (SUPPL. 2), 7–13. https://doi.org/10.1093/nar/gkq291.

(90) Johnson, L. S.; Eddy, S. R.; Portugaly, E. Hidden Markov Model Speed Heuristic and Iterative HMM Search Procedure. *BMC Bioinformatics* **2010**, *11* (431), 1471–2105. https://doi.org/10.1186/1471-2105-11-431.

(91) Edgar, R. C. Search and Clustering Orders of Magnitude Faster than BLAST. *Bioinformatics* **2010**, *26* (19), 2460–2461. https://doi.org/10.1093/bioinformatics/btq461.

(92) Buchfink, B.; Xie, C.; Huson, D. H. Fast and Sensitive Protein Alignment Using DIAMOND. *Nat. Methods* **2015**, *12* (1), 59–60. https://doi.org/10.1038/nmeth.3176.

(93) Foster, W.; Raoult, A. Early Descriptions of Antibiosis. *J. R. Coll. Gen. Pract.* **1974**, *24* (149), 889–894.

(94) Van Epps, H. L. René Dubos: Unearthing Antibiotics. *J. Exp. Med.* **2006**, *203* (2), 259. https://doi.org/10.1084/jem.2032fta.

(95) Clardy, J.; Fischbach, M. A.; Currie, C. R. The Natural History of Antibiotics. *Current Biology*. June 9, 2009. https://doi.org/10.1016/j.cub.2009.04.001.

(96) Fischbach, M. A.; Walsh, C. T. Antibiotics for Emerging Pathogens. *Science*. 2009, pp 1089–1093. https://doi.org/10.1126/science.1176667.

(97) Cheng, K.; Rong, X.; Pinto-Tomás, A. A.; Fernández-Villalobos, M.; Murillo-Cruz, C.; Huang, Y. Population Genetic Analysis of Streptomyces Albidoflavus Reveals Habitat

Barriers to Homologous Recombination in the Diversification of Streptomycetes. *Appl. Environ. Microbiol.* **2015**, *81* (3), 966–975. https://doi.org/10.1128/AEM.02925-14.

(98) Lemetre, C.; Maniko, J.; Charlop-Powers, Z.; Sparrow, B.; Lowe, A. J.; Brady, S. F. Bacterial Natural Product Biosynthetic Domain Composition in Soil Correlates with Changes in Latitude on a Continent-Wide Scale. *Proc. Natl. Acad. Sci. U. S. A.* **2017**, *114* (44), 11615–11620. https://doi.org/10.1073/pnas.1710262114.

(99) Borsetto, C.; Amos, G. C. A.; da Rocha, U. N.; Mitchell, A. L.; Finn, R. D.; Laidi, R. F.; Vallin, C.; Pearce, D. A.; Newsham, K. K.; Wellington, E. M. H. Microbial Community Drivers of PK/NRP Gene Diversity in Selected Global Soils. *Microbiome* **2019**, *7* (1), 78. https://doi.org/10.1186/s40168-019-0692-8.

(100) Kautsar, S. A.; Blin, K.; Shaw, S.; Navarro-Muñoz, J. C.; Terlouw, B. R.; van der Hooft, J. J. J.; van Santen, J. A.; Tracanna, V.; Suarez Duran, H. G.; Pascal Andreu, V.; et al. MIBiG 2.0: A Repository for Biosynthetic Gene Clusters of Known Function. *Nucleic Acids Res.* **2019**. https://doi.org/10.1093/nar/gkz882.

(101) Blin, K.; Shaw, S.; Steinke, K.; Villebro, R.; Ziemert, N.; Lee, S. Y.; Medema, M. H.; Weber, T. AntiSMASH 5.0: Updates to the Secondary Metabolite Genome Mining Pipeline. *Nucleic Acids Res.* **2019**, *47* (W1), W81–W87. https://doi.org/10.1093/nar/gkz310.

(102) Adamek, M.; Alanjary, M.; Ziemert, N. Applied Evolution: Phylogeny-Based Approaches in NPs Research. *Natural Product Reports*. Royal Society of Chemistry September 1, 2019, pp 1295–1312. https://doi.org/10.1039/c9np00027e.

(103) Elfeki, M.; Alanjary, M.; Green, S. J.; Ziemert, N.; Murphy, B. T. Assessing the Efficiency of Cultivation Techniques To Recover Natural Product Biosynthetic Gene Populations from Sediment. *ACS Chem. Biol.* **2018**, *13* (8), 2074–2081.

https://doi.org/10.1021/acschembio.8b00254.

(104) Braesel, J.; Crnkovic, C. M.; Kunstman, K. J.; Green, S. J.; Maienschein-Cline, M.; Orjala, J.; Murphy, B. T.; Eustáquio, A. S. Complete Genome of *Micromonospora*Sp. Strain B006 Reveals Biosynthetic Potential of a Lake Michigan Actinomycete. *J. Nat. Prod.* **2018**, *81* (9), 2057–2068. https://doi.org/10.1021/acs.jnatprod.8b00394.

(105) Schmelz, S.; Naismith, J. H. Adenylate-Forming Enzymes. *Current Opinion in Structural Biology*. December 2009, pp 666–671. https://doi.org/10.1016/j.sbi.2009.09.004.

(106) Chen, A.; Re, R. N.; Burkart, M. D. Type II Fatty Acid and Polyketide Synthases: Deciphering Protein-Protein and Protein-Substrate Interactions. *Natural Product Reports*. Royal Society of Chemistry October 1, 2018, pp 1029–1045. https://doi.org/10.1039/c8np00040a.

(107) Du, D.; Katsuyama, Y.; Shin-ya, K.; Ohnishi, Y. Reconstitution of a Type II Polyketide Synthase That Catalyzes Polyene Formation. *Angew. Chemie* **2018**, *130* (7), 1972–1975. https://doi.org/10.1002/ange.201709636.

(108) Bentley, S. D.; Chater, K. F.; Cerdeño-Tárraga, A. M.; Challis, G. L.; Thomson, N. R.; James, K. D.; Harris, D. E.; Quail, M. A.; Kieser, H.; Harper, D.; et al. Complete Genome Sequence of the Model Actinomycete Streptomyces Coelicolor A3(2). *Nature* **2002**, *417* (6885), 141–147. https://doi.org/10.1038/417141a.

(109) Scholz, M.; Ward, D. V.; Pasolli, E.; Tolio, T.; Zolfo, M.; Asnicar, F.; Truong, D. T.; Tett, A.; Morrow, A. L.; Segata, N. Strain-Level Microbial Epidemiology and Population Genomics from Shotgun Metagenomics. *Nat. Methods* **2016**, *13* (5), 435–438. https://doi.org/10.1038/nmeth.3802.

(110) Renner, M. K.; Shen, Y. C.; Cheng, X. C.; Jensen, P. R.; Frankmoelle, W.; Kauffman, C.

A.; Fenical, W.; Lobkovsky, E.; Clardy, J. Cyclomarins A-C, New Antiinflammatory Cyclic Peptides Produced by a Marine Bacterium (Streptomyces Sp.). *J. Am. Chem. Soc.* **1999**, *121* (49), 11273–11276. https://doi.org/10.1021/ja992482o.

(111) Wagman, G. A.; Waitz, J. A.; Marquez, J.; Murawski, A.; Oden, E. M.; Testa, R. T.; Weinstein, M. J. A New Micromonospora-Produced Macrolide Antibiotic, Rosamicin. *J. Antibiot. (Tokyo).* **1972**. https://doi.org/10.7164/antibiotics.25.641.

(112) Oh, D. C.; Poulsen, M.; Currie, C. R.; Clardy, J. Sceliphrolactam, a Polyene Macrocyclic Lactam from a Wasp-Associated Streptomyces Sp. *Org. Lett.* **2011**, *13* (4), 752–755. https://doi.org/10.1021/ol102991d.

(113) Smith, P. A.; Koehler, M. F. T.; Girgis, H. S.; Yan, D.; Chen, Y.; Chen, Y.; Crawford, J. J.; Durk, M. R.; Higuchi, R. I.; Kang, J.; et al. Optimized Arylomycins Are a New Class of Gram-Negative Antibiotics. *Nature* **2018**, *561* (7722), 189–194. https://doi.org/10.1038/s41586-018-0483-6.

(114) Kodani, S.; Bicz, J.; Song, L.; Deeth, R. J.; Ohnishi-Kameyama, M.; Yoshida, M.; Ochi, K.; Challis, G. L. Structure and Biosynthesis of Scabichelin, a Novel Tris-Hydroxamate Siderophore Produced by the Plant Pathogen Streptomyces Scabies 87.22. *Org. Biomol. Chem.* **2013**, *11* (28), 4686–4694. https://doi.org/10.1039/c3ob40536b.

(115) Bruns, H.; Crüsemann, M.; Letzel, A. C.; Alanjary, M.; McInerney, J. O.; Jensen, P. R.; Schulz, S.; Moore, B. S.; Ziemert, N. Function-Related Replacement of Bacterial Siderophore Pathways. *ISME J.* **2018**, *12* (2), 320–329. https://doi.org/10.1038/ismej.2017.137.

(116) Cox, C. D.; Adams, P. Siderophore Activity of Pyoverdin for Pseudomonas Aeruginosa. *Infect. Immun.* **1985**, *48* (1), 130–138.

(117) Challis, G. L.; Ravel, J. Coelichelin, a New Peptide Siderophore Encoded by the *Streptomyces Coelicolor* Genome: Structure Prediction from the Sequence of Its Non-Ribosomal Peptide Synthetase. *FEMS Microbiol. Lett.* **2000**, *187* (2), 111–114. https://doi.org/10.1111/j.1574-6968.2000.tb09145.x.

(118) Yunt, Z.; Reinhardt, K.; Li, A.; Engeser, M.; Dahse, H.-M.; Gütschow, M.; Bruhn, T.; Bringmann, G.; Piel, J. Cleavage of Four Carbon−Carbon Bonds during Biosynthesis of the Griseorhodin A Spiroketal Pharmacophore. *J. Am. Chem. Soc.* **2009**, *131* (6), 2297–2305. https://doi.org/10.1021/ja807827k.

(119) Ninomiya, A.; Katsuyama, Y.; Kuranaga, T.; Miyazaki, M.; Nogi, Y.; Okada, S.; Wakimoto, T.; Ohnishi, Y.; Matsunaga, S.; Takada, K. Biosynthetic Gene Cluster for Surugamide A Encompasses an Unrelated Decapeptide, Surugamide F. *ChemBioChem* **2016**, *17* (18), 1709–1712. https://doi.org/10.1002/cbic.201600350.

(120) Dunshee, B. R.; Leben, C.; Keitt, G. W.; Strong, F. M. The Isolation and Properties of Antimycin A. *J. Am. Chem. Soc.* **1949**. https://doi.org/10.1021/ja01175a057.

(121) Liu, J.; Zhu, X.; Kim, S. J.; Zhang, W. Antimycin-Type Depsipeptides: Discovery, Biosynthesis, Chemical Synthesis, and Bioactivities. *Natural Product Reports*. 2016. https://doi.org/10.1039/c6np00004e.

(122) Umezawa, K.; Nakazawa, K.; Uemura, T.; Ikeda, Y.; Kondo, S.; Naganawa, H.; Kinoshita, N.; Hashizume, H.; Hamada, M.; Takeuchi, T.; et al. Polyoxypeptin Isolated from Streptomyces: A Bioactive Cyclic Depsipeptide Containing the Novel Amino Acid 3-Hydroxy-3-Methylproline. *Tetrahedron Lett.* **1998**. https://doi.org/10.1016/S0040-4039(98)00031-8.

(123) Newton, R. J.; Jones, S. E.; Eiler, A.; McMahon, K. D.; Bertilsson, S. A Guide to the Natural

History of Freshwater Lake Bacteria. *Microbiol. Mol. Biol. Rev.* **2011**. https://doi.org/10.1128/mmbr.00028-10.

(124) Costa, M. S.; Clark, C. M.; Ómarsdóttir, S.; Sanchez, L. M.; Murphy, B. T. Minimizing Taxonomic and Natural Product Redundancy in Microbial Libraries Using MALDI-TOF MS and the Bioinformatics Pipeline IDBac. *J. Nat. Prod.* **2019**. https://doi.org/10.1021/acs.jnatprod.9b00168.

(125) Mullowney, M. W.; Hwang, C. H.; Newsome, A. G.; Wei, X.; Tanouye, U.; Wan, B.; Carlson, S.; Barranis, N. J.; Ó hAinmhire, E.; Chen, W.-L.; et al. Diaza-Anthracene Antibiotics from a Freshwater-Derived Actinomycete with Selective Antibacterial Activity toward *Mycobacterium Tuberculosis*. *ACS Infect. Dis.* **2015**, *1* (4), 168–174. https://doi.org/10.1021/acsinfecdis.5b00005.

(126) Shaikh, A. F.; Elfeki, M.; Landolfa, S.; Tanouye, U.; Green, S. J.; Murphy, B. T. Deuteromethylactin B from a Freshwater-Derived Streptomyces Sp. *Nat. Prod. Sci.* **2015**, *21* (4). https://doi.org/10.20307/nps.2015.21.4.261.

(127) Wang, Q.; Garrity, G. M.; Tiedje, J. M.; Cole, J. R. Naïve Bayesian Classifier for Rapid Assignment of RRNA Sequences into the New Bacterial Taxonomy. *Appl. Environ. Microbiol.* **2007**, *73* (16), 5261–5267. https://doi.org/10.1128/AEM.00062-07.

(128) CLARKE, K. R. Non-Parametric Multivariate Analyses of Changes in Community Structure. *Austral Ecol.* **1993**, *18* (1), 117–143. https://doi.org/10.1111/j.1442-9993.1993.tb00438.x.

(129) DeSantis, T. Z.; Hugenholtz, P.; Larsen, N.; Rojas, M.; Brodie, E. L.; Keller, K.; Huber, T.; Dalevi, D.; Hu, P.; Andersen, G. L. Greengenes, a Chimera-Checked 16S RRNA Gene Database and Workbench Compatible with ARB. *Appl. Environ. Microbiol.* **2006**, *72* (7),

5069–5072. https://doi.org/10.1128/AEM.03006-05.

(130) Ludwig, W.; Strunk, O.; Westram, R.; Richter, L.; Meier, H.; Yadhukumar, A.; Buchner, A.; Lai, T.; Steppi, S.; Jacob, G.; et al. ARB: A Software Environment for Sequence Data. *Nucleic Acids Res.* **2004**, *32* (4), 1363–1371. https://doi.org/10.1093/nar/gkh293.

(131) Ronquist, F.; Huelsenbeck, J. P. MrBayes 3: Bayesian Phylogenetic Inference under Mixed Models. *Bioinformatics* **2003**, *19* (12), 1572–1574. https://doi.org/10.1093/bioinformatics/btg180.

(132) Liang, Y.; Si, J.; Nikolic, M.; Peng, Y.; Chen, W.; Jiang, Y.; Lee, S.-H.; Ka, J.-O.; Cho, J.-C.; Lagomarsino, A.; et al. MEGA5: Molecular Evolutionary Genetics Analysis Using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. *Soil Biol. Biochem.* **2008**, *19* (1), 406–425. https://doi.org/10.1016/j.ejsobi.2007.09.004.

(133) Lee, J. Y.; Lee, J. Y.; Jung, H. W.; Hwang, B. K. Streptomyces Koyangensis Sp. Nov., a Novel Actinomycete That Produces 4-Phenyl-3-Butenoic Acid. *Int. J. Syst. Evol. Microbiol.* **2005**, *55* (1), 257–262. https://doi.org/10.1099/ijs.0.63168-0.

(134) Neu, H. C. The Crisis in Antibiotic Resistance. *Science (80-.).* **1992**, *257* (5073), 1064–1073. https://doi.org/10.1126/science.257.5073.1064.

(135) Witte, W. Medical Consequences of Antibiotics Use in Agriculture. *Science*. February 13, 1998, pp 996–997. https://doi.org/10.1126/science.279.5353.996.

(136) Schwartz, T.; Kohnen, W.; Jansen, B.; Obst, U. Detection of Antibiotic-Resistant Bacteria and Their Resistance Genes in Wastewater, Surface Water, and Drinking Water Biofilms. *FEMS Microbiol. Ecol.* **2003**, *43* (3), 325–335. https://doi.org/10.1111/j.1574-6941.2003.tb01073.x.

(137) Norrby, S. R.; Nord, C. E.; Finch, R. Lack of Development of New Antimicrobial Drugs:

A Potential Serious Threat to Public Health. *Lancet Infect. Dis.* **2005**, *5* (2), 115–119. https://doi.org/10.1016/s1473-3099(05)01283-1.

(138) Mathew, A. G.; Cissell, R.; Liamthong, S. Antibiotic Resistance in Bacteria Associated with Food Animals: A United States Perspective of Livestock Production. *Foodborne Pathogens and Disease*. June 2007, pp 115–133. https://doi.org/10.1089/fpd.2006.0066.

(139) Yang, Y.; Li, B.; Zou, S.; Fang, H. H. P.; Zhang, T. Fate of Antibiotic Resistance Genes in Sewage Treatment Plant Revealed by Metagenomic Approach. *Water Res.* **2014**, *62*, 97–106. https://doi.org/10.1016/j.watres.2014.05.019.

(140) Rodriguez-Mozaz, S.; Chamorro, S.; Marti, E.; Huerta, B.; Gros, M.; Sànchez-Melsió, A.; Borrego, C. M.; Barceló, D.; Balcázar, J. L. Occurrence of Antibiotics and Antibiotic Resistance Genes in Hospital and Urban Wastewaters and Their Impact on the Receiving River. *Water Res.* **2015**, *69*, 234–242. https://doi.org/10.1016/j.watres.2014.11.021.

(141) Hopwood, D. A. Highlights of Streptomyces Genetics. *Heredity*. Nature Publishing Group July 1, 2019, pp 23–32. https://doi.org/10.1038/s41437-019-0196-0.

(142) Demain, A. L.; Sanchez, S. Microbial Drug Discovery: 80 Years of Progress. *Journal of Antibiotics*. January 2009, pp 5–16. https://doi.org/10.1038/ja.2008.16.

(143) Moffitt, M. C.; Neilan, B. A. Evolutionary Affiliations Within the Superfamily of Ketosynthases Reflect Complex Pathway Associations. https://doi.org/10.1007/s00239-002-2415-0.

(144) Stachelhaus, T.; Mootz, H. D.; Marahiel, M. A. The Specificity-Conferring Code of Adenylation Domains in Nonribosomal Peptide Synthetases. *Chem. Biol.* **1999**, *6* (8), 493–505. https://doi.org/10.1016/S1074-5521(99)80082-9.

(145) Bloudoff, K.; Schmeing, T. M. Structural and Functional Aspects of the Nonribosomal

Peptide Synthetase Condensation Domain Superfamily: Discovery, Dissection and Diversity. *Biochimica et Biophysica Acta - Proteins and Proteomics*. Elsevier B.V. November 2017, pp 1587–1604. https://doi.org/10.1016/j.bbapap.2017.05.010.

(146) Libis, V.; Antonovsky, N.; Zhang, M.; Shang, Z.; Montiel, D.; Maniko, J.; Ternei, M. A.; Calle, P. Y.; Lemetre, C.; Owen, J. G.; et al. Uncovering the Biosynthetic Potential of Rare Metagenomic DNA Using Co-Occurrence Network Analysis of Targeted Sequences. *Nat. Commun.* **2019**, *10* (1). https://doi.org/10.1038/s41467-019-11658-z.

(147) Piel, J.; Hui, D.; Fusetani, N.; Matsunaga, S. Targeting Modular Polyketide Synthases with Iteratively Acting Acyltransferases from Metagenomes of Uncultured Bacterial Consortia. *Environ. Microbiol.* **2004**, *6* (9), 921–927. https://doi.org/10.1111/j.1462-2920.2004.00531.x.

(148) Wawrik, B.; Kerkhof, L.; Zylstra, G. J.; Kukor, J. J. Identification of Unique Type II Polyketide Synthase Genes in Soil. *Appl. Environ. Microbiol.* **2005**, *71* (5), 2232–2238. https://doi.org/10.1128/AEM.71.5.2232-2238.2005.

(149) Cimermancic, P.; Medema, M. H.; Claesen, J.; Kurita, K.; Wieland Brown, L. C.; Mavrommatis, K.; Pati, A.; Godfrey, P. A.; Koehrsen, M.; Clardy, J.; et al. Insights into Secondary Metabolism from a Global Analysis of Prokaryotic Biosynthetic Gene Clusters. *Cell* **2014**, *158* (2), 412–421. https://doi.org/10.1016/j.cell.2014.06.034.

(150) Frasch, H. J.; Medema, M. H.; Takano, E.; Breitling, R. Design-Based Re-Engineering of Biosynthetic Gene Clusters: Plug-and-Play in Practice. *Current Opinion in Biotechnology*. December 2013, pp 1144–1150. https://doi.org/10.1016/j.copbio.2013.03.006.

(151) Ren, H.; Shi, C.; Zhao, H. Computational Tools for Discovering and Engineering Natural Product Biosynthetic Pathways. *iScience* **2019**, 100795.

https://doi.org/10.1016/j.isci.2019.100795.

(152) Medema, M. H.; Fischbach, M. A. Computational Approaches to Natural Product Discovery. *Nat. Chem. Biol.* **2015**, *11* (9), 639–648. https://doi.org/10.1038/nchembio.1884.

(153) Khaldi, N.; Seifuddin, F. T.; Turner, G.; Haft, D.; Nierman, W. C.; Wolfe, K. H.; Fedorova, N. D. SMURF: Genomic Mapping of Fungal Secondary Metabolite Clusters. *Fungal Genet. Biol.* **2010**, *47* (9), 736–741. https://doi.org/10.1016/j.fgb.2010.06.003.

(154) Li, M. H. T.; Ung, P. M. U.; Zajkowski, J.; Garneau-Tsodikova, S.; Sherman, D. H. Automated Genome Mining for NPs. *BMC Bioinformatics* **2009**, *10*. https://doi.org/10.1186/1471-2105-10-185.

(155) Chevrette, M. G.; Aicheler, F.; Kohlbacher, O.; Currie, C. R.; Medema, M. H. SANDPUMA: Ensemble Predictions of Nonribosomal Peptide Chemistry Reveal Biosynthetic Diversity across Actinobacteria. *Bioinformatics* **2017**, *33* (20), 3202–3210. https://doi.org/10.1093/bioinformatics/btx400.

(156) Starcevic, A.; Zucko, J.; Simunkovic, J.; Long, P. F.; Cullum, J.; Hranueli, D. ClustScan: An Integrated Program Package for the Semi-Automatic Annotation of Modular Biosynthetic Gene Clusters and in Silico Prediction of Novel Chemical Structures. *Nucleic Acids Res.* **2008**, *36* (21), 6882–6892. https://doi.org/10.1093/nar/gkn685.

(157) Camacho, C.; Coulouris, G.; Avagyan, V.; Ma, N.; Papadopoulos, J.; Bealer, K.; Madden, T. L. BLAST+: Architecture and Applications. *BMC Bioinformatics* **2009**, *10*, 1–9. https://doi.org/10.1186/1471-2105-10-421.

(158) Nichols, D.; Cahoon, N.; Trakhtenberg, E. M.; Pham, L.; Mehta, A.; Belanger, A.; Kanigan, T.; Lewis, K.; Epstein, S. S. Use of Ichip for High-Throughput in Situ Cultivation of

"uncultivable Microbial Species▽. *Appl. Environ. Microbiol.* **2010**, *76* (8), 2445–2450. https://doi.org/10.1128/AEM.01754-09.

(159) Mahler, L.; Wink, K.; Beulig, R. J.; Scherlach, K.; Tovar, M.; Zang, E.; Martin, K.; Hertweck, C.; Belder, D.; Roth, M. Detection of Antibiotics Synthetized in Microfluidic Picolitre-Droplets by Various Actinobacteria. *Sci. Rep.* **2018**, *8* (1). https://doi.org/10.1038/s41598-018-31263-2.

(160) Zang, E.; Brandes, S.; Tovar, M.; Martin, K.; Mech, F.; Horbert, P.; Henkel, T.; Figge, M. T.; Roth, M. Real-Time Image Processing for Label-Free Enrichment of Actinobacteria Cultivated in Picolitre Droplets. *Lab Chip* **2013**, *13* (18), 3707–3713. https://doi.org/10.1039/c3lc50572c.

(161) Azman, A. S.; Othman, I.; Velu, S. S.; Chan, K. G.; Lee, L. H. Mangrove Rare Actinobacteria: Taxonomy, Natural Compound, and Discovery of Bioactivity. *Frontiers in Microbiology*. Frontiers Media S.A. 2015. https://doi.org/10.3389/fmicb.2015.00856.

(162) Pettit, R. K. Culturability and Secondary Metabolite Diversity of Extreme Microbes: Expanding Contribution of Deep Sea and Deep-Sea Vent Microbes to Natural Product Discovery. *Marine Biotechnology*. February 2011, pp 1–11. https://doi.org/10.1007/s10126-010-9294-y.

(163) Gerea, A. L.; Branscum, K. M.; King, J. B.; You, J.; Powell, D. R.; Miller, A. N.; Spear, J. R.; Cichewicz, R. H. Secondary Metabolites Produced by Fungi Derived from a Microbial Mat Encountered in an Iron-Rich Natural Spring. *Tetrahedron Lett.* **2012**, *53* (32), 4202–4205. https://doi.org/10.1016/j.tetlet.2012.05.156.

(164) Hoffmann, T.; Krug, D.; Bozkurt, N.; Duddela, S.; Jansen, R.; Garcia, R.; Gerth, K.; Steinmetz, H.; Müller, R. Correlating Chemical Diversity with Taxonomic Distance for

Discovery of NPs in Myxobacteria. *Nat. Commun.* **2018**, *9* (1). https://doi.org/10.1038/s41467-018-03184-1.

(165) Schorn, M. A.; Alanjary, M. M.; Aguinaldo, K.; Korobeynikov, A.; Podell, S.; Patin, N.; Lincecum, T.; Jensen, P. R.; Ziemert, N.; Moore, B. S. Sequencing Rare Marine Actinomycete Genomes Reveals High Density of Unique Natural Product Biosynthetic Gene Clusters. *Microbiology (United Kingdom)*. Microbiology Society December 2016, pp 2075–2086. https://doi.org/10.1099/mic.0.000386.

(166) Van Bloois, E.; Torres Pazmiño, D. E.; Winter, R. T.; Fraaije, M. W. A Robust and Extracellular Heme-Containing Peroxidase from Thermobifida Fusca as Prototype of a Bacterial Peroxidase Superfamily. *Appl. Microbiol. Biotechnol.* **2010**, *86* (5), 1419–1430. https://doi.org/10.1007/s00253-009-2369-x.

(167) Dittmann, E.; Gugger, M.; Sivonen, K.; Fewer, D. P. Natural Product Biosynthetic Diversity and Comparative Genomics of the Cyanobacteria. *Trends in Microbiology*. Elsevier Ltd October 2015, pp 642–652. https://doi.org/10.1016/j.tim.2015.07.008.

(168) Rice, P.; Longden, L.; Bleasby, A. EMBOSS: The European Molecular Biology Open Software Suite. *Trends Genet.* **2000**, *16* (6), 276–277. https://doi.org/10.1016/S0168-9525(00)02024-2.

(169) Rajendhran, J.; Gunasekaran, P. Microbial Phylogeny and Diversity: Small Subunit Ribosomal RNA Sequence Analysis and Beyond. *Microbiol. Res.* **2011**, *166* (2), 99–110. https://doi.org/10.1016/j.micres.2010.02.003.

(170) Woo, P. C. Y.; Lau, S. K. P.; Teng, J. L. L.; Tse, H.; Yuen, K. Y. Then and Now: Use of 16S RDNA Gene Sequencing for Bacterial Identification and Discovery of Novel Bacteria in Clinical Microbiology Laboratories. *Clin. Microbiol. Infect.* **2008**, *14* (10), 908–934.

https://doi.org/10.1111/j.1469-0691.2008.02070.x.

(171) Caporaso, J. G.; Kuczynski, J.; Stombaugh, J.; Bittinger, K.; Bushman, F. D.; Costello, E. K.; Fierer, N.; Peña, A. G.; Goodrich, J. K.; Gordon, J. I.; et al. QIIME Allows Analysis of High- Throughput Community Sequencing Data Intensity Normalization Improves Color Calling in SOLiD Sequencing. *Nat. Publ. Gr.* **2010**, *7* (5), 335–336. https://doi.org/10.1038/nmeth0510-335.

(172) O'Leary, N. A.; Wright, M. W.; Brister, J. R.; Ciufo, S.; Haddad, D.; McVeigh, R.; Rajput, B.; Robbertse, B.; Smith-White, B.; Ako-Adjei, D.; et al. Reference Sequence (RefSeq) Database at NCBI: Current Status, Taxonomic Expansion, and Functional Annotation. *Nucleic Acids Res.* **2016**, *44* (D1), D733--45. https://doi.org/10.1093/nar/gkv1189.

(173) Huson, D. H.; Mitra, S.; Ruscheweyh, H.-J.; Weber, N.; Schuster, S. C. Integrative Analysis of Environmental Sequences Using MEGAN4. *Genome Res.* **2011**, *21* (9), 1552–1560. https://doi.org/10.1101/gr.120618.111.

(174) Bolyen, E.; Rideout, J. R.; Dillon, M. R.; Bokulich, N. A.; Abnet, C. C.; Al-Ghalith, G. A.; Alexander, H.; Alm, E. J.; Arumugam, M.; Asnicar, F.; et al. Reproducible, Interactive, Scalable and Extensible Microbiome Data Science Using QIIME 2. *Nature Biotechnology*. Nature Publishing Group August 1, 2019, pp 852–857. https://doi.org/10.1038/s41587-019-0209-9.

(175) Callahan, B. J.; McMurdie, P. J.; Rosen, M. J.; Han, A. W.; Johnson, A. J. A.; Holmes, S. P. DADA2: High-Resolution Sample Inference from Illumina Amplicon Data. *Nat. Methods* **2016**, *13* (7), 581–583. https://doi.org/10.1038/nmeth.3869.

(176) Pfaffl, M. W.; Tichopad, A.; Prgomet, C.; Neuvians, T. P. Determination of Stable Housekeeping Genes, Differentially Regulated Target Genes and Sample Integrity:

BestKeeper - Excel-Based Tool Using Pair-Wise Correlations. *Biotechnol. Lett.* **2004**, *26* (6), 509–515. https://doi.org/10.1023/B:BILE.0000019559.84305.47.

(177) Ahmed, E.; Holmström, S. J. M. Siderophores in Environmental Research: Roles and Applications. *Microb. Biotechnol.* **2014**, *7* (3), 196–208. https://doi.org/10.1111/1751-7915.12117.

APPENDICES

APPENDIX A. Supporting Information for Chapter 2

Supplementary Experimental Procedures.

Collection of sediment samples.

Sediment samples H054 and NC68 were collected in Lake Huron at a surface depth of 134.9 m and 17.3 m, respectively during a research expedition aboard the Environmental Protection Agency's (EPA's) Lake Guardian Research Vessel. The samples were collected using a PONAR grab in summer 2012 from Georgian Bay and the Northern Channel (Figure 18). The top layer of sediment was homogenized, and two aliquots were placed into two sterile 50 mL conical tubes containing 20% glycerol. They were stored in cryogenic vials in a Dewar.

Cultivating sediment bacteria on nutrient agar.

Conical tubes were thawed and homogenized, and aliquots of two sediment samples were individually collected and placed in 4 mL vials for duplicates processing (sample A and sample B). Samples were diluted with filter-sterilized deionized (DI) water to a 1/10th concentration and incubated in a 57 °C water bath for 15 minutes. A 50 µL aliquot of the sediment dilution was spread onto the surface of an agar plate. Six different media types were used to make nutrient agar diversity plates: A1, 1/10th dilution of A1 (M1), ISP2, 1/10th dilution of ISP2 (DISP2), minimal agar media (LWA), and chitin (Table X).

Bioinformatic method validation using reference sequences from the MIBiG database.

Gene entries from the MIBiG database were downloaded in FASTA format as amino acid sequences. The standalone HMMER tool (http://hmmer.org/) was used to extract KS, KSα, and A domain subsequences from the amino acid sequences. The HMM models used for this extraction were the same pre-built generic detection models downloaded from antiSMASH v3.0.5, used for

our BGC data (see manuscript Methods section). The models used were: PKS_KS.hmm for PKS type I, AMP-binding and A-OX for AD, and t2ks and t2pks2 for PKS type II. Subsequences were back translated into nucleotide sequences using EMBOSS's backtranseq tool.[168] For clustering, we followed USEARCH v10's pipeline[91] as follows: fastx_uniques was used to find unique sequence reads and abundances of reads; reads were then sorted by length using the *-sortedbylength* command, and the sorted reads were clustered at 80, 85, and 90% using USEARCH v10's UCLUST cluster_fast greedy algorithm via the cluster_fast command. The chemical product of the subsequences within each cluster were analyzed (See Supplementary Experimental Procedures section below). Sequence reads belonging to the same molecular class clustered best at 85% and were used for subsequent analysis.

*Choosing a similarity threshold for clustering KS, KS**α**, and A domain sequence data.*

(a) Extraction of KS, KSα, and A domain sequences from the MIBiG database.

As of 12/20/17, there are 1,424 entries in the MIBiG[68] repository consisting of 24,085 full or partial FASTA sequence entries. Each NRP biosynthetic gene cluster entry contains one or more A domains. Similarly, each polyketide entry may contain one or more KS and/or KSα domains. We used HMM models downloaded from Antismash[76] to extract a) full sequences that contained A, KS, and KSα HMM domain hits and b) A, KS, and KSα domain hits (*e.g.*, A, KS, and KSα domains extracted via HMM using the HMM envelope coordinates). Table VIII shows the number of a) full sequences present in MIBiG containing one or more A, KS, and KSα domains and b) the number of A, KS, and KSα domains present in all sequences in MIBiG.

(b) Clustering of KS, KSα, and A domain sequences extracted from the MIBiG database.

KS, KSα, and A domains belonging to similar chemical structures can exhibit slightly different sequence homology, therefore we clustered the sequences according to percent similarity in an attempt to create compound class groupings. Three different percentages were tested for clustering the subsequences extracted from MIBiG.[68] The clustering methods are described in the main article's methods section under the "Bioinformatic analyses of BGC data" subsection.

(c) Validation of percentage used to cluster BGC sequences

We selected 12 common antibiotics representing 8 antibiotic classes to test the accuracy of our clustering threshold: ansamysin (rifamycin and geldanamycin), macrolide (erythromycin), and tetracycline antibiotics (chlortetracycline and oxytetracyline) for type I KS domains; aromatic polyketides such as the benzoisochromanequinone compounds (actinorhodin) and type II tetracycline antibiotics (tetracenomycin) for KSα domains; and streptogramins (pristinamycin and virginiamycin), lipopeptide (daptomycin), non-ribosomal cyclic peptide (bacillibactin), and glycopeptide (vancomycin) antibiotics for NRPS A domains. KS, KSα, and A domains were extracted from MIBiG.

Since PKS and NRPS clusters usually contain more than one KS/ KSα and A domain, respectively, in most cases multiple KS, KSα and A domains were extracted from each MIBiG[68] entry. For example, we clustered sequences associated with the ansamycin antibiotic rifamycin at 80%, and obtained a total of 30 KS domains, which clustered into 10 OBUs representing 9 compounds. Only 2 of these candidate sequences belonged to the ansamycin class of antibiotics (rifamycin and rubradirin). Additionally, we clustered the sequences at 90% similarity and obtained a total of 26 KS domains, which grouped into 10 OBUs and represented only 3 compounds: rifamycin, naphthomycin, and chaxamycin analogues A/B/C/D (each produced by the same cluster and

modified post-translationally) – all of which belong to the ansamycin class. Finally, we clustered the sequences at 85% similarity, and obtained a total of 40 KS domains, which clustered into 4 OBUs that represented 4 molecules: rifamycin, rubradirin, naphthomycin, and chaxamycin analogues A/B/C/D. This analysis was repeated for all the aforementioned antibiotic classes. We found that the optimal clustering threshold fluctuates and is dependent on the specific compound class. However, since the optimal thresholds ranged from 80 to 90%, we selected 85% as the most suitable for our purposes.

Table V. 16S rRNA gene amplicon sequence number, OTU number, and Shannon index for individual samples at different rarefaction depths.

These data are divided into three tables (VA to VC). OTU clustering was performed at the 97% similarity threshold, though such a clustering approach can underestimate microbial diversity.[169,170] Clustering resulted in 664,384 sequences clustered into 31,076 OTUs. The number of sequences per sample was rarified to the fewest sequence reads present in any sample: 7,047. OTUs on nutrient agar plates with less than 0.1% of sequence reads (7 sequences) were not considered cultivated bacteria and were therefore dropped from subsequent analysis. This resulted in an uneven number of sequences per sample. This led to a second round of rarefaction of the dataset to 5,834. The latter was used to compute the Shannon index. The Shannon index was computed using the scikit-bio's diversity calculation via QIIME.[171] The Shannon (aka Shannon-Wiener) index is defined as:

$$H = -\sum_{i=1}^{s}(p_i \, log_2 p_i)$$

Where s is the number of OTUs and p_i is the proportion of the community represented by OTU i. The Shannon indices reported are that of the second rarefaction of dataset (from 7,047 to 5,834). This data was then rarified further stepwise by 500 reads from 5,500 until 3,500. The Shannon indices from the latter rarefactions were averaged and reported in Table V.

APPENDIX A (continued)

VA. The number of sequences, OTUs, and Shannon index for all samples at a rarefaction depth of 7,047, after removing OTUs with less than 7 sequences, and after a second rarefaction of the dataset to 5,834 sequences per sample.

	Before removing OTUs with < 7 sequences		After removing OTUs with < 7 sequences		After second rarefaction			Shannon, averaged after rarefaction at multiple depths
Samples	Sequence count	OTU count	Sequence count	OTU count	Sequence count	OTU count	Shannon	
H054.Sediment	**7047**	**2412**	**7047**	**2412**	**5834**	**2088**	**9.03**	**8.91**
H054A.A1	7047	862	5981	34	5834	34	3.03	3.02
H054A.Chitin	7047	382	6568	20	5834	20	2.27	2.28
H054A.DISP2	7047	597	6298	21	5834	21	3.04	3.04
H054A.ISP2	7047	493	6456	11	5834	11	0.92	0.92
H054A.LWA	7047	541	6368	48	5834	48	3.67	3.67
H054A.M1	7047	806	6064	28	5834	28	2.55	2.55
H054B.A1	7047	909	5973	43	5834	43	3.27	3.27
H054B.Chitin	7047	326	6617	21	5834	21	1.48	1.48
H054B.DISP2	7047	752	6157	27	5834	27	1.76	1.76
H054B.ISP2	7047	678	6225	34	5834	34	3.43	3.43
H054B.LWA	7047	294	6685	17	5834	17	1.70	1.70
H054B.M1	7047	919	5958	39	5834	39	3.09	3.09
NC68.Sediment	**7047**	**2482**	**7047**	**2482**	**5834**	**2157**	**9.04**	**8.94**
NC68A.A1	7047	895	5938	36	5834	36	2.67	2.66
NC68A.Chitin	7047	518	6391	40	5834	40	3.52	3.52
NC68A.DISP2	7047	946	5834	41	5834	41	2.69	2.68
NC68A.ISP2	7047	651	6105	40	5834	40	1.40	1.39
NC68A.LWA	7047	611	6264	58	5834	58	3.90	3.89
NC68A.M1	7047	807	6109	46	5834	46	3.51	3.52
NC68B.A1	7047	817	6027	31	5834	31	2.68	2.68
NC68B.Chitin	7047	629	6263	39	5834	39	2.72	2.72
NC68B.DISP2	7047	698	6197	54	5834	54	3.72	3.72
NC68B.ISP2	7047	644	6159	36	5834	36	1.55	1.54
NC68B.LWA	7047	864	5963	31	5834	31	2.45	2.45
NC68B.M1	7047	867	6011	45	5834	45	3.39	3.39

APPENDIX A (continued)

VB. The percentage of sequences and OTUs belonging to different taxonomic groups averaged for sediment and nutrient agar samples.

	Nutrient agar	Sediment
% OTU Firmicutes	72.40	7.87
% OTU Actinobacteria	20.48	9.84
% OTU Proteobacteria	3.98	22.2
% OTU *Streptomyces*	5.92	0.17
% OTU *Micromonospora*	8.99	0.47
% OTU *Bacillus*	41.22	4.59
% Sequence reads Firmicutes	69.20	14.2
% Sequence reads Actinobacteria	27.35	16.0
% Sequence reads Proteobacteria	1.77	23.2
% Sequence reads *Streptomyces*	4.94	0.07
% Sequence reads *Micromonospora*	19.47	2.36
% Sequence reads *Bacillus*	54.87	9.70

APPENDIX A (continued)

VC. The number of OTUs as a function of select phyla and genera for all sediment and nutrient agar samples.

Sample	#OTU	#OTU, Firmicutes	#OTU, Actinobacteria	#OTU, Proteobacteria	#OTU, *Streptomyces*	#OTU, *Micromonospora*	#OTU, *Bacillus*
H054.SED	**2088**	**13**	**196**	**492**	**4**	**4**	**5**
H054A.A1	34	34	0	0	0	0	22
H054A.Chitin	20	12	6	1	6	0	2
H054A.DISP2	21	16	3	1	3	0	13
H054A.ISP2	11	10	0	0	0	0	8
H054A.LWA	48	23	22	2	2	12	6
H054A.M1	28	24	3	0	3	0	12
H054B.A1	43	43	0	0	0	0	20
H054B.Chitin	21	8	10	2	6	0	2
H054B.DISP2	27	23	2	1	2	0	15
H054B.ISP2	34	22	1	8	1	0	14
H054B.LWA	17	7	8	1	3	0	4
H054B.M1	39	36	1	1	1	0	15
NC68.SED	**2157**	**326**	**222**	**451**	**16**	**3**	**193**
NC68A.A1	36	33	1	1	1	0	22
NC68A.Chitin	40	20	20	0	9	5	6
NC68A.DISP2	41	37	2	1	2	0	17
NC68A.ISP2	40	38	0	1	0	0	38
NC68A.LWA	58	20	30	7	6	13	3
NC68A.M1	46	41	3	1	3	0	16
NC68B.A1	31	30	0	0	0	0	20
NC68B.Chitin	39	9	29	1	7	18	4
NC68B.DISP2	54	15	36	2	8	16	5
NC68B.ISP2	36	34	0	1	0	0	31
NC68B.LWA	31	25	4	1	2	2	16
NC68B.M1	45	37	6	1	5	0	16

APPENDIX A (continued)

Table VI. Detailed breakdown of sequence reads by phylum in sediment.

#OTU ID	% sequence reads
Proteobacteria	23.2
Actinobacteria	16.0
Firmicutes	14.2
Acidobacteria	9.16
Chloroflexi	7.76
Planctomycetes	5.13
Bacteroidetes	3.57
Nitrospirae	3.33
Unassigned	3.31
Verrucomicrobia	3.27
Thaumarchaeota	2.97
Rokubacteria	1.44
Cyanobacteria	1.38
Latescibacteria	1.21
Other	4.00

These taxa are represented in Figure 3 in Chapter 2. Several of the above phyla were condensed into “Other” to simplify the figure. The percentages represent the average number of sequence reads in sediment belonging to different phyla after rarefaction analysis of sequence data.

APPENDIX A (continued)

Table VII. Bray-Curtis analysis to compare similarity between duplicate samples.

We tested bacterial and BGC community differences using analysis of similarity (ANOSIM). Rarefied sequence data were used to generate the Bray-Curtis resemblance matrix. First, sequence data for each sample was standardized by total, and then the Bray-Curtis similarity scores were generated based on family-level taxonomic classification for 16S sequence data and at 85% OBU-level for BGC data using Primer6 (Primer-E, version 6.1.13, United Kingdom). The Bray-Curtis similarity compares samples across all the taxa in each sample and takes into account the relative abundance of each taxon/OBU. The higher the Bray Curtis value, the more similar two samples are.

VIIA. Bray-Curtis similarity for 16S sequence data on nutrient agar.

Sample	Bray-Curtis score
H054.A1	81.66
H054.Chitin	84.30
H054.DISP2	66.13
H054.ISP2	51.20
H054.LWA	11.31
H054.M1	90.25
NC68.A1	96.73
NC68.Chitin	54.39
NC68.DISP2	20.50
NC68.ISP2	91.64
NC68.LWA	18.27
NC68.M1	85.29

APPENDIX A (continued)

VIIB. Bray-Curtis similarity for KS domain sequence data on nutrient agar.

Sample	Bray-Curtis score
H054.A1	13.44
H054.Chitin	9.08
H054.DISP2	NA
H054.ISP2	10.64
H054.LWA	70.16
H054.M1	84.56
NC68.A1	NA
NC68.Chitin	21.20
NC68.DISP2	5.02
NC68.ISP2	12.85
NC68.LWA	46.01
NC68.M1	NA

VIIC. Bray-Curtis similarity for KSα domain sequence data on nutrient agar.

Sample	Bray-Curtis score
H054A.A1	2.17
H054A.Chitin	1.13
H054A.DISP2	0.85
H054A.ISP2	6.43
H054A.LWA	0.66
H054.M1	NA
NC68A.A1	2.08
NC68A.Chitin	6.05
NC68A.DISP2	25.24
NC68A.ISP2	2.84
NC68A.LWA	2.36
NC68.M1	2.55

VIID. Bray-Curtis similarity for A domain sequence data on nutrient agar.

Sample	Bray-Curtis score
H054A.A1	12.46
H054A.Chitin	0
H054A.DISP2	2.43
H054A.ISP2	0
H054A.LWA	48.09
H054.M1	8.36
NC68A.A1	5.44
NC68A.Chitin	14.11
NC68A.DISP2	25.89
NC68A.ISP2	NA
NC68A.LWA	6.43
NC68.M1	1.23

Table VIII. Sequences extracted from MIBiG.

The "Full sequences" column corresponds to the number of NRP sequence entries containing A domains, the number of PKS entries containing KS domains, and the number of PKS type II entries containing KSα domains. These entries were extracted using the HMM models provided in antiSMASH,[76] specified in the table. The number of "subsequences" corresponds to the number of A, KS, and KSα domains present in the MIBiG[68] database.

HMM model	**Full sequences**	**Subsequences**
A-OX	1,331	4,667
AMP-binding	1,294	2,414
PKS_KS	1,692	2,980
t2ks	1,682	3,239
t2ks2	1,702	3,050

APPENDIX A (continued)

Table IX. MIBiG subsequences clustered at different percentages.

The subsequences extracted from MIBiG[68] were subjected to USEARCH's clust_fast[91] greedy algorithm. The numbers in the table indicate the number of OBUs created at the given percentage.

Clustering %	**80%**	**85%**	**90%**	**95%**
A domain	123	256	330	370
KS domain	111	233	379	558
KSα domain	110	232	379	557

APPENDIX A (continued)

Table X. Nutrient agar composition.

Nutrient agar was prepared by adding 7.5 g of agar per liter of water.

Media Ingredient / 1 L	***A1***	***1/10th A1 (M1)***	***ISP2***	***1/10th ISP2 (DISP2)***	***LWA***	***Chitin***
Peptone	2 g	0.2 g	0 g	0 g	0 g	0 g
Yeast Extract	4 g	0.4 g	4 g	0.4 g	0 g	0 g
Soluble Starch	10 g	1 g	0 g	0 g	0 g	0 g
Chitin	0 g	0 g	0 g	0 g	0 g	4 g
Dextrose	0 g	0 g	4 g	0.4 g	0 g	0 g
Malt Extract	0 g	0 g	10 g	0.10 g	0 g	0 g
K_2HPO_4	0 g	0 g	0 g	0 g	0 g	0.7 g
KH_2PO_4	0 g	0 g	0 g	0 g	0 g	0.3 g
$MgSO_4 \cdot 7H_2O$	0 g	0 g	0 g	0 g	0 g	0.5 g
$FeSO_4 \cdot 7H_2O$	0 g	0 g	0 g	0 g	0 g	0.01 g
dH2O	0.5 L	0.5 L	0.5 L	0.5 L	0.5 L	0.5 L
Filtered Lake Water	0.5 L	0.5 L	0.5 L	0.5 L	0.5 L	0.5 L

APPENDIX A (continued)

Table XI. Accession Codes.

XIA. Accession codes for 16S rRNA data.

Accession	Sample Name
SAMN09061457	H054A_A1
SAMN09061458	H054A_Chitin
SAMN09061459	H054A_DISP2
SAMN09061460	H054A_ISP2
SAMN09061461	H054A_LWA
SAMN09061462	H054A_M1
SAMN09061463	H054B_A1
SAMN09061464	H054B_Chitin
SAMN09061465	H054B_DISP2
SAMN09061466	H054B_ISP2
SAMN09061467	H054B_LWA
SAMN09061468	H054B_M1
SAMN09061469	H054B_SED
SAMN09061470	NC68A_A1
SAMN09061471	NC68A_Chitin
SAMN09061472	NC68A_DISP2
SAMN09061473	NC68A_ISP2
SAMN09061474	NC68A_LWA
SAMN09061475	NC68A_M1
SAMN09061476	NC68A_SED
SAMN09061477	NC68B_A1
SAMN09061478	NC68B_Chitin
SAMN09061479	NC68B_DISP2
SAMN09061480	NC68B_ISP2
SAMN09061481	NC68B_LWA
SAMN09061482	NC68B_M1

APPENDIX A (continued)

XIB. Accession codes for BGC data.

Accession	Sample Name
SAMN09205062	H054A_A1_Adomain
SAMN09205063	H054A_Chitin_Adomain
SAMN09205064	H054A_DISP2_Adomain
SAMN09205065	H054A_ISP2_Adomain
SAMN09205066	H054A_LWA_Adomain
SAMN09205067	H054A_M1_Adomain
SAMN09205068	H054B_A1_Adomain
SAMN09205069	H054B_Chitin_Adomain
SAMN09205070	H054B_DISP2_Adomain
SAMN09205071	H054B_ISP2_Adomain
SAMN09205072	H054B_LWA_Adomain
SAMN09205073	H054B_M1_Adomain
SAMN09205074	H054B_SED_Adomain
SAMN09205075	NC68A_A1_Adomain
SAMN09205076	NC68A_Chitin_Adomain
SAMN09205077	NC68A_DISP2_Adomain
SAMN09205078	NC68A_ISP2_Adomain
SAMN09205079	NC68A_LWA_Adomain
SAMN09205080	NC68A_M1_Adomain
SAMN09205081	NC68A_SED_Adomain
SAMN09205082	NC68B_A1_Adomain
SAMN09205083	NC68B_Chitin_Adomain
SAMN09205084	NC68B_DISP2_Adomain
SAMN09205085	NC68B_ISP2_Adomain
SAMN09205086	NC68B_LWA_Adomain
SAMN09205087	NC68B_M1_Adomain
SAMN09205088	H054A_A1_KSdomain
SAMN09205089	H054A_Chitin_KSdomain
SAMN09205090	H054A_DISP2_KSdomain
SAMN09205091	H054A_ISP2_KSdomain
SAMN09205092	H054A_LWA_KSdomain
SAMN09205093	H054A_M1_KSdomain

APPENDIX A (continued)

SAMN09205094	H054B_A1_KSdomain
SAMN09205095	H054B_Chitin_KSdomain
SAMN09205096	H054B_DISP2_KSdomain
SAMN09205097	H054B_ISP2_KSdomain
SAMN09205098	H054B_LWA_KSdomain
SAMN09205099	H054B_M1_KSdomain
SAMN09205100	H054B_SED_KSdomain
SAMN09205101	NC68A_A1_KSdomain
SAMN09205102	NC68A_Chitin_KSdomain
SAMN09205103	NC68A_DISP2_KSdomain
SAMN09205104	NC68A_ISP2_KSdomain
SAMN09205105	NC68A_LWA_KSdomain
SAMN09205106	NC68A_M1_KSdomain
SAMN09205107	NC68A_SED_KSdomain
SAMN09205108	NC68B_A1_KSdomain
SAMN09205109	NC68B_Chitin_KSdomain
SAMN09205110	NC68B_DISP2_KSdomain
SAMN09205111	NC68B_ISP2_KSdomain
SAMN09205112	NC68B_LWA_KSdomain
SAMN09205113	NC68B_M1_KSdomain
SAMN09205114	H054A_A1_KSalphadomain
SAMN09205115	H054A_Chitin_KSalphadomain
SAMN09205116	H054A_DISP2_KSalphadomain
SAMN09205117	H054A_ISP2_KSalphadomain
SAMN09205118	H054A_LWA_KSalphadomain
SAMN09205119	H054A_M1_KSalphadomain
SAMN09205120	H054B_A1_KSalphadomain
SAMN09205121	H054B_Chitin_KSalphadomain
SAMN09205122	H054B_DISP2_KSalphadomain
SAMN09205123	H054B_ISP2_KSalphadomain
SAMN09205124	H054B_LWA_KSalphadomain
SAMN09205125	H054B_M1_KSalphadomain
SAMN09205126	H054B_SED_KSalphadomain
SAMN09205127	NC68A_A1_KSalphadomain
SAMN09205128	NC68A_Chitin_KSalphadomain
SAMN09205129	NC68A_DISP2_KSalphadomain

APPENDIX A (continued)

SAMN09205130	NC68A_ISP2_KSalphadomain
SAMN09205131	NC68A_LWA_KSalphadomain
SAMN09205132	NC68A_M1_KSalphadomain
SAMN09205133	NC68A_SED_KSalphadomain
SAMN09205134	NC68B_A1_KSalphadomain
SAMN09205135	NC68B_Chitin_KSalphadomain
SAMN09205136	NC68B_DISP2_KSalphadomain
SAMN09205137	NC68B_ISP2_KSalphadomain
SAMN09205138	NC68B_LWA_KSalphadomain
SAMN09205139	NC68B_M1_KSalphadomain

Figure 13. Formula used to calculate 16S and BGC percent recovery from sediment.

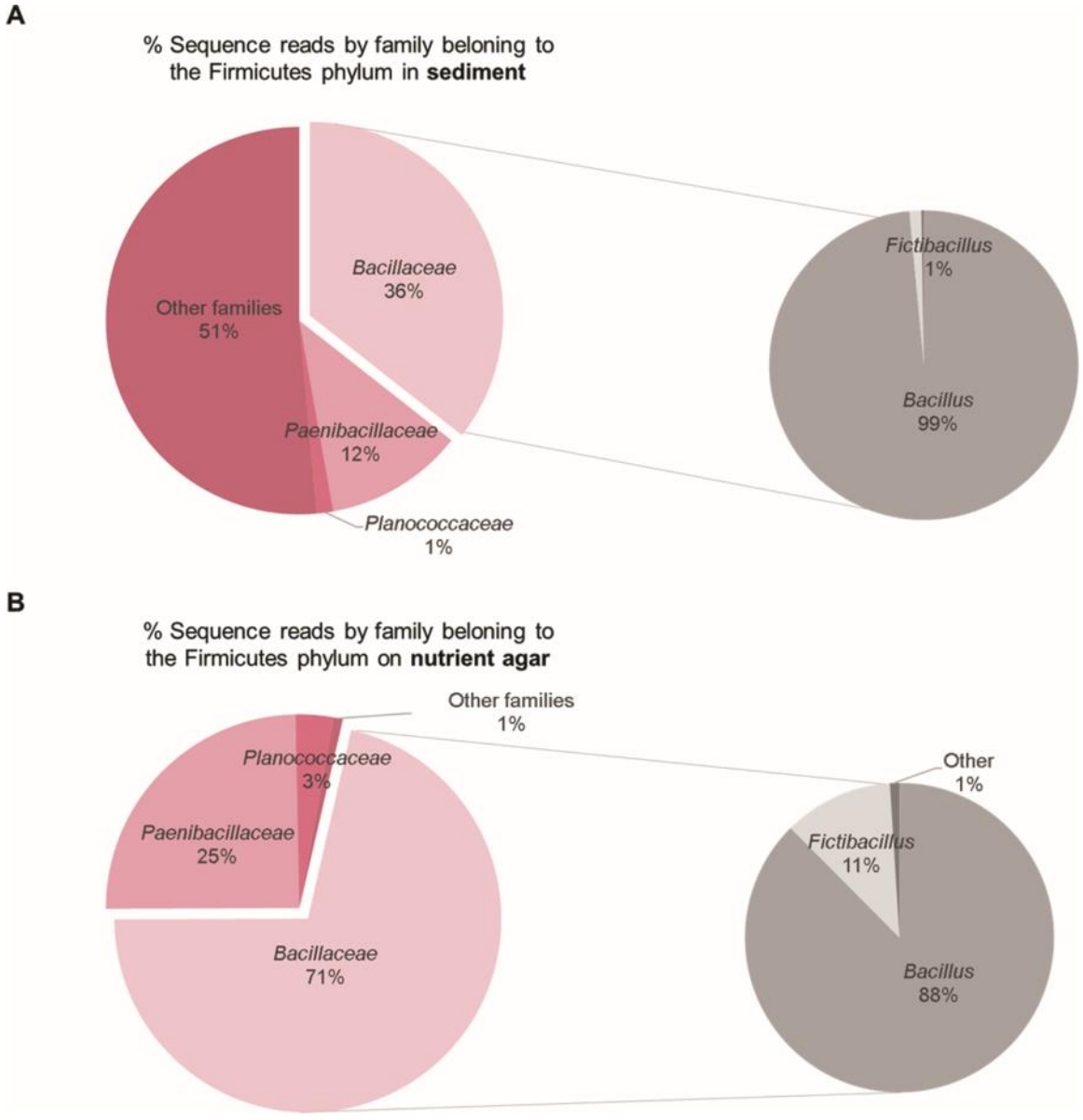
The percent recovery of 16S rRNA, KS, KSα, and A domain OBUs is calculated as follows:

$$100 \times \frac{\#\ of\ observed\ OBUs\ present\ on\ nutrient\ agar}{(\#\ of\ observed\ OBUs\ in\ sediment\ +\ \#\ of\ observed\ OBUs\ on\ nutrient\ agar - \#\ of\ overlapping\ OBUs)}$$

We observed minimal overlap between sequences detected in sediment and sequences observed on nutrient agar; this was likely due to our selection of spore forming bacteria on nutrient agar, which were present at low concentrations and just below our sequencing detection limit in sediment. This may also be due to lower sampling depth in sediment sequencing in comparison to nutrient agar, or the low yield of genomic DNA extracted from spores. We presumed that the sequences detected on nutrient agar also existed in sediment, so adding the two totals together would afford us greater accuracy when calculating the estimated OTU and OBU recovery. The numbers reported represent a conservative estimate of the bacterial and NP recovery. A deeper sequencing of sediment microbial communities and their NP may yield smaller percent recovery than reported.

APPENDIX A (continued)

Figure 14. Firmicutes sequence reads on sediment and on nutrient agar.

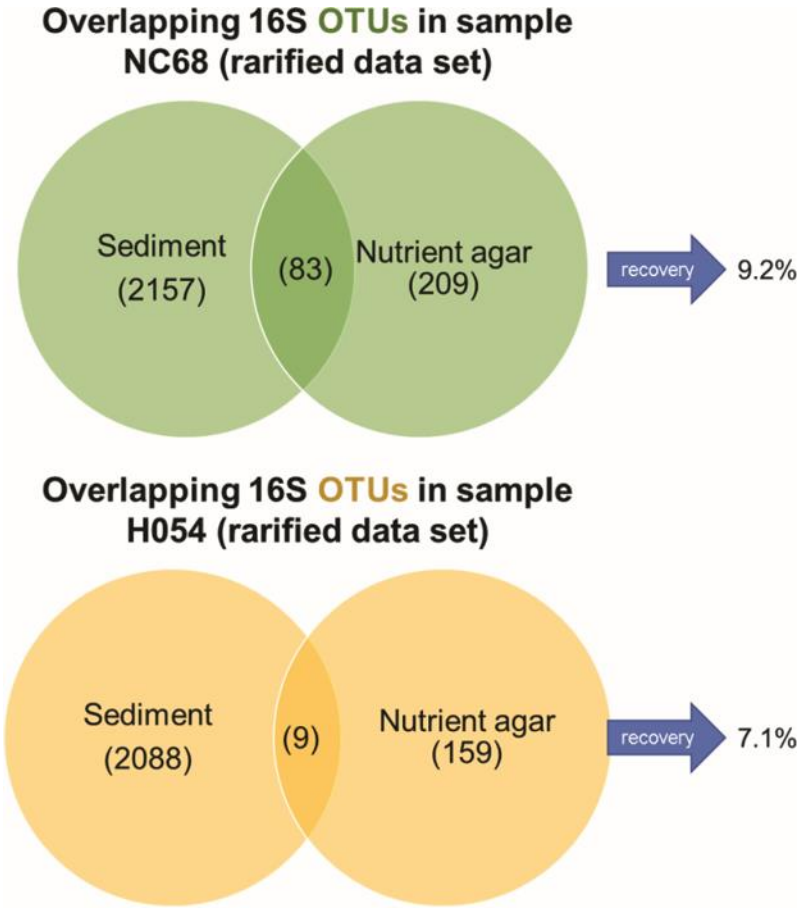


APPENDIX A (continued)

Figure 15. Overlap of OTUs in sediment and on nutrient agar.

We calculated the overlap in terms of OTUs and OBUs per sample. The few OTUs and OBUs detected on nutrient media exhibited minimal overlap with corresponding sediment. One way to explain this disparity is that several bacterial genera were not present/abundant on cultivation media due to our selection techniques for spore-forming bacteria.

The recovery percentages were calculated using the formula in Figure 13.



APPENDIX A (continued)

Figure 16. Overlap of OBUs in sediment and on nutrient agar.

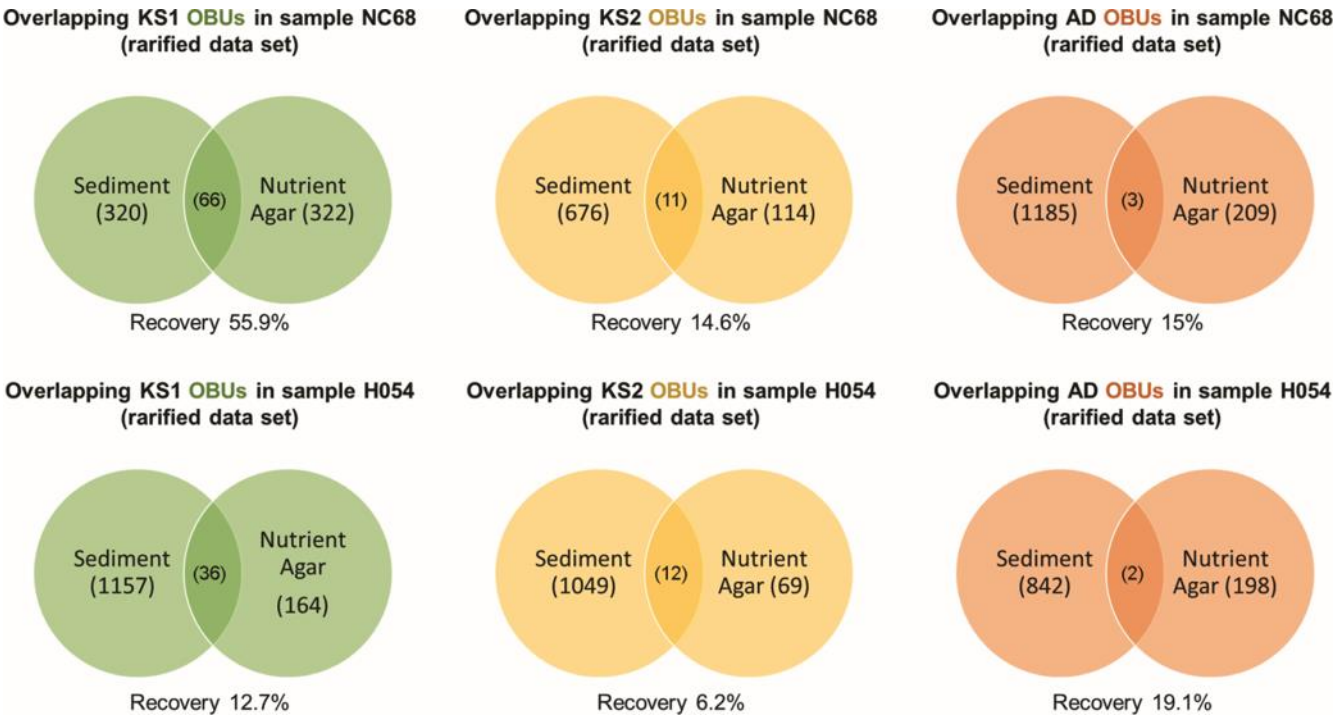
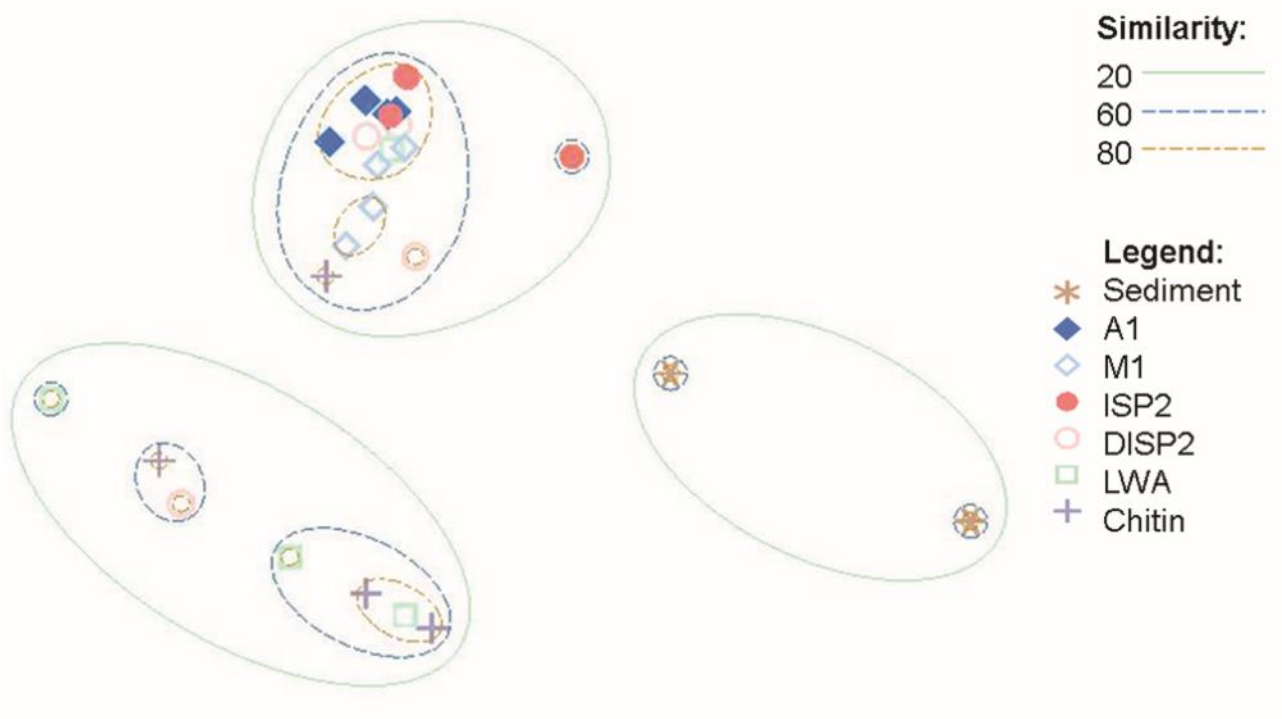


Figure 17. Bacterial and BGC community differences in sediment and on nutrient agar.

To visualize the difference in microbial community between sediment and the different nutrient agars, we used the biological observation matrix (BIOM) to generate the following non-metric multidimensional scaling (NMDS) plots. First, sequence data for each sample was standardized by total, and then the Bray-Curtis similarity scores were computed. The Bray-Curtis similarity is an algorithm that compares samples to each other across all taxa/OBUs and it considers the relative abundance of each taxon/OBU. The higher the Bray-Curtis value, the more similar the two samples. The Bray-Curtis similarity scores were generated based on family-level taxonomic classification for 16S sequence data and at 85% OBU-level for BGC data using Primer6 (Primer-E, version 6.1.13, United Kingdom). The Bray-Curtis resemblance matrix generated (see Table VII) was used to generate NMDS plots for 16S sequence data at family-level and at 85% OBU-level for BGC data using Primer6 (Primer-E, version 6.1.13, United Kingdom). The NMDS takes the Bray-Curtis data and transforms it into a ranking system (the most similar pair of samples = 1, the second most similar pair =2, and so on). Then the data are plotted in n-1-dimensional space and compressed into (in this case) two dimensions. The stress factor represents how much information is lost by this compression (less than 0.2 is considered acceptable).

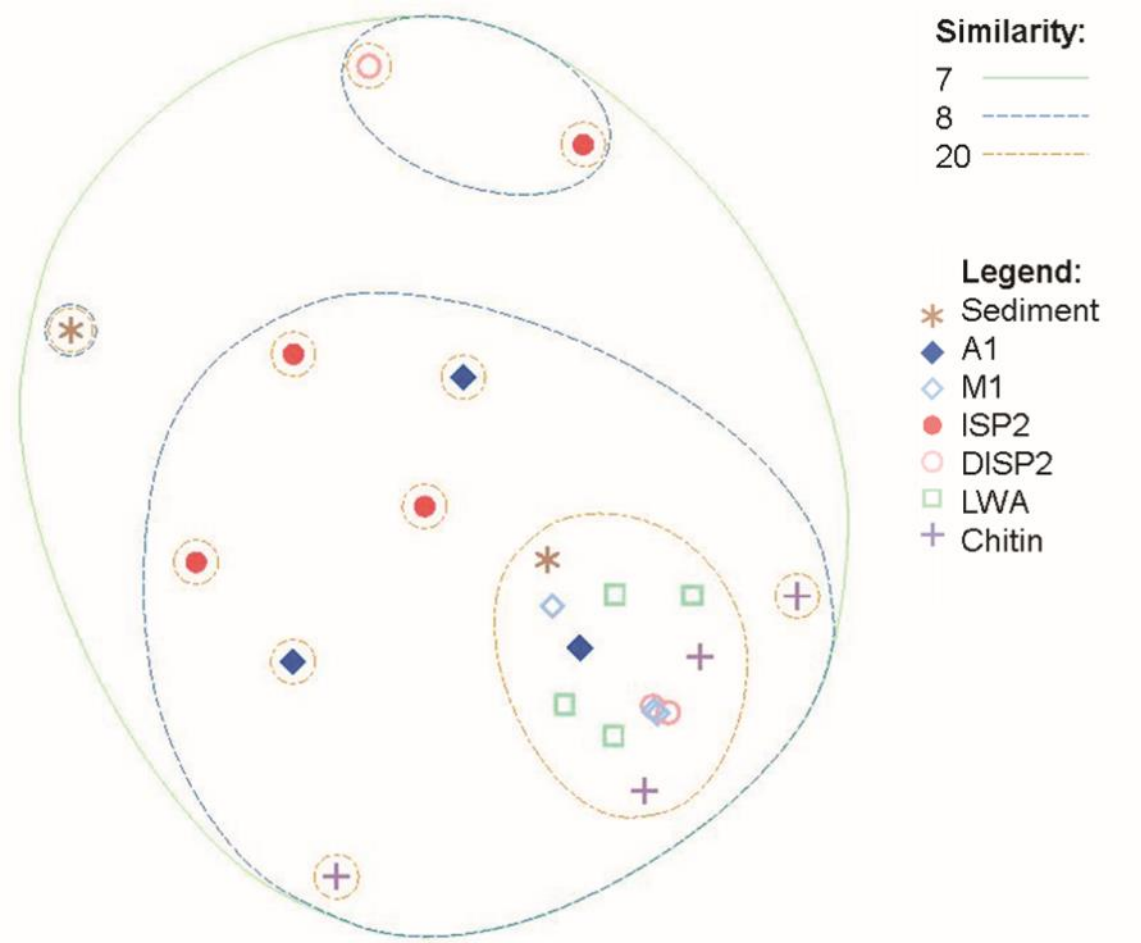
APPENDIX A (continued)

17A. Bray-Curtis similarity between 16S sequence data in sediment and on nutrient agar samples at the family-level (2D stress = 0.09).



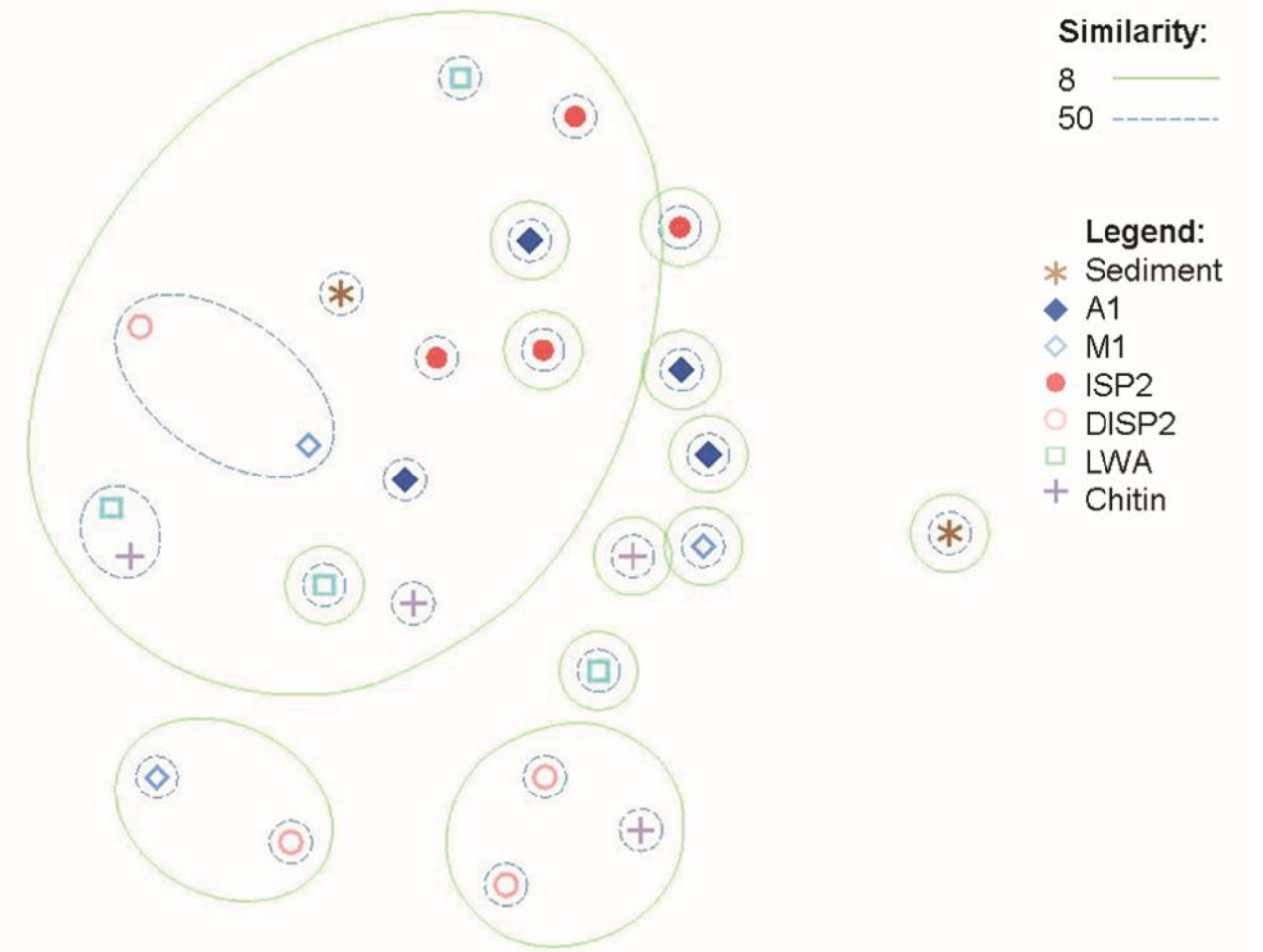
APPENDIX A (continued)

17B. Bray-Curtis similarity between KS domain sequence data in sediment and on nutrient agar at 85% OBU-level (2D stress = 0.13).



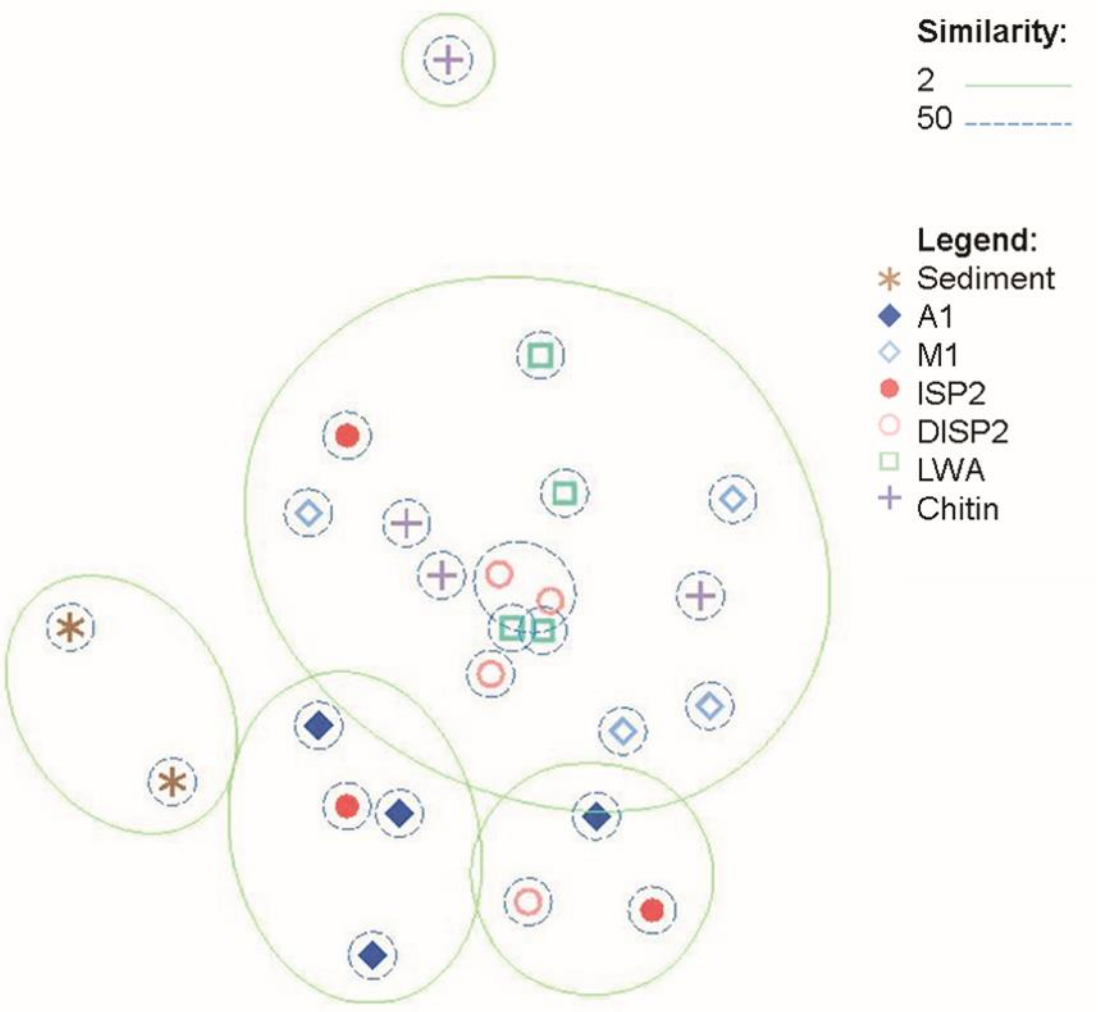
APPENDIX A (continued)

17C. Bray-Curtis similarity between KSα domain sequence data in sediment and on nutrient agar samples at 85% OBU-level (2D stress = 0.22).



APPENDIX A (continued)

17D. Bray-Curtis similarity between A domain sequence data in sediment and on nutrient agar samples at 85% OBU-level (2D stress = 0.17).



APPENDIX A (continued)

Figure 18. Lake Huron sediment collection sites H054 (green) and NC68 (pink).

The sediment samples H054 (green) and NC68 (pink) were collected using PONAR. Sample H054 was collected at a depth of 134.9 m at longitude of -83.402917 and a latitude of 45.633767. Sample NC68 was collected at a depth of 17.3 m at longitude of -83.853633 and a latitude of 46.054600.

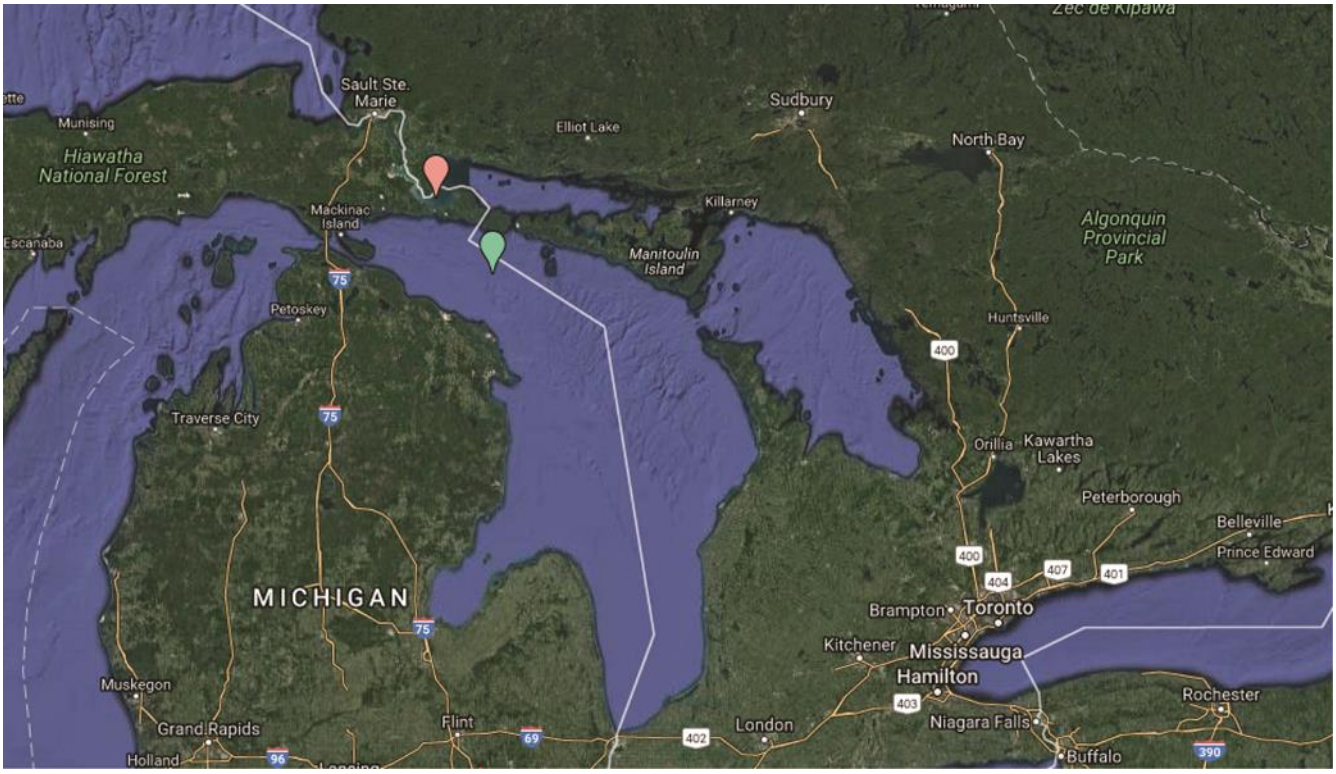


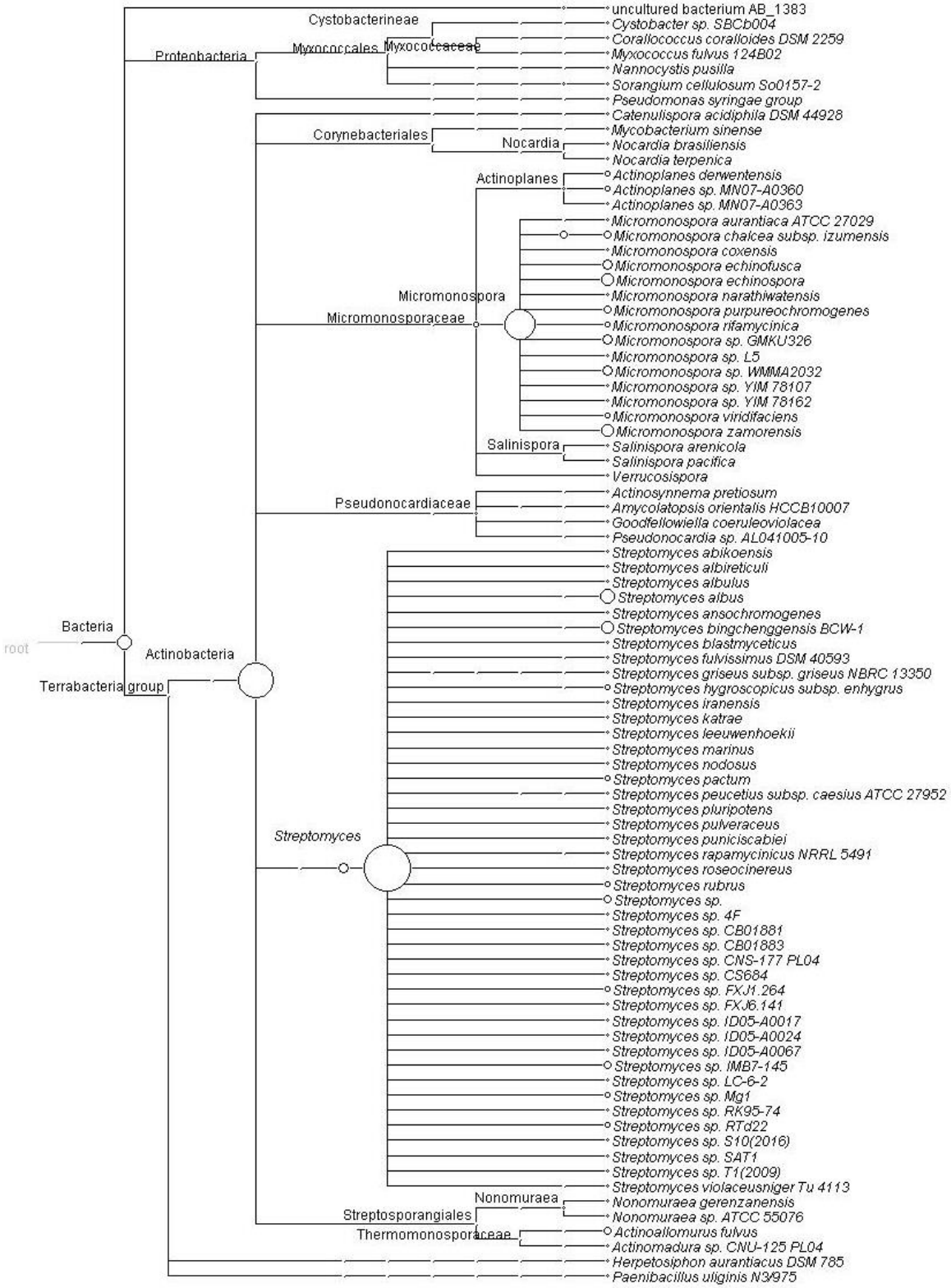
Figure 19. MEGAN taxonomy assignments of BGC sequence reads.

In order to check whether BGC sequences originating from nutrient agar samples belong to characterized/cultivated organisms or whether they belong to uncultivated organisms, we aligned OBU representative sequences against the NCBI nucleotide database (NCBI-NT)[172] using the software package DIAMOND.[92] Although OBU representative sequences aligned to NCBI-NT sequences at different percentages and e-values, all results were analyzed without *a priori* thresholds. Data were visualized using the software package MEGAN.[173] The input file to MEGAN contains the accession numbers of the sequences to which each OBU representative sequence was aligned. MEGAN[173] parses and analyzes the input file to estimate the taxonomic content ("species profile") of the nutrient agar sample from which each OBU representative sequence read was collected. It uses different algorithms to place reads into a given taxonomy by assigning each read to a taxon at some level in the NCBI hierarchy, based on their hits to known sequences, as recorded in the alignment file. The resulting cladogram shows the taxon and the number of reads assigned to the taxon. These cladograms are not weighted by the number of sequences in each OBU. The size of a node is scaled logarithmically to represent the number of assigned OBUs. These results likely represent an underestimation of the taxonomic diversity of the source of the sequence reads, as only one representative sequence from each OBU was used for this analysis. It is possible that each OBU contains reads that can belong to multiple different taxonomic groups. The resulting cladograms show that the majority of OBUs on plates originate from the phyla Actinobacteria (predominantly genera *Streptomyces* and *Micromonospora*), Proteobacteria and some Firmicutes. However, there are still OBUs that originate from

understudied genera. This analysis is only predictive and is limited by the size of the NCBI NT database.

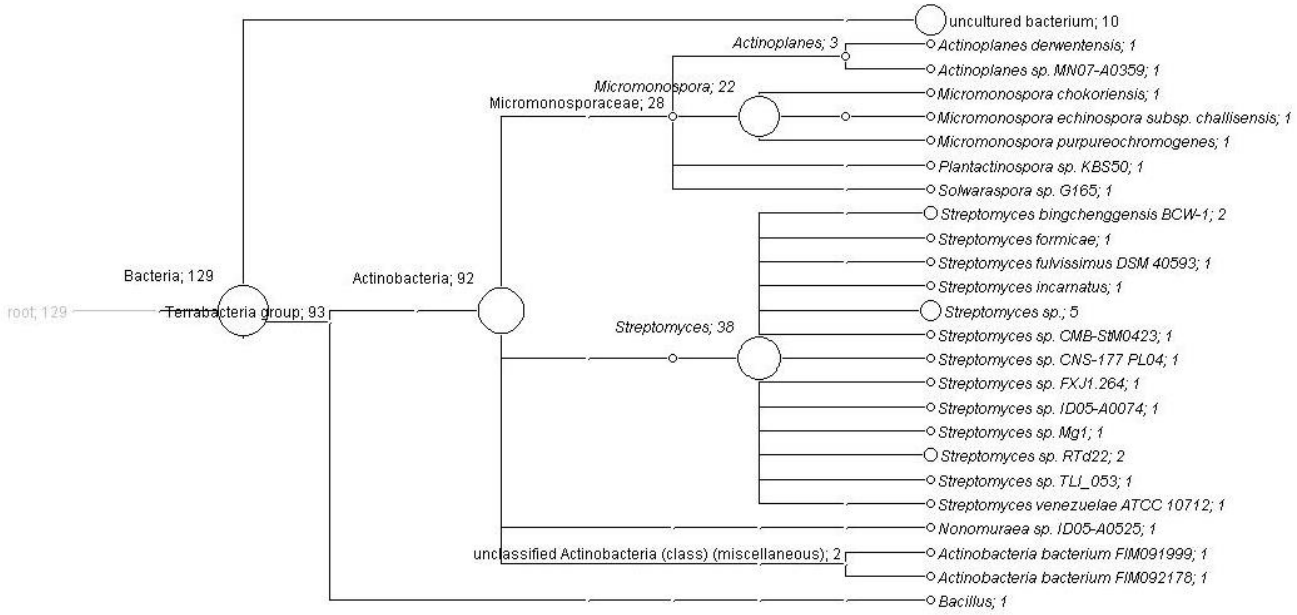
APPENDIX A (continued)

19A. Taxonomic assignment of KS domain OBUs originating from nutrient agar samples.



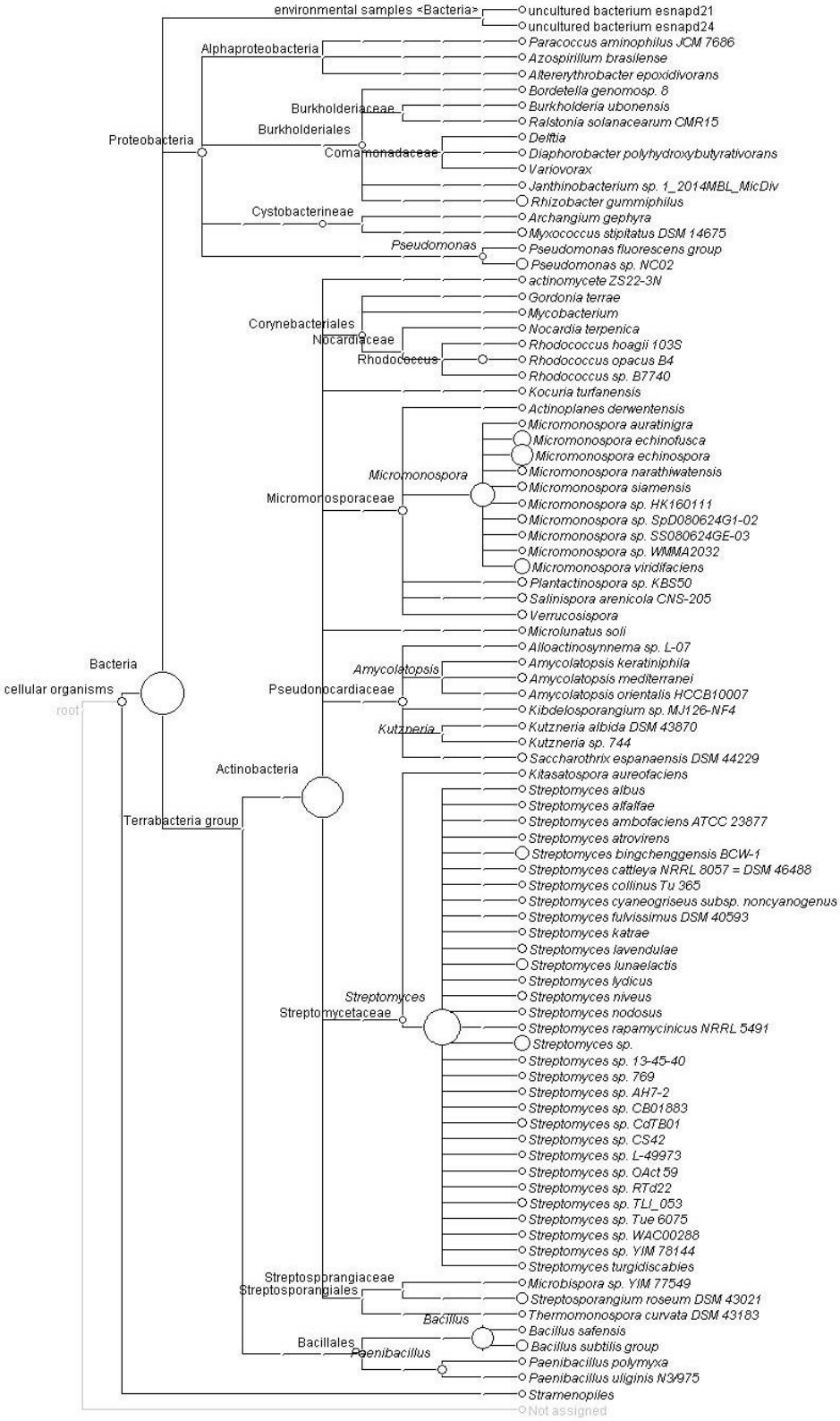
APPENDIX A (continued)

19B. Taxonomic assignment of KSα domain OBUs originating from nutrient agar samples.



APPENDIX A (continued)

19C. Taxonomic assignment of A domain OBUs originating from nutrient agar samples.



Supplementary Discussion

Charlop-Powers Z. et al. (2014) compared biosynthetic gene richness and diversity from 96 different sediment samples located throughout the southwestern and northeastern regions of the United States.[78] The Chao1 diversity metric estimated the presence of 1,000 to greater than 7,000 OBUs clustered at 95% sequence identity per soil microbiome. Moreover, based on our calculations from their published data, only 30% of these OBUs identified to KS and A domain fragments found in functionally characterized gene clusters. In a separate study, Charlop-Powers Z. et al. (2015) compared NP biosynthetic potential of soil samples from a diverse array of environmental microbiomes.[42] They showed that 185 biomes predicted greater than 350,000 OBUs for each of the two studied domains, KS and A, with rarefaction analysis suggesting that the sequence space had not yet been saturated. In addition, the authors found that only 5-10% of the total KS and A domain sequences originating from all 185 biomes were confidently assigned to known gene clusters using the eSNaPD algorithm. Our results corroborate these efforts. We observed approximately 3.5-, 12-, and 5.4-fold greater KS, KSα, and A domain OBUs in sediment compared to those on nutrient agar. Moreover, the Shannon[80] diversity index was found to be significantly greater in sediment compared with nutrient agar. These results highlight the disparity in NP biosynthetic diversity between the sediment and the nutrient agar populations.

APPENDIX B. Supporting Information for Chapter 3

Supplementary Experimental Procedures.

16S rRNA Gene Amplification and Sequencing

The V4 region of small subunit rRNA genes (16S rRNA) was PCR-amplified from genomic DNA using a two-stage PCR protocol, as described previously.[84] Primers 515F (5′-GTGCCAGCMGCCGCGGTAA-3′) and 806R (5′-GGACTACHVGGGTWTCTAAT-3′) were synthesized with 5′ linker sequences CS1 (forward primer; ACACTGACGACATGGTTCTACA) and CS2 (reverse primer; TACGGTAGCAGAGACTTGGTCT).[85] Each 25 μL PCR reaction mixture consisted of 0.5 μL of DNA, 0.8 μL of 10 μM of 515F, 10 μM of 806R, 12.5 μL KAPA Taq 2X ReadyMix (Kapa Biosystems), and 10.4 μL of deionized (DI) water. The thermal cycling conditions were set to a denaturation step at 95 °C for 5 min, 28 cycles of 95 °C for 30 s, 45 °C for 60 s, and 68 °C for 90 s, and a final elongation step at 68 °C for 7 min. Amplification products were observed by agarose gel electrophoresis and purified using Qiagen's QIAquick PCR cleanup kit, according to the manufacturer's protocol (Qiagen, Inc.). Subsequently, a second PCR amplification was performed to incorporate Illumina sequencing adapters and a sample-specific barcode into the amplicons. Each reaction received a separate primer pair with a unique 10-base barcode, obtained from the Access Array Barcode Library for Illumina (Fluidigm, South San Francisco, CA). In addition to Illumina adapter sequences and sample-specific barcodes, these "Access Array" primers contained the Illumina CS1 and CS2 linker primers at the 3′ ends of the oligonucleotides. Cycling conditions were as follows: 95 °C for 5 min, followed by 8 cycles of 95 °C for 30 min, 60 °C for 30 min, and 72 °C for 60 min. The pooled libraries, with a 20% phiX spike-in, were loaded onto MiSeq V2 flow cells, and sequenced. Fluidigm sequencing primers, targeting the CS1 and CS2 linker regions, were used to initiate paired-end 2 × 250 base read

sequencing. Library preparation, pooling, and sequencing were performed at the University of Illinois at Chicago Sequencing Core (UICSQC).

Bioinformatic Analyses of 16S rRNA Sequence Data

Approximately 6.5 million 16S rRNA sequencing reads were obtained for 59 sediment samples in duplicate. All sequence data generated from the Illumina MiSeq sequencer were first pre-processed using the QIIME-1.9.7 pipeline[86] at the UIC Sequencing Core. Bar-coded 16S rRNA gene sequences were demultiplexed, primers and chimeras were removed, and the reads were filtered according to Phred quality scores. Forward and reverse reads were merged and labeled according to sample source. Samples were then processed according to the "Moving Pictures" tutorial in Qiime2[174] using the DADA2 option for sequence quality control and feature table construction. The resulting analysis generated 141,078 amplicon sequence variants (ASVs) – "a higher-resolution analogue of the traditional OTU table".[175] A sequence representative was extracted from each ASV and was classified using the Silva_128 database.[87] A taxon-by-sample abundance matrix (a feature table or biological observation matrix, BIOM)[88] file was then created.

APPENDIX B (Continued)

Table XII. Sediment sample collection data for Lake Huron expedition.

Sample name	Longitude	Latitude
GB01	-80.85639333	44.71783667
GB03	-80.61701	44.72527
GB04	-80.16726167	44.64574667
GB05	-80.24311167	44.796915
GB06	-80.435975	44.73815167
GB09	-79.9675	44.87164167
GB12	-80.87475333	44.92021167
GB17	-80.87422167	45.24485
GB29	-81.08299167	45.58357
GB35	-81.670485	45.52572833
GB36	-81.620125	45.70816833
GB39	-81.25839833	45.87294667
GB42	-81.59540667	45.91245667
H001	-83.61419167	43.937425
H002	-83.33244667	44.12494167
H006	-82.01849667	43.52649333
H012	-82.11304667	43.900655
H027	-82.50245667	44.09988833
H032	-82.35962333	44.35418333
H037	-82.78362833	44.76185333
H038	-82.20237833	44.75069333
H048	-82.59118667	45.26139333
H054	-83.402845	45.63384
H061	-83.91640833	45.74978833
H096	-82.83258	44.33275
H101	-82.33487667	43.26900667
H102	-82.403855	43.70586833
H103	-82.22092167	44.14485833
H104	-81.83796	44.37196167

APPENDIX B (Continued)

H107	-82.554065	44.61541667
H108	-83.05021	44.557415
H109	-83.000015	44.150185
H110	-83.80368833	43.77230833
H118	-83.165955	44.91682333
H119	-82.81068167	45.39766833
H121	-83.403945	45.81889667
H123	-83.90591	45.93646167
H124	-84.42156833	45.85121
HTXD	-82.33345	43.33989
HTXM	-82.46681	43.33977
HTXS	-82.49911667	43.33974333
HXSG	-82.49911667	43.33974333
NC68	-83.85360333	46.04127
NC70	-83.671975	46.13648
NC71	-83.74624	46.23346833
NC73	-83.35517833	46.18685167
NC76	-83.43291167	46.00034
NC77	-83.19770833	45.97041667
NC79	-82.886655	46.12299667
NC82	-82.7588	45.93686333
NC83	-82.5497	45.99998167
NC84	-82.55644167	46.09173833
NC87	-82.197085	46.06112167
NC88	-81.999815	46.05529667
NC89	45.91649	-82.16171167
TB01	-83.14963667	44.89958667
TB02	-83.240505	44.93872833
TB03	-83.277	44.95524667
TB04	-83.03529444	44.15244444

APPENDIX B (Continued)

Table XIII. List of molecular classes that KSα and A domain sequences aligned to in the MIBiG 2.0 database.[100]

XIIIA. List of identified A domain hits.

Molecular class	Molecular class detected in x samples	# of sequences belonging to molecular class
Pyoverdin	39	89
Scabichelin	13	26
Salinichelins	11	31
Albachelin	6	6
Polyoxypeptin	5	18
Cyclomarin D	5	15
Coelichelin	5	5
RP-1776	4	7
Arylomycin	4	5
Phthoxazolin	4	4
Thaxteramide A1/A2/B1/B2	3	11
Sarpeptin A/B	3	9
Anikasin	3	6
Aurantimycin A	3	6
Microtermolide A	3	6
Erythrochelin	3	5
Antimycin	3	3
Ficellomycin	3	3
Mycobactin	3	3
Taromycin A	2	8
Surugamide A/D	2	6
Tolaasin A	2	6
Coelibactin	2	5

APPENDIX B (Continued)

Clorobiocin	2	4
Pyxipyrrolone A/B	2	4
UK-68,597	2	3
Viscosin	2	3
Balhimycin	2	2
BE-43547 A1/A2/B1/B2/B3/C1/C2	2	2
GacamideA	2	2
Rakicidin A/B	2	2
Telomycin	2	2
Cadaside A/B	1	3
CDA 1b/2a/2b/3a/3b/4a/4b	1	2
Lokisin	1	2
Malonomycin	1	2
Massetolide A	1	2
Myxoprincomide-c506	1	2
OxalomycinB	1	2
Rhodochelin	1	2
A-47934	1	1
Colistin A/B	1	1
Cyphomycin	1	1
Cystothiazole A	1	1
Delftibactin A/B	1	1
Friulimicin A/B/C/D	1	1
Griseoviridin fijimycin A	1	1
Heterobactin A/S2	1	1
Myxochelin A/B	1	1
Nunapeptin nunamycin	1	1
Octapeptin C4	1	1
Polymyxin	1	1
Syringomycin	1	1
Thaxteramide C	1	1
Virginiamycin S1	1	1
Weishanmycin	1	1

APPENDIX B (Continued)

XIIIB. List of identified KSα domain hits.

Molecular class	Molecular class detected in x many samples	# of sequences belonging to molecular class
Griseorhodin A	39	78
Spore pigment	33	66
Rosamicin (salinipyrone A/pacificanone A)	9	25
Meridamycin	7	118
Rifamycin	6	32
Chaxamycin A/B/C/D	3	5
Sceliphrolactam	3	4
Epothilone B	2	16
Glycopeptidolipid	2	2
Rakicidin A/B	2	23
Tiacumicin B	2	14
7-deoxypactamycin	1	1
A83543A	1	1
Borrelidin	1	1
ECO-02301	1	1
Lydicamycin	1	1
Methylatedalkyl-resorcinol/Methylatedacyl-phloroglucinol	1	1
Piericidin A1	1	1
Streptovaricin	1	3
Tautomycetin	1	3
Tylactone	1	2

APPENDIX B (Continued)

Table XIV. Correlation coefficients between OTU/OBU groups.

In order to examine the correlation between the presence/absence and abundance between different OBUs and between OBUs and OTUs, the correlation coefficient between different groups was calculated using the following formula:[176]

$$Correl(X,Y) = \frac{\sum(x - x_{ave})(y - y_{ave})}{\sqrt{\sum(x - x_{ave})^2 \sum(y - y_{ave})^2}}$$

Where $\boldsymbol{x_{ave}}$ and $\boldsymbol{y_{ave}}$ are the sample means.

A correlation coefficient calculates the relationship between two OBU/OTU groups. A correlation coefficient of -1 denotes an absolute negative relationship, 0 denotes a lack of relationship, and 1 denotes a positive correlation. For example, a perfect negative relationship between two OBUs indicates that OBU1 is only present when OBU2 is not present. In contrast, perfect positive relationship indicates that OBU1 is only present when OBU2 is also present. The correlation between groups tested are reported in Tables S3A-C. All numbers were rounded up to display two decimals.

APPENDIX B (Continued)

XIVA. Correlation coefficients between select (and all) KSα and A domain OBUs and select (and all).

	KSα OBUs	A Domain OBUs	Siderophores	Antibiotics	Other bioactive NPs
16S OTUs	-0.10	-0.05	-0.01	-0.02	-0.25
Actinobacteria	0.18	-0.09	0.07	0.18	0.39
Proteobacteria	0.13	0.16	0.25	0.19	0.04

In general, there was no correlation between KSα domain OBUs and 16S OTUs or A domain OBUs and 16S OTUs. The presence/absence of siderophores, antibiotics, or other bioactive NPs did not correlate with the presence/absence of Actinobacteria or Proteobacteria.

APPENDIX B (Continued)

XIVB. Correlation coefficients between KSα domain OBUs.

To test for co-occurrence patterns, correlation coefficients were calculated for the most abundant KSα domain OBUs against each other. This resulted in the correlation matrix below. The rows and columns indicate the KSα OBUs in order of most to least abundance. One notable correlation observed was between the most abundant KSα domain OBU (KSα_1) and the fourteenth most abundant KSα domain OBU (KSα_14). The correlation coefficient for these OBUs was 0.999987 (reported as 1 in the table). To ensure that these OBUs were not nearly identical, the sequence representative for these OBUs were aligned against each other using BLAST. This yielded an identity of 68.93%. This suggests that (1) either these belong to the same BGC, but one of them is the starter KS domain which tends to separate from other KSs within the BGC, or (2) that these OBUs may co-occur in the environment, providing evidence of either phylogenetic or ecological forces that drive regional NP distribution.

	KSα_1	**KSα_2**	**KSα_3**	**KSα_4**	**KSα_5**	**KSα_6**	**KSα_7**	**KSα_8**	**KSα_9**	**KSα_10**	**KSα_11**	**KSα_12**	**KSα_13**	**KSα_14**	**KSα_15**	**KSα_16**	**KSα_17**	**KSα_18**	**KSα_19**	**KSα_20**
KSα_1		-0.04	-0.05	-0.01	0.07	-0.05	-0.06	-0.03	-0.05	-0.06	-0.05	-0.05	-0.02	1.00	-0.04	-0.03	-0.04	-0.03	-0.03	-0.05
KSα_2			0.13	-0.08	0.18	-0.07	-0.10	-0.07	0.02	0.38	-0.05	0.29	0.86	-0.04	0.78	-0.04	-0.08	-0.05	-0.06	0.69
KSα_3				0.09	0.33	0.15	-0.04	-0.16	0.19	0.08	-0.13	0.06	-0.10	-0.05	-0.07	0.02	-0.04	0.00	-0.16	-0.07
KSα_4					0.39	0.24	0.16	-0.07	0.02	-0.01	0.24	0.06	-0.08	-0.01	-0.10	0.34	-0.14	0.34	-0.11	0.21
KSα_5						0.16	-0.03	-0.11	0.01	0.36	-0.09	0.11	-0.05	0.07	-0.04	0.47	-0.15	0.51	-0.03	0.11
KSα_6							-0.06	-0.21	-0.17	-0.13	0.09	0.09	-0.05	-0.05	-0.01	0.02	0.09	0.00	-0.07	0.20
KSα_7								-0.04	0.07	-0.06	0.08	0.40	-0.06	-0.06	-0.08	0.14	0.53	0.13	-0.09	-0.01
KSα_8									-0.04	-0.05	-0.05	-0.06	-0.03	-0.03	-0.01	-0.02	-0.04	0.02	-0.03	-0.05

APPENDIX B (Continued)

KSα_9										0.11	0.01	-0.11	-0.05	-0.05	0.16	0.36	-0.07	0.18	-0.07	-0.01
KSα_10											-0.04	-0.05	0.16	-0.06	0.09	0.12	-0.10	0.05	0.14	0.10
KSα_11												-0.11	-0.05	-0.05	-0.06	-0.07	-0.05	-0.07	-0.07	0.03
KSα_12													0.38	-0.05	0.35	-0.06	0.51	-0.06	-0.07	0.43
KSα_13														-0.02	-0.04	-0.03	-0.04	-0.03	-0.03	-0.05
KSα_14															-0.04	-0.03	-0.04	-0.03	-0.03	-0.05
KSα_15																0.00	-0.09	0.00	-0.05	0.75
KSα_16																	-0.01	0.92	-0.04	0.18
KSα_17																		-0.05	0.05	-0.04
KSα_18																			-0.04	0.14
KSα_19																				-0.08
KSα_20																				

APPENDIX B (Continued)

XIVC. Correlation coefficients between A domain OBUs.

Similarly, co-occurrence patterns were examined by calculating correlation coefficients for the most abundant A domain OBUs against each other. This resulted in the correlation matrix below. The rows and columns indicate the KSα OBUs in order of most to least abundance. One notable correlation observed was between the twelfth most abundant A domain OBU (A_12) and the twentieth most abundant A domain OBU (A_20). The correlation coefficient for these OBUs was 0.94. To ensure that these OBUs were not nearly identical, the sequence representative for these OBUs were aligned against each other using BLAST. This yielded an identity of 92.00%. This provides additional evidence for cooccurrence patterns.

	A_1	**A_2**	**A_3**	**A_4**	**A_5**	**A_6**	**A_7**	**A_8**	**A_9**	**A_10**	**A_11**	**A_12**	**A_13**	**A_14**	**A_15**	**A_16**	**A_17**	**A_18**	**A_19**	**A_20**
A_1		-0.089	0.611	0.32	0.626	0.158	0.1	0.078	0.232	-0.14	0.22	-0.26	0.44	0.245	0.49	-0.1	0.092	0.727	-0.116	-0.23
A_2			0.076	0.02	-0.31	0.523	-0.2	0.616	-0.06	0.41	0.39	-0.31	0.1	-0.05	0.28	0.76	0.023	-0.31	-0.336	-0.25
A_3				0.62	0.202	0.575	-0.2	0.359	0.158	0.19	0.21	-0.5	0.58	0.125	0.56	0.18	0.461	0.295	-0.283	-0.44
A_4					0.164	0.172	-0.1	0.02	0.066	0.28	0.09	-0.38	0.15	0.062	0.1	0.11	0.474	-0	-0.21	-0.3
A_5						-0.11	0.33	-0.19	0.205	-0.36	-0.22	0.368	0.17	0.241	0.25	-0.3	-0.24	0.867	0.038	0.426
A_6							-0.1	0.782	0.01	0.26	0.27	-0.36	0.51	-0.02	0.66	0.57	0.275	-0.01	-0.337	-0.3

APPENDIX B (Continued)

A_7								-0.32	0.374	-0.16	-0.06	0.234	-0.17	0.379	-0.2	0	-0.17	0.306	0.415	0.252
A_8									-0.16	0.18	0.32	-0.37	0.59	-0.17	0.7	0.58	0.267	-0.11	-0.337	-0.33
A_9										-0.12	-0.07	-0.1	0.08	0.955	-0.05	-0	-0.09	0.362	-0.058	-0.09
A_10											0.32	-0.28	-0.03	-0.14	-0.03	0.56	0.369	-0.37	-0.189	-0.24
A_11												-0.33	0.13	-0.08	0.17	0.5	0.221	-0.13	-0.084	-0.27
A_12													-0.36	-0.09	-0.35	-0.4	-0.42	0.122	0.353	0.94
A_13														0.078	0.76	0.13	0.486	0.32	-0.261	-0.36
A_14															-0.01	-0	-0.11	0.384	-0.073	-0.07
A_15																0.25	0.247	0.394	-0.357	-0.31
A_16																	0.209	-0.33	-0.329	-0.29
A_17																		-0.16	-0.224	-0.36
A_18																			0.011	0.15
A_19																				0.189
A_20																				

Table XV. Shannon index and OBU count for individual samples before and after rarefaction.

The number of sequences per sample was rarified to the fewest sequence reads present in any sample (15 sequences for KSα domain OBUs and 3,487 for A domain OBUs). It was computed using the scikit-bio's diversity calculation via QIIME.[171] The Shannon (aka Shannon-Wiener) index is defined as:

$$H = -\sum_{i=1}^{s}(p_i \, log_2 p_i)$$

Where *s* is the number of OBUs and p_i is the proportion of the community represented by OTU *i*. The Shannon indices reported are for KSα and A domain OBUs before and after rarefaction. Both data was included because the fewest sequence reads present in KSα domain samples was too low (15 sequences) for significant conclusions. The Shannon indices are reported in Tables XIVA-B.

APPENDIX B (Continued)

XVA. Shannon index for KSα domain OBUs before and after rarefaction.

Before rarefaction			**After rarefaction**	
Sample	**Shannon**	**OBU count**	**Shannon**	**OBU count**
GB01	5.45	57	3.46	12
GB03	5.33	42	3.77	14
GB04	4.29	40	3.24	11
GB05	3.83	23	2.87	10
GB06	4.63	39	3.13	11
GB09	5.46	49	3.91	15
GB12	5.18	64	3.37	12
GB17	4.66	26	3.91	15
GB29	4.84	41	3.77	14
GB35	4.63	37	3.46	12
GB36	5.51	55	3.91	15
GB39	5.30	53	3.32	11
GB42	4.83	38	3.77	14
H001	4.28	23	3.51	12
H002	2.92	23	2.17	7
H006	5.02	40	3.06	10
H012	3.67	20	2.74	9
H027	4.84	43	3.46	12
H032	4.51	33	3.37	11
H037	3.16	19	2.61	7
H038	4.10	23	3.46	12
H048	4.09	36	3.14	10

APPENDIX B (Continued)

H054	**1.52**	**28**	**1.56**	**5**
H061	3.50	13	3.46	12
H096	3.79	30	3.46	12
H101	4.83	35	3.64	13
H102	4.10	31	3.51	12
H103	2.27	14	1.77	5
H104	3.73	15	3.37	12
H107	3.05	13	2.56	8
H108	3.77	24	2.68	8
H109	2.68	8	2.68	8
H110	3.19	55	1.55	4
H118	1.69	16	1.74	5
H119	3.95	25	3.06	9
H121	4.24	37	3.19	10
H123	6.28	114	3.64	13
H124	5.23	63	3.77	14
HTXD	4.26	26	3.19	10
HTXM	5.55	64	3.91	15
HTXS	4.77	49	3.46	12
HXSG	5.51	68	3.91	15
NC68	N/A	N/A	N/A	N/A
NC70	N/A	N/A	N/A	N/A
NC71	N/A	N/A	N/A	N/A
NC73	N/A	N/A	N/A	N/A
NC76	N/A	N/A	N/A	N/A

APPENDIX B (Continued)

NC77	**N/A**	**N/A**	**N/A**	**N/A**
NC79	5.14	51	3.51	12
NC82	6.71	135	3.77	14
NC83	5.38	99	2.00	7
NC84	5.43	67	3.77	14
NC87	6.01	91	3.51	12
NC88	5.69	55	3.77	14
NC89	5.55	53	3.91	15
TB01	5.74	79	3.64	13
TB02	6.50	97	3.91	15
TB03	5.80	77	3.64	13
TB04	4.18	21	3.77	14

APPENDIX B (Continued)

XVB. Shannon index for A domain OBUs before and after rarefaction.

Before rarefaction			**After rarefaction**	
Sample	**Shannon**	**OBU count**	**Shannon**	**OBU count**
GB01	8.95	1326	9.76	5956
GB03	9.47	1478	10.40	7158
GB04	9.27	1357	10.05	5621
GB05	9.34	1539	10.17	6914
GB06	8.44	1260	9.23	5430
GB09	10.03	1975	10.03	1975
GB12	9.64	1748	10.56	6741
GB17	9.30	1494	9.98	4783
GB29	9.54	1532	10.38	6569
GB35	9.77	1666	10.64	6797
GB36	9.27	1589	10.17	5979
GB39	9.06	1388	9.75	5845
GB42	9.34	1422	10.18	5557
H001	9.30	1576	10.12	5936
H002	9.22	1431	9.98	5677
H006	9.25	1465	10.10	6030
H012	9.06	1382	9.80	5580
H027	9.36	1576	10.13	6407
H032	8.96	1399	9.88	6165
H037	6.13	919	6.55	2508

APPENDIX B (Continued)

H038	**9.38**	**1494**	**10.29**	**6472**
H048	9.21	1404	9.91	5052
H054	8.89	1175	9.50	4467
H061	8.85	1213	9.32	3468
H096	9.19	1367	9.83	4138
H101	9.17	1382	9.98	5836
H102	9.41	1478	10.15	5342
H103	8.92	1304	9.67	5574
H104	9.42	1472	10.12	4769
H107	9.18	1418	9.96	5235
H108	9.11	1432	9.92	6175
H109	9.08	1348	9.92	5393
H110	9.16	1362	9.99	6244
H118	9.95	1783	10.88	7084
H119	9.83	1684	10.69	6735
H121	9.36	1459	10.14	5248
H123	9.63	1606	10.55	6806
H124	9.13	1392	10.04	7168
HTXD	9.60	1559	10.42	5848
HTXM	9.68	1617	10.69	7125
HTXS	9.77	1524	10.61	6173
HXSG	9.77	1486	10.55	5964

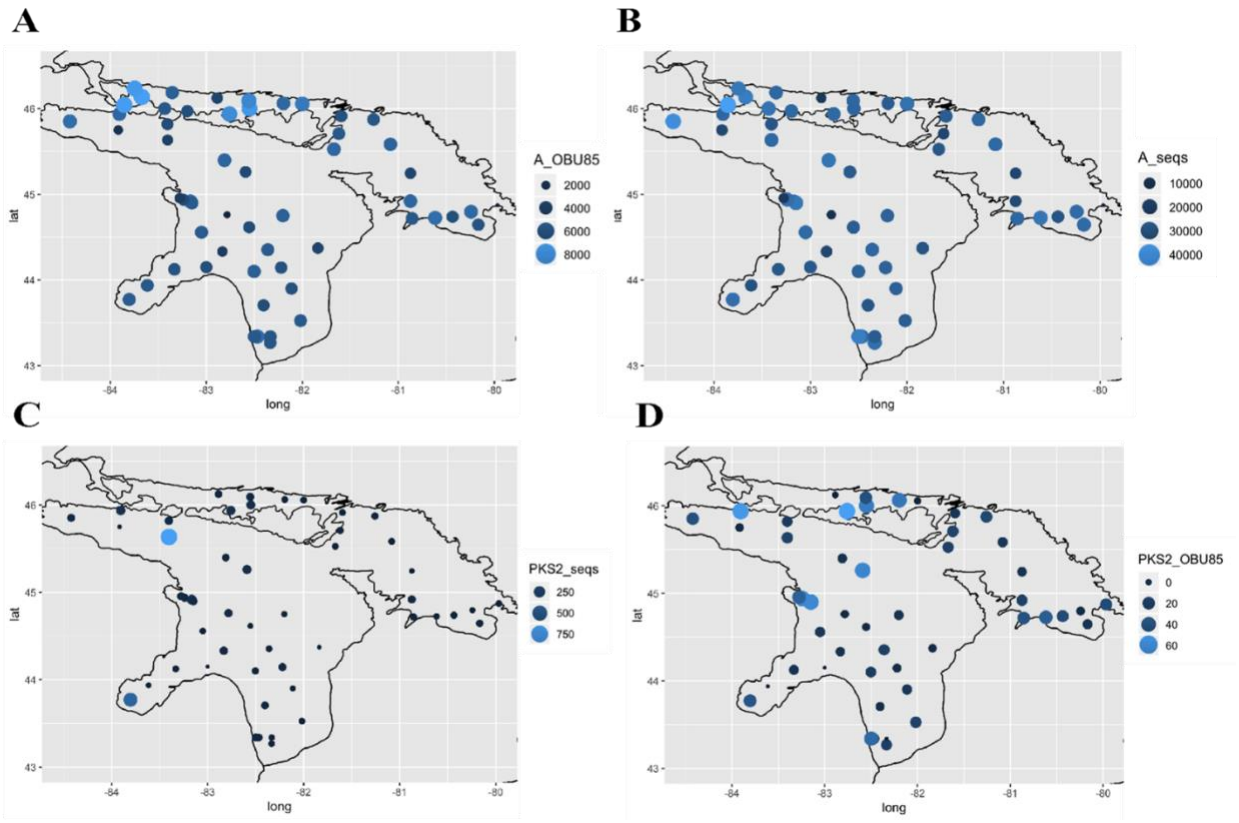
APPENDIX B (Continued)

NC68	**9.76**	**1728**	**10.77**	**9532**
NC70	9.65	1721	10.82	9611
NC71	9.80	1745	10.94	9350
NC73	8.87	1391	9.67	6408
NC76	8.87	1329	9.64	5979
NC77	9.05	1307	9.84	5450
NC79	9.62	1717	10.34	4791
NC82	9.71	1775	10.80	8694
NC83	9.69	1768	10.88	9247
NC84	9.51	1639	10.55	8432
NC87	9.32	1501	10.25	6832
NC88	9.36	1548	10.22	7740
NC89	N/A	N/A	N/A	N/A
TB01	9.52	1457	10.32	6129
TB02	9.65	1417	10.31	5227
TB03	8.99	1328	9.70	4152
TB04	8.97	1156	9.49	4105

APPENDIX B (Continued)

Figure 20. A and KSα domain OBU and sequence abundances.

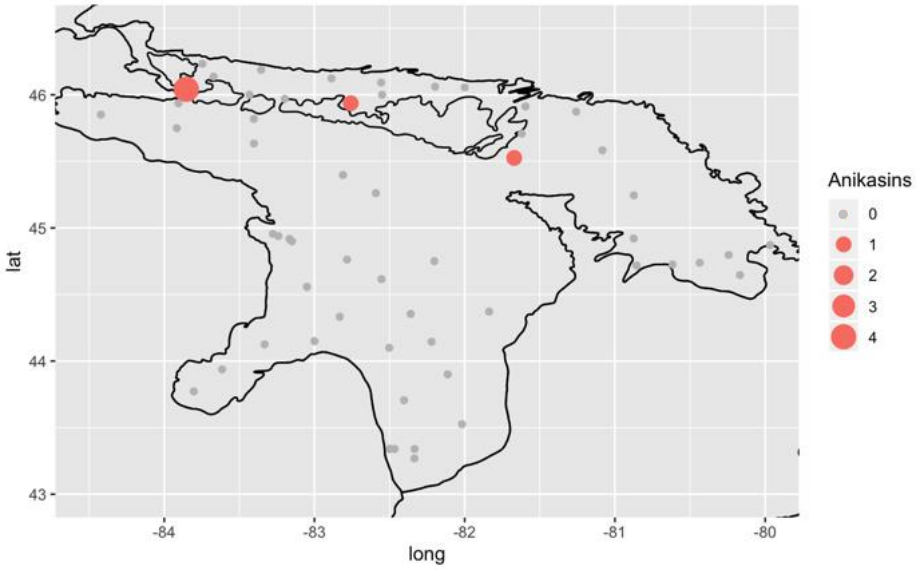
The sequence read abundance at each collection site was mapped and represented as different sized circles. A-D show the relative abundances of A domain OBUs clustered at 85% (A), of A domain sequences (B), KSα domain OBUs clustered at 85% (C), and KSα domain sequences. (D), respectively.



APPENDIX B (Continued)

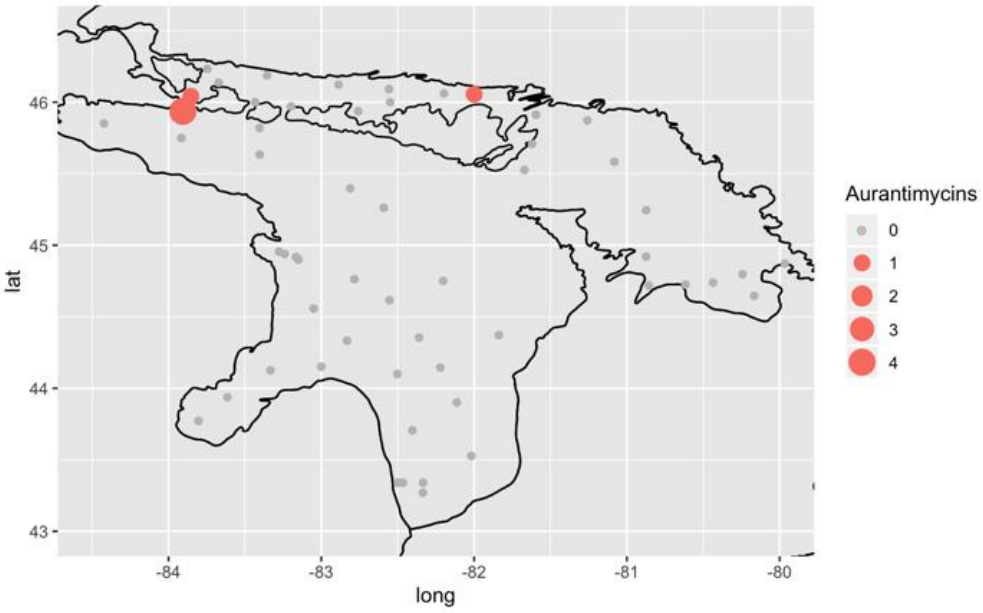
Figure 21. Occurrence of all detected antibiotics in Lake Huron sediment.

21A. Occurrence of anikasins in Lake Huron sediment.



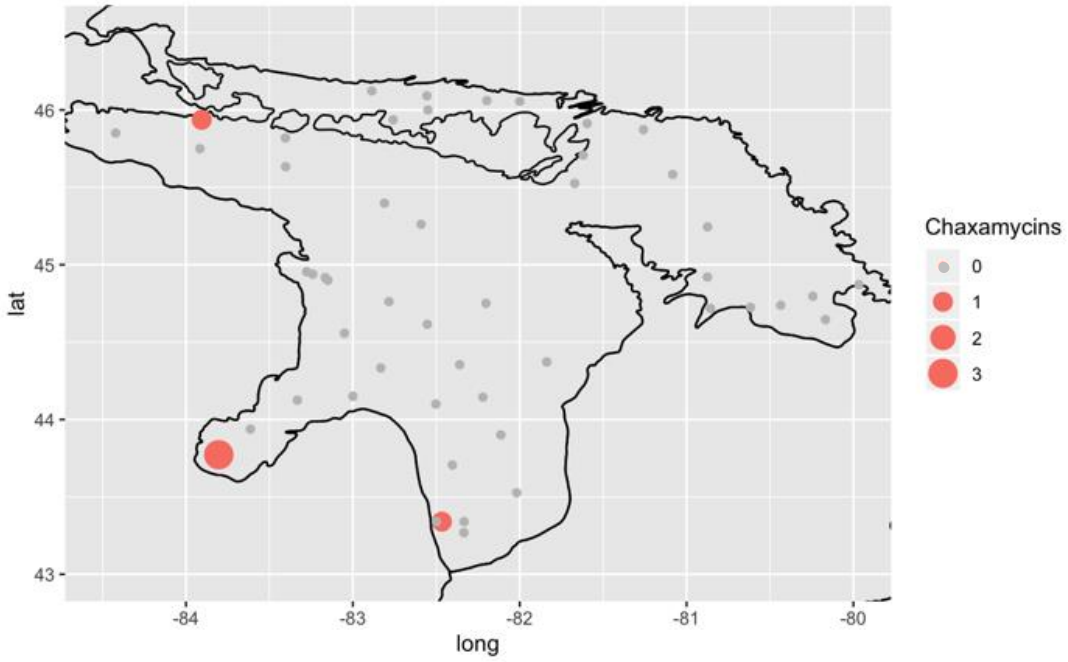
APPENDIX B (Continued)

21B. Occurrence of aurantimycins in Lake Huron sediment.



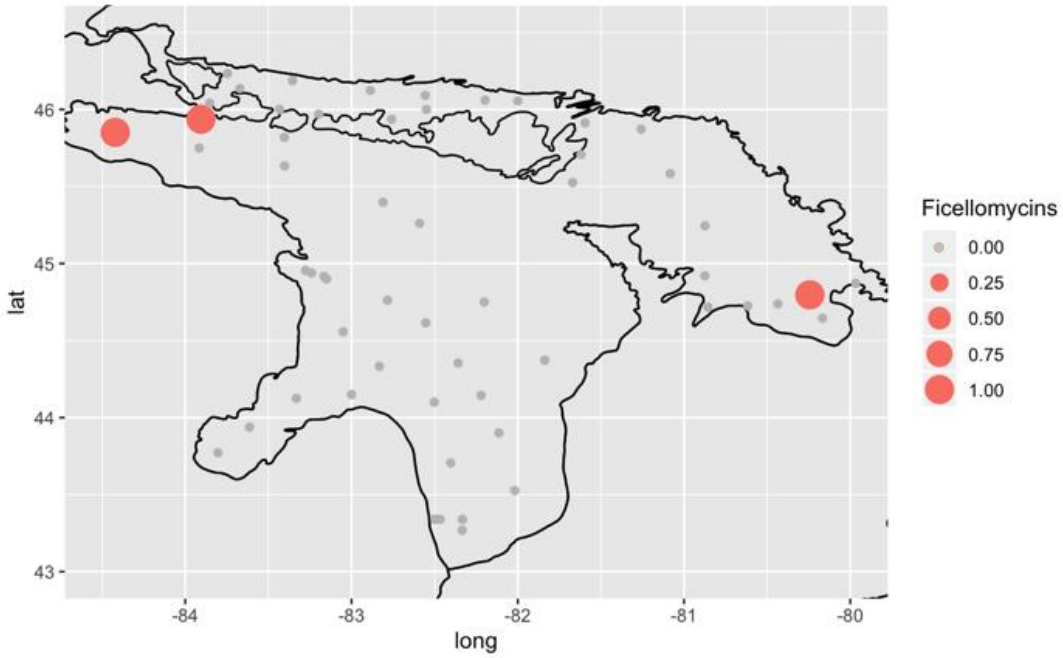
APPENDIX B (Continued)

21C. Occurrence of chaxamycins in Lake Huron sediment.

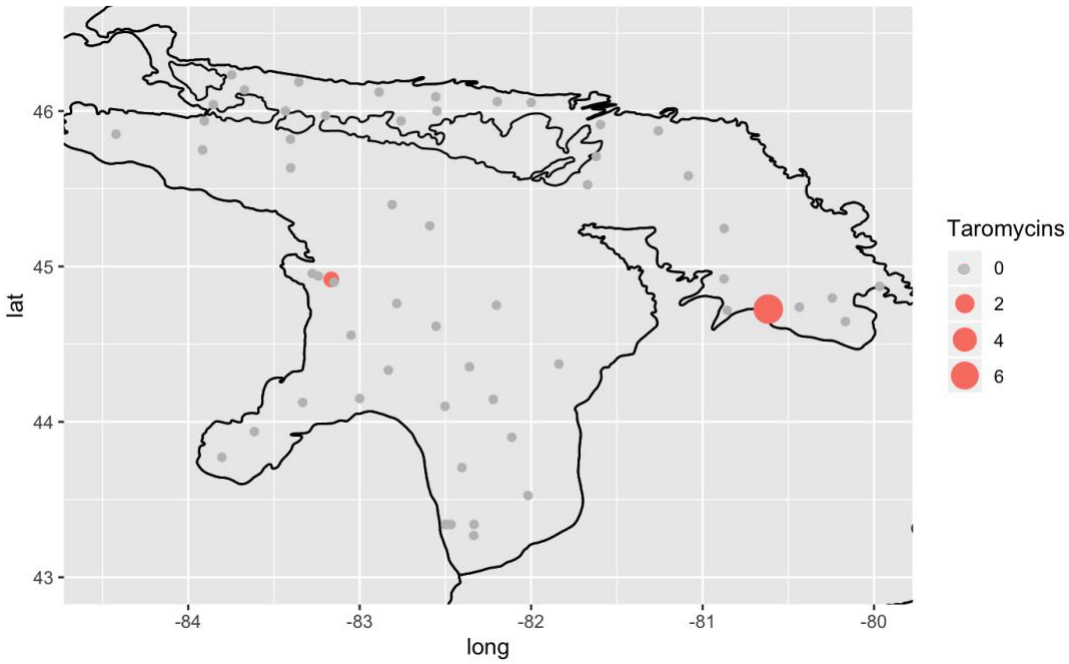


APPENDIX B (Continued)

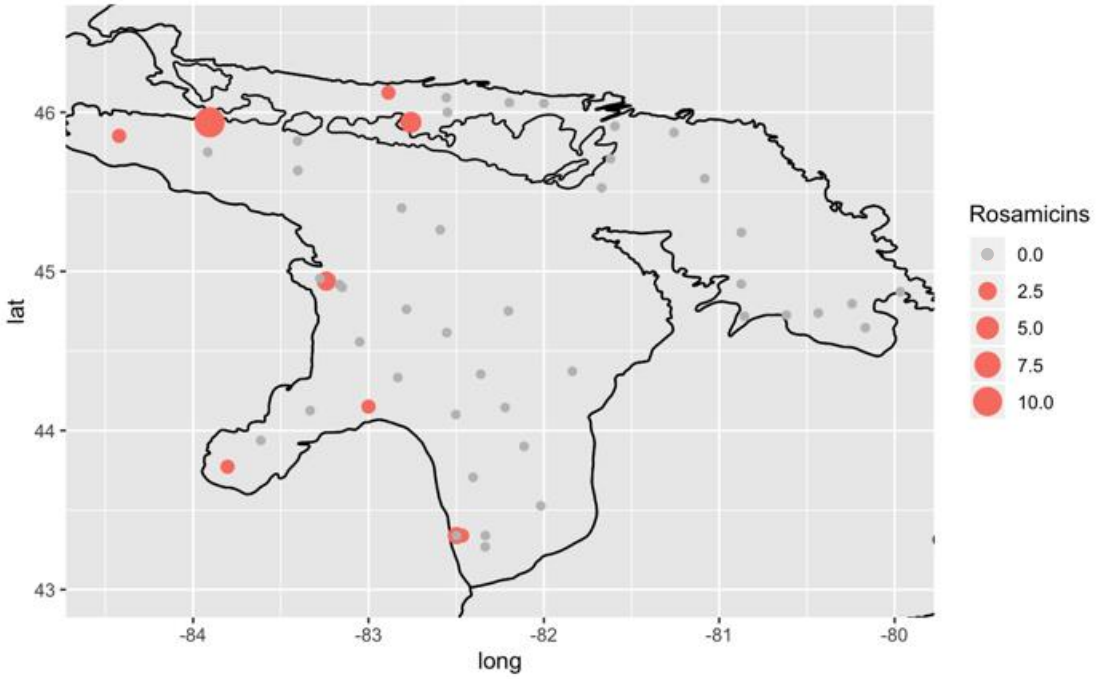
21D. Occurrence of ficellomycins in Lake Huron sediment.



21E. Occurrence of taromycins in Lake Huron sediment.



21F. Occurrence of rosamycins in Lake Huron sediment.



APPENDIX B (Continued)

21G. Occurrence of rifamycins in Lake Huron sediment.

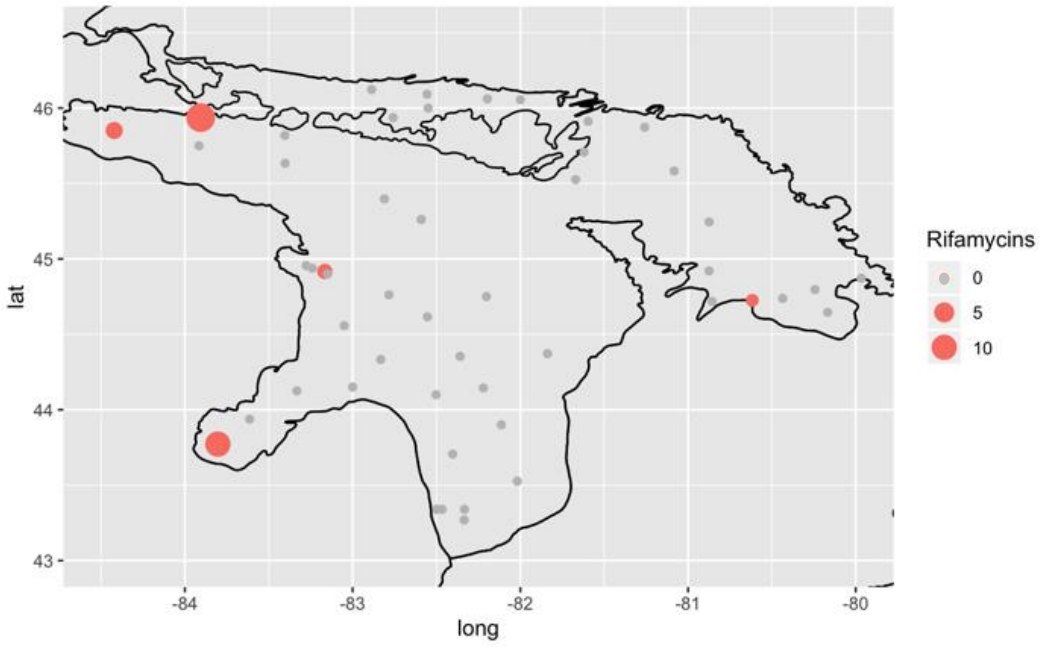
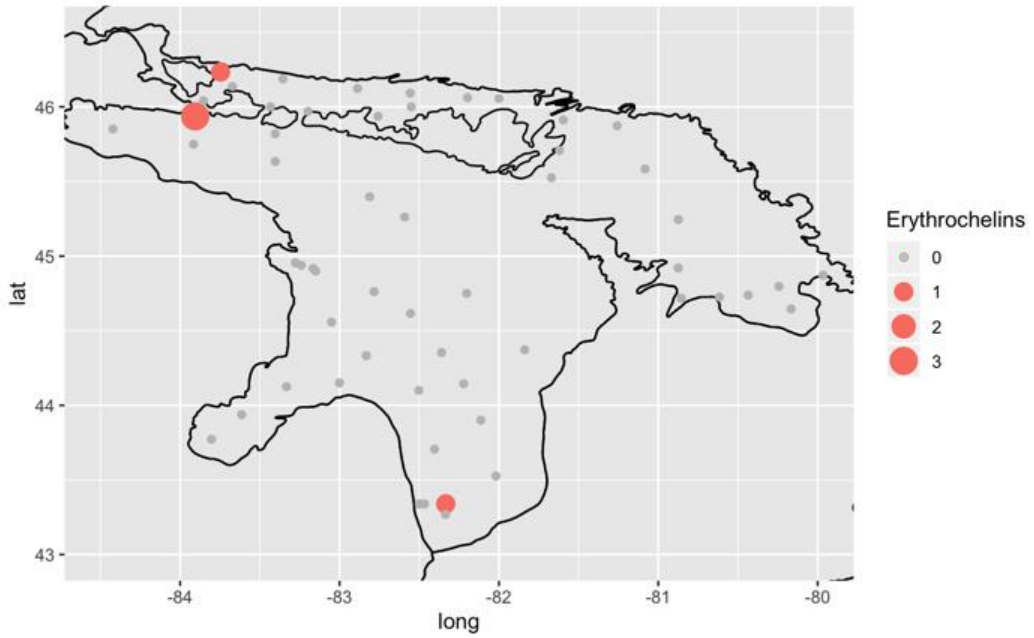
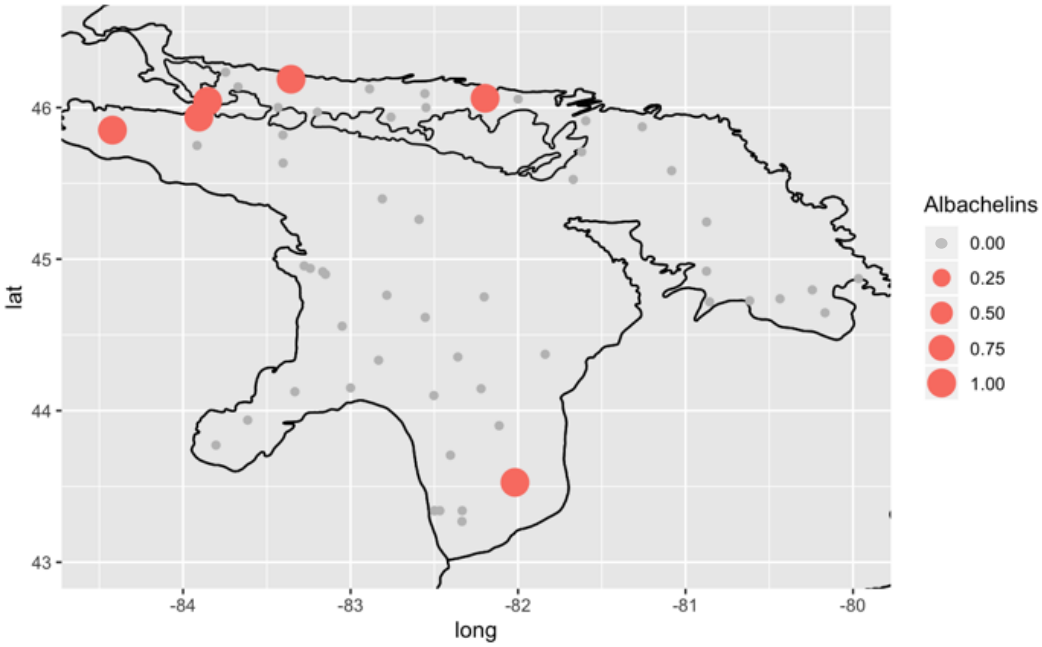


Figure 22. Occurrence of all detected siderophores in Lake Huron sediment.

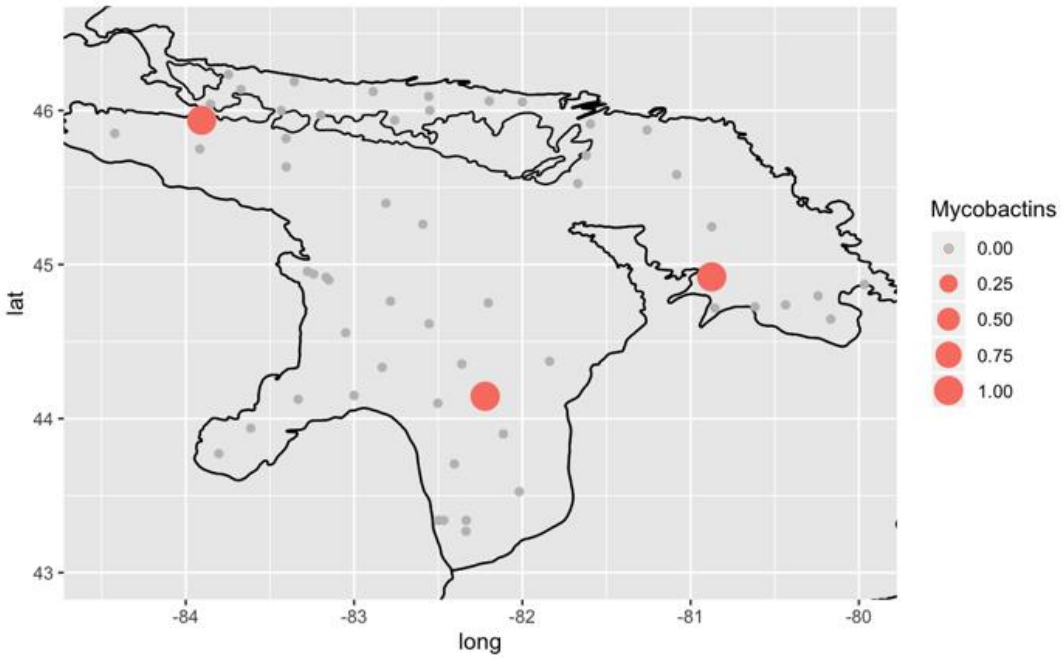
22A. Occurrence of erythrochelins in Lake Huron sediment.



22B. Occurrence of albachelins in Lake Huron sediment.



22C. Occurrence of mycobactins in Lake Huron sediment.



22D. Occurrence of coelibactins in Lake Huron sediment.

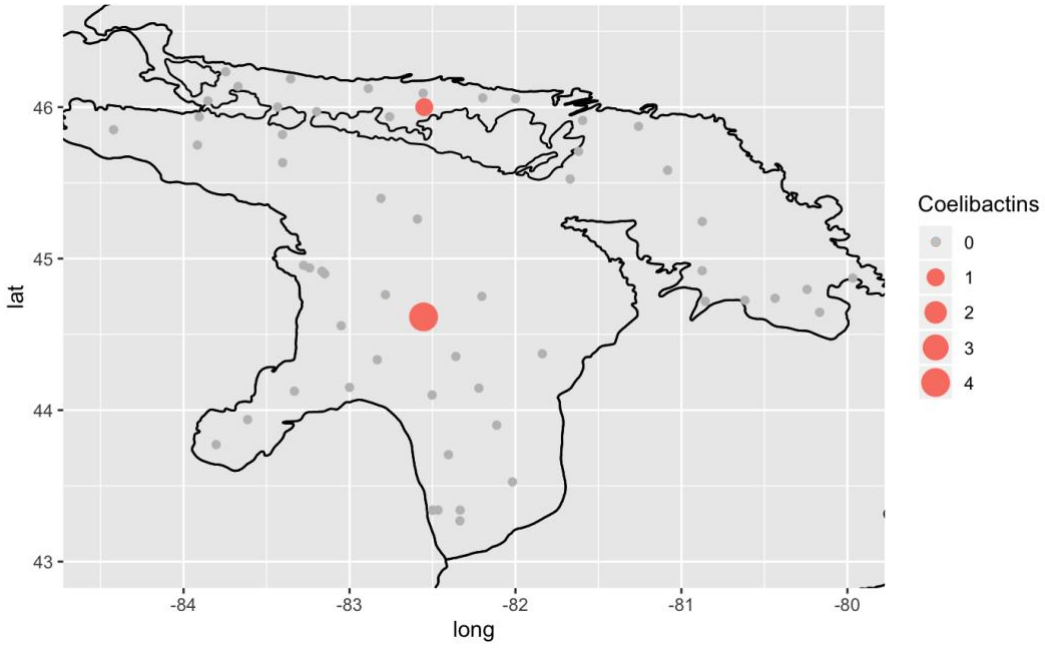
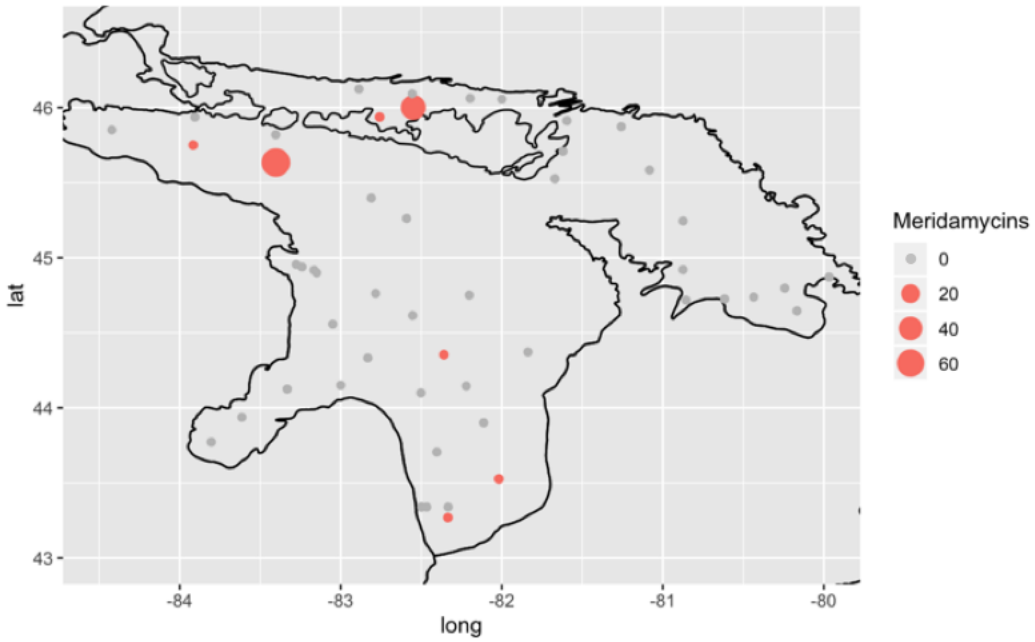
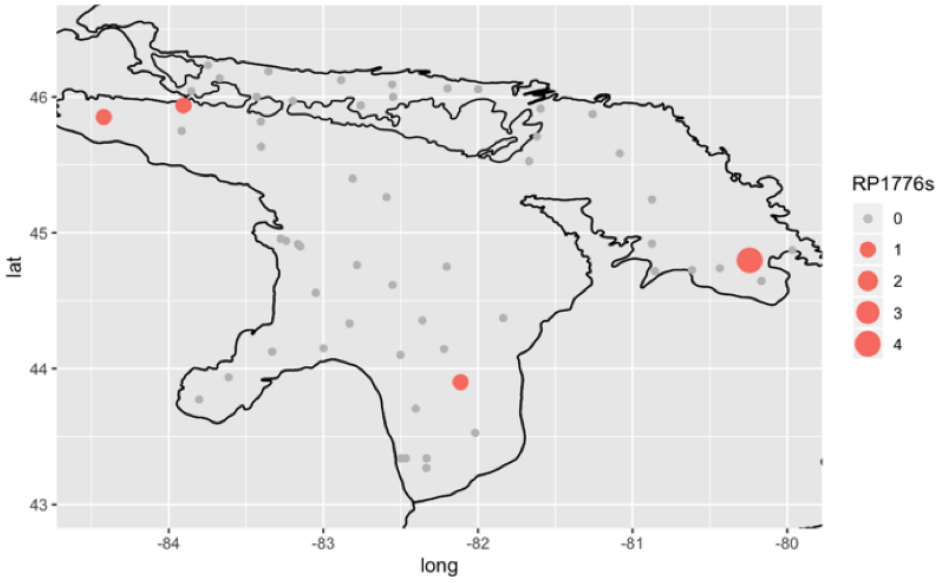


Figure 23. Occurrence of all other detected bioactive NPs in Lake Huron sediment.

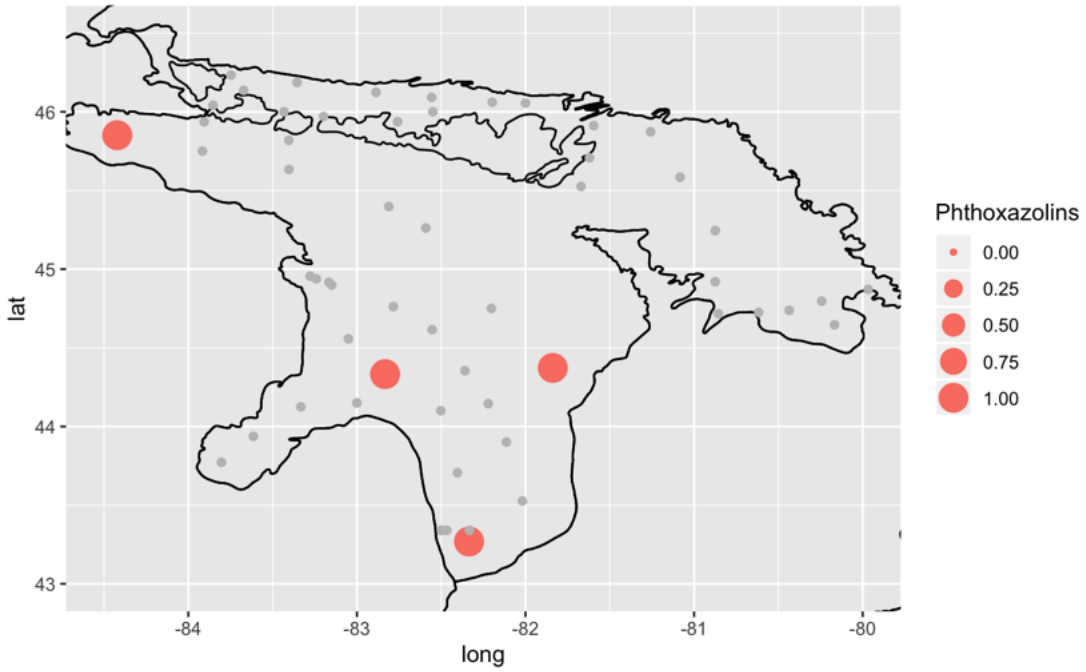
23A. Occurrence of meridamycins in Lake Huron sediment.



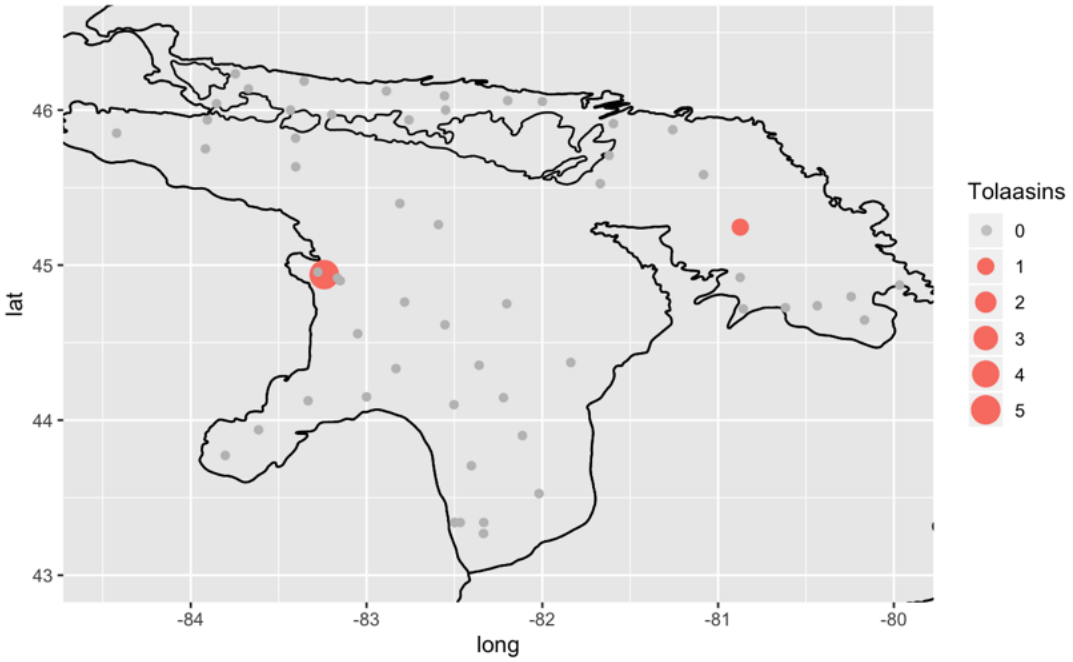
23B. Occurrence of RP1776-like compounds in Lake Huron sediment.



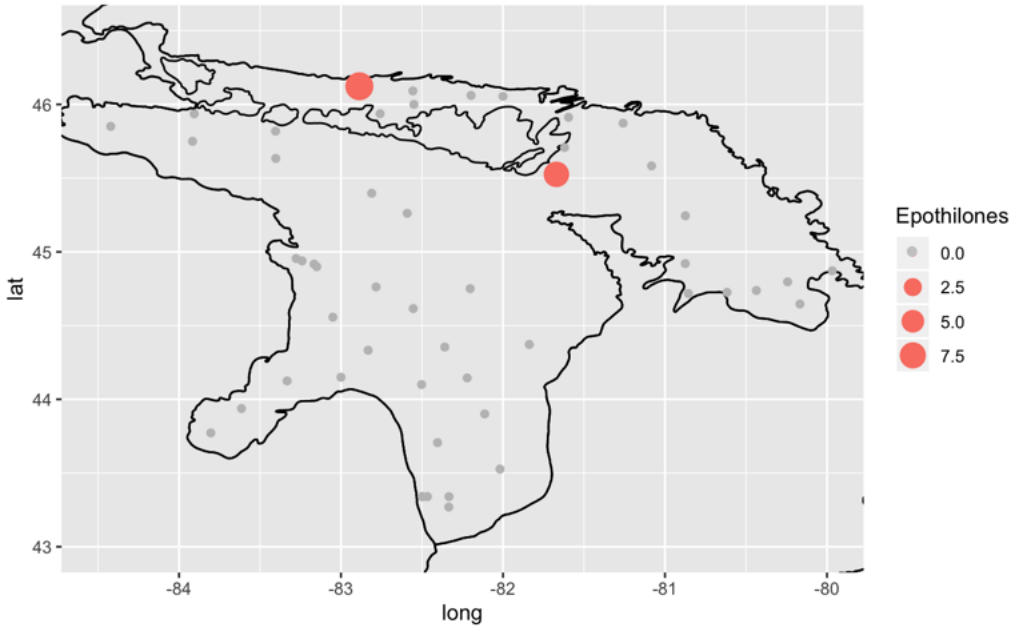
23C. Occurrence of phthoxazolins in Lake Huron sediment.



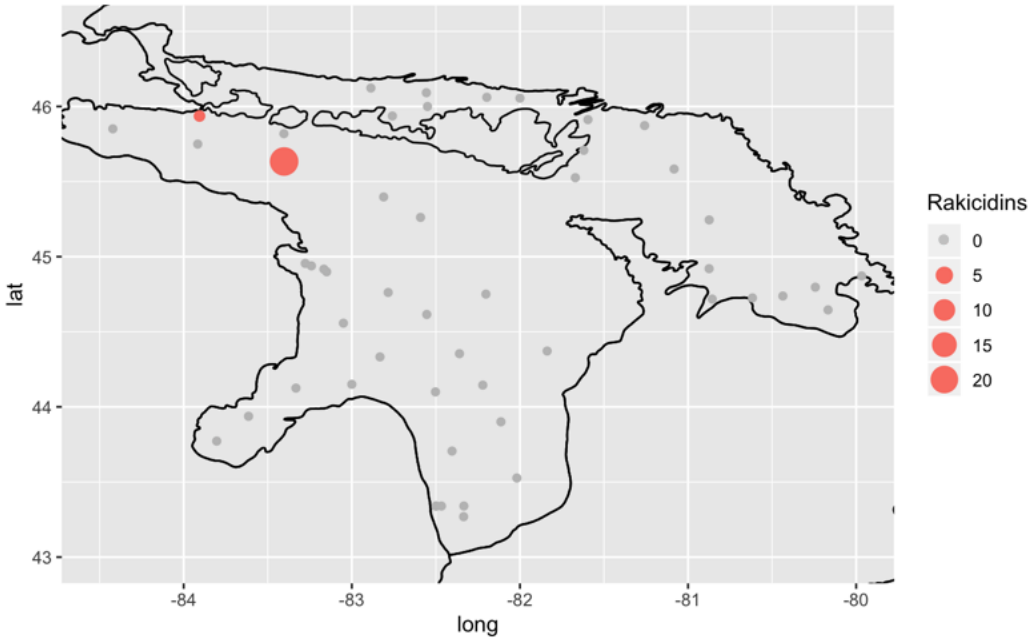
23D. Occurrence of phthoxazolins in Lake Huron sediment.



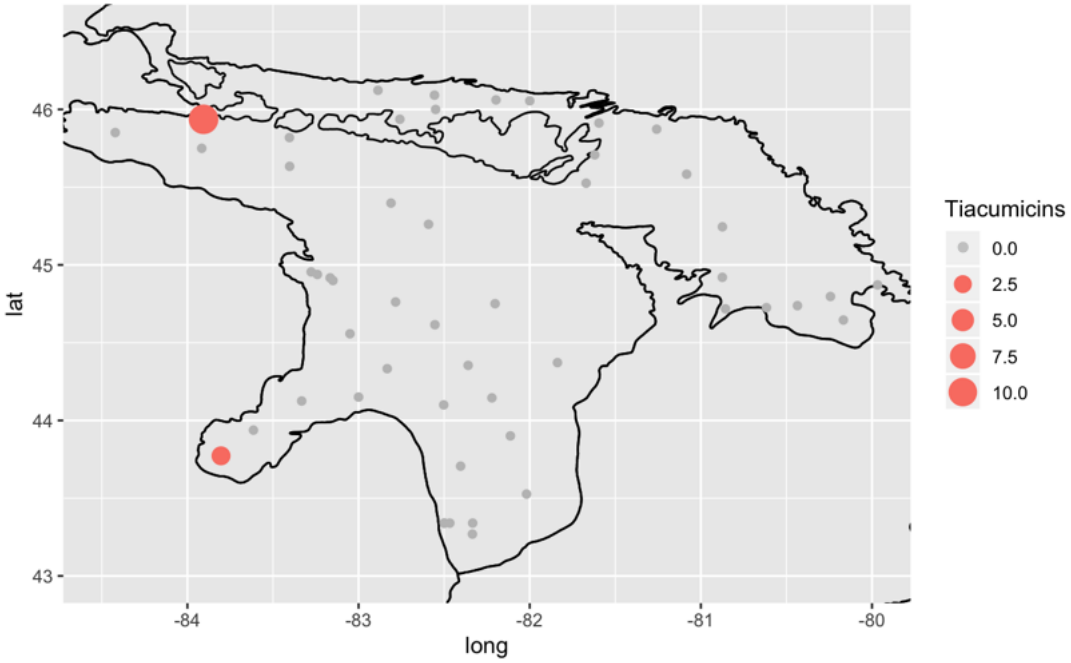
23E. Occurrence of epothilones in Lake Huron sediment.



23F. Occurrence of rakicidins in Lake Huron sediment.



23G. Occurrence of tiacumicins in Lake Huron sediment.



Supplementary Discussion.

OBUs corresponding to antibiotics and other bioactive compounds were scarce in comparison to siderophores; on average, four sequence reads and six sequence reads were detected per location for antibiotics and other bioactive compounds, respectively. In contrast, ten sequence reads were detected per location for siderophores. One potential explanation for this is the necessity of ecological function of these NP classes. Siderophores are essential for a microbe's survival: they chelate iron and thereby make it available for use in processes such as oxygen metabolism, and DNA and RNA syntheses.[116,177] Conversely, antibiotics and other bioactives are only indirectly linked to survival. The discrepancies in abundance of different NP classes can also be attributed to biases associated with primer degeneracies and database annotation, as they are both biased towards the gene sequences of strains relevant to NPs drug discovery. Strains that are relevant for the field of NP drug discovery are present in undetectable amounts in sediment.[103] This might be the reason a large proportion of the OBUs (99.98% A domain OBUs and 93.5% KSα domain OBUs, respectively) failed to match any of the compounds available in the MIBiG database, further preventing the observation of discernable patterns of NP occurrence. It is also worth noting there was no observed correlation between OBU presence/abundance and OTU presence/abundance (Table XIV).

APPENDIX C. Permissions to Reuse Published Materials.

9/24/2019 Rightslink® by Copyright Clearance Center

Copyright Clearance Center RightsLink®

Home Account Info Help LIVE CHAT

SPRINGER NATURE

Title: Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking

Author: Mingxun Wang, Jeremy J Carver, Vanessa V Phelan, Laura M Sanchez, Neha Garg et al.

Publication: Nature Biotechnology

Publisher: Springer Nature

Date: Aug 9, 2016

Copyright © 2016, Springer Nature

Logged in as:
Maryam Elfeki
University of Illinois at Chicago

LOGOUT

Author Request

If you are the author of this content (or his/her designated agent) please read the following. If you are not the author of this content, please click the Back button and select no to the question "Are you the Author of this Springer Nature content?".
Ownership of copyright in original research articles remains with the Author, and provided that, when reproducing the contribution or extracts from it or from the Supplementary Information, the Author acknowledges first and reference publication in the Journal, the Author retains the following non-exclusive rights:
To reproduce the contribution in whole or in part in any printed volume (book or thesis) of which they are the author(s).
The author and any academic institution, where they work, at the time may reproduce the contribution for the purpose of course teaching.
To reuse figures or tables created by the Author and contained in the Contribution in oral presentations and other works created by them.
To post a copy of the contribution as accepted for publication after peer review (in locked Word processing file, of a PDF version thereof) on the Author's own web site, or the Author's institutional repository, or the Author's funding body's archive, six months after publication of the printed or online edition of the Journal, provided that they also link to the contribution on the publisher's website.
Authors wishing to use the published version of their article for promotional use or on a web site must request in the normal way.

If you require further assistance please read Springer Nature's online author reuse guidelines.

For full paper portion: Authors of original research papers published by Springer Nature are encouraged to submit the author's version of the accepted, peer-reviewed manuscript to their relevant funding body's archive, for release six months after publication. In addition, authors are encouraged to archive their version of the manuscript in their institution's repositories (as well as their personal Web sites), also six months after original publication.

v1.0

BACK CLOSE WINDOW

Copyright © 2019 Copyright Clearance Center, Inc. All Rights Reserved. Privacy statement. Terms and Conditions.
Comments? We would like to hear from you. E-mail us at customercare@copyright.com

APPENDIX C (continued)

9/24/2019 Ρ ιγ ητσθινκ→βψ Χοπψριγ ητ Χλεαρανχε Χεντερ

Copyright Clearance Center RightsLink®

Home Create Account Help LIVE CHAT

ACS Publications Most Trusted. Most Cited. Most Read.

Title: Assessing the Efficiency of Cultivation Techniques To Recover Natural Product Biosynthetic Gene Populations from Sediment

Author: Maryam Elfeki, Mohammad Alanjary, Stefan J. Green, et al

Publication: ACS Chemical Biology

Publisher: American Chemical Society

Date: Aug 1, 2018

Copyright © 2018, American Chemical Society

LOGIN

If you're a copyright.com user, you can login to RightsLink using your copyright.com credentials.

Already **a RightsLink user** or want to learn more?

PERMISSION/LICENSE IS GRANTED FOR YOUR ORDER AT NO CHARGE

This type of permission/license, instead of the standard Terms & Conditions, is sent to you because no fee is being charged for your order. Please note the following:

- Περμ ισσιον ισγραντεδ φορ ψουρ ρεθυεστ ιν βοτη πριντ ανδ ελεχτρονιχ φορμ ατς, ανδ τρανσλατιονς.
- Ιφφιγρεσ ανδ/ορ τβλεσ ωερε ρεθυεστεδ, πλεασ μαψ βε αδαπτεδ φορ υσεδ ιν παρτ.
- Πλεασε πριντ τηισ παγε φορ ψουρ ρεχορδσ ανδ σενδ α χοπψ οφ ιτ το ψουρ πυβλισηερ/γραδυατε σχηοολ.
- Αππροπριατε χρεδιτ φορ τηε ρεθυεστεδ ματεριαλ σηουλδ βε γιπεν ασ φολλοωσ: \"Ρεπριντεδ (αδαπτεδ) ωιτη περμ ισσιον φρομ (ΧΟΜΠΛΕΤΕ ΡΕΦΕΡΕΝΧΕ ΧΙΤΑΤΙΟΝ). Χοπψριγ ητ (ΨΕΑΡ) Αμ ερχαν Χηεμ ιχαλ Σοχιετψ\". Ινσερτ αππροπριατε ινφορμ ατιον ιν πλαχε οφ τηε χοπιτολιζεδ ωορδς.
- Ονε-τιμ ε περμ ισσιον ισγραντεδ ονλψ φορ τηε υσε σπεχιφιε ιν ψουρ ρεθυεστ. Νο αδδιτιοναλ υσεσ αρε γραντεδ (συχη ασδεριπατιβε ωορκσ ορ τρανσλατιονς). Φορ ανψ οτηερ υσες, πλεασε συβμ ιτ α νεω ρεθυεστ.

BACK CLOSE WINDOW

Copyright © 2019 Copyright Clearance Center, Inc. All Rights Reserved. Privacy statement. Terms and Conditions.
Comments? We would like to hear from you. E-mail us at customercare@copyright.com

APPENDIX C (continued)

Home | Account Info | Help | LIVE CHAT

Title: Antibiotic resistance genes show enhanced mobilization through suspended growth and biofilm-based wastewater treatment processes

Author: Petrovich, Morgan; Chu, Binh

Publication: FEMS Microbiology Ecology

Publisher: Oxford University Press

Date: 2018-03-09

Copyright © 2018, Oxford University Press

Logged in as:
Maryam Elfeki
University of Illinois at Chicago

LOGOUT

Order Completed

Thank you for your order.

This Agreement between University of Illinois at Chicago -- Maryam Elfeki ("You") and Oxford University Press ("Oxford University Press") consists of your license details and the terms and conditions provided by Oxford University Press and Copyright Clearance Center.

Your confirmation email will contain your order number for future reference.

printable details

License Number	4675451462463
License date	Sep 24, 2019
Licensed Content Publisher	Oxford University Press
Licensed Content Publication	FEMS Microbiology Ecology
Licensed Content Title	Antibiotic resistance genes show enhanced mobilization through suspended growth and biofilm-based wastewater treatment processes
Licensed Content Author	Petrovich, Morgan; Chu, Binh
Licensed Content Date	Mar 9, 2018
Licensed Content Volume	94
Licensed Content Issue	5
Type of Use	Thesis/Dissertation
Requestor type	Author of this OUP content
Format	Print and electronic
Portion	Text Extract
Number of pages requested	3
Will you be translating?	No
Title	Occurrence of bacterial natural product production genes across an environment
Institution name	University of Illinois at Chicago
Expected presentation date	Nov 2019
Portions	Introductory paragraph, figure 4 and its description (caption and usage within results and discussion section), part of the methods section detection of ARGs and APGs.
Requestor Location	University of Illinois at Chicago 900 S. Ashland ave. Room 3118 CHICAGO, IL 60608 United States Attn: University of Illinois at Chicago
Publisher Tax ID	GB125506730
Total	0.00 USD

ORDER MORE CLOSE WINDOW

Copyright © 2019 Copyright Clearance Center, Inc. All Rights Reserved. Privacy statement. Terms and Conditions.
Comments? We would like to hear from you. E-mail us at customercare@copyright.com

This page is available in the following languages:

Creative Commons License Deed

Attribution-NonCommercial 3.0 Unported (CC BY-NC 3.0)

This is a human-readable summary of (and not a substitute for) the license.

You are free to:

Share — copy and redistribute the material in any medium or format

Adapt — remix, transform, and build upon the material

The licensor cannot revoke these freedoms as long as you follow the license terms.

Under the following terms:

Attribution — You must give appropriate credit, provide a link to the license, and indicate if changes were made. You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.

NonCommercial — You may not use the material for commercial purposes.

No additional restrictions — You may not apply legal terms or technological measures that legally restrict others from doing anything the license permits.

Notices:

You do not have to comply with the license for elements of the material in the public domain or where your use is permitted by an applicable exception or limitation.

No warranties are given. The license may not give you all of the permissions necessary for your intended use. For example, other rights such as publicity, privacy, or moral rights may limit how you use the material.

VITA

NAME

Maryam Elfeki

EDUCATION

University of Illinois at Chicago, Chicago, IL.

PhD, Pharmaceutical Sciences 2013 – 2020

B.Sc. Biochemistry 2010 – 2012

B.Sc. Applied Mathematics 2010 – 2012

RESEARCH EXPERIENCE

Graduate Research Assistant 2010 – present

University of Illinois at Chicago

Department of Medicinal Chemistry and Pharmacognosy

Dr. Brian T Murphy

Visiting Research Assistant 2015 – 2015

University of Tübingen

Institute of Microbiology and Infection Medicine

Dr. Nadine Ziemert

PROFESSIONAL EXPERIENCES

Aspire Capital 2019 – 2020

Biotechnology Analyst Intern

Office of Technology Management 2016 – 2019

Commercialization Analyst

Wright Times 2018 – 2019

Arts and culture editor

AWARDS & HONORS

1. Myron Goldsmith Scholarship 2018
2. CCTS/COP Pre-doctoral Education for Clinical and Translational Scientists Fellowship 2015
3. American Society of Pharmacognosy (ASP) student travel award 2015
4. W.E. van Doren Scholar Award 2015
5. Caterpillar Award for Undergraduate Research 2011
6. ASP Undergraduate Research Award 2011
7. Associated Collegiate Press, Best Editorial/Commentary, Second Place 2010

PUBLICATIONS

1. **Elfeki, M.**; Mantri, S.; Naqib, A.; Green, S. J.; Ziemert, N.; Murphy, B. T. Evaluating distribution of bacterial natural product biosynthetic gene clusters in Lake Huron sediment. In preparation.
2. **Elfeki, M.**; Alanjary, M.; Green, S. J.; Ziemert, N.; Murphy, B. T. Assessing the efficiency of cultivation techniques to recover natural product biosynthetic gene populations from sediment. ACS Chemical Biology 2018, 13, 8, 2074-2081.
3. Petrovich, M.; Chu, B.; Wright, D.; Griffin, J.; **Elfeki, M.;** Murphy, B. T.; Poretsky, R.; Wells, G. Antibiotic resistance genes show enhanced mobilization through suspended growth and biofilm-based wastewater treatment processes. FEMS Microbiol. Ecol. 2018, 94, fiy041.

4. Wang, M.; Carver, J.; Phelan, V. V.; Sanchez, L. M.; Garg, N.; Peng, Y.; Watrous, J.; Nguyen, D. D.; Kapono, C. A.; Luzzatto-Knaan, T.; Porto, C; Bouslimani, A.; Melnik, A. V.; Meehan, M. J.; Liu, W-T.; Crüsemann, M.; Boudreau, P. D.;' Esquenazi, E.; Sandoval-Calderón, M.; Kersten, R. D.; Pace, L. A.; Quinn, R. A.; Duncan, K. R.; Hsu, C-C.; Floros, D. J.; Gavilan, R. G.; Kleigrewe, K.; Northen, T.; Dutton, R. J.; Parrot, D.; Carlson, E. E.; Aigle, B.; Michelsen, C. F.; Jelsbak, L.; Sohlenkamp, C.; Pevzner, P.; Edlund, A.; McLean, J.; Piel, J.; Murphy, B. T.; Gerwick, L.; Liaw, C-C.; Yang, Y-L.; Humpf, H-U.; Maansson, M.; Keyzers, R. A.; Sims, A. C.; Johnson, A. R.; Sidebottom, A. M.; Sedio, B. E.; Klitgaard, A.; Larson, C. B.; Boya, C. A.; Torres-Mendoza, D.; Gonzalez, D. J.; Silva, D. B.; Marques, L. M. M.; Demarque, D. P.; Pociute, E.; O'Neill, E. C.; Briand, E.; Helfrich, E. J. N.; Granatosky, E. A.; Glukhov, E.; Ryffel, F.; Houson, H.; Mohimani, H.; Kharbush, J. J.; Zeng, Y.; Vorholt, J. A.; Kurita, K. L.; Charusanti, P.; McPhail, K. L.; Nielsen, K. F.; Vuong, L.; **Elfeki, M.**; Traxler, M. F.; Engene, N.; Koyama, N.; Vining, O. B.; Baric, R.; Silva, R. R.; Mascuch, S. J.; Tomasi, S.; Jenkins, S.; Macherla, V.; Hoffmann, T.; Agarwal, V.; Williams, P. G.; Dai, J.; Neupane, R.; Gurr, J.; Rodríguez, A. M. C.; Lamsa, A.; Zhang, C.; Dorrestein, K.; Duggan, B. M.; Almaliti, J.; Allard, P-M.; Wolfender, J-L.; Peryea, T.; Nguyen, D-T.; VanLeer, D.; Shinn, P.; Jadhav, A.; Müller, R.; Waters, K. M.; Shi, W.; Liu, X.; Zhang, L.; Knight, R.; Jensen, P. R.; Palsson, B. O.; Pogliano, K.; Linington, R. G.; Gutiérrez, M.; Lopes, N. P.; Gerwick, W. H.; Moore5, B. S.; Dorrestein, P. C.; Bandeira, N. Sharing and community curation of mass spectra by GNPS. Nat. Biotechnol., 34: 828-837, 2016.
5. Shaikh, A., **Elfeki, M**., Landolpha, S., Tanouye, U., Green, S. J., Murphy, B. T. Deuteromethylactin B from a freshwater-derived Streptomyces sp. Nat. Prod. Sci. 21(4): 1-7 (2015)

ORAL PRESENTATIONS

1. Society for Industrial Microbiology and Biotechnology, August 2018; "Assessing the efficiency of cultivation techniques to recover natural product biosynthetic gene populations from sediment".

2. Center for Bimolecular Sciences Seminar Series, College of Pharmacy, UIC, December 2016; “Assessing the biosynthetic capacity of the cultivatable microbiome from Lake Huron sediment”.
3. Medicinal Chemistry Meeting-in-Miniature, April 2017; “Recovery of natural product biosynthetic capacity from Lake Huron sediment”.
4. German Centre for Infection Research (DZIF), Dept. Microbiology and Biotechnology, University of Tübingen, April 2016; "The great lakes as a source for drug discovery”.
5. Center for Biopharmaceutical Sciences Seminar Series, College of Pharmacy, UIC, February 2015; “Investigating the taxonomic diversity and chemical potential of Lake Huron actinomycetes”.

POSTER PRESENTATIONS

1. **Elfeki, M.;** Alanjary, M.; Nakib, A.; Green S.J.; Ziemert, N.; Murphy, B.T.; “Assessing the Natural Product Biosynthetic Capacity of the Cultivatable Microbiome from 2 Lake Huron Sediment Samples” Research Day at the College of Pharmacy, University of Illinois at Chicago, IL, February 2018.
2. **Elfeki, M.;** Nakib, A.; Green S.J.; Murphy, B.T.; “Prospecting Freshwater Sources for Drug-lead Discovery.” American Society of Pharmacognosy annual meeting, Copper Mountain, CO, July 2015.
3. **Elfeki, M.;** Nakib, A.; Green S.J.; Murphy, B.T.; “Metagenomic studies of actinomycete populations In Great Lakes sediment.” Minnesota-Iowa-Kansas-Illinois Medicinal Chemistry Meeting, Chicago, IL, April 2014.
4. **Elfeki, M.;** Nakib, A.; Green S.J.; Murphy, B.T.; “Metagenomic studies of actinomycete populations In Great Lakes sediment.” Marine NPs Gordon Research Conference, Ventura, CA, March 2014.
5. **Elfeki, M.;** Nakib, A.; Green S.J.; Murphy, B.T.; “Metagenomic studies of actinomycete populations In Great Lakes sediment.” Marine NPs Gordon Research Seminar, Ventura, CA, March 2014

6. **Elfeki, M.,** Tanouye, U., Wei, X., and Murphy, B.T. "Exploring the potential of marine actinomycetes to inhibit the growth of pathogenic bacteria." UIC Forum, Chicago, IL, April 2011.
7. **Elfeki, M.,** Tanouye, U., Wei, X., and Murphy, B.T. "Exploring the potential of marine actinomycetes to inhibit the growth of pathogenic bacteria." Chicago Area Undergraduate Research Symposium (CAURS), Chicago, IL, March 2011.

PROFESSIONAL MEMBERSHIPS

- American Society of Microbiology
- American Society of Pharmacognosy
- American Association for Pharmaceutical Scientists
- Graduate Women in Science
- Medicinal Chemistry & Pharmacognosy Graduate Student Organization
- Egyptians Abroad for Development
- Founding member of Egyptians Abroad for Development – Chicago Chapter
- Alpha Theta Kappa Honors Society

OUTREACH

- Boys and Girls Club monthly science experiments volunteer 2016 – 2018
- Expand Your Horizons Chicago – Workshop leader 2016 – 2017
- Catalyst Leadership Retreat 2015
- Organizing committee of the 52nd Annual Minnesota-Iowa-Kansas-Illinois regional medicinal chemistry meeting 2014
- International Club, Sports Activities Coordinator 2011– 2014
- International Club, Social Activities Coordinator 2012– 2013

- International student Orientation Leader 2012– 2013
- Volunteer sports activist for the Office of Student with Disabilities at UIC 2013 – 2017