METAGENOME MINING TO EXPLORE NOVEL REGIONS OF NATURAL PRODUCTS CHEMICAL SPACE

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät
der Eberhard Karls Universität Tübingen
zur Erlangung des Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat.)

vorgelegt von
Shrikant Subhash Mantri
aus Daund, Pune (India)

Tübingen
2021

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation:	14.02.2022
Dekan:	Prof. Dr. Thilo Stehle
1. Berichterstatterin:	Prof. Dr. Nadine Ziemert
2. Berichterstatter:	Prof. Dr. Daniel Huson

Abstract

Genomics has accelerated discovery in biology in an unprecedented way. Still, we are far from solving grand challenges facing humanity. The challenge to combat antimicrobial resistance requires us to accelerate natural product discovery by several orders of magnitude. We are already running out of our existing arsenal of antibiotics and novel approaches are needed to accelerate the pace of their discovery and development. Quick screening of natural product biosynthesis potential via metagenome mining holds new hope to revive the antibiotic discovery pipeline. Thanks to recent advancements in next generation sequencing technologies and big data mining, now we can hope to rationally survey the diverse ecosystem metagenomes to discover novel secondary metabolites.

In this thesis we have presented our developed metagenome data mining pipeline and approaches to explore novel regions of natural products chemical space. We present our results and insights from multiple ecosystem metagenome surveys. Novel biosynthesis genes, domains, cluster sequences and comparative patterns from the surveyed ecosystem are highlighted in separate chapters. Metagenome mining patterns from following diverse ecosystems were studied: 1) Different horizons of soil sampled from three sites in close vicinity from the Schoenbuch forest; 2) Lake Huron sediments; 3) human gut microbiome and 4) the Tuebingen actinomycetes strain collection.

The insights gained from this thesis will be helpful to the natural products research community to accelerate metagenome based novel natural products discovery and revive the antibiotics discovery pipeline.

Zusammenfassung

Die Genomik hat die Entdeckung in der Biologie auf beispiellose Weise beschleunigt. Dennoch sind wir weit davon entfernt, die großen Herausforderungen der Menschheit zu lösen. Die Herausforderung, antimikrobielle Resistenzen zu bekämpfen, erfordert, dass wir die Entdeckung von Naturstoffen um mehrere Größenordnungen beschleunigen. Unser vorhandenes Antibiotika-Arsenal geht uns bereits aus, und es werden neue Ansätze benötigt, um das Tempo ihrer Entdeckung und Entwicklung zu beschleunigen. Ein schnelles Screening des Biosynthesepotenzials von Naturstoffen durch Metagenom-Mining birgt neue Hoffnung, die Pipeline zur Entdeckung von Antibiotika wiederzubeleben. Dank der jüngsten Fortschritte bei den Sequenzierungstechnologien der nächsten Generation und dem Big-Data-Mining können wir nun hoffen, die vielfältigen Metagenome des Ökosystems rational zu untersuchen, um neue Sekundärmetaboliten zu entdecken.

In dieser Dissertation haben wir unsere entwickelte Metagenom-Data-Mining-Pipeline und Ansätze zur Erforschung neuartiger Regionen des chemischen Raums von Naturstoffen vorgestellt. Wir präsentieren unsere Ergebnisse und Erkenntnisse aus mehreren Ökosystem-Metagenom-Untersuchungen. Neuartige Biosynthesegene, Domänen, Clustersequenzen und Vergleichsmuster aus dem untersuchten Ökosystem werden in separaten Kapiteln beleuchtet. Es wurden Metagenom-Mining-Muster ausfolgenden verschiedenen Ökosystemen untersucht: 1) Verschiedene Bodenhorizonte, die von drei Standorten in unmittelbarer Nähe des Schönbucher Waldes beprobt wurden; 2) Sedimente des Huron Sees; 3) menschliches Darmmikrobiom und 4) die Tübinger Stammsammlung.

Die aus dieser Dissertation gewonnenen Erkenntnisse werden der Naturstoffforschungsgemeinschaft helfen, die Metagenom-basierte Entdeckung neuer Naturstoffe zu beschleunigen und die Entdecker- Pipeline von Antibiotika wiederzubeleben.

ACKNOWLEDGEMENT

I would like to thank:

Nadine Ziemert, for the opportunity of working in exciting applied natural product metagenome mining project and international collaboration, for her continued guidance, encouragement and support throughout, for her enthusiastic hosting of the “Genes to Molecules” seminars, I learned a lot from our every interactions and group meetings, for the great fun we had during lab outings and parties.

Daniel Huson, for agreeing to be my second supervisor, for evaluating my thesis, for introducing me to the exciting field of “Metagenome mining” through the theory and practicals of “BIOINF4399: Microbiome Analysis Course”.

ZiemertLab colleagues (Martina Adamek, Timo Negri, Ahmed Desouky, Mehmet Direnç Mungan, Franziska Höhn, Athina Gavriilidou, Bita Pourmohsenin, Patrick Weiss, Monica Daneliuc, Davide Paccagnella, Helena Sales-Ortells, Mark Polster, Simon Edenhart, Daniel Männle, Marita Wurm, Caner Bağcı) and all the members of CMFI, IMIT, IBMI and “Microbiome SuperGroup”, for research help and scientific discussions.

Marnix Medema, Vittorio Traccana, Brian Murphy, Maryam Elfeki, Harald Neidhardt, Angel Angelov, Eric Kemen, Chamber Hughes, Sven Nahnsen, Marc Chevrette, Nelly Selem-Mojica, Francisco Barona-Gómez Ewa Maria Musiol-Kroll, for exciting collaborations and sharing their expertise.

bwForCluster BinAC and de.NBI, for providing the computing resources, for technical support to configure bioinformatics applications on cloud and cluster.

DZIF, CMFI, and DFG for funding and structural support.

QBiC and NCCT for help in planning and sequencing the metagenome samples.

NABI for study leave to pursue Ph.D.

Vasvi for making this journey fun, exciting and pleasant, for being my pillar of strength, for being the sunshine that colors my life.

My parents, my sisters and my family for their constant motivation, encouragement and love, for their understanding and sacrifices.

"What we know is surely considerable, but it is dwarfed by what we still have to learn."

— John Craig Venter

Dedication

I dedicate this work to my family and

to the research spirit of explorer polymath Alexander von Humboldt

CONTENTS

CHAPTER 1: INTRODUCTION **4**

1.1 The quest for understanding the fundamental secret of life. 4
1.2 Pre antibiotics era, antibiotics era and the discovery void 5
1.3 Post antibiotics era and Antimicrobial Resistance 6
1.4 Natural products, secondary metabolites, specialised metabolites, antibiotics: What are they? 6
1.5 Traditional routes to discovery 9
1.6 Hidden microbial dark matter and genome mining 9
1.7 BigData and metagenome mining potential 9
1.8 Research problem 11
1.9 Objectives 12
1.10 Thesis Outline 12
1.11 References 13

CHAPTER 2: BIOLOGICAL BACKGROUND **15**

2.1 Microbes as source of natural products 15
2.2 Bacterial Diversity: How much do we know? 15
2.3 16S rRNA based profiling of diversity of bacteria 16
2.4 Primary and secondary metabolism pathways 18
2.5 Natural products chemical space and atlas 18
2.6 Biosynthesis of bacterial natural products 19
2.6.1 Polyketide Biosynthesis Pathway *19*
2.6.2 Nonribosomal Peptide Biosynthesis Pathway *21*
2.7 Next generation sequencing and genome mining revolution 23
2.8 Biosynthesis domain diversity profiling via amplicon sequencing. 24
2.9 Metagenome mining for estimating biosynthesis potential 25
2.10 References 26

CHAPTER 3: TECHNICAL BACKGROUND **28**

3.1 Microbial Community Diversity Profiling Methods 28
3.2 Natural Products Biosynthesis Domain exploration methods 29
3.2.1 NaPDoS *29*
3.2.2 BiG-MEx *29*
3.2.3 Dom2BGC *30*
3.3 Natural Products Biosynthesis Cluster exploration methods 31
3.3.1 CLUSEAN *31*
3.3.2 antiSMASH *31*
3.3.3 BIGSCAPE *32*
3.3.4 deepBGC *32*
3.4 Metagenome Assembly and Natural products biosynthesis potential exploration methods 33
3.4.1 metaSPADES *33*
3.4.2 CloudSPADES *33*
3.4.3 TELL-Link *33*
3.5 Tools worth exploring in future 34
3.5.1 Miscellaneous tools *34*

3.6 Reference 34

CHAPTER 4: MBEZ: EASY BIOSYNTHETIC POTENTIAL EXPLORATION METAGENOME MINING PIPELINE..... 36

4.1 Introduction 38
4.2 Material and Methods 38
4.2.1 Microbial community diversity exploration pipeline: 38
4.2.2 BGC domain diversity exploration pipeline: 40
4.2.3 Biosynthesis potential exploration pipeline: 40
4.2.4 Implementation 40
4.3 Conclusion 40
4.4 References 41

CHAPTER 5: METAGENOMIC SEQUENCING OF MULTIPLE SOIL HORIZONS AND SITES IN CLOSE VICINITY REVEALED NOVEL SECONDARY METABOLITE DIVERSITY 42

5.1 Introduction 44
5.2 Results 48
5.2.1 Amplicon-seq mining revealed major differences in bacterial diversity and their biosynthetic potential in the different soils and their horizon 48
5.2.2 Shotgun metagenome mining further uncovered microbial diversity and identified novel BGCs 56
5.2.3 Comparative analysis highlights the advantage of long reads to capture biosynthetic potential. 60
5.3 Discussion 61
5.4 Conclusion 66
5.5 Methods 67
5.5.1 Soil sampling, physico-chemical parameters characterization. 67
5.5.2 Metagenome sequencing 69
5.5.3 Shotgun-seq Analysis. 73
5.5.4 Amplicon-seq Analysis. 74
5.6 References 76

CHAPTER 6: METAGENOMIC BIG-DATA EXPLORATIONS OF NATURAL PRODUCTS DIVERSITY IN DIVERSE ECOSYSTEMS 83

Part I: Evaluating the Distribution of Bacterial Natural Product Biosynthetic Genes Across Lake Huron Sediment. 84
6.1 Introduction 85
6.2 Results and Discussion 87
6.2.1 . Characterization of BGC Domain Sequence Diversity in Sediment 87
6.2.2 .Analysis of Characterized NP BGC Distribution in Lake Sediment 89
6.2.3 Analysis of Uncharacterized NP BGC Distribution in Lake Sediment 94
6.3 Conclusion 98
6.4 Methods 99
6.4.1 Collection of Sediment Samples, Cultivation of Sediment Bacteria on Nutrient Agar 99
6.4.2 Genomic DNA Isolation from Sediment and Nutrient Agar 99
6.4.3 KSα and A Domain Amplification and Sequencing 99
6.4.4 Bioinformatic Analyses of BGC Data 100
6.5 Reference 102
Part II: Dynamics of the human gut secondary metabolome during antibiotic treatment. 106
6.6 Introduction 106
6.7 Methods 107
6.8 Results and Discussion 108
6.9 References 113

Part III : Using linked reads and long reads to recover biosynthetic gene clusters from Tuebingen actinomycetes strain collections. *114*
6.10 Overview and motivation 114
6.11 Methods 115
6.11.1 Library preparation and NGS sequencing *115*
6.11.2 Bioinformatics Analysis: *116*
6.11.3 TELL-Seq analysis *117*
6.12 Results and Discussion 117
6.13 Conclusion and Outlook 119
6.14 References 119

CHAPTER 7: GENERAL CONCLUSION 121

7.1 Concluding remarks on results 121
7.2 Impact and applications of developed approaches and methods 123
7.3 Future challenges 124
7.4 References 124

ANNEXURE A: PUBLICATION: THE CONFLUENCE OF BIG DATA AND EVOLUTIONARY GENOME MINING FOR THE DISCOVERY OF NATURAL PRODUCTS **..... 125**

ANNEXURE B: SUPPLEMENTAL INFORMATION : EVALUATING DISTRIBUTION OF BACTERIAL NATURAL PRODUCT BIOSYNTHETIC GENES ACROSS LAKE HURON SEDIMENT **..... 143**

ANNEXURE C: LIST OF PUBLICATIONS AND MANUSCRIPTS **..... 181**

ANNEXURE D: DECLERATION OF CONTRIBUTION **..... 182**

ANNEXURE E: LIST OF ABBREVEATIONS **..... 184**

Chapter 1: Introduction

1.1 The quest for understanding the fundamental secret of life.

Understanding the fundamental secrets of life is the overarching motivation of all human endeavours. Tuebingen has witnessed the unravelling of such secrets, most fundamental discoveries in research areas of biology, chemistry and cosmology. Johannes Friedrich Miescher in 1869 became the first scientist to isolate nucleic acid in the lab of Felix Hoppe-Seyler at University of Tuebingen (Dahm, 2008). This discovery truly marked the beginning of the genomics era. During the subsequent period in the twentieth century our understanding of biology and genetics improved many folds. Julius Lothar Meyer discovered the early version of the periodic table and ushered the chemistry revolution (Pulkkinen, 2020). In the seventeenth century, Johannes Kepler discovered the laws of planetary motion (Voelkel, 2001); Wilhelm Schikard invented a mechanical calculator (Hanisch et al., 2000). Meyer, Kepler and Schikard have all been associated with the University of Tuebingen. More recently in 1995 Christiane Nüsslein-Volhard was awarded the Nobel Prize in Medicine for her research on the genetic control of embryonic development (Nüsslein-Volhard, 2012). The knowledge — generated during these quests — has become the holy grail for overcoming human suffering and disease. Established in 1477 Eberhard Karls Universität Tübingen, since then has attracted the best minds and nurtured their talents to find answers to the fundamental secrets of life. This makes Tuebingen truly a place of scientific pilgrimage for the researchers seeking ultimate knowledge via the path of science. While frontiers of understanding consciousness, overcoming disease, extending longevity and making life transplanetary are the challenging goals of our generation, this thesis is a small step in the humble pursuit of accelerating the novel natural products discovery through mining the metagenomic data.

In the introductory chapter, firstly I have given an historical account of the Antibiotics era and discussed the problem of antimicrobial resistance. Next, I have described how genome mining revolutionised and revived the discovery pipelines of antibiotics. Subsequently, I have discussed why we need to go beyond studying single genomes and start exploring metagenomes to uncover the hidden microbial diversity and unlock the treasure of vast natural product chemical space. Then I have described the research problem that I have studied during this dissertation project and how I have solved certain aspects of this vast topic. Then I have given an overview of the structure of my thesis and brief contents of the chapters.

1.2 Pre antibiotics era, antibiotics era and the discovery void

Before the discovery of antibiotics (pre antibiotics era) even a small injury was equivalent to a death sentence. Due to infection post injury only a lucky few survived who recovered miraculously, the remaining just died often painfully. This situation changed on the arrival of antibiotics (the present, antibiotics era). Serendipitous discovery of Penicillin by Alexander Fleming in 1928 ushered in the beginning of the antibiotics era (Droog, 2015). First few decades of the beginning of this era were considered a golden epoch as numerous antibiotics were discovered. Selman Waksman, Albert Schatz, Yellapragada Subbarow were the most successful pioneers of this field who discovered the widely used clinical antibiotics (Samanta and Bandyopadhyay, 2019). Most of the discovered antibiotics were from the bacteria isolated from soil samples and rediscovery of these known antibiotics during subsequent discovery expeditions became the next big challenge. The big pharma industry even started to close the discovery units due to the challenges and limited profits (Shlaes, 2010). This led to a discovery void and now there is dearth of drugs to treat new infectious diseases.

1.3 Post antibiotics era and Antimicrobial Resistance

We are currently on the threshold of the post antibiotics era due to the widely developed antimicrobial resistance (AMR) and we might reach a stage where again microbial infections can become untreatable. We — as human species — are truly living in a microbial world. It is estimated that there are more than about a trillion bacterial species, of which we have so far discovered and studied only a few hundreds of thousands (Locey and Lennon, 2016). These bacterial species are constantly in an arms race amongst each other. They keep competing for resources, food, survival, and evolution. Some of the species have evolved mechanisms to biosynthesize natural chemical products to kill other species. Penicillin is one such compound produced by Penicillium mould that kills bacteria by inhibiting the cell wall synthesis. It is interesting to raise several questions here. Bacteria also evolve to combat this chemical attack by several mechanisms viz: export/efflux the antibiotic, degrade/catabolize the antibiotic, target modification etc(Reygaert, 2018). These resistance mechanisms make these evolved bacteria more dreadful and are responsible for the havoc we are currently experiencing and we term it as "Antimicrobial resistance (AMR)"

1.4 Natural products, secondary metabolites, specialised metabolites, antibiotics: What are they?

All these terms are mostly used interchangeably by the research community. These small molecules are naturally produced by microbes hence this broadest term of “natural products” is frequently used. As these compounds are mostly distinct from the primary metabolites produced by the particular species, the term “secondary metabolites” is also used. Natural products definition encompasses both primary and secondary metabolites. Since the word “secondary” has the possibility of diluting the importance of these biomolecules, some researchers prefer the term “specialised metabolites” as being more appropriate. Generally these biomolecules function as inhibitors of bacterial growth (bacteriostatic) or sometimes even kill the other organisms (bactericidal), so the term antibiotics is used.

Chemically these small molecules can be classified according to the biosynthetic pathway they follow. Major biosynthetic pathways include those encoding for polyketide synthases (PKS), non-ribosomal peptide synthetases (NRPS), ribosomally synthesised and post translationally synthesized peptides (RiPPs), terpenes, saccharides etc. Clinically used secondary metabolites have diverse pharmacological actions. Antibiotics, antifungal, anti-cancerous, immuno-modulatory, antiviral, antimalarial, antipsychotic, antiobesity etc. are some of the pharmacological actions shown by the diverse natural products produced by microbes and plants. Some of these natural products' chemical structures are shown in Figure 1.

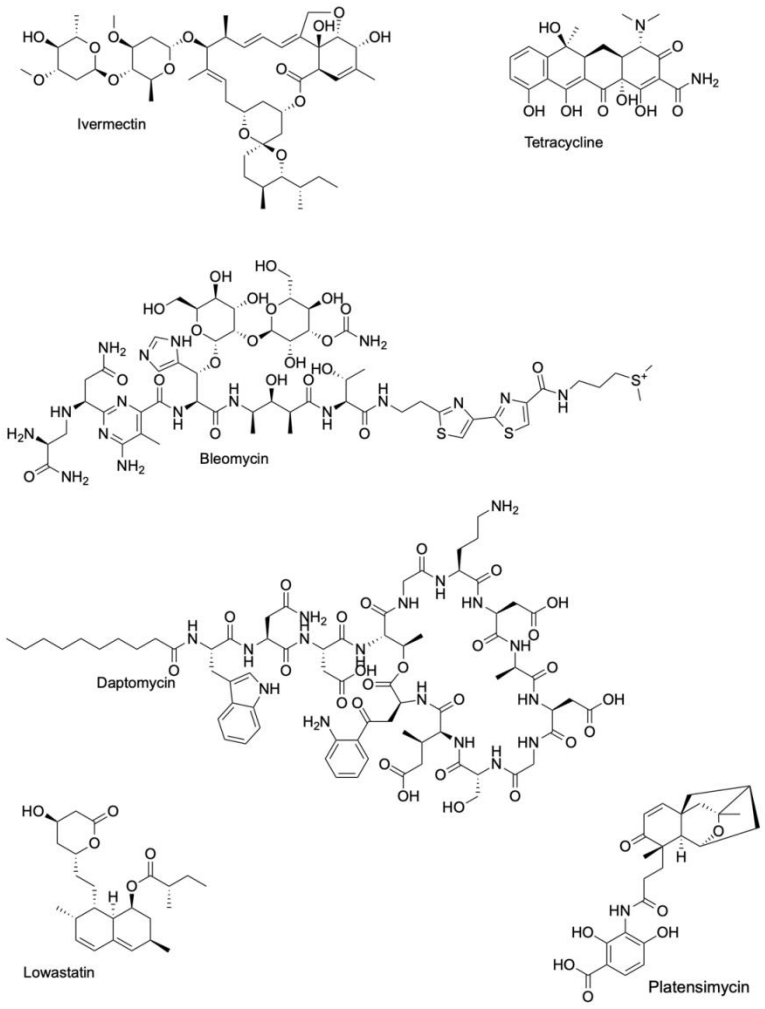


Figure 1

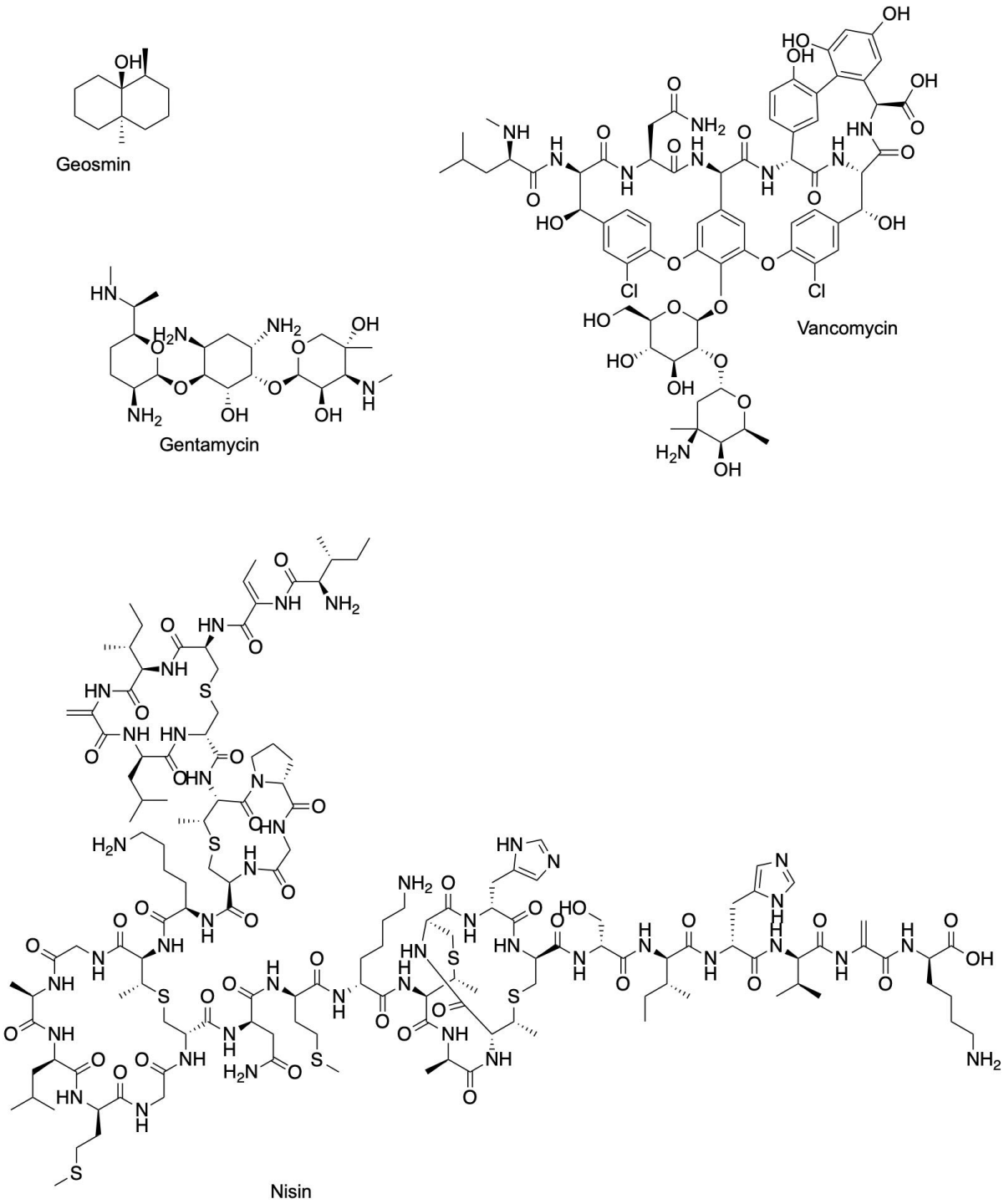


Figure 1: Popular natural products structures. Biosynthetic class these drug belong to and their pharmacological actions is mentioned in the brackets in the caption. Ivermectin (PKS:antiviral and antiparasitic), Tetracycline (PKS: antibiotics), bleomycin (PKS-NRPS: anticancer), daptomycin (NRP: antibiotics), Lovastatin (PKS:anti-cholesterol), Platensimycin (Terpene: antibiotic), Geosmin (Terpene: volatile odour), Nisin (RiPPs: antibiotic),

Gentamycin (Saccharide-PKS: antibiotics), Vancomycin (Glycopeptide: antibiotic). Image Source: Chemical structures from PubChem https://pubchem.ncbi.nlm.nih.gov/

1.5 Traditional routes to discovery

Historically most of the clinically used antibiotic drugs are produced by microbes isolated from soil. Culturing requirement was of prime importance for exploring the antibiotics potential. After growing the isolated species in fermenters, classical chemistry methods came in handy to extract and isolate the potential molecules that show activity. Bioassays measuring the antibiotic activity of isolated compounds against collection of pathogenic organisms were performed. Sophisticated structure elucidation methods are then used to get the exact chemical structure of the biomolecule.

1.6 Hidden microbial dark matter and genome mining

The next generation sequencing (NGS) technologies ushered in the genomics revolution to the extent that it has now become a routine to sequence microbial genomes. Genome databases have become treasure trove that can be mined for novel genes. As these genomes were annotated and studied, it also started becoming evident that the genomes of the biosynthetically gifted streptomyces contained more than one biosynthesis gene cluster (BGC), in some cases several BGCs (Baltz, 2021). Most of the remaining BGCs were silent and were not expressed. Additionally, it is also reported that we are successful in culturing only a fraction (few percent) of the total microbial species present in a particular soil sample (Stewart, 2012). Due to this limitation the large fraction of microbial diversity remains hidden and there is dire need for research methods that can help in uncovering this diversity and realise the full biosynthetic potential.

1.7 BigData and metagenome mining potential

Genomic and metagenomic databases (listed in Annexure A: Table 1) contain huge amounts of genomes and shotgun metagenomic data. Using novel approaches and tools this BigData

can be mined for discovering biosynthetic gene clusters and biosynthetic diversity patterns. The observed biosynthetic pattern when correlated with relevant metadata about the geographical coordinates, horizons (in case of soil samples), treatment conditions, taxonomic diversity, can become useful in rationally answering the long-standing questions about 1) Where the metagenome sampling studies should be conducted to maximise the chances of discovery on novel BGC? 2) How evolutionary patterns shape the microbial communities in a particular ecosystem? Chemical databases (listed in Annexure A: Table 2) along with genomic databases (listed in Annexure A: Table 1) contain rich sources of information that can be mined to discover novel natural products and evolutionary patterns. Some of these datasets can also be used to train the machine learning algorithms that can further help in predicting structural and functional aspects of known and unknown natural products. Crude patterns observed from the meta analysis should be taken with a grain of salt and more standardised sampling procedures supported with detailed documentation would be necessary to claim definite patterns and mechanisms. While it is extremely challenging to discover novel natural products via culture independent approaches using metagenomics, the field has recently tasted success through discovery of Malacidin and Cadaside (Figure 2) (Hover et al., 2018; Wu et al., 2019). Both these natural products have been discovered through metagenomic surveys of soil samples. They show activity against multi-drug resistant pathogens.

The field is currently ripe with standardised tools (listed in Annexure A: Table 3) for genome mining of BGCs and comparative analysis to infer meaningful patterns. Some of the analysis approaches can be directly applied to developing metagenome mining workflows, for some minor adaptation might be necessary. In most of the cases these algorithms run smoothly for smaller metagenome data sizes, but as the sample sizes increase or if the data volume per sample increases, the existing tools run into problems. Sometimes these problems are specific to hardware requirements and can be handled by increasing the resource size (example by increasing the Random Access Memory of the workstations or the storage space). There is definitely a tremendous scope of developing new optimised

algorithms which can accelerate the metagenome mining and analysis. For more details, insights on the topic of big data and evolutionary genome mining for discovery of novel natural products, refer to our published review [See Annexure A].

Figure 2

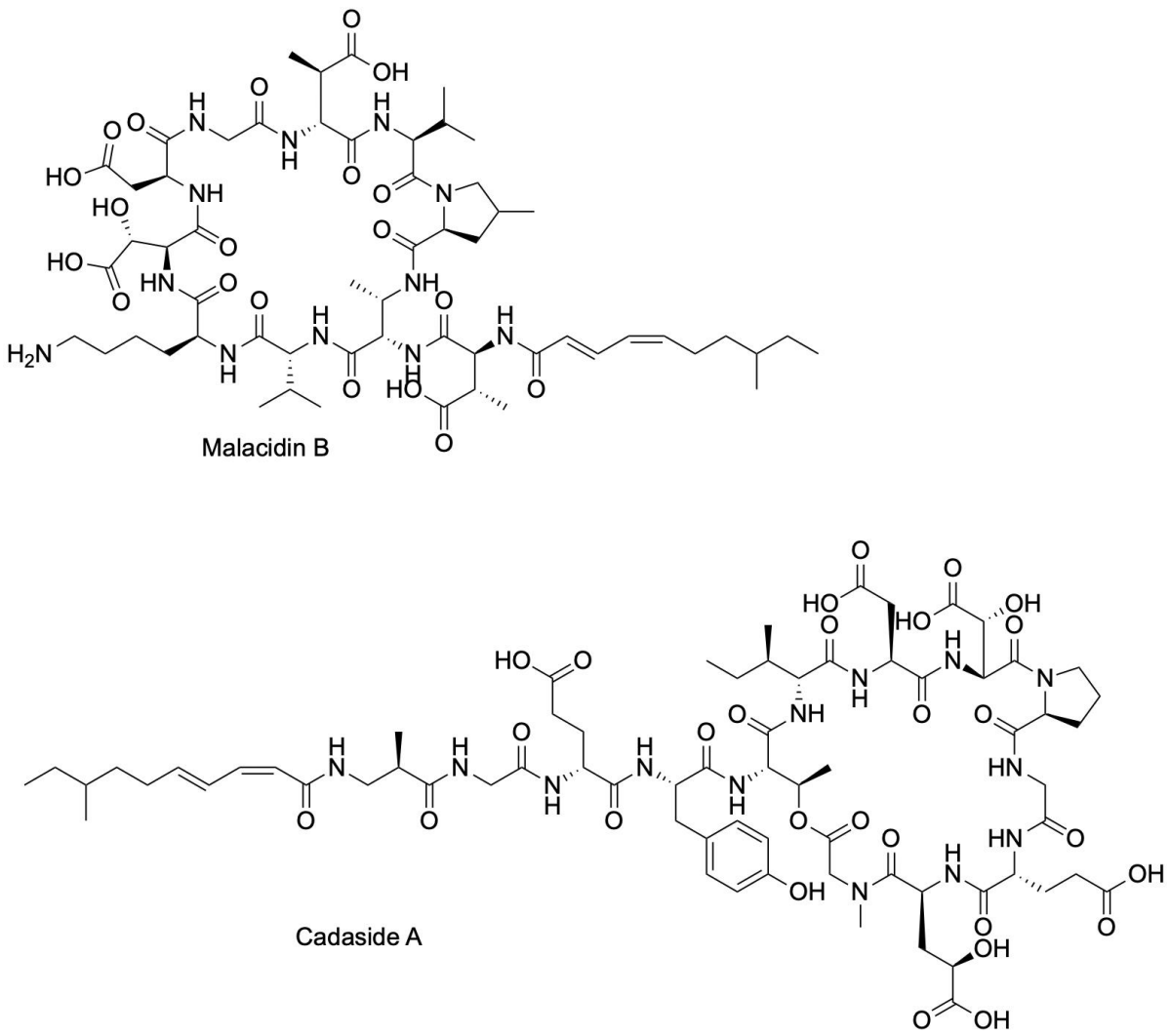


Figure 2: Metagenome derived natural products structures: Malacidin and cadaside. Image Source: Chemical structures from PubChem https://pubchem.ncbi.nlm.nih.gov/

1.8 Research problem

Almost all of the low hanging fruits — the clinically used antibiotics drugs discovered so far — have been picked up by the traditional methods. After tasting the success of discovering novel BGCs from the genomes of culturable microbial species, now the focus has shifted to

mine the hidden dark matter of unculturable microbial diversity. Metagenome sequencing and subsequent metagenome mining method has the potential to uncover novel regions of natural products chemical space. Following factors make adoption of metagenome mining methods challenging: 1) the huge amount of microbial diversity present in diverse ecosystems; this makes it difficult to decide which ecosystems, geographical locations should be surveyed and sampled to maximise the chances of discovery of natural products 2) High sequencing costs to capture complete metagenomes; this makes the method accessible to limited generously funded laboratories 3) Unavailability of easy to use metagenome mining methods and tools; the natural product chemist and microbial ecologists who are interested in using metagenomics methods find it challenging to use the existing command line bioinformatics tools and might need additional skills of using cloud and cluster computing to handle the huge memory space and computation requirements.

1.9 Objectives

Relevant metagenomic approaches, methods and tools are required to be developed using genomics and computational biology techniques to harness the biosynthetic potential. Following objectives were taken up during this dissertation project.

1) Development of bioinformatics pipeline for exploration of natural products chemical space.
2) Comparative metagenomic exploration of soil horizons from multiple sites to identify domain and BGC diversity patterns and correlations.
3) Metagenomic exploration of diverse ecosystems viz. human gut, lake sediments, and strain collections, to identify patterns and generate new hypotheses.

1.10 Thesis Outline

The biological concepts and topics crucial for understanding the basics involved in the area of natural products genome mining and metagenome mining have been described in

Chapter 2. Briefly, this covers topics about microbial diversity, natural products chemical space, biosynthesis pathways of natural products, next generation sequencing technologies and metagenome mining.

Chapter 3 covers technical background. The algorithms and databases involved in genome and metagenome mining have been surveyed and reviewed in this chapter. Microbial community diversity profiling methods, natural products domain exploration methods, natural products biosynthetic cluster exploration methods, metagenome assembly and biosynthesis potential exploration method, and easy to use tools and techniques have been covered.

In Chapter 4 MBEZ pipeline and scripts are described. Analysis steps, workflows, required inputs and generated results output formats have been described.

The Chapter 5 discusses the collaborative pilot project results of a survey of microbial community diversity and biosynthetic diversity of different horizons of Schoenbuch forest soil. Results from the amplicon sequencing, shotgun sequencing (short reads), and shotgun sequencing (long reads) methods have been described. Comparative advantages of each of the methods is also highlighted.

Chapter 6 covers the biosynthesis potential survey of diverse ecosystems, specifically gut microbiomes, lake sediments metagenome and Tuebingen strain collection were studied.

In the final Chapter 7 overall conclusion, expected future impact of the developed methods and approaches, and future challenges in the field are discussed.

1.11 References

Baltz, R.H., 2021. Genome mining for drug discovery: progress at the front end. J. Ind. Microbiol. Biotechnol. https://doi.org/10.1093/jimb/kuab044

Dahm, R., 2008. Discovering DNA: Friedrich Miescher and the early years of nucleic acid research. Hum. Genet. 122, 565–581. https://doi.org/10.1007/s00439-007-0433-0

Droog, S., 2015. Commentary on “Penicillin.” J. R. Nav. Med. Serv. 101. https://doi.org/10.1136/jrnms-101-31

Hanisch, F., Eberhardt, B., Nill, B., 2000. Reconstruction and virtual model of the Schickard calculator. J. Cult. Herit. 1, 335–340. https://doi.org/10.1016/S1296-2074(00)01090-6

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., Perlin, D.S., Brady, S.F., 2018. Culture-independent discovery of the malacidins as calcium-dependent antibiotics with activity against multidrug-resistant Gram-positive pathogens. Nat. Microbiol. 3, 415–422. https://doi.org/10.1038/s41564-018-0110-1

Locey, K.J., Lennon, J.T., 2016. Scaling laws predict global microbial diversity. Proc. Natl. Acad. Sci. U. S. A. 113, 5970–5975. https://doi.org/10.1073/pnas.1521291113
Nüsslein-Volhard, C., 2012. The zebrafish issue of Development. Development 139, 4099–4103. https://doi.org/10.1242/dev.085217
Pulkkinen, K., 2020. Values in the Development of Early Periodic Tables. Ambix 67, 174–198. https://doi.org/10.1080/00026980.2020.1747325
Reygaert, W.C., 2018. An overview of the antimicrobial resistance mechanisms of bacteria. AIMS Microbiol. 4, 482–501. https://doi.org/10.3934/microbiol.2018.3.482
Samanta, I., Bandyopadhyay, S., 2019. Antimicrobial Resistance in Agriculture: Perspective, Policy and Mitigation. Academic Press.
Shlaes, D.M., 2010. Antibiotics: The perfect storm. Antibiot. Perfect Storm 1–106. https://doi.org/10.1007/978-90-481-9057-7
Stewart, E.J., 2012. Growing unculturable bacteria. J. Bacteriol. 194, 4151–4160. https://doi.org/10.1128/JB.00345-12
Voelkel, J.R., 2001. Johannes Kepler and the New Astronomy. Oxford University Press, USA.
Wu, C., Shang, Z., Lemetre, C., Ternei, M.A., Brady, S.F., 2019. Cadasides, Calcium-Dependent Acidic Lipopeptides from the Soil Metagenome That Are Active against Multidrug-Resistant Bacteria. J. Am. Chem. Soc. 141, 3910–3919. https://doi.org/10.1021/jacs.8b12087

Chapter 2: Biological Background

Bacterial species and communities were the focus during this dissertation project and in this chapter I have described the necessary biological background that will be helpful in understanding the subsequent chapters.

2.1 Microbes as source of natural products

Microbes are microscopic organisms and are ubiquitously present in all ecosystems. They constitute archaea, bacteria and fungi. They are the most primitive life forms that arose billions of years ago. Some studies have dated this to be precisely in the range of somewhere around 3.4-3.9 billions of years ago (Betts et al. 2018). Bacterial taxonomy comprises following ranks: phylum, family, class, order, genus, species, strain. Bacterial genomes are circular in structure with densely packed genes. Some bacteria also harbour plasmid genomes. The genome sizes range from a few hundred nucleotide kilobase pairs to few megabase pairs. Most of the clinically used antibiotics to treat infectious disease are produced by bacteria isolated from soil. Antibiotics are the small molecules biosynthesized naturally by bacteria to help them survive.

2.2 Bacterial Diversity: How much do we know?

It won't be an understatement to say that we are living on the microbial planet. According to some estimates there are a trillion bacterial species that live on this planet (Locey and Lennon 2016). Only a minute fraction of these have been studied so far. Figure 1 gives a glimpse of the exponential increase of numbers of genomes deposited over several years.

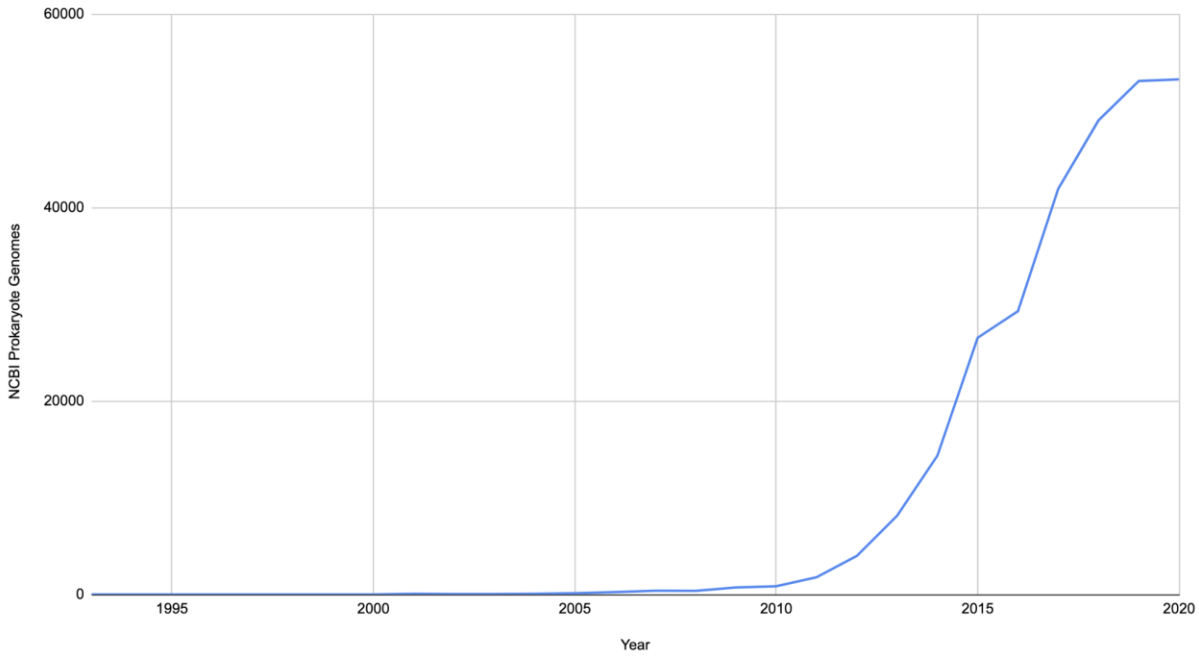


Figure 1. Exponential increase in the number of prokaryotic genomes in NCBI over the course of years.

Every gram of soil contains approximately thousands of bacterial species and of these generally only a few percent (around 2%) of species we can isolate and culture in the lab(Bubnoff 2006). The remaining 98 % we cannot isolate as these require complex media and culture conditions. Practically it is difficult or often impossible to create such conditions artificially in the lab or to create complex growth media that meets the growth needs of unculturable bacteria. Methods that can help in exploring this hidden microbial diversity have the potential to expedite the natural products discovery rate.

2.3 16S rRNA based profiling of diversity of bacteria

16s rRNA gene is conserved across bacteria and is used as a marker for taxonomic labelling (Johnson et al. 2019). This gene contains 1542 nucleotides and constitute several variable regions which help in differentiating the bacteria (Figure 2). These variable regions can be PCR amplified and sequenced via NGS (Yarza et al. 2014). Amplicon based metagenome profiling method has become widely used to profile diversity of a particular metagenomic

sample. Primers covering the variable v3 region have been used in the projects described later in this thesis.

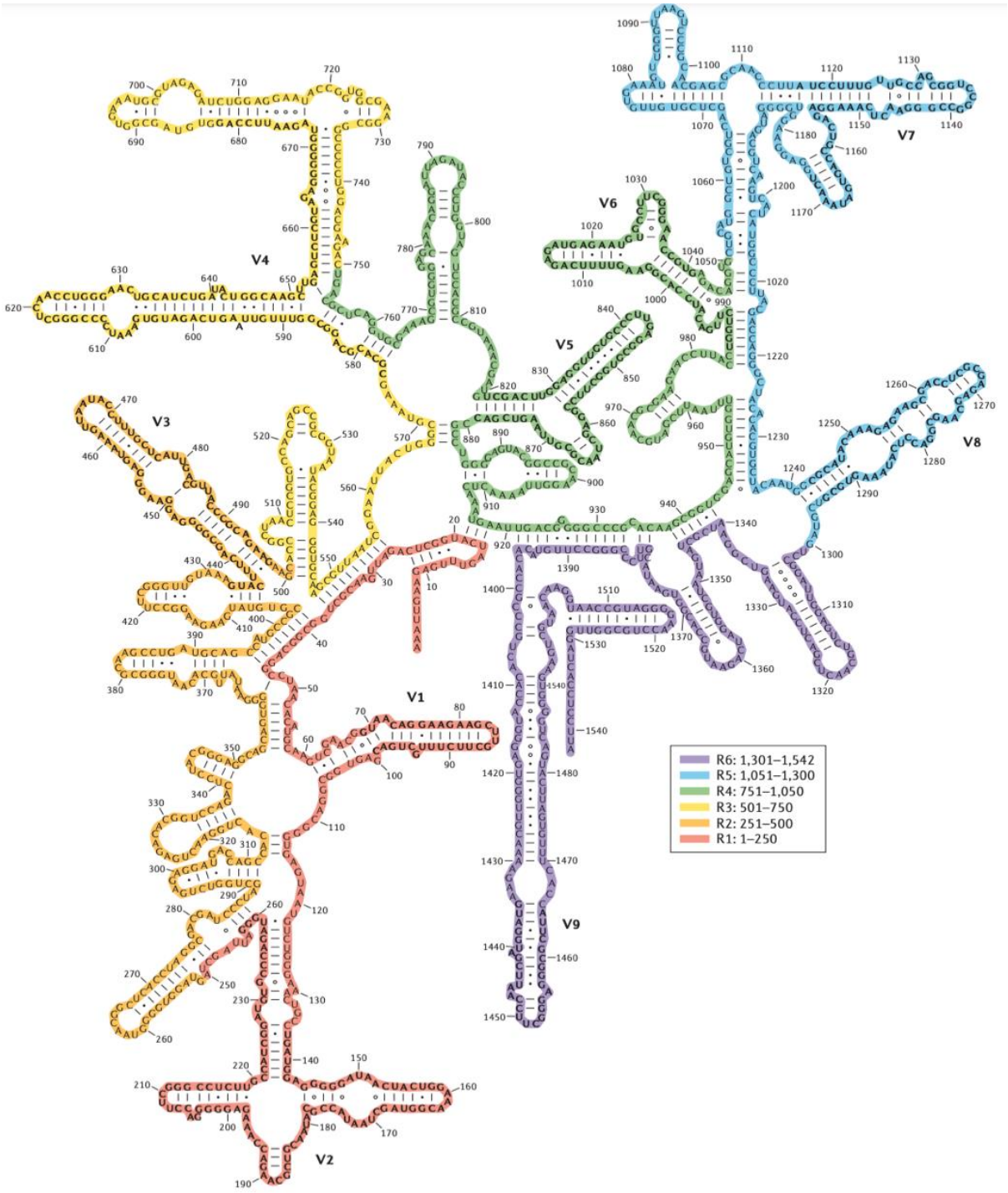


Figure 2. E.coli 16S ribosomal RNA secondary structure showing variable regions. Nucleotide positions: V3 (433-497) and V4 (576-682). Reprinted by permission from [Springer Nature Customer Service Centre GmbH]: Yarza, P., et al. (2014). Nature Reviews Microbiology, 12(9), 635-645. https://doi.org/10.1038/nrmicro3330 . © 2014

2.4 Primary and secondary metabolism pathways

Primary metabolic pathways in bacteria produce the basic biological metabolites necessary for development and growth of bacteria. These involve carbohydrates, proteins, nucleic acids, and lipid metabolic pathways. Apart from these primary metabolic pathways bacteria also produce specialized metabolites via secondary metabolic pathways (Craney, Ahmed, and Nodwell 2013; Davies 2013). These involve biosynthesis of polyketides, NRP, RiPP, terpenes. Biosynthetic pathways of a few antibiotics are described later in this chapter.

2.5 Natural products chemical space and atlas

How many structural families of natural products are biosynthesised by bacteria that are known so far? One indirect way to get a rough estimate would be to subtract archaeal and fungal GCFs from the total ~30k GCFs from BIGFAM (Kautsar et al. 2021). This would give an idea about the potential but the exact structure that these GCFs/BGCs encode cannot be known until each of these BGCs are experimentally characterised. Alternatively, if we only cluster the known BGCs from MiBiG database then this would also be an underestimation as many NP structures have not been connected to their respective BGCs. So the best way to get an estimate would be to classify the NP atlas containing structures (includes bacterial and fungal compounds) (van Santen et al. 2019) using a NP Classifier (Kim et al. 2020). Currently, 653 classes (including plants, marine organisms, fungi, and microorganisms) have been assigned under 7 major chemical pathways by NP Classifier. Pathways include amino acids/peptides, fatty acids, carbohydrates, polyketides, shikimates-phenylpropanoids, terpenoids, and alkaloids. For the latest Natural Products atlas [29,006 total (11,264 bacterial) compounds] there are 466 total (313 bacterial) unique classes predicted by NP Classifier. NP Atlas also reports structural similarity based clustering of microbial compounds based on Dice similarity scoring (0.75 cutoff) and Morgan fingerprinting (radius= 2). Specifically for bacterial compounds this resulted in 3297 clusters and 2487 nodes (based

on atom pairs fingerprinting and Dice similarity scoring (0.7 cutoff)). If we compare these cluster numbers to the number of GCF from BIGFAM (25,667 bacterial GCFs; applying taxon filter), it becomes evident that only about 10 percent of biosynthetic potential has been structurally characterised and studied.

2.6 Biosynthesis of bacterial natural products

2.6.1 Polyketide Biosynthesis Pathway

Nature has devised a strategy analogous to the famous assembly line production process of the automobile industry, to make a class of molecules called the polyketide antibiotics. As a representative example for the class of polyketides, I have briefly described the biosynthesis of a popular antibiotic Erythromycin. Erythromycin is produced by the bacterial species *Aeromicrobium erythreum*. The BGC sequence length of erythromycin is 61845 nucleotides (Figure 3). The key intermediate synthesised during the biosynthesis of Erythromycin is called 6-Deoxyerythronolide B (Figure 4) (Musiol-Kroll and Wohlleben 2018). It is synthesised by 6-Deoxyerythronolide B Synthase (DEBS). DEBS is made up of 3 very large proteins consisting of multiple domains. DEBS1 is a homodimer consisting of two modules (Module 1 and Module 2). Likewise DEBS2 contains modules 3 and 4; and DEBS3 contains modules 5 and 6. Each of the domains does a specific chemical modification function. Polyketide synthase domains (PKS) namely acyltransferase (AT), acyl carrier protein(ACP), ketosynthase(KS), dehydrogenase (DH); ketoreductase(KR), enoyl reductase (ER) and thioesterase (TE) are present in the DEBS. DEBS assembly line uses multiple precursors viz. propionyl coenzyme A and methylmalonyl coenzyme A (Cortes et al. 1990; Donadio et al. 1991). Subsequently, incremental addition of precursors via each module leads to production of 6-Deoxyerythronolide B. The actual enzymatic chemistry that is happening on each of the modules involves the synchronous functioning of catalytic domains present in each of the modules. Translocation of precursor (the growing polyketide) from ACP of upstream module is moved to KS domain (in active site) and is bound to KS domain through

thioester linkage. Next event in the catalytic cycle is acyl transfer. Acyltransferase governs which precursor would be used for the acyl transfer reaction. This is followed by a chain elongation step where the polyketide chain elongation takes place. A two dimensional reaction scheme depicted in Figure 4 is a cartoon representation of the biosynthesis pathway, simplified for ease of understanding. Three dimensional structural details and complexity of a few of these steps have been solved using X-ray crystallography and nuclear magnetic resonance.

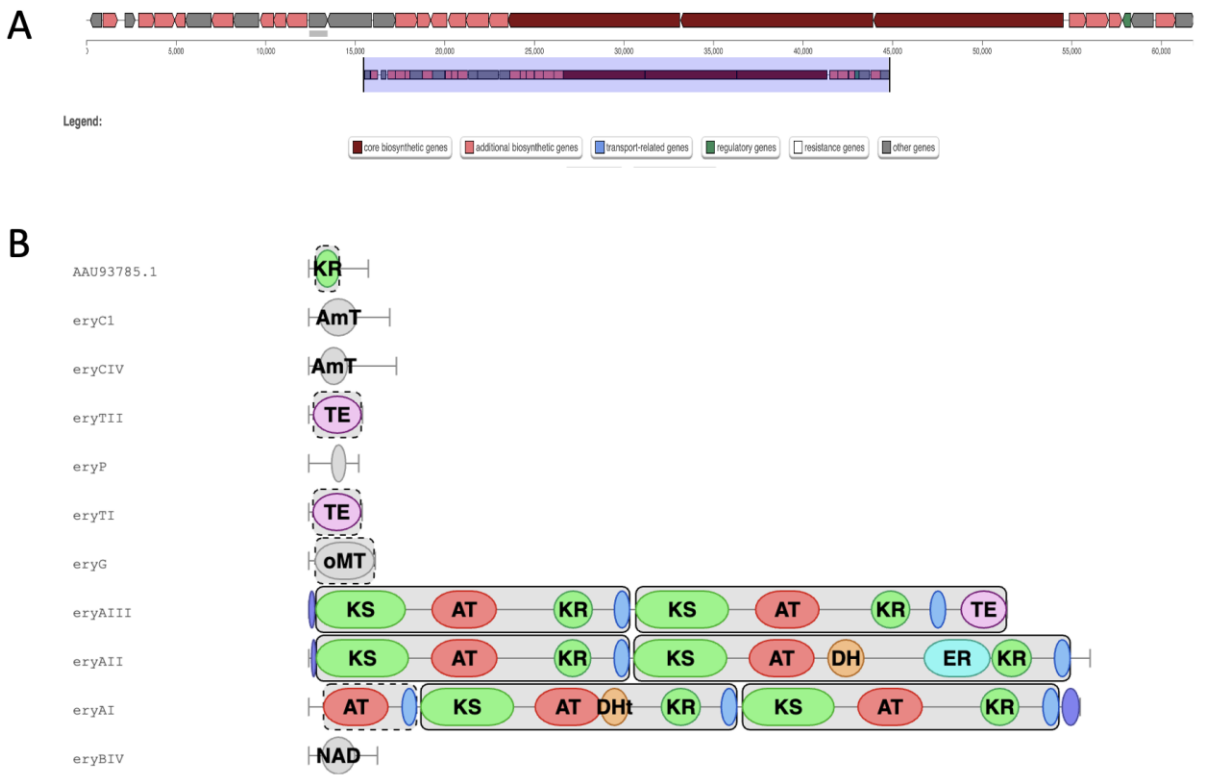


Figure 3: Biosynthetic gene cluster of erythromycin. A) Map showing the position and lengths of genes. B) Modules domain view of the erythromycin biosynthesis genes. Image Source: https://mibig.secondarymetabolites.org/repository/BGC0000054/index.html#r1c1

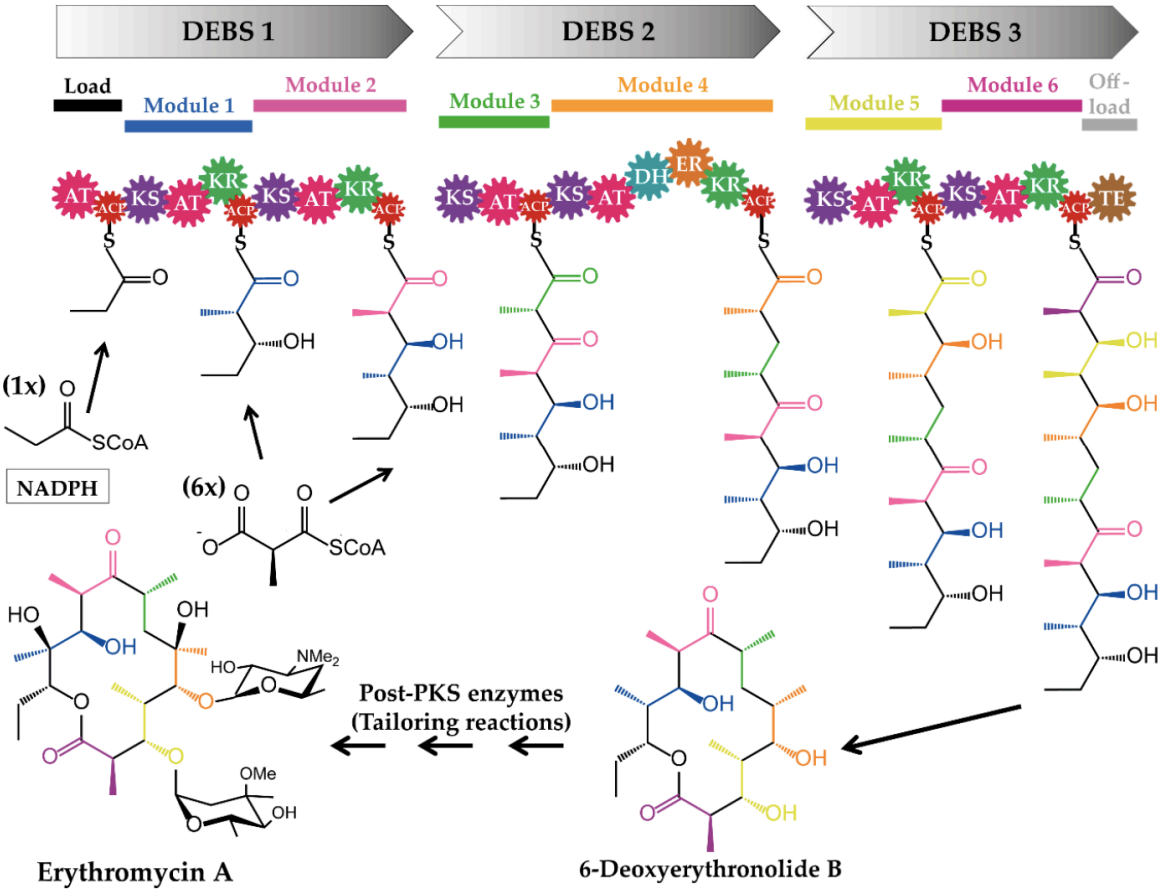


Figure 4. Biosynthesis pathway of erythromycin (DEBS). The assembly line consists of modular PKS machinery. Three subunits DEBS 1, DEBS 2 and DEBS 3 are organized into multiple modules. Every module catalyzes one elongation step using the multiple PKS domains present within these modules (PKS domains: AT, acyltransferase; ACP, acyl carrier protein; KS, ketosynthase; DH, dehydrogenase; KR, ketoreductase; ER, enoylreductase; TE, thioesterase.). Image Source: Musiol-Kroll, E. M., & Wohlleben, W. (2018). Antibiotics, 7(3), 62. https://doi.org/10.3390/antibiotics7030062 , Copyright: Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).

2.6.2 Nonribosomal Peptide Biosynthesis Pathway

As a representative example for the class of NRPs, I have briefly described the biosynthesis of a Kistamicin. Kistamicin is a glycopeptide antibiotic produced by *Actinomadura parvosata* subsp. Kistnae (Nonomuraea sp. ATCC55076). The biosynthetic gene cluster is of length of around 60 kbp (Figure 5A) (Greule et al. 2019). Biosynthesis pathway of glycopeptide

consists of 1) NRPS biosynthetic system 2) Oxidative cyclization cascade. NRPS comprises of multiple biosynthetic domains, namely adenylation domain, condensation domain, epimerization domain, peptidyl carrier protein, thioesterase and X-domain. These domains catalyze the assembly line like linear chain elongation reaction to synthesize the heptapeptide precursor. The chain elongation steps occur in linear fashion as shown in Figure 5B on different modules one after another.

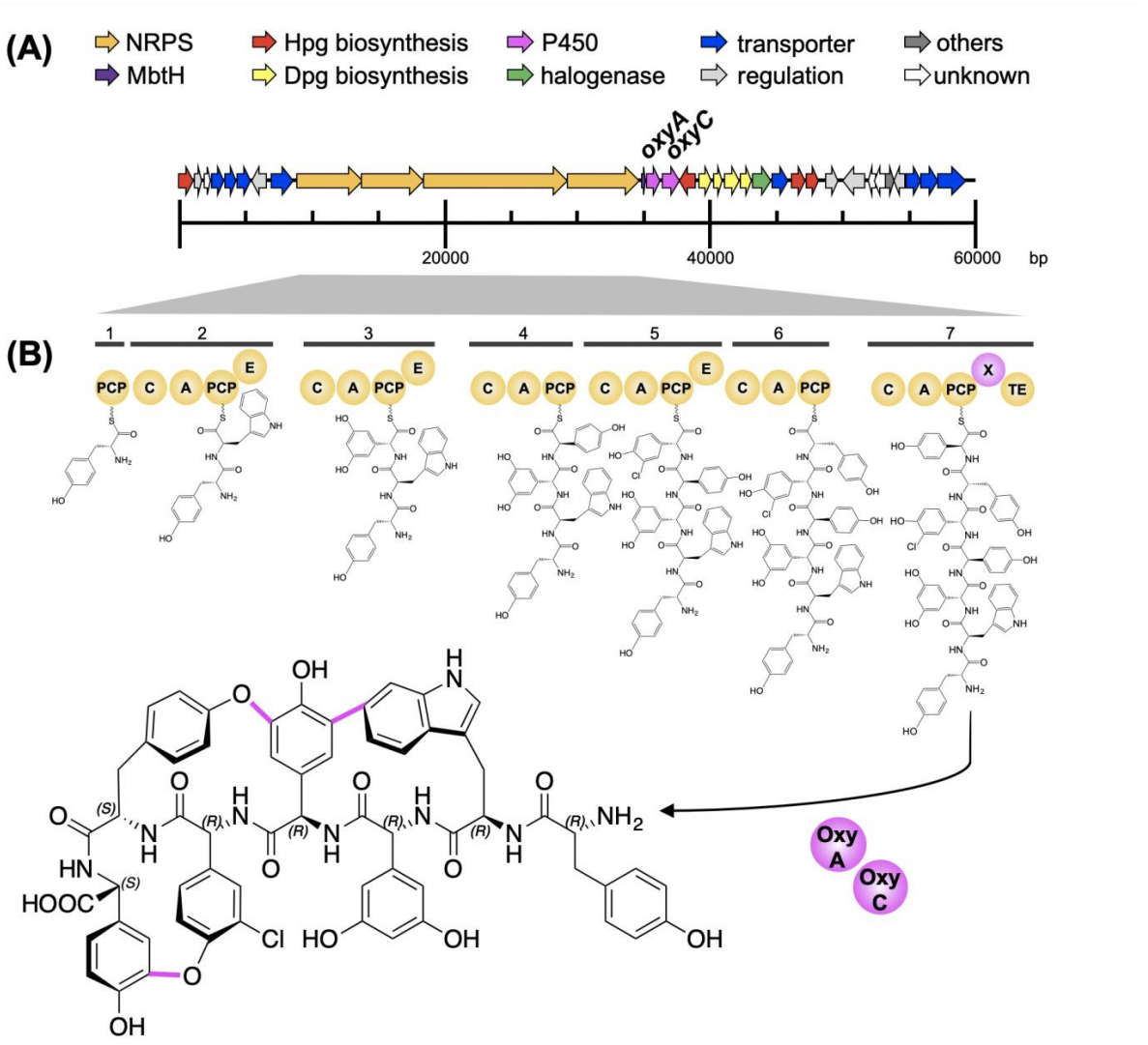


Figure 5. Kistamicin biosynthetic gene cluster and the biosynthesis pathway.

(A) Map showing following genes: 4 nonribosomal peptide synthetase, MbtH protein, biosynthesis genes of non-proteinogenic amino acids 4- hydroxyphenylglycine (Hpg) and 3,5-dihydroxyphenylglycine (Dpg), 2 Cytochrome P450 encoding genes – oxyA and oxyC, FAD-type halogenase, transporter genes, genes encode regulatory proteins,

additional genes and genes with unknown functions. BGC Source organism: Actinomadura parvosata subsp. Kistnae (Nonomuraea sp. ATCC55076); Cluster length: around 60 kb; MIBiG link: https://mibig.secondarymetabolites.org/repository/BGC0001635/index.html#r1c1 (B) Seven NRPS modules (module 1-7) consisting of multiple biosynthesis domains are shown. Heptapeptide precursor linear biosynthesis steps happening at each module are elucidated serially under each module. Final structure of kistamicin is produced via three crosslinking reactions catalysed by the Oxy enzymes and the X domain present in the last module. NRPS domain abbreviation : A, adenylation domain; C, condensation domain; E, epimerization domain; PCP, peptidyl carrier protein; TE, thioesterase; X, Oxy recruiting domain. Image Source: Greule, A. et al. Nat Commun 10, 2613 (2019). https://doi.org/10.1038/s41467-019-10384-w, Copyright: Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).

2.7 Next generation sequencing and genome mining revolution

Sanger sequencing was the workhorse of the human genome project that was completed in 2001(Venter et al. 2001). NIH and Celera were the joint winners of the race to decipher our genome. At that time billions of dollars were spent to accomplish this goal and technological revolution (genome sequencing) was hoped for making the fruits of the human genome available for the masses. Expected sequencing costs had to be reduced by several orders of magnitude. Even the throughput had to be scaled up to achieve accessibility of these methods in clinics and research labs.

Pyrosequencing based sequencing technology achieved by 454 (later acquired by Roche) headed by Jonathan Rothberg truly ushered in the next generation sequencing era (Margulies et al. 2005). Early generations of 454 sequencers could churn out several hundreds of megabases of nucleotide sequences. This was followed by Illumina which further achieved greater throughput and was a more economical method. Paired end short read sequencing method has become the routine method for sequencing bacterial genomes. Short reads have the limitation that they produce many contigs upon assembly, and very

rarely one can get a complete genome as a single contig. This limitation was subsequently removed by long read technologies such as Oxford Nanopore and PacBio (Amarasinghe et al. 2020). Sequencing of several kilobase of DNA fragments, sometimes even the reads in megabase lengths became possible. The limitation of this method is the low quality of bases and sequencing costs are high as compared to the short read technology.

Using the Illumina or nanopore only data or hybrid data assembled genome as an input to algorithms for genome mining of BGCs has become a starting step in any endeavours hoping for discovering novel natural products. antiSMASH uses rule-based logic to annotate and find the BGC in the genome. Briefly, it first finds the genes using prodigal and annotates the genes using pHMMs from PFAM and custom HMM models. Subsequently, clusters are annotated based on collection of cluster rules which comprise the composition and order of genes and domains as previously found in known clusters (Medema et al. 2011).

2.8 Biosynthesis domain diversity profiling via amplicon sequencing.

16S rRNA gene amplicon sequencing gives a glimpse into microbial diversity present in a particular sample. This method is economical and standard protocols, primers and analysis tools are available, which makes it a widely used method. Amplicon sequencing has also been used for studying diversity of biosynthetic domains involved in biosynthesis of secondary metabolites. Degenerate primers capable of amplifying the biosynthetic domains or genes have been reported (Ginolhac et al. 2004; Pimentel-Elardo et al. 2012). These include ketosynthase domain, adenylation domain. KS and A domain primers were also used in one of the projects described later. This domain diversity can be used as a proxy for inferring the biosynthetic potential of the particular sample.

2.9 Metagenome mining for estimating biosynthesis potential

Metagenome sequencing methods decipher total DNA present in the sample. After extracting the DNA, sequencing libraries can be prepared according to the chosen short read technology or the long-read technology protocol. Subsequently the sequencing is done using the next generation sequencers. After platform specific image processing and base calling, the fastq files containing reads are produced. This data can be assembled in metagenomic contigs using appropriate metagenome assembler. Assemblers used in this thesis are briefly described in chapter 3. How much metagenome data should be generated for a particular sample to capture the total metagenomic content? This is a crucial question that should be considered before sequencing any sample. Following factors affect the decision: 1) Desired goals of the particular project 2) Available budget 3) Accessible sequencers 4) Sample alpha diversity estimates.

The assembled metagenome-assembled genomes and contigs give access to the biosynthetic gene clusters which can be detected by antiSMASH. Clusters having high similarity with clusters from MiBiG database can be considered known clusters, while the remaining as unknown ones, which could be harbouring novel biomolecules.

Metagenomes from diverse ecosystems have been profiled so far. These include soil, animal and human gut, different body sites of humans, marine sources, lakes and plants. Studies reporting metagenomic surveys from soils from different sites and covering different scales of land are available in literature. Microbial diversity patterns at continent wide scale, in grassland meadows and even in urban green spaces show the immense diversity that is present in the different soils (Bahram et al. 2018; Crits-Christoph et al. 2020; Delgado-Baquerizo and Eldridge 2019; Thompson et al. 2017; Wang et al. 2018). Massive sequencing efforts would be needed to capture the diversity pattern and develop rational approaches that can guide the future survey.

After the annotation of BGCs derived from the MAGs, the clustered are prioritised for their further wet-lab exploration. Based on the BGCs taxonomic phylogenetic proximity, a suitable

host is chosen for heterologous expression. Currently, due to limited availability of suitable hosts from distant phylum, the required amount of optimisation based on transcriptional and translational regulatory conditions, it is extremely challenging to reach the stage of successful production of novel natural products. These limitations hinder the realization of biosynthetic potential harboured by promising samples and environments. Still metagenomes sequencing followed by BGC mining analysis gives a comprehensive glimpse into the biosynthetic potential.

2.10 References

Amarasinghe, Shanika L., Shian Su, Xueyi Dong, Luke Zappia, Matthew E. Ritchie, and Quentin Gouil. 2020. “Opportunities and Challenges in Long-Read Sequencing Data Analysis.” *Genome Biology* 21 (1): 30. https://doi.org/10.1186/s13059-020-1935-5.

Bahram, Mohammad, Falk Hildebrand, Sofia K Forslund, Jennifer L Anderson, Nadejda A Soudzilovskaia, Peter M Bodegom, Johan Bengtsson-Palme, et al. 2018. “Structure and Function of the Global Topsoil Microbiome.” *Nature* 560 (7717): 233–37. https://doi.org/10.1038/s41586-018-0386-6.

Betts, Holly C., Mark N. Puttick, James W. Clark, Tom A. Williams, Philip C. J. Donoghue, and Davide Pisani. 2018. “Integrated Genomic and Fossil Evidence Illuminates Life’s Early Evolution and Eukaryote Origin.” *Nature Ecology & Evolution* 2 (10): 1556–62. https://doi.org/10.1038/s41559-018-0644-x.

Bubnoff, Andreas von. 2006. “Seeking New Antibiotics in Nature’s Backyard.” *Cell* 127 (5): 867–69. https://doi.org/10.1016/j.cell.2006.11.021.

Cortes, Jesus, Stephen F. Haydock, Gareth A. Roberts, Debra J. Bevitt, and Peter F. Leadlay. 1990. “An Unusually Large Multifunctional Polypeptide in the Erythromycin-Producing Polyketide Synthase of Saccharopolyspora Erythraea.” *Nature* 348 (6297): 176–78. https://doi.org/10.1038/348176a0.

Craney, Arryn, Salman Ahmed, and Justin Nodwell. 2013. “Towards a New Science of Secondary Metabolism.” *The Journal of Antibiotics* 66 (7): 387–400. https://doi.org/10.1038/ja.2013.25.

Crits-Christoph, Alexander, Matthew R. Olm, Spencer Diamond, Keith Bouma-Gregson, and Jillian F. Banfield. 2020. “Soil Bacterial Populations Are Shaped by Recombination and Gene-Specific Selection across a Grassland Meadow.” *ISME Journal* 14 (7): 1834–46. https://doi.org/10.1038/s41396-020-0655-x.

Davies, Julian. 2013. “Specialized Microbial Metabolites: Functions and Origins.” *The Journal of Antibiotics* 66 (7): 361–64. https://doi.org/10.1038/ja.2013.61.

Delgado-Baquerizo, Manuel, and David J Eldridge. 2019. “Cross-Biome Drivers of Soil Bacterial Alpha Diversity on a Worldwide Scale.” *Ecosystems* 22 (6): 1220–31. https://doi.org/10.1007/s10021-018-0333-2.

Donadio, S., M. J. Staver, J. B. McAlpine, S. J. Swanson, and L. Katz. 1991. “Modular Organization of Genes Required for Complex Polyketide Biosynthesis.” *Science (New York, N.Y.)* 252 (5006): 675–79. https://doi.org/10.1126/science.2024119.

Ginolhac, Aurélien, Cyrille Jarrin, Benjamin Gillet, Patrick Robe, Petar Pujic, Karine Tuphile, Hélène Bertrand, et al. 2004. “Phylogenetic Analysis of Polyketide Synthase I Domains from Soil Metagenomic Libraries Allows Selection of Promising Clones.” *Applied and Environmental Microbiology* 70 (9): 5522–27. https://doi.org/10.1128/AEM.70.9.5522-5527.2004.

Greule, Anja, Thierry Izore, Dumitrita Iftime, Julien Tailhades, Melanie Schoppet, Yongwei Zhao, Madeleine Peschke, et al. 2019. “Kistamicin Biosynthesis Reveals the Biosynthetic Requirements for Production of Highly Crosslinked Glycopeptide Antibiotics.” *Nature Communications* accepted. https://doi.org/10.1038/s41467-019-10384-w.

Johnson, Jethro S., Daniel J. Spakowicz, Bo-Young Hong, Lauren M. Petersen, Patrick Demkowicz, Lei Chen, Shana R. Leopold, et al. 2019. “Evaluation of 16S RRNA Gene Sequencing for

Species and Strain-Level Microbiome Analysis." *Nature Communications* 10 (1): 5029. https://doi.org/10.1038/s41467-019-13036-1.

Kautsar, Satria A, Kai Blin, Simon Shaw, Tilmann Weber, and Marnix H Medema. 2021. "BiG-FAM: The Biosynthetic Gene Cluster Families Database." *Nucleic Acids Research* 49 (D1): D490–97. https://doi.org/10.1093/nar/gkaa812.

Kim, Hyunwoo, Mingxun Wang, Christopher Leber, Louis-Felix Nothias, Raphael Reher, Kyo Bin Kang, Justin J. J. van der Hooft, Pieter Dorrestein, William Gerwick, and Garrison Cottrell. 2020. "NPClassifier: A Deep Neural Network-Based Structural Classification Tool for Natural Products," August. https://doi.org/10.26434/chemrxiv.12885494.v1.

Locey, Kenneth J., and Jay T. Lennon. 2016. "Scaling Laws Predict Global Microbial Diversity." *Proceedings of the National Academy of Sciences* 113 (21): 5970–75. https://doi.org/10.1073/pnas.1521291113.

Margulies, Marcel, Michael Egholm, William E. Altman, Said Attiya, Joel S. Bader, Lisa A. Bemben, Jan Berka, et al. 2005. "Genome Sequencing in Microfabricated High-Density Picolitre Reactors." *Nature* 437 (7057): 376–80. https://doi.org/10.1038/nature03959.

Medema, Marnix H., Kai Blin, Peter Cimermancic, Victor de Jager, Piotr Zakrzewski, Michael A. Fischbach, Tilmann Weber, Eriko Takano, and Rainer Breitling. 2011. "AntiSMASH: Rapid Identification, Annotation and Analysis of Secondary Metabolite Biosynthesis Gene Clusters in Bacterial and Fungal Genome Sequences." *Nucleic Acids Research* 39 (suppl_2): W339–46. https://doi.org/10.1093/nar/gkr466.

Musiol-Kroll, Ewa Maria, and Wolfgang Wohlleben. 2018. "Acyltransferases as Tools for Polyketide Synthase Engineering." *Antibiotics (Basel, Switzerland)* 7 (3): E62. https://doi.org/10.3390/antibiotics7030062.

Pimentel-Elardo, Sheila Marie, Lubomir Grozdanov, Sebastian Proksch, and Ute Hentschel. 2012. "Diversity of Nonribosomal Peptide Synthetase Genes in the Microbial Metagenomes of Marine Sponges." *Marine Drugs* 10 (6): 1192–1202. https://doi.org/10.3390/md10061192.

Santen, Jeffrey A. van, Grégoire Jacob, Amrit Leen Singh, Victor Aniebok, Marcy J. Balunas, Derek Bunsko, Fausto Carnevale Neto, et al. 2019. "The Natural Products Atlas: An Open Access Knowledge Base for Microbial Natural Products Discovery." *ACS Central Science* 5 (11): 1824–33. https://doi.org/10.1021/acscentsci.9b00806.

Thompson, Luke R, Jon G Sanders, Daniel McDonald, Amnon Amir, Joshua Ladau, Kenneth J Locey, Robert J Prill, et al. 2017. "A Communal Catalogue Reveals Earth's Multiscale Microbial Diversity." *Nature* 551 (7681): 457–63. https://doi.org/10.1038/nature24621.

Venter, J. Craig, Mark D. Adams, Eugene W. Myers, Peter W. Li, Richard J. Mural, Granger G. Sutton, Hamilton O. Smith, et al. 2001. "The Sequence of the Human Genome." *Science* 291 (5507): 1304–51. https://doi.org/10.1126/science.1058040.

Wang, Haitao, Minying Cheng, Melissa Dsouza, Pamela Weisenhorn, Tianling Zheng, and Jack A Gilbert. 2018. "Soil Bacterial Diversity Is Associated with Human Population Density in Urban Greenspaces." *Environmental Science and Technology* 52 (9): 5115–24. https://doi.org/10.1021/acs.est.7b06417.

Yarza, Pablo, Pelin Yilmaz, Elmar Pruesse, Frank Oliver Glöckner, Wolfgang Ludwig, Karl Heinz Schleifer, William B Whitman, Jean Euzéby, Rudolf Amann, and Ramon Rosselló-Móra. 2014. "Uniting the Classification of Cultured and Uncultured Bacteria and Archaea Using 16S RRNA Gene Sequences." *Nature Reviews Microbiology* 12 (9): 635–45. https://doi.org/10.1038/nrmicro3330.

Chapter 3: Technical Background

3.1 Microbial Community Diversity Profiling Methods

While studying microbiomes from any ecosystem the following two questions are of prime importance: 1) Which microbial species are present in the sample? 2) What are the particular species doing in the sample? The first question can be answered by using two methods. Firstly, by studying the sequence diversity of 16S rRNA amplicons. 16S ribosomal RNA gene is highly conserved across the species of bacteria and archaea and is used as a phylogenetic marker. The SILVA 16S rRNA gene database is used for taxonomic annotation for amplicon based methods (Quast et al. 2013). QIIME2 a microbiome data science platform implements the multi-step microbial community diversity profiling workflow for amplicon data (Bolyen et al. 2019). Briefly, the steps include 1) Data quality filtration and preprocessing 2) DADA2 based denoising and chimera filtration to construct the Amplicon sequence variants (ASV) 3) OTU construction and taxonomic analysis 4) Rarefaction, alpha, beta diversity analysis 5) Correlation and association analysis (Callahan et al. 2016). Profiling and tracking of particular strains is difficult while using the 16S rRNA based amplicon sequencing dataset.

Secondly, by using the shotgun metagenome sequencing data and annotating it with the non-redundant protein database. Such protein annotation can be accelerated by using BLASTX and DIAMOND and the classification based on NCBI or SILVA taxonomy can be visualised using MEGAN (Camacho et al. 2009; Buchfink, Xie, and Huson 2014; Huson et al. 2016). Recent methods available for taxonomic classification a) Kraken2 is based on k-mer matches b) kaiju is based on Maximum Exact Matches

3.2 Natural Products Biosynthesis Domain exploration methods

Protein domain regions present in the genes responsible for multi-step biosynthesis pathways of natural products, can be used to study the secondary metabolite gene diversity. Ketosynthase (KS), Adenylation (A), Condensation (C) have been studied extensively so far. In this section we review the methods available for natural products biosynthesis domain annotation and diversity analysis.

3.2.1 NaPDoS

Natural Products Domain Seeker is the web based tool for automated computation of biosynthetic gene diversity analysis (Ziemert et al. 2012). Currently polyketide synthase (PKS) and non-ribosomal peptide synthetase (NRPS) genes can be analysed using NaPDoS. Specifically, KS and C domain annotation and phylogenetic placement analysis is catered to. Input data types that are accepted include PCR products, genome sequence and metagenomic reads. Output results include Hidden Markov Model (HMM) search using KS and C domain HMM models. BLAST annotation against the 459 KS and 190 C domains is performed to assign the biosynthesis pathway related information.

3.2.2 BiG-MEx

Biosynthetic Gene cluster MEtagenomic eXploration toolbox (BiG-MEx) can be used for annotation of numerous BGC protein domains (Pereira 2020). Presently, annotation of 150 BGC domains covering major secondary metabolite biosynthesis pathways is possible (Table) . Input data types compatible with this tool are BGC domain amplicon-seq datasets, shotgun metagenomic datasets. Results include UProC based domain annotation and abundance statistics. Diversity analysis includes computation of Shannon alpha diversity index. Five docker containers (bgc_dom_annot,bgc_dom_amp_div,bgc_dom_meta_div,bgc_dom_merge_div, bgc_class_pred) are available from DockerHub (https://hub.docker.com/u/epereira) and the source code in available from GitHub (https://github.com/pereiramemo/BiG-MEx).

AMP-binding	LE-LAC481	PF13575	cypI	novK
ATd	LE-MER+2PEP	PKS_AT	cypemycin	phosphonates
AfsA	Lactococcin	PKS_KS	dmat	phytoene_synt
Antimicrobial14	Lactococcin_972	PP-binding	ectoine_synt	phzB
Antimicrobial17	Lant_dehyd_C	TIGR03601	fabH	prnB
Antimicrobial18	Lant_dehyd_N	TIGR03602	fom1	pur6
Autoind_synth	LcnG-beta	TIGR03603	frbD	pur10
A-OX	Linocin_M18	TIGR03604	ft1fas	salQ
BLS	LipM	TIGR03605	fung_ggpps	skfc
BacteriocIIc_cy	LipU	TIGR03651	fung_ggpps2	spcDK_like_cou
Bacteriocin_II	LipV	TIGR03678	glycocin	spcDK_like_glyc
Bacteriocin_IIc	LmbU	TIGR03693	goadsporin_like	spcFG_like
Bacteriocin_IId	Lycopene_cycl	TIGR03731	hglD	strH_like
Bacteriocin_IIi	L_biotic_typeA	TIGR03793	hglE	strK_like1
CAS	MA-2PEPA	TIGR03795	indsynth	strK_like2
CaiA	MA-DUF	TIGR03798	lasso	strepbact
Chal_sti_synt_C	MA-EPI	TIGR03882	mcjC	strep_PEQAXS
Chal_sti_synt_N	MA-LAC481	TIGR03975	melC	sublancin
Cloacin	MA-NIS	TIGR04363	micJ25	subtilosin
Condensation	MA-NIS+EPI	Terpene_synth	mitE	t2clf
DOIS	MGT	Terpene_synth_C	mmyO	t2fas
DUF692	MoeO5	ToyB	moeGT	t2ks
DUF1205	NAD_binding_4	TunD	mvd_mst	t2ks2
Gallidermin	NapT7	YcaO	mvnA	tabtoxin
Glycos_transf_1	Neocarzinostat	bacilysin	neoL_like	terpene_cyclase
Glycos_transf_2	PF00067	bcpB	nikJ	thiostrepton
Glyco_transf_28	PF00106	botH	nikO	thuricin
lucA_lucC	PF02441	bt1fas	novH	trichodiene_synth
LANC_like	PF04820	cyanobactin_synth	novI	valA_like
LE-DUF	PF13561	cycdipepsynth	novJ	vlmB

Table: List of BGC domains that can be analysed using BiG-MEx. Source:

https://github.com/pereiramemo/BiG-MEx/blob/master/data/150_uproc_bgc_dom.list

3.2.3 Dom2BGC

dom2BGC is a pipeline tool helpful in analysing the functional amplicons that target BGC domains (Tracanna et al. 2021). AMP binding domain involved in non-ribosomal peptide synthetase biosynthesis pathway can be analysed using dom2BGC. MIBiG and antiSMASH-

DB contain the comprehensive collection of AMP binding domains present in BGCs sequences (Blin, Pascal Andreu, et al. 2019). dom2BGC annotates the sample amplicons based on sequence similarity to this largest collection of AMP binding domains. Using co-occurrence network dom2BGC also detects groups of amplicons that jointly originate from the same BGCs from multiple samples.

3.3 Natural Products Biosynthesis Cluster exploration methods

3.3.1 CLUSEAN

Bacterial secondary metabolites are small molecules having diverse functions. Some of these have antimicrobial and cytostatic actions and are used as drugs to fight infections and cancerous diseases. These molecules are biosynthesized in an assembly line-like multi-step process by multimodular megaenzymes. These megaenzyme genes are often clustered in the genome. CLUSEAN (CLUster SEquence ANalyzer) helps in detecting and analyzing such gene clusters (Weber et al. 2009). It uses BLAST and HMMER for annotating the functional domains.

3.3.2 antiSMASH

Subsequent to CLUSEAN, antiSMASH (antibiotics & Secondary Metabolite Analysis Shell) was developed by the research group led by University of Tuebingen and was released in 2011 (Medema et al. 2011). Since then it has become a popular tool and is being updated continuously to improve the analysis of existing BGC classes and add newer classes of BGCs. Recently version 6 of antiSMASH has been released (Blin, Shaw, et al. 2019). The pipeline annotates 70 BGC types (https://docs.antismash.secondarymetabolites.org/glossary/) covering major secondary metabolite compound classes: polyketides, non-ribosomal peptides, lantibiotics, bacteriocins, nucleosides, beta-lactams, terpenes, aminoglycosides, aminocoumarins, indolocarbazoles, butyrolactones, siderophores, melanins and others. At the heart of the

antiSMASH is the rule-based detection of BGCs using the signature profile Hidden Markov Models (pHMMs) of proteins or protein domains.

3.3.3 BIGSCAPE

Due to the next generation sequencing revolution and availability of an easy to use antiSMASH tool, more and more genomes and BGCs became publicly available and it started becoming challenging to draw suitable inferences without comprehensive analysis of these numerous BGCs together. Through BIGSCAPE a new informatics workflow was created that made scaling up the mining of entire microbiome and strain collection comprising hundreds or even thousands of bacteria (Navarro-Muñoz et al. 2020). Using BIGSCAPE it is possible to create a sequence similarity network of BGCs and gene cluster families. Further using CORASON (Core analysis of syntenic orthologs to prioritize natural products cluster), the phylogenetic relationship across the BGCs can be studied. It is also possible to include the MIBiG clusters during the BIGSCAPE analysis. This helps in knowing which sample clusters are having similarity to known clusters present in the MIBiG database.

3.3.4 deepBGC

Machine learning methods have recently become extremely popular and have been widely used to improve the prediction accuracies and precisions of numerous bioinformatics algorithms. As long as high quality data is abundantly available for training the machine learning algorithms, this method has a potential to revolutionize the complete landscape of research and development. deepBGC uses deep learning supplemented with random forest classifier to identify BGCs and predict their compound classes and potential chemical activity (Hannigan et al. 2019). Previously undetected BGCs have been shown to be identified by deepBGC.

3.4 Metagenome Assembly and Natural products biosynthesis potential exploration methods

3.4.1 metaSPADES

Metagenome assembly is challenging in terms of huge magnitude of data and also requires an extensive amount of computational resources. High amounts of RAM, CPU cores and processing time, and storage is a must for performing metagenome assemblies. metaSPADES constructs the de Bruijn graph using all the metagenomic sequence reads (Nurk et al. 2017). After transformation and creation of the assembly graph it reconstructs paths that belong to longer genomic contigs. It can accept both short and long reads and can also perform hybrid assembly of such metagenomic sequence data.

3.4.2 CloudSPADES

Low cost Illumina short reads and high cost PacBio or Oxford Nanopore long reads are generally both needed for de novo assembly of the genome. Synthetic long reads technology is useful in generating low cost contiguous de novo assemblies. 10X Genomics and TELL-Seq methods have been recently introduced that cater the synthetic long reads data (Chen et al. 2020). CloudSPADE uses the sets of collections of substrings in a cloud containing a set of all the substring (Tolstoganov et al. 2019). Barcoded reads are assembled into contigs which are subsequently used to create clouds based on the set of contigs that a synthetic long read is mapped to. Using the assembly graphs the correct order and orientation of the contigs is deduced.

3.4.3 TELL-Link

Transposase enzyme linked long reads sequencing library technology generates barcode linked reads for genome and metagenome scale sequencing application (Chen et al. 2020). TELL-Link is as barcode aware assembly pipeline that assembles contigs and creates scaffolds. It takes as the input the processed FASTQ data processed through the TELL-Read pipeline. K-mer based assembly graphs are constructed and the barcode information

is used to resolve complex structures. The reads that share the same barcode are used to reconstruct the local assembly. Chosen k-mer sizes affect the assembly results and the pipeline provides options to specify global assembly graph and local assembly graph k-mer sizes.

3.5 Tools worth exploring in future

NextFlow enables reproducible scientific workflow pipeline deployment on both clouds and clusters. It caters Docker and Singularity containers and makes the pipelines portable on diverse computational platforms.

GECCO (GEne Cluster prediction with COnditional random fields; https://gecco.embl.de) uses conditional random fields (CRFs) for identifying BGCs in genomic and metagenomic data (Carroll et al. 2021). A recent preprint describing GECCO claims a significant increase in identification of BGCs than the traditional rule-based approach.

3.5.1 Miscellaneous tools

Numerous technical tools were used in the projects described in this thesis. Some of these tools and technologies make command-line agnostic researchers' lives easy. These include Docker containers, CONDA and Pip package management system, Jupyter Notebooks and VIM editor.

3.6 Reference

Blin, Kai, Victòria Pascal Andreu, Emmanuel L C De Los Santos, Francesco Del Carratore, Sang Yup Lee, Marnix H Medema, and Tilmann Weber. 2019. “The AntiSMASH Database Version 2: A Comprehensive Resource on Secondary Metabolite Biosynthetic Gene Clusters.” *Nucleic Acids Research* 47 (D1): D625–30. https://doi.org/10.1093/nar/gky1060.

Blin, Kai, Simon Shaw, Katharina Steinke, Rasmus Villebro, Nadine Ziemert, Sang Yup Lee, Marnix H Medema, and Tilmann Weber. 2019. “AntiSMASH 5.0: Updates to the Secondary Metabolite Genome Mining Pipeline.” *Nucleic Acids Research* 47 (W1): W81–87. https://doi.org/10.1093/nar/gkz310.

Bolyen, Evan, Jai Ram Rideout, Matthew R. Dillon, Nicholas A. Bokulich, Christian C. Abnet, Gabriel A. Al-Ghalith, Harriet Alexander, et al. 2019. “Reproducible, Interactive, Scalable and Extensible Microbiome Data Science Using QIIME 2.” *Nature Biotechnology* 37 (8): 852–57. https://doi.org/10.1038/s41587-019-0209-9.

Buchfink, Benjamin, Chao Xie, and Daniel H Huson. 2014. “Fast and Sensitive Protein Alignment Using DIAMOND.” *Nature Methods* 12 (1): 59–60. https://doi.org/10.1038/nmeth.3176.

Callahan, Benjamin J, Paul J McMurdie, Michael J Rosen, Andrew W Han, Amy Jo A Johnson, and Susan P Holmes. 2016. “DADA2: High-Resolution Sample Inference from Illumina Amplicon

Data.” *Nature Methods* 13 (7): 581. https://doi.org/10.1038/nmeth.3869.
Camacho, Christiam, George Coulouris, Vahram Avagyan, Ning Ma, Jason Papadopoulos, Kevin Bealer, and Thomas Madden. 2009. “{BLAST+}: Architecture and Applications.” *BMC Bioinformatics* 10 (1): 421. https://doi.org/10.1186/1471-2105-10-421.
Carroll, Laura M., Martin Larralde, Jonas Simon Fleck, Ruby Ponnudurai, Alessio Milanese, Elisa Cappio, and Georg Zeller. 2021. “Accurate de Novo Identification of Biosynthetic Gene Clusters with GECCO.” https://doi.org/10.1101/2021.05.03.442509.
Chen, Zhoutao, Long Pham, Tsai-Chin Wu, Guoya Mo, Yu Xia, Peter L. Chang, Devin Porter, et al. 2020. “Ultralow-Input Single-Tube Linked-Read Library Method Enables Short-Read Second-Generation Sequencing Systems to Routinely Generate Highly Accurate and Economical Long-Range Sequencing Information.” *Genome Research* 30 (6): 898–909. https://doi.org/10.1101/gr.260380.119.
Hannigan, Geoffrey D, David Prihoda, Andrej Palicka, Jindrich Soukup, Ondrej Klempir, Lena Rampula, Jindrich Durcak, et al. 2019. “A Deep Learning Genome-Mining Strategy for Biosynthetic Gene Cluster Prediction.” *Nucleic Acids Research* 47 (18): e110–e110. https://doi.org/10.1093/nar/gkz654.
Huson, Daniel H, Sina Beier, Isabell Flade, Anna Górska, and Mohamed El-hadidi. 2016. “MEGAN Community Edition - Interactive Exploration and Analysis of Large-Scale Microbiome Sequencing Data,” 1–12. https://doi.org/10.1371/journal.pcbi.1004957.
Medema, Marnix H., Kai Blin, Peter Cimermancic, Victor de Jager, Piotr Zakrzewski, Michael A. Fischbach, Tilmann Weber, Eriko Takano, and Rainer Breitling. 2011. “AntiSMASH: Rapid Identification, Annotation and Analysis of Secondary Metabolite Biosynthesis Gene Clusters in Bacterial and Fungal Genome Sequences.” *Nucleic Acids Research* 39 (suppl_2): W339–46. https://doi.org/10.1093/nar/gkr466.
Navarro-Muñoz, Jorge C, Nelly Selem-Mojica, Michael W Mullowney, Satria A Kautsar, James H Tryon, Elizabeth I Parkinson, Emmanuel L C De Los Santos, et al. 2020. “A Computational Framework to Explore Large-Scale Biosynthetic Diversity.” *Nature Chemical Biology* 16 (1): 60–68. https://doi.org/10.1038/s41589-019-0400-9.
Nurk, Sergey, Dmitry Meleshko, Anton Korobeynikov, and Pavel A Pevzner. 2017. “MetaSPAdes: A New Versatile Metagenomic Assembler.” *Genome Research* 27 (5): 824–34. https://doi.org/10.1101/gr.213959.116.
Pereira, Emiliano. 2020. “Improvements in Natural Product Biosynthetic Gene Clusters Research and Functional Trait-Based Approaches in Metagenomics.” Jacobs University Bremen. http://nbn-resolving.org/urn:nbn:de:gbv:579-opus-1008965.
Quast, Christian, Elmar Pruesse, Pelin Yilmaz, Jan Gerken, Timmy Schweer, Pablo Yarza, Jörg Peplies, and Frank Oliver Glöckner. 2013. “The {SILVA} Ribosomal {RNA} Gene Database Project: Improved Data Processing and Web-Based Tools.” *Nucleic Acids Research* 41 (D1): D590–96. https://doi.org/10.1093/nar/gks1219.
Tolstoganov, Ivan, Anton Bankevich, Zhoutao Chen, and Pavel A Pevzner. 2019. “CloudSPAdes: Assembly of Synthetic Long Reads Using de Bruijn Graphs.” *Bioinformatics* 35 (14): i61–70. https://doi.org/10.1093/bioinformatics/btz349.
Tracanna, Vittorio, Adam Ossowicki, Marloes L. C. Petrus, Sam Overduin, Barbara R. Terlouw, George Lund, Serina L. Robinson, et al. 2021. “Dissecting Disease-Suppressive Rhizosphere Microbiomes by Functional Amplicon Sequencing and 10× Metagenomics.” *MSystems* 6 (3): e01116-20. https://doi.org/10.1128/mSystems.01116-20.
Weber, T, C Rausch, P Lopez, I Hoof, V Gaykova, D H Huson, and W Wohlleben. 2009. “CLUSEAN: A Computer-Based Framework for the Automated Analysis of Bacterial Secondary Metabolite Biosynthetic Gene Clusters.” *Journal of Biotechnology* 140 (1–2): 13–17. https://doi.org/10.1016/j.jbiotec.2009.01.007.
Ziemert, Nadine, Sheila Podell, Kevin Penn, Jonathan H Badger, Eric Allen, and Paul R Jensen. 2012. “The Natural Product Domain Seeker NaPDoS: A Phylogeny Based Bioinformatic Tool to Classify Secondary Metabolite Gene Diversity.” *PLoS ONE* 7 (3): 1–9. https://doi.org/10.1371/journal.pone.0034064.

Chapter 4: MBEZ: Easy biosynthetic potential exploration metagenome mining pipeline.

Status: Manuscript (*In preparation*)

Citation: Shrikant Mantri, Nadine Ziemert *(2021): MBEZ: Easy biosynthetic potential exploration metagenome mining pipeline.*

Own contribution: Conceived research (with co-authors)

Implemented the pipeline and scripts

Analysed and interpreted the data

Wrote the manuscript (with co-authors)

Abstract

Motivation: With the increasing threat of antibiotic resistant pathogens, reemerging infectious diseases and high cancer rates, there is an urgent need for new therapeutics. The majority of drugs has been, and continues to be, developed from chemical scaffolds produced by living organisms, so called natural products. A large portion of these natural products is produced as secondary metabolites by microbes. Next generation sequencing methods and the enormous amount of available DNA data has shifted drug discovery efforts from traditional bioactivity guided screening methods towards genome-based approaches. Genome mining, heterologous expression, and genetic engineering offer the unique opportunity to discover the huge untapped potential hidden in environmental data. Shotgun metagenomic DNA sequencing and meta-barcoding approaches have revealed the expansive biodiversity of bacteria and their secondary metabolites that have been missed by traditional culture-based drug discovery methods. However, the complex nature of metagenomic data and the highly repetitive structure of natural product biosynthetic pathways makes the analysis challenging.

Results: MBEZ contains a collection of easy to use pipelines for microbial community profiling, biosynthetic domain abundance and diversity profiling, and biosynthesis potential exploration. It allows easy screening of the shotgun and amplicon metagenomic data for known and novel natural products. This pipeline enables natural product chemists, microbiologists and microbial ecologists to mine their metagenomic data fast, efficiently and without a deeper knowledge about natural product biosynthesis or bioinformatic analyses.

Availability and implementation: https://github.com/thinkgenome/MBEZ

4.1 Introduction

Most of the currently used drugs to fight infections are secondary metabolites (SM) produced by microbial species (Patridge et al. 2016). Biosynthesis of these SMs involves assembly line-like multi step pathways (Helfrich and Piel 2016; Walsh 2016; Ray and Moore 2016). Polyketide, non ribosomal peptides (NRP), ribosomally synthesized and post-translationally modified peptides (RiPPs), terpenes are the biosynthetic classes in which these natural products (NPs) are generally classified. Genes of enzymes and proteins involved in biosynthesis of these NPs are found in clusters in the genomes of the respective producer organism. Some of the enzymes are composed of multiple domains that function in tandem i.e ketosynthase (KS), adenylation (A), condensation (C) domains.

As the sequencing costs have dropped and next generation sequencing (NGS) throughput has tremendously improved, huge amounts of metagenomic data is currently publicly available from databases (Chevrette et al. 2021). This data can be mined to discover novel biosynthetic domains, genes and clusters. Currently few tools or pipelines are available to explore these metagenomic datasets to explore the biosynthetic diversity and potential. We are presenting here MBEZ, a collection of pipelines that we have developed to help in exploring the metagenomic datasets for facilitating the discovery of biosynthesis genes, domains .

4.2 Material and Methods

4.2.1 Microbial community diversity exploration pipeline:

The inputs to this pipeline can be 16S rRNA gene amplicons or shotgun seq datasets (Figure 1). For 16S rRNA amplicons the pipeline uses QIIME2 tool and accomplishes multi-step analysis involving raw data quality control, denoising, amplicon sequence variants computation, taxonomic annotations, correlation analysis, rarefaction analysis (Bolyen et al. 2019). Bash script and Jupyter notebook of these pipelines is also made available for the users to run the analysis with custom threshold parameters. For shotgun seq datasets the

pipeline uses Diamond to accelerate the annotation of reads against non redundant protein database of NCBI followed by taxonomic classification using NCBI taxonomy (Buchfink, Xie, and Huson 2014).

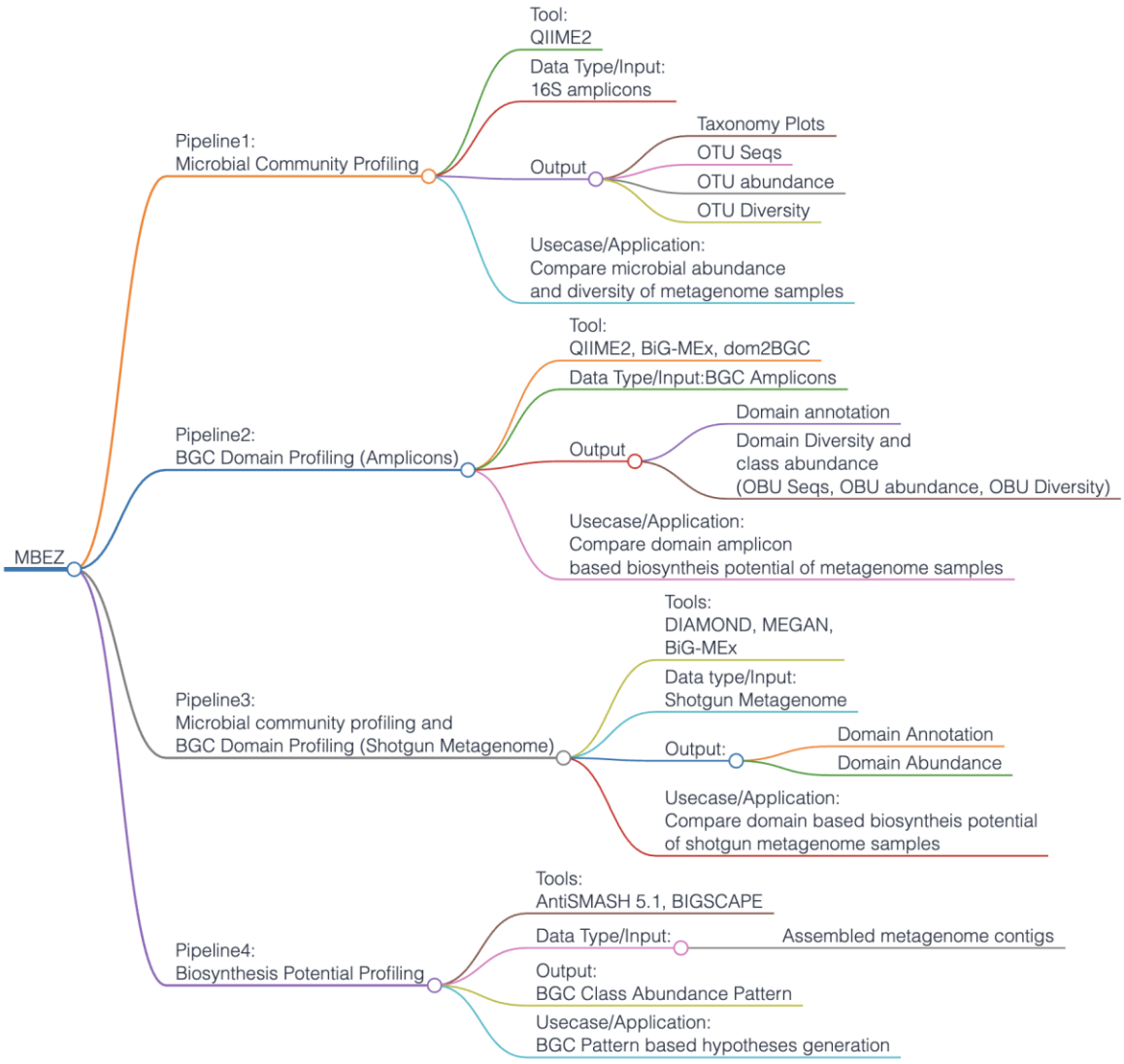


Figure 1: Markmap showing the details of MBEZ pipelines. Integrated tools, requisite inputs, resulting outputs, and Use Case/Applications of all the implemented pipelines is depicted. Markmap developed using https://markmap.js.org/

4.2.2 BGC domain diversity exploration pipeline:

The inputs to this pipeline can be KS, A domain amplicons or shotgun seq datasets. For BGC amplicon analysis the pipeline can be run using QIIME2, dom2BGC and BiG-MEx (Pereira 2020; Pereira-Flores et al. 2021; Tracanna et al. 2021; Bolyen et al. 2019). For domain diversity analysis using shotgun seq dataset, the pipeline can be used to profile +100 domains using BiG-MEx.

4.2.3 Biosynthesis potential exploration pipeline:

The inputs to this pipeline should be the assembled metagenomic contigs. BGC annotation is performed separately for each sample using antiSMASH. Sequence similarity network of the predicted clusters is performed using BiGSCAPE to quantify frequency of occurrence for each BGC class sample (Navarro-Muñoz et al. 2020). Diversity results are displayed as boxplots for each BGC class.

4.2.4 Implementation

MBEZ pipelines are written as bash scripts and can also be run in stepwise manner using the available jupyter notebook for each pipeline. Conda, Python and Docker availability is a prerequisite for running MBEZ pipelines. QIIME2, BiG-MEx, dom2BGC, MEGAN, antiSMASH, BiGSCAPE, DIAMOND, and HMMER have been integrated into different pipelines of MBEZ (Huson et al. 2016). Detailed manual and help documentation is available in the GitHub repo. For each pipeline, a demo dataset is made available for the ease of testing and interpretation. Existing pipelines can also be customised using the bash scripts and jupyter notebooks.

4.3 Conclusion

MBEZ fills the gap that was previously there due to unavailability of easy to use metagenome mining pipeline for exploring natural products diversity. For advanced users,

the bash scripts and Jupyter notebooks will be helpful in running the pipeline with custom parameters. Overall, MBEZ will facilitate and accelerate the metagenome mining analysis, explore natural products domains, BGC diversity and assess biosynthesis potential.

4.4 References

Bolyen, Evan, Jai Ram Rideout, Matthew R. Dillon, Nicholas A. Bokulich, Christian C. Abnet, Gabriel A. Al-Ghalith, Harriet Alexander, et al. 2019. “Reproducible, Interactive, Scalable and Extensible Microbiome Data Science Using QIIME 2.” *Nature Biotechnology* 37 (8): 852–57. https://doi.org/10.1038/s41587-019-0209-9.

Buchfink, Benjamin, Chao Xie, and Daniel H Huson. 2014. “Fast and Sensitive Protein Alignment Using DIAMOND.” *Nature Methods* 12 (1): 59–60. https://doi.org/10.1038/nmeth.3176.

Chevrette, Marc G., Athina Gavrilidou, Shrikant Mantri, Nelly Selem-Mojica, Nadine Ziemert, and Francisco Barona-Gómez. 2021. “The Confluence of Big Data and Evolutionary Genome Mining for the Discovery of Natural Products.” *Natural Product Reports*, August. https://doi.org/10.1039/D1NP00013F.

Helfrich, Eric J N, and Jörn Piel. 2016. “Biosynthesis of Polyketides by Trans-AT Polyketide Synthases.” *Natural Product Reports* 33 (2): 231–316. https://doi.org/10.1039/c5np00125k.

Huson, Daniel H, Sina Beier, Isabell Flade, Anna Górska, and Mohamed El-hadidi. 2016. “MEGAN Community Edition - Interactive Exploration and Analysis of Large-Scale Microbiome Sequencing Data,” 1–12. https://doi.org/10.1371/journal.pcbi.1004957.

Navarro-Muñoz, Jorge C, Nelly Selem-Mojica, Michael W Mullowney, Satria A Kautsar, James H Tryon, Elizabeth I Parkinson, Emmanuel L C De Los Santos, et al. 2020. “A Computational Framework to Explore Large-Scale Biosynthetic Diversity.” *Nature Chemical Biology* 16 (1): 60–68. https://doi.org/10.1038/s41589-019-0400-9.

Patridge, Eric, Peter Gareiss, Michael S. Kinch, and Denton Hoyer. 2016. “An Analysis of FDA-Approved Drugs: Natural Products and Their Derivatives.” *Drug Discovery Today* 21 (2): 204–7. https://doi.org/10.1016/j.drudis.2015.01.009.

Pereira, Emiliano. 2020. “Improvements in Natural Product Biosynthetic Gene Clusters Research and Functional Trait-Based Approaches in Metagenomics.” Jacobs University Bremen. http://nbn-resolving.org/urn:nbn:de:gbv:579-opus-1008965.

Pereira-Flores, Emiliano, Marnix Medema, Pier Luigi Buttigieg, Peter Meinicke, Frank Oliver Glöckner, and Antonio Fernández-Guerra. 2021. “Mining Metagenomes for Natural Product Biosynthetic Gene Clusters: Unlocking New Potential with Ultrafast Techniques.” https://doi.org/10.1101/2021.01.20.427441.

Ray, Lauren, and Bradley S. Moore. 2016. “Recent Advances in the Biosynthesis of Unusual Polyketide Synthase Substrates.” *Natural Product Reports* 33 (2): 150–61. https://doi.org/10.1039/C5NP00112A.

Tracanna, Vittorio, Adam Ossowicki, Marloes L. C. Petrus, Sam Overduin, Barbara R. Terlouw, George Lund, Serina L. Robinson, et al. 2021. “Dissecting Disease-Suppressive Rhizosphere Microbiomes by Functional Amplicon Sequencing and 10× Metagenomics.” *MSystems* 6 (3): e01116-20. https://doi.org/10.1128/mSystems.01116-20.

Walsh, Christopher T. 2016. “Insights into the Chemical Logic and Enzymatic Machinery of NRPS Assembly Lines.” *Natural Product Reports* 33 (2): 127–35. https://doi.org/10.1039/C5NP00035A.

Chapter 5: Metagenomic sequencing of multiple soil horizons and sites in close vicinity revealed novel secondary metabolite diversity

Status: Manuscript (*Under review*)

Citation: *Shrikant Mantri, Timo Negri, Helena Sales-Ortells, Angel Angelov, Silke Peter, Harald Neidhardt, Yvonne Oelmann, Nadine Ziemert, (2021), Metagenomic sequencing of multiple soil horizons and sites in close vicinity revealed novel secondary metabolite diversity.*

Own contribution: Conceived research (with co-authors)

Implemented the pipeline and scripts

Analysed and interpreted the data (with co-authors)

Wrote the manuscript (with co-authors)

Abstract: Discovery of novel antibiotics is crucial for combating rapidly spreading antimicrobial resistance and new infectious diseases. Most of the clinically used antibiotics are natural products, secondary metabolites produced by soil microbes that can be cultured in the lab. Rediscovery of these secondary metabolites during discovery expeditions costs both time and resources. Metagenomics approaches can overcome this challenge by capturing both culturable and unculturable hidden microbial diversity. To be effective, such an approach should address questions like: Which sequencing method is better at capturing the microbial diversity and biosynthesis potential? What part of soil should be sampled? Can patterns and correlations from such big data explorations guide future novel natural products discovery surveys? Here we address these questions by a paired amplicon and shotgun metagenomic sequencing survey of samples from soil horizons of multiple forest sites very close to each other. Metagenome mining identified numerous novel biosynthetic gene clusters (BGC), and enzymatic domain sequences. Hybrid assembly of both long reads and short reads improved the metagenomic assembly and resulted in better BGC annotations. A higher percentage of novel domains was recovered from shotgun metagenome datasets than amplicon datasets. Overall, in addition to revealing the biosynthetic potential of soil microbes, our results suggest the importance of sampling not only different soils but also their horizons to capture microbial and biosynthetic diversity and highlight the merits of metagenome sequencing methods.

Importance: This study helped uncover the biosynthesis potential of forest soils via exploration of shotgun metagenome and amplicon sequencing methods and showed that both methods are needed to expose the full microbial diversity in soil. Based on our metagenome mining results, we suggest revising the historical strategy of sampling soils from far-flung places as we found a significant amount of novel and diverse BGCs and domains even from different soils that are very close to each other. Furthermore, sampling of different soil horizons can reveal the additional diversity that remains often hidden and is mainly caused by differences in environmental key parameters such as soil pH and nutrient contents. This paired metagenomic survey identified diversity patterns and correlations, a step towards developing a rational approach for future natural products discovery surveys.

5.1 Introduction

One of the major driving forces of the medical revolution in the twentieth century was the discovery of antibiotics, which are often derived from secondary metabolites produced by microorganisms (Davies and Davies, 2010; Wohlleben et al., 2016). These natural products can be categorized based on their biosynthesis pathways. Major biosynthetic classes are polyketides (PKS), non-ribosomal peptides (NRPS), ribosomally synthesized and post translationally modified peptides (RiPPs), terpenes and saccharides. In bacteria the genes that encode these biosynthetic pathways are clustered together in the genome, popularly termed as biosynthetic gene clusters (BGC). The genes in some of these BGCs encode modular domains and enzymes that function in an assembly line-like fashion to produce complex biomolecules. Ketosynthase (KS) and Adenylation (A) domains, which have been the focus of this study, are involved in the biosynthesis of PKS and NRPS classes of secondary

metabolites in bacteria. Studying the gene sequence diversity of these domains aids in predicting the chemical structures encoded by BGCs that contain such domains(Ziemert et al., 2012). Based on the understanding of the biosynthetic chemical logic of these natural products, novel strategies have been developed not only to chemically synthesise analogous or derivative molecules, but also to accelerate their discovery via genome and metagenome mining methods (Chu et al., 2020; Sugimoto et al., 2019; Zhang et al., 2017).

Many natural products have been discovered as well as studied and a collection of more than 400,000 of such biomolecules is freely available on publicly accessible repositories (Mouncey et al., 2019; Sorokina and Steinbeck, 2020). These biomolecules show diverse pharmacological functions such as antibacterial, antifungal, anticancer, immuno-modulatory and antiviral activity (Boufridi and Quinn, 2018). Less characterized is their ecological function. Multiple hypotheses and theories have been proposed about the role of secondary metabolites in the lives of the microbes that produce them. Some of these bioactive molecules are deployed in the arms race against other species in a particular microbial community; others might serve as intra-, inter-species, or even inter-kingdom, signalling and communication agents or regulate developmental processes (Tyc et al., 2017).

Most of the antibiotics discovered so far have been isolated from soil microbes, specifically those that could be cultured in the lab. As research groups around the world started to extensively survey random soils to identify novel antibiotics, they experienced the challenge of rediscovering previously characterized antibiotics (Baltz, 2008; Silver, 2011). The use of 16S rRNA gene based metagenome profiling unveiled the extent of the hidden microbial diversity as only about 1-2 % of all the species present in a particular soil sample could be cultured in the lab (Bodor et al.,

2020; Yarza et al., 2014). The subsequent revolution in next generation sequencing technologies made it possible not only to easily sequence the isolated species genomes, but also to capture the unculturable microbial diversity using metagenome sequencing approaches (Bahram et al., 2018; Delgado-Baquerizo et al., 2018; Handelsman, 2004). More recently, long read sequencing technologies, namely Oxford Nanopore and PacBio sequencing, have enabled significant improvements in assembly of shotgun metagenomes into long contigs. These are a prerequisite for the identification of the often very large biosynthetic clusters encoding secondary metabolites. One study even reported comparable results by only using MinION nanopore sequencing for recovering multiple complete bacterial genomes from complex microbial communities within a bioreactor (Arumugam et al., 2019).

The metagenomic soil surveys reported so far aimed at identifying microbial community diversity and patterns, and covered areas spanning from urban green spaces, grassland meadows, up to continent-wide scale soil analyses (Bahram et al., 2018; Crits-Christoph et al., 2018; Delgado-Baquerizo et al., 2018; Thompson et al., 2017; Wang et al., 2018). Few of them also aimed at identifying the biosynthetic domain composition of bacterial natural products but using exclusively amplicon sequencing approaches (Borsetto et al., 2019; Crits-Christoph et al., 2020; Elfeki et al., 2018; Lemetre et al., 2017; Reddy et al., 2012; Sharrar et al., 2019). Those studies were able to identify diversity patterns and correlations between natural product diversity and environmental features, thus improving our understanding of ecological and evolutionary pressures that drive the distribution of natural products across different geographical scales. However, little is known about how sampling strategies can be optimized for improved discovery of diverse natural products. Those studies that addressed these issues identified distribution patterns of PKS and

NRPS based on biomes, types and characteristics of the soil (composition, pH, temperature, etc.), as well as geographic distance (Charlop-Powers et al., 2016, 2015; Morlon et al., 2015; Reddy et al., 2012). However, they analyzed the soil in either similar or different ecosystems on a global scale. Moreover, while Morlon identified plant community composition as the main driver of natural product diversity, Charlop-Powers showed that geographic proximity was more important. In fact, soil types and associated soil properties may largely vary even at a local scale (i.e. decimeters) due to differences such as in the geological parental material, (micro-)relief, or plant community. Also, soil properties may considerably vary vertically, as different soil horizons may largely differ in physico-chemical properties (e.g. pH, available nutrients, redox conditions, water content) due to pedogenetic processes (FAO and IUSS, 2015). As a consequence of such highly diverse micro-environments, microbial diversity was shown to generally vary by soil depth being accompanied by decreasing abundances (Eilers et al., 2012; Fierer et al., 2003; Will et al., 2010). Therefore, we speculated that analysis of different soil samples from different ecosystems in the same geographical area could provide more insight into the fine scale distribution of secondary metabolites and how sampling strategies can affect natural product discovery.

Here we report results from our metagenomics study of different horizons of soil sampled from various sites within the Schönbuch Forest, a nature reserve area in Southern Germany, using both Nanopore and Illumina NGS sequencing technology. Major objectives of this pilot project were a) to compare the natural product domains and biosynthesis cluster diversity of different soils and their horizons; b) to recover longer metagenome assembled contigs via hybrid assembly of short and long reads facilitating discovery of biosynthesis gene clusters; c) to compare the amplicon

sequencing and shotgun metagenome sequencing methods; d) to assess correlation between microbial community diversity and physico-chemical properties of different soils. Our findings indicate that natural product diversity is high in different soils even in close proximity to each other, and that sampling the different soil horizons also makes a difference. Mining of metagenomic reads led to the detection of many known and novel domains involved in the biosynthesis of polyketide and non-ribosomal peptides. Hybrid assembly of short and long reads led to the identification of biosynthesis gene clusters that could have never been detected by short read sequencing alone.

5.2 Results

5.2.1 Amplicon-seq mining revealed major differences in bacterial diversity and their biosynthetic potential in the different soils and their horizon

In order to understand how the diversity of secondary metabolites changes with the type of soil and its horizons, we identified a study area located in the Schönbuch forest nature reserve, which is part of the South German Scarplands region (Einsele, 1986). Soils in this area are characterized by a high diversity due to a variety of geological material and landscape morphology. Samples were collected from three soil pits representing three characteristic but highly diverse soil types, named Cambisol, Podzol, and Stagnosol. All soil pits are located near to each other in a straight line within some 150m from each other (Fig. 1A). Soil analysis have shown that these soils are heavily layered with very different parameters in each layer, studies have shown that the bacterial diversity is changing greatly but no one knows about the secondary metabolite diversity (Eilers et al., 2012). In order to get an overview of the actual domain diversity of the three different soils, all three soils and

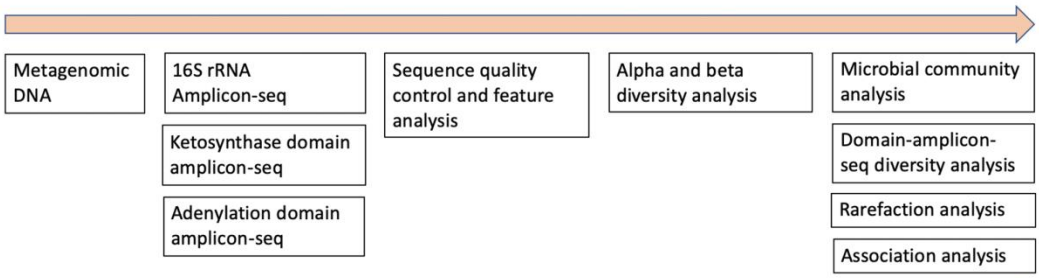
their respective horizons were sampled, metagenomic DNA was isolated and subsequently sequenced using Illumina amplicon as well as shotgun sequencing methods. Additionally, Oxford nanopore sequencing was used to sequence one sample. Sample details, study outline, sequencing yields and analysis workflow are summarised in Figure 1 and Table S1a-c (see Table S1a-c at https://doi.org/10.5281/zenodo.5195507).

A

B

Sampling Site Name		Cambisol	Podzol	Stagnosol
Sampling Year	2016	2019	2019	2019
Soil Horizons	-	O	O	O
	Ah	Ah	Ah	Ah
	-	Bw1	-	-
Shotgun Seq (Illumina)	✓	✓	✓	✓
Shotgun Seq (Nanopore)	✓	-	-	-
Amplicon Seq (Illumina)	-	✓	✓	✓

C



D

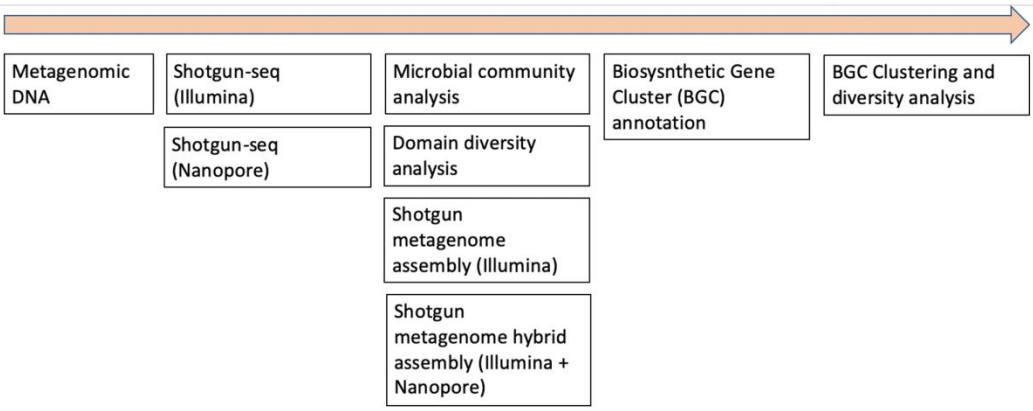


Figure 1: Geographic location, study outline and analysis workflow. A) Sampling site geographic location map of Tuebingen, Germany (Map Data ©2021 Google) . Multiple soil horizons from three sites were sampled. Photo depicting the 3 horizons of Cambisol soil. B) Sample and sequencing information, See Table S10a (see Table S10a at

https://doi.org/10.5281/zenodo.5195507) for details about soil names and profile description. C) Amplicon sequencing and analysis workflow, D) Shotgun sequencing and analysis workflow}

Amplicon analysis of specific genes of interest has proven to be an efficient and cost-effective strategy for metagenomic analysis. Amplifying specific genes of interest allows high coverage of these genes without extensive sequencing. Therefore, in a first approach, we explored the microbial diversity and natural products domain diversity by sequencing the 16S rRNA gene, A domain and KS domain amplicons (biosynthetic diversity indicators) using an Illumina paired end sequencing approach.

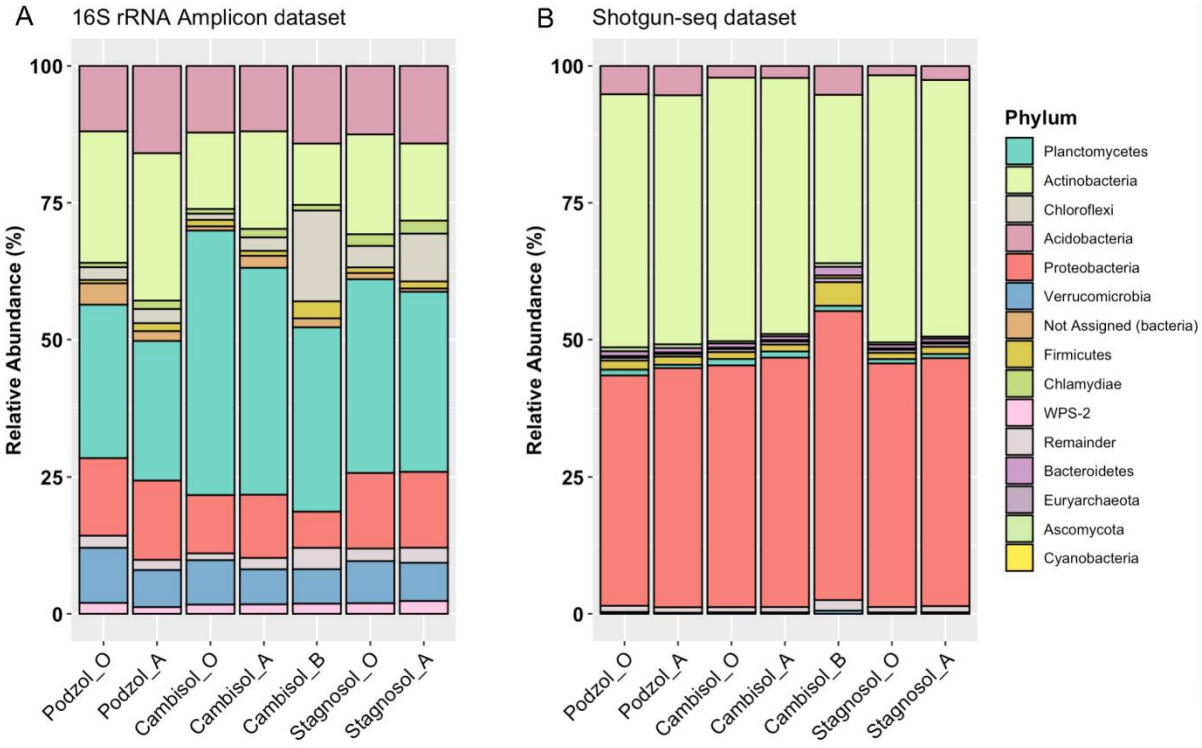
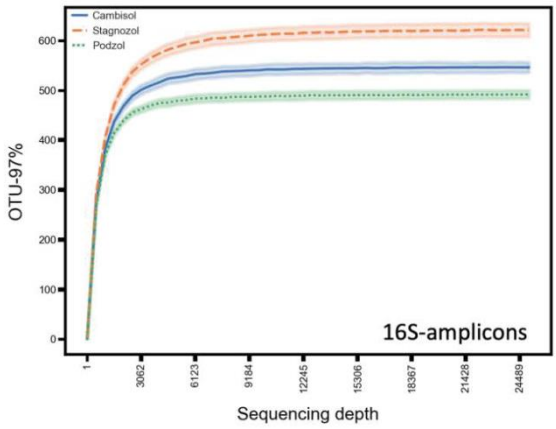


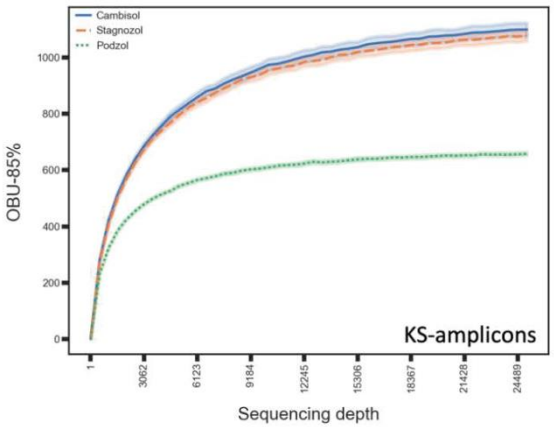
Figure 2: Microbial composition across 3 sampling sites (Podzol, Stagnosol and Cambisol) and 3 soil horizons (O, A and B). A) Bar plot showing taxonomic profile for 16S rRNA amplicon dataset. B) Bar plot showing taxonomic profile for Shotgun-seq dataset. Taxonomic profile at phylogenetic rank of phyla is shown. Top ten phyla are depicted in different colours and remaining phyla are grouped as category of

"Remainder" depicted in grey colour. Same colours for each phyla are used for side-by-side visualisation. SILVA rRNA database was used for classifying amplicons and maxikraken2 database was used for classifying shotgun-seq reads.

Taxonomic annotation of the Illumina based 16S rRNA gene Amplicon Sequence Variants (ASVs) using the Silva taxonomic database showed that all soil samples have a very diverse bacterial composition as expected (Fig.2 and see Table S4a at https://doi.org/10.5281/zenodo.5195507). Comparing the taxonomic composition of all samples revealed that not only the three different soils but also their various horizons differed in their bacterial composition, even on the relatively wide phylum level (Fig. 2). Planctomycetes was the most abundant phylum in all three soil samples and all horizons. The Chloroflexi phylum was most abundant in the Cambisol B horizon with a relative frequency double than that of other soils. By comparing the number of ASVs and clustering them to OTUs (operational taxonomic units), we noticed that the highest number of OTUs was present in the A horizon of Cambisol, which represents the second layer below the surface (see Table S6a at https://doi.org/10.5281/zenodo.5195507). In contrast, in Podzol and Stagnosol the number of OTUs in the O horizons was higher as compared to the respective A horizons. The lowest number of OTUs was found in the Cambisol B horizon, indicating that Cambisol contained the most but also the least bacterial diversity of the three different soils depending on the horizon. In order to classify A domain- and KS domain amplicons into groups that represent distinct chemical classes and biosynthetic gene clusters (BGCs), we clustered these amplicons into operational biosynthetic units (OBU) (as previously described (Elfeki et al., 2018)). Rarefaction curve analysis for both classes of OBUs showed that the curves are still ascending, indicating that the full biosynthetic diversity hasn´t been captured yet, in contrast to

the taxonomic diversity represented by the 16s rRNA amplicons (Fig 3). Comparing the domain diversity of the different soils and their horizons showed that unique KS and A domains (ASV clustered at 97% similarity, see Fig 4) were at a maximum in the Cambisol B horizon, the soil with the lowest number of OTUs (see Table S6a at https://doi.org/10.5281/zenodo.5195507). In order to uncover any possible correlation between biosynthetic diversity and taxonomic diversity, we compared various alpha diversity indices of KS and A domains with the 16S diversity. The OTU alpha diversity, Faith PD, Shannon and Evenness showed high correlation across 16S and A domain amplicons (see Table S6a at https://doi.org/10.5281/zenodo.5195507), whereas there was no clear correlation for KS domains, and even negative correlation between evenness of 16S and KS domains was observed.





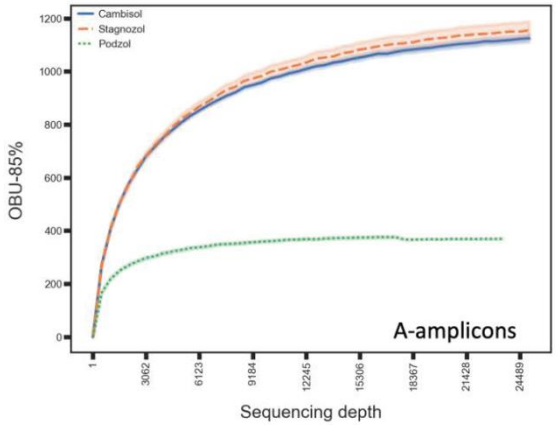


Figure 3: Rarefaction curves for 16S rRNA gene amplicons, A domain amplicons and KS domain amplicons. The bold curve shows mean value of OTU/OBU at a particular sequencing depth for all horizons of a particular site. The faint colour area around each curve shows the confidence interval of 67 %.

In order to disclose any overlap between the different soils, we compared 16S as well as KS and A domain amplicons in the different samples using Upset plots (Fig. 4). This analysis revealed that, while there was an overlap of 42 16S amplicons across all the 7 samples, no such degree of sequence similarity was observed for KS and A domains. ASVs of these domains were only conserved between samples of different horizons of the same site.

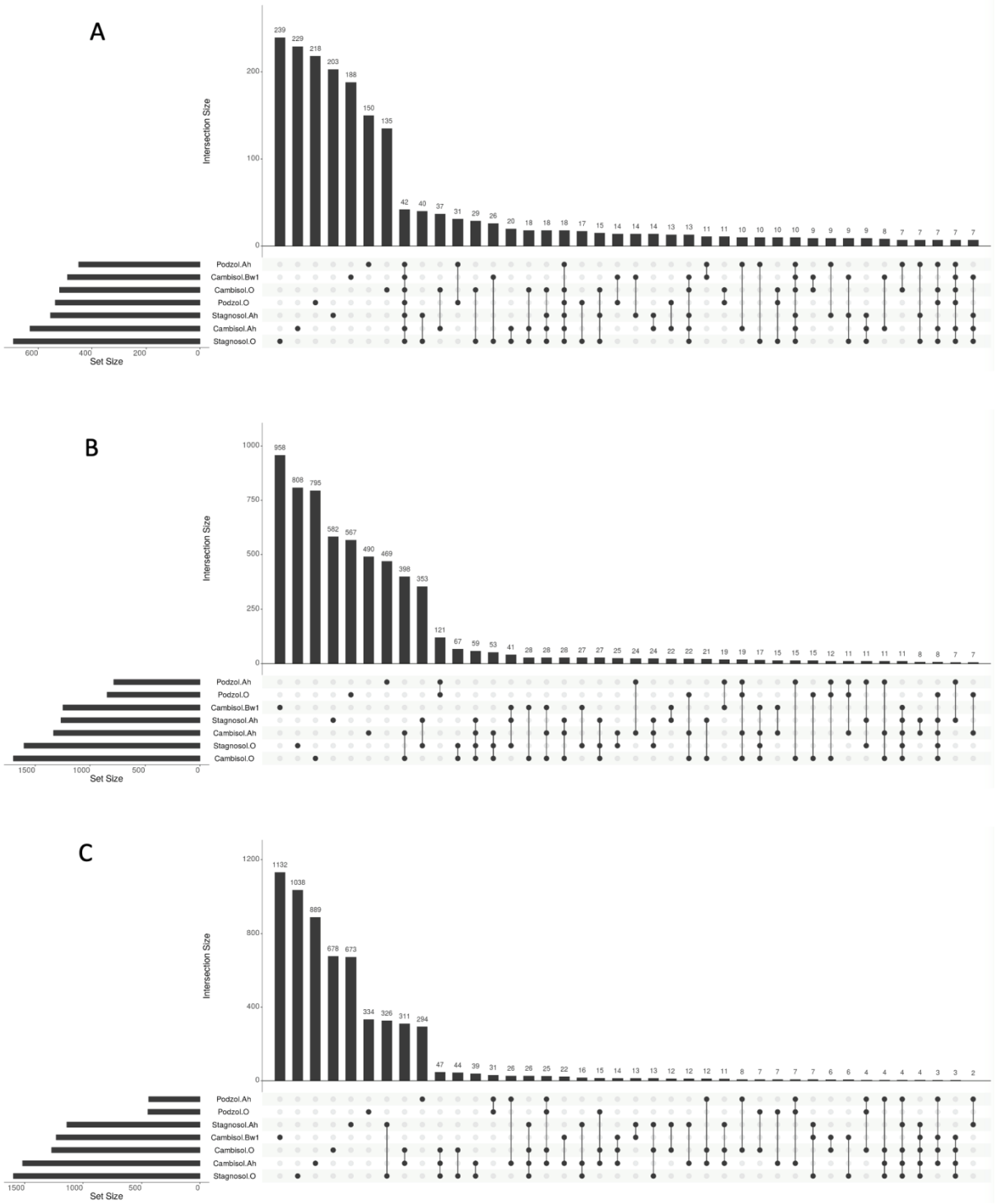

A
Intersection Size
239 229 218 203 188 150 135 42 40 37 31 29 26 20 18 18 18 17 15 14 14 14 13 13 11 11 10 10 10 10 9 9 9 9 8 7 7 7 7 7
Podzol.Ah
Cambisol.Bw1
Cambisol.O
Podzol.O
Stagnosol.Ah
Cambisol.Ah
Stagnosol.O
600 400 200 0
Set Size
B
Intersection Size
958 808 795 582 567 490 469 398 353 121 67 59 53 41 28 28 28 27 27 25 24 24 22 22 21 19 19 17 15 15 15 12 11 11 11 11 8 8 7 7
Podzol.Ah
Podzol.O
Cambisol.Bw1
Stagnosol.Ah
Cambisol.Ah
Stagnosol.O
Cambisol.O
1500 1000 500 0
Set Size
C
Intersection Size
1132 1038 889 678 673 334 326 311 294 47 44 39 31 26 26 25 22 16 15 14 13 13 12 12 12 11 8 7 7 7 7 6 6 4 4 4 4 3 3 2
Podzol.Ah
Podzol.O
Stagnosol.Ah
Cambisol.Bw1
Cambisol.O
Cambisol.Ah
Stagnosol.O
1500 1000 500 0
Set Size

Figure 4

Figure 4: Intersections and distribution of A) 16S, B) KS domain and C) A domain (ASVs clustered at 97% similarity). The bar plot (top) in each panel shows intersection size (the number of ASVs) in the combinatorial sets of relevant samples. The matrix below the bar plot indicates sets of samples that are represented by each bar.

To see if the differences in taxonomic diversity and biosynthetic potential of the different soil samples were correlated with the unique soil physico-chemical parameters, we calculated alpha diversity (16S and Domains) correlations with the soil parameters (see Table S6b at https://doi.org/10.5281/zenodo.5195507). Although, we were able to detect some correlations between biosynthetic potential - pH showed a close correlation between KS domain alpha diversity measures (shannon r=0.75, p=0.05 and evenness, r=0.78,p=0.03), we think that more data are needed in order to interpret these results properly.

5.2.2 Shotgun metagenome mining further uncovered microbial diversity and identified novel BGCs

Amplicon sequencing based studies of the metagenome diversity are an economical approach, however, its limitations became evident when we performed shotgun metagenome sequencing using Illumina short reads and Nanopore long reads for the same samples and compared both methods.

Table 1: Taxonomic annotation summary (Tool: kraken2, database: maxikraken2) of shotgun-seq Illumina metagenomes

Name	#raw paired end reads	Classified reads %	Unclassified reads %	Microbial reads %	Bacterial reads %	Viral reads %
Podzol-O	113,350,452	43.90	56.10	43.80	42.90	0.01
Podzol-A	86,440,710	45.80	54.20	45.80	44.90	0.01
Cambisol-O	82,298,268	51.60	48.40	51.60	50.70	0.01
Cambisol-A	71,637,596	50.30	49.70	50.20	49.40	0.01
Cambisol-B	75,654,703	35.30	64.70	35.30	34.40	0.01
Stagnosol-O	64,281,069	52.50	47.50	52.50	51.50	0.01
Stagnosol-A	53,255,349	49.90	50.10	49.90	49	0.01

We used the Kraken2 algorithm in order to annotate the shotgun metagenomes, which led to an average of 47.04 percent of classified reads and an average of 52.95 percent unclassified reads (Table 1). Interestingly, Proteobacteria and Actinobacteria were the top 2 annotated phyla amongst all the metagenomes (Figure 2), a result which differs greatly from the 16S rRNA gene amplicon annotations. Using the unassembled metagenomes, we also used the BiG-MEx software for annotations of BGC domains and the diversity analysis. BiG-MEx was able to annotate 150 BGC

domains (see Table S5b at https://doi.org/10.5281/zenodo.5195507), most of them as A-domains. By performing comparative analysis of KS and A domains captured via amplicon and shotgun metagenome sequencing, we found that more than 90 percent of domains detected in shotgun metagenomes could not be detected using amplicon sequencing. More precisely, sequence similarity analysis between domains identified via amplicon sequencing and shotgun metagenome sequencing revealed the presence of domains unique to each of the methods. 638 KS Amplicon-seq amplicons did not show similarity to any of the KS shotgun-seq OBUs, whereas 1571 A-domain amplicon-seq amplicons did not show similarity to any of the 181,324 A-domain shotgun-seq OBUs (see Table S9 at https://doi.org/10.5281/zenodo.5195507). The alpha diversity comparisons between microbial community diversity and biosynthetic domain diversity showed a diverse pattern for each domain. We also found no concurrence of these diversity correlations between amplicon-seq and shotgun-seq datasets.

In a next step, we assembled the shotgun metagenome data to recover full biosynthetic gene cluster sequences and thus obtain more valuable information about the encoded compounds. The metaSPADEs based assembly of Illumina reads of all the metagenomic samples led to a total of more than 2 million contigs longer than 1kb. The total length of all the contigs exceeded 9 Giga bases, with the largest contig of about 3,5 Mega bases. The assembled contigs longer than 10kb were analyzed for the presence of BGCs using antiSMASH (version 5). A total of 1102 BGCs were identified. The detailed biosynthetic class wise breakup of the BGC annotation is provided in figure 5. Again, the highest number of BGCs was annotated as belonging to the class of NRPSs followed by 262 RiPPs (see Table S7a at https://doi.org/10.5281/zenodo.5195507). Podzol O horizon contained a maximum

number of 470 BGCs followed by Podzol A horizon with 315 BGCs (Figure 5). In contrast to the domain analysis, Podzol samples displayed the maximum number of clusters as compared to other sites. However, this might be due to the better assembly of Podzol samples as a result of the highest number of reads being generated from the O and A horizon of Podzol soil (see Table S1a at https://doi.org/10.5281/zenodo.5195507). BiG-SCAPE clustering of the dataset composed of Illumina only assembled contigs helped investigate the overlap of clusters across the soil. While most of the BGCs were unique to each sample, we found only a single Gene Cluster Family (GCF) containing BGCs from each of the seven samples. This GCF belongs to the class of terpenes.

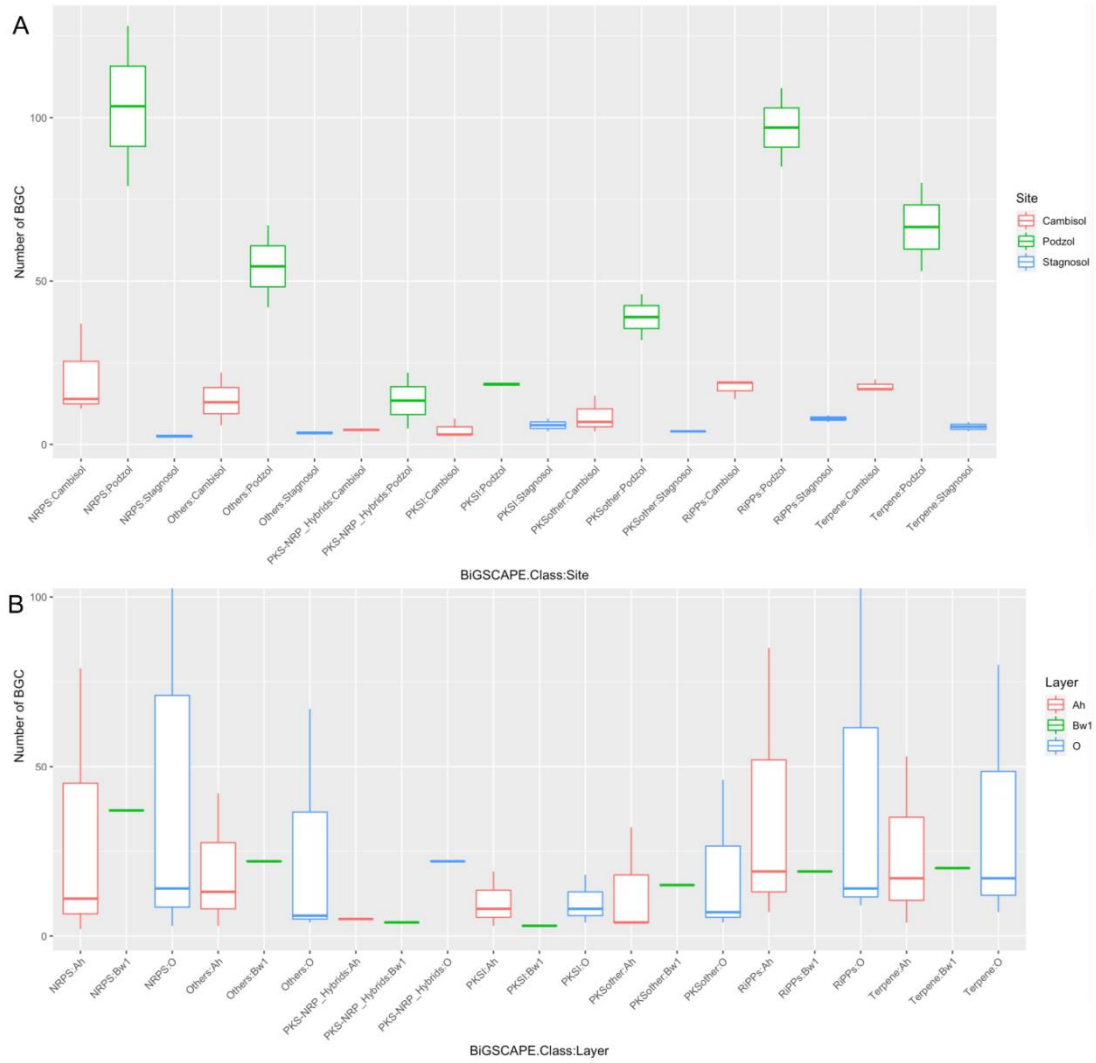


Figure 5: {Biosynthetic gene cluster abundance distribution. A) BGC abundance distribution across soil sampling sites.(grouped according to BiG-SCAPE class). B) BGC abundance distribution across soil horizons

Apart from antiSMASH based BGC discovery, we also explored the machine learning based method for novel cluster discovery and annotation. We found around 22194 putative BGCs in the metagenomic contigs using the DeepBGC tool. For 7295 of these BGCs the biosynthesis class could be predicted. Biological activity could be predicted in 17032 putative BGCs (see Table S8 at https://doi.org/10.5281/zenodo.5195507). While the number of the detected BGCs is several fold higher than that annotated via antiSMASH, it will be interesting to see the wet-lab validation of these clusters in future studies. Although absolute numbers of predicted BGCs differ between antiSMASH and DeepBGC, highest number of BGCs were predicted in Podzol samples by both these tools.

5.2.3 Comparative analysis highlights the advantage of long reads to capture biosynthetic potential.

The assembly statistics of the short-read shotgun data helped appreciate its advantages and limits. Subsequently, as we were interested in assessing how long reads Nanopore data would improve the recovery of BGCs, we performed a metaSPADES based hybrid assembly of Illumina and Nanopore reads of the Cambisol A metagenome. The hybrid assembly substantially enhanced the overall length of the contigs and the number of longer contigs. We found seven times more hybrid contigs of length greater than 50 kb as compared to the Illumina only contigs of same length. The largest hybrid contig was of 598,670 bases (see Table S3a and S3b at https://doi.org/10.5281/zenodo.5195507). AntiSMASH analysis resulted in the

annotation of 169 BGCs among the hybrid contigs longer than 10 Kb. This is more than double the number of BGCs that were found in Illumina only contigs. A total of 1026 BGCs were even annotated in the hybrid contigs with lengths greater than 1 kb. Comparison of metagenomic contig length (Illumina only versus hybrid data) revealed substantial improvements with the hybrid assembly approach (see Table S3a and S3b at https://doi.org/10.5281/zenodo.5195507. In several instances hybrid assembly enabled the extension of Illumina contigs containing BGCs, thus making it possible to determine whether resistance markers or regulator-encoding genes were present within the clusters. We found more than two fold more BGCs in hybrid contigs that were not on contig-borders as compared to illumina only contigs detected via antiSMASH annotation. We also performed BiG-SCAPE clustering of all BGCs from Illumina and hybrid metagenomes to identify BGCs that were detected in multiple samples. This analysis led to the identification of 1803 GCFs. 1625 GCFs contained only single members (see Table S7b at https://doi.org/10.5281/zenodo.5195507).

5.3 Discussion

Soil formation is a slow process: depending on climatic conditions; it might take several hundred years to form just a centimeter layer of soil. While most of the antibiotics discovered so far have been largely isolated from culturable microbes in random sampling of topsoils, the immense metabolic diversity of unculturable microbial dark matter in both, topsoils and deeper soil horizons, has remained largely hidden (Durand et al., 2019). As the depth of soil increases, the organic and inorganic chemical constituents and morphology of soil change drastically creating micro-environments that can accelerate the evolution of novel microbial species

(Wilpiszeski et al., 2019). To capture the biosynthetic novelty of all such microbes, those that were born due to serendipitous events and those that survived the so called microbial arms race, we decided to broaden the soil surveys not only to include soils from different sites but also to cover sampling of diverse soil horizons (Hao et al., 2021). Our study is also unique in that it used both, amplicon sequencing and shotgun metagenome sequencing of the same soil samples to determine the biosynthetic potential that a particular site and ecosystem hold, and to discover novel natural products domains and BGCs (Figure 1).

Although few species were ubiquitously present across all the sites and all the soil horizon layers, a significantly higher proportion of OTUs/species were seen to be unique to individual samples (Figure 4a). BGC domain diversity and distribution observed across all the samples indicate higher overlap within a particular sampling site than across sites (Figure 4b and 4c). Our survey of multiple soil horizons from multiple sites helped appreciate the presence of high vertical diversity (differences between O, A and B horizons of each soil type) emphasizing the importance of sampling not only different geographical sites but also the vertical diversity present in different soil horizons. This is in line with previous findings based on 16S rRNA analysis (Eilers et al., 2012). The reasons behind such a great diversity across sites could be attributed to the variable environmental conditions (Will et al., 2010). For example, Podzol is an extreme nutrient-poor, acidic and water-scarce environment where microbial decomposition of the tree litter is so much hampered that a thick organic litter layer sits on top of the topsoil (i.e. A-horizon); in the Stagnosol's A and B horizon, instead, the water dynamics can entirely fall dry during summer, changing the redox from reducing to oxic.

Drastic deviations in estimating microbial composition via both 16S amplicon-seq and shotgun-seq have been previously reported (Brumfield et al., 2020; Jovel et al., 2016). In our study, Planctomycetes emerged as the major phylum in the amplicon-seq analysis while Proteobacteria and Actinobacteria were the predominant phyla in the shotgun-seq analysis (Figure 2). This deviation could be attributed to primer and PCR bias of the 16S amplicon method (Brumfield et al., 2020; Jovel et al., 2016) and to the different bioinformatics workflows (Balvočiūtė and Huson, 2017). Also, the sequencing depth in studying the microbial composition via 16S amplicon sequencing appeared to be sufficient and saturating as per the rarefaction curves (Figure 3). Subsequent shotgun metagenome sequencing analysis of the same samples revealed that amplicon-based analysis underestimated the alpha diversity of the samples.

Although we hoped to find unique patterns of correlations between microbial community diversity and biosynthetic diversity, our results of both amplicon-seq and shotgun-seq datasets only revealed few correlations with few biosynthetic gene domains. We speculate that these patterns would become more evident as more optimised amplicon primers, capable of amplifying additional biosynthetic genes and their domains, would become available. In case of shotgun-seq datasets, higher depth of sequencing of the samples would not only help in recovering more full length BGCs but also help in revealing biosynthesis domain diversity patterns. Better software tools capable of handling such high volume of data would be required to mine the biosynthetic diversity patterns.

Assembly of shotgun-seq Illumina reads followed by antiSMASH annotation led to the discovery of 1102 BGCs. Proteobacteria, Acidobacteria and Actinobacteria were the major phyla to which many of these BGCs were taxonomically annotated.

Distribution patterns of BGC classes across the sampling sites and soil horizons, show that the Podzol site has the maximum BGCs (Figure 5). BGC abundance distribution was observed more in sampling site-wise comparison than soil layer-wise comparison. BGC clustering analysis also revealed how different the various samples and horizons are as only a single BGC was found to be present across all the samples (see Figure S1 at https://doi.org/10.5281/zenodo.5195507). Hybrid assembly of Illumina short reads with nanopore long reads led to the recovery of complete BGCs in some cases, enabling the identification of the regulatory genes and resistance genes in the vicinity of the identified BGCs. Such proximity analysis can be helpful in prioritizing the BGCs for e.g., the characterization of the encoded compounds in heterologous expression systems (Mungan et al., 2020). Machine learning based annotation of assembled contigs using DeepBGC led to identification of even more putative BGCs. For many of them, however, the biosynthesis class and activity could not be predicted, likely as a consequence of the low similarity between these novel BGCs and those used for DeepBGC training.

Amplicon sequencing and shotgun metagenome sequencing both are important when aiming for novel domain discovery as we observed unique domain sequences with each of the methods (see Table S9 at https://doi.org/10.5281/zenodo.5195507). For both KS and A domains, 90 percent more domain sequences were identified in shotgun datasets as compared to amplicon datasets, highlighting the immense biosynthesis potential that has yet to be discovered. As the costs of shotgun metagenomic sequencing are still prohibitive and make these methods accessible to only a few, our shotgun results will be useful to design domain sequence-based primers that are not biased to a particular genus and can be used for massive, amplicon-based diversity surveys.

Our study helped capture the snapshot of microbial diversity and metabolic novelty from the soils sampled on a single day. However, the limited number of samples, made it hard to draw meaningful biological conclusions from the observed correlations between the diversity of BGCs and soil physico-chemical parameters. Large-scale and more systematic sampling across changing weather or seasons will be necessary to capture the true dynamics and complete diversity. We were not able to recover metagenome assembled genomes (MAGs) due to sequencing volume limitation. Considering the massive diversity present in soil, hundreds if not thousands of gigabases would be required to reach a stage to claim complete coverage of all the species genome in a particular metagenome sample (Rodriguez-R et al., 2018). Reaching terabase scales (10^{12}) is not only a current economical bottleneck, but also it calls for better metagenome assembly algorithms that are both space and time efficient. Alternatively, novel methods that uses live-FISH (fluorescence in situ hybridization) combined with FACS (fluorescence-activated cell sorting) has been reported to be capable of isolating live bacteria solely based on their 16S rRNA gene sequence (Batani et al., 2019). In future, using such novel methods it will become possible to accelerate the BGC discovery from candidate or novel phyla present in densely rich soil samples.

Some of the BGCs discovered in this study are currently being explored for further heterologous expression and structure elucidation in our laboratory. All the data resources generated here have been shared in the public domain to facilitate further experiments and analysis by the natural products research community. It will be a herculean task to explore and map the complete chemical space that natural products cover on the entire earth. Our metagenomic data give a glimpse of the immense microbial and biosynthetic diversity that exists even in the next door soils.

5.4 Conclusion

Overall, this study helped uncover the biosynthesis potential of the Schönbuch forest soil by combining metagenome and amplicon sequencing. This paired strategy helped identify more novel BGC domains than it would have been possible with only either of the sequencing methods. Our analysis also confirmed the limitations of amplicon sequencing, which is extremely powerful in providing a glimpse of the microbial and biosynthetic diversity in soil samples, but this is biased to sequences that are abundant in the samples and to the chosen primers. We show that a shotgun metagenome approach is able to overcome these limitations and better as compared to the amplicon-based approach at capturing the microbial diversity. The additional use of Nanopore sequencing data for one of the soil samples allowed us to improve metagenome assembly and to recover novel BGCs. Nonetheless, long read sequencing remains too costly to be routinely used in soil surveys of microbial and BGCs diversity. Physico-chemical parameters that correlate with the domains or BGC diversity will help develop a rationale to guide such explorative surveys. In the future, sequencing terabases of metagenomes might become feasible and economical. At such sequencing depths we might then only be limited by heterologous expression and functional validation of novel natural products. Probably such a foreseeable future is just a decade away. Until then, the approaches and rationale developed here will help fuel the drug discovery pipeline to combat antimicrobial resistance.

5.5 Methods

5.5.1 Soil sampling, physico-chemical parameters characterization.

The sampled Schönbuch forest soils developed from Lower and Middle Triassic Keuper sequences, which locally comprise thin sequences of sandstones and evaporitic marlstones, as well as aeolian (loess), colluvial, and alluvial deposits (Einsele, 1986; Grathwohl et al., 2013) . The soils were described and classified according to the classification system of the Food and Agriculture Organization of the United Nations (Jahn et al., 2006) and IUSS Working Group WRB (FAO and IUSS, 2015). Differences concerning the geochemistry (i.e. pH and CaCO3 concentrations) of the geological soil parent material resulted in highly different soil types, which were explicitly taken into account in this study. The first soil pit, located at the top-slope of a south-exposed slope was classified as a Podzol, which has developed from a sandstone outcrop. The second soil was classified as a Cambisol, which has developed from sandstone mixed with aeolian deposits (loess). The third soil was a Stagnosol, which has formed from a clay-rich marl. See Table S10a for further details on the soil profiles (see Table S10a at https://doi.org/10.5281/zenodo.5195507). Sampling was carried out horizon-wise. Bulk samples were taken from the soil genetic horizons for geochemical analyses, comprising the mineral topsoil (A horizon) and mineral subsoil (B horizon). For simplification, the organic litter layers (Oi and Oe) that cover the mineral soil horizons were combined as one bulk sample per site. Carbon and Nitrogen measurements: Dried (40°C) litter and fine soil (<2mm) samples were homogenized with a planetary ball mill (Pulverisette 5, Fritsch Idar-Oberstein, Germany). Total C and N concentrations were measured by a CNS elemental analyser (Vario EL III, Elementar

Analyse systeme GmbH, Langenselbold, Germany). For details regarding detection limits and quality controls, see Table S10b (see Table S10b at https://doi.org/10.5281/zenodo.5195507).

X-ray fluorescence: To determine the major element concentrations in fine mineral soil samples of A and B horizons, glass beads of a homogenized mixture of 1.5 g dried and powdered sample material and 7.5 g lithium tetraborate were fused at 1050 °C for 30 min. On Bruker AXS Pioneer S4, glass beads were analyzed by wavelength dispersive X-ray fluorescence (XRF).

ICP-OES: To determine concentrations of major and trace elements in O horizon soils, litter samples were dissolved by an acid pressure digestion system (Loftfield PDS-6, Loftfield Analytical Solutions, Neu Eichenberg, Germany). Therefore, homogenized sample material (target weight: 0.05g) was transferred into Teflon pressure beakers before adding 4mL HNO3 conc. (65%, Merck KGaA, p.a. ≥ 98%). After heating for seven hours at 180°C, digestion solutions were filtered (MN 619 G¼ Ø185mm, Macherey-Nagel, Düren, Germany) and diluted with Millipore water (Synergy UV ultrapure, Millipore) to a final volume of 50 mL. The digests were finally analysed by an inductively coupled plasma optical emission spectrometer (ICP-OES Optima 5300 DV, PerkinElmer, Wellesley USA) according to EN ISO 11885. To check for accuracy and precision of the digestions, the two certified reference materials BCR-129 (hay powder) and BCR-141 (plankton) were used. Based on the measured average concentration values and the target values, recovery rates were calculated for each element (see Table S10c at https://doi.org/10.5281/zenodo.5195507). Despite a good reproducibility (RSD of 5 to 11%), most major and trace elements in BCR-129 and 141 were systematically underestimated (up to 30%, see Table S10c at

https://doi.org/10.5281/zenodo.5195507), which is why correction factors were calculated and applied to the other samples. Additional analytical information is provided in Table S10c (see Table S10c at https://doi.org/10.5281/zenodo.5195507). All vessels used were soaked in 10% HCl overnight and rinsed with Millipore water prior use.

Soil sampling for Nanopore/Illumina sequencing: The A horizon of the soil type Cambisol used for high molecular weight (HMW) DNA isolation for subsequent sequencing was sampled from the Schönbuch forest in November 2016, transported to the lab and stored at -20 °C.

Soil sampling for Illumina and amplicon sequencing of 7 soil samples: The O and A horizon of the soil types Podzol and , Stagnosol as well as the O, A and B horizon of Cambisol soil were sampled from the Schönbuch forest on May 3, 2019. Samples were collected using a soil probe, transported to the lab and stored at -20°C. To obtain the fine soil fraction, all soil samples were passed through a coarse mesh screen (1.2 x 1.2 cm) and subsequently a fine mesh screen (2 x 2 mm) prior to metagenomic DNA isolation.

5.5.2 Metagenome sequencing.

Isolation of HMW DNA from the A horizon of Cambisol for Nanopore sequencing run 1: HMW DNA was isolated from thawed fine soil samples using a published protocol (Brady 2007) with the following modification to increase the purity of the isolated DNA: After electroelution of the DNA out of the gel and into the dialysis bag, the dialysis bag was incubated in 0.5X TE buffer overnight before following the next steps of the protocol. Library preparation and Nanopore sequencing of the isolated DNA was performed by genXONE Inc. on a GridION device.

Isolation of HMW DNA from the A horizon of Cambisol for Illumina sequencing: For Illumina sequencing the above described DNA sample was further purified using the spin columns of the PowerLyzer PowerSoil DNA Isolation Kit (MO BIO Laboratories, Inc., #12855-100) and following an alternative protocol that was provided by MO BIO: The DNA sample isolated for Nanopore sequencing run 1 was filled up to 650 µl with H2O, and 650 µl of solution C4 and 650 µl of 100% ethanol were added. 650 µl of the mixture was loaded at a time on a MO BIO spin column and DNA was bound in three steps by centrifugation. The membrane was washed with 650 µl of 100% ethanol and subsequently with 500 µl of solution C5. The spin column was dried by centrifugation for 2 min at full speed and transferred to a clean tube. DNA was eluted with H2O. Library preparation (TrueSeq DNA PCR-Free) and Illumina sequencing was performed by CeGaT GmbH on a NovaSeq 6000 PE150.

Isolation of HMW DNA from the A horizon of Cambisol for Nanopore sequencing run 2: HMW DNA was isolated from 6 x 5g of thawed fine soil using a published protocol (Verma, Singh et al. 2017) with the following modifications to increase DNA yield and purity: After dissolving the dried pellets in 1 ml of 1X TE buffer, 1 µl of RNase I was added and incubated for 30 min at 37 °C before following the next steps of the protocol. In addition to precipitating the DNA with 0.7 volumes of isopropanol, 0.1 volumes of 5 M sodium acetate were added. After completing the protocol, the DNA was further gel purified as described by Brady (Brady, 2007) with adding a dialysis step in 0.5X TE overnight after electroelution of the DNA out of the gel and into the dialysis bag. Library preparation (native ligation sequencing kit, SQK-LSK109) and sequencing was performed by the NGS Competence Center Tübingen (NCCT) on a PromethION device.

Isolation of metagenomic DNA from 7 soil samples for Illumina sequencing: Metagenomic DNA was isolated from the O and A horizon of the Podzol, Cambisol and Stagnosol using the PowerLyzer PowerSoil DNA Isolation Kit (MO BIO Laboratories, Inc., #12855-100) and following an alternative protocol that was provided by MO BIO: 250 mg of each thawed fine soil sample was added to dry glass bead tubes and 500 µl of bead solution and 200 µl of phenol/chloroform/isoamyl alcohol were added followed by 60 µl of solution C1. Cells were opened using a Precellys 24 device (6500 rpm, 2 cycles of 20 seconds with 5 seconds pause) followed by centrifugation to the pellet. The supernatant was transferred to a new tube and 5 µl of RNase A were added as an additional step not mentioned in the protocol. 250 µl of solution C2, followed by 100 µl of solution C3 were added and mixed. The mixture was incubated for 5 min at 4 °C and subsequently centrifuged to the pellet. The supernatant was transferred to a new tube and 650 µl of solution C4 and 650 µl of 100% ethanol were added. 650 µl of the mixture was loaded at a time on a MO BIO spin column and DNA was bound in three steps by centrifugation. The membrane was washed with 650 µl of 100% ethanol and subsequently with 500 µl of solution C5 in case of non-stained membranes. In the case of brown membranes, a mixture of 300 µl solution C4 and 370 µl 100% ethanol were used to wash the membrane before washing with 100% ethanol and solution C5. The spin column was dried by centrifugation for 2 min at full speed and transferred to a clean tube. DNA was eluted with H2O. Metagenomic DNA from the B horizon of the Cambisol was isolated following the protocol of Verma, Singh et al. (Verma et al., 2017) with the above mentioned modifications. Library preparation (TrueSeq DNA PCR-Free) and Illumina sequencing was performed by CeGaT GmbH on a NovaSeq 6000 PE150.

Amplicon sequencing: Isolated metagenomic DNA of the 7 soil samples and published degenerate primers that recognize conserved regions in NRPS A domains (Adom_fw:GCSTACSYSATSTACACSTCSGG; Adom_rv:SASGTCVCCSGTSCGGTAS) (Pimentel-Elardo et al., 2012), PKSI KSI domains (KSI_fw:CCSCAGSAGCGCSTSYTSCTSGA; KSI_rv:GTSCCSGTSCCGTGSGYSTCSA) (Ginolhac et al., 2004) and 16S rRNA genes (16S_fw:CCTACGGGNGGCWGCAG; 16S_rv:GACTACHVGGGTATCTAATCC) (Klindworth et al., 2013) were used to generate amplicons via PCR. Concentrations of the DNA extracted from each of the 7 soil samples was measured using Qubit 3.0 Fluorometer and adjusted to 1.5 ng/ µl. PCR was performed using the Q5 High-Fidelity DNA Polymerase Kit (NEB) with the following reaction setup for a 25 µl reaction: 5 µl of 5X Q5 Reaction Buffer, 0.5 µl of 10 mM dNTPs, 0.5 µl of 10 µ Fw/Rv Primer, 3 µl of template DNA, 0.25 µl of Q5 High-Fidelity DNA Polymerase, 5 µl of 5X Q5 High GC Enhancer and 10.25 µl of nuclease-free water. The following thermocycling conditions were used: 98 °C for 30 sec followed by 30 cycles of 98 °C for 10 sec, 58.5 °C (A domain) or 68 °C (KSI domain, 16S rRNA gene) for 30 sec, 72 °C for 20 sec and a final step with 72 °C for 2 min. For each soil and primer pair, four 25 µl reactions were performed. 5 µl of each was analyzed via agarose gel electrophoresis and the remaining volume of the samples (20 µl each) were pooled. Pooled A domain and pooled 16S rRNA gene amplicons for each soil were purified using the QIA quick PCR purification Kit (50) following the manufacturer's instructions. Pooled KSI domain amplicons were gel purified using the QIAquick Gel Extraction Kit (QIAGEN) following the manufacturer's instructions. Sequencing was performed by the NGS Competence Center Tübingen (NCCT) on a MiSeq System.

5.5.3 Shotgun-seq Analysis.

Shotgun metagenome analysis: The shotgun Illumina and Nanopore reads were checked for sequence quality and adapter sequences using FastQC tool. To assess the advantages of using both short and long reads for recovering metagenomic BGCs, we performed both individual technology specific reads assembly as well as hybrid assembly. Illumina reads were assembled using metaSPADES (version 3.11.1) using default parameters (Nurk et al., 2017) . Hybrid assembly of Illumina and Nanopore reads were performed using metaSPADES (De Maio et al., 2019). Assembly comparisons were performed using the QUAST tool (Gurevich et al., 2013). Taxonomic Annotation and abundance estimation analysis was performed both on reads and assembled contigs. Accelerated BlastX annotations against NCBI non redundant proteins database was done using Diamond (version 0.9.24) (Buchfink et al., 2014). Alignment free fast taxonomic annotation tool Kraken2 with maxikraken2 database (available from https://lomanlab.github.io/mockcommunity/mc_databases.html) was also used to annotate the taxonomy of reads and assembled metagenomes (Wood et al., 2019).

Natural products biosynthesis domains and cluster annotation and diversity analysis: Using the BiG-MEx tool, we performed the BGC domain annotation and diversity analysis (Pereira, 2020). Annotation of 150 domains involved in biosynthesis of natural products was done.The assembled contigs with length greater than 10 kb were run through a local installation of the antiSMASH pipeline (version 5) for identifying the BGCs (Blin et al., 2019). For more focused annotations of KS and C domains, NaPDoS online server was used (Ziemert et al., 2012). BGCs were clustered using BiG-SCAPE with default parameters(Navarro-Muñoz et al., 2020). GCFs containing MIBiG (version 2.0) BGCs were considered closer to known BGC

products (Kautsar et al., 2020). The assembled contigs were also annotated using DeepBGC tool to predict novel BGCs based machine learning method (Hannigan et al., 2018).

5.5.4 Amplicon-seq Analysis.

Amplicon Analysis (Microbial Abundance and Diversity): The QIIME2 (version 2019.4) "Moving Pictures" tutorial steps were mostly followed for 16S Amplicons analysis (Bolyen et al., 2019). DADA2 was used to process both sequencing reads, leading to longer Amplicon Sequence Variants (ASV) (Callahan et al., 2016). DADA2 pipeline performed quality filtering, denoising and chimera detection (see Table S2 at https://doi.org/10.5281/zenodo.5195507). The ASVs were clustered into OTU by vsearch plugin available in QIIME2 at 97% identity by the de-novo clustering method. OTUs were classified using Naive Bayes classifier with the Silva database (version 132) (Quast et al., 2013). Subsequently, the maftt based multiple sequence alignment of features was performed which was used for phylogenetic tree construction via FastTree (Price et al., 2010). Q2-diversity plugin based alpha diversity and beta diversity analysis was performed to compute Shannon, Faith PD, OTU, Evenness alpha diversity indices and Jaccard, Bray-curtis, and UniFrac beta diversity distances .

Amplicon Analysis (BGC Domain Abundance): Amplicons of AMP-binding domain and KS domain were analyzed using QIIME2 pipeline steps described above for 16S amplicon analysis with modifications as described in the following text. Only read1 sequences were used as there was no overlap with read 2 and the relative quality of read 2 was bad. HMM search was performed using domain specific HMM models available via antiSMASH tool. Only the features matching the HMM models at default

thresholds were further analysed. ASVs were clustered at 97% identity using q2-diversity plugin. KS domain sequence amplicons were further annotated using NaPDoS to identify putative pathway products. Domains matching with NaPDoS database domains with less than 85 % identity were considered to be putative novel domains.

Comparison of amplicon-seq and shotgun-seq identified BGC domains: All the shotgun-seq domains identified for each sample after the BiG-MEx analysis, were concatenated. Using Dedupe script from BBTools (version 37.62), domains were deduplicated at 85 percent identity. Amplicon-seq domains were mapped on the deduplicated domains from shotgun-seq using BWA and SAMtools to identify common and unique domains.

Statistical analysis. Spearman rank correlation was computed between alpha diversity indices of 16S, A domain and KS domains. Similarly, correlation was also computed between alpha diversity indices and soil physico-chemical parameters. R version 3.6.2 and Rstudio were used to compute the statistical significance and correlation. The ggplot2 package was used to develop the boxplots (Wickham, 2011). Upset plots were developed using online UpSetR Shiny App webserver (Lex et al., 2014). qiime2R package, Pavian (Breitwieser and Salzberg, 2020) and Seaborn python visualisation library were used to plot the taxonomic profile and rarefaction curve.

Acknowledgements

The authors thank Christina Engesser for the excellent technical support, Sabine Flaiz and Rita Mögenburg for the CNS and ICP-OES analyses, and Heinrich Taubald for the XRF analysis. Further thanks to Franziska Höhn for soil sampling; Thomas Scholten and Peter Kühn for information on the soils; Libera Lo Presti and Martina

Adamek for scientific discussion. The authors acknowledge support by the High Performance and Cloud Computing Group at the Zentrum für Datenverarbeitung of the University of Tübingen, the state of Baden-Württemberg through bwHPC and the German Research Foundation (DFG) through grant no INST 37/935-1 FUGG. Computational resources provided by de.NBI cloud (https://www.denbi.de/cloud) and BinAC HPC Cluster (https://www.binac.uni-tuebingen.de/) were used for accelerating metagenome analysis. Authors also thank Cluster of Excellence - Control of Microorganisms to Fight Infections (CMFI), for structural support.

Funding: SM is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2124 – 390838134. SM and HSO were funded by the German Center for Infection biology TI06.903. NZ is funded by the German Center for Infection Research (TTU09.716). TN is funded by the State Postgraduate Fellowship Programme (Landesgraduiertenförderung).

Availability of data and materials: The sequencing data generated during the current study were submitted to NCBI's the Sequence Read Archive (SRA) and is accessible with BioProject ID: PRJNA717813 and PRJNA717918 . All relevant metadata, assembled contigs, annotated BGCs, clustering results are available for download from: https://doi.org/10.5281/zenodo.4644371.

Competing interests: The authors declare that they have no competing interests.

Authors' contributions: Study Conception: NZ, SM,TN; Soil Sampling: TN, HSO,SM; Soil Analysis: HN,YO; DNA isolation, library preparation and sequencing: TN, SP, AA; Bioinformatics Analysis: SM; Manuscript writing: SM, TN, NZ; Manuscript editing and approval to final draft: All; Funding acquisition: NZ

5.6 References

Arumugam, K., Bağcl, C., Bessarab, I., Beier, S., Buchfink, B., Górska, A., Qiu, G., Huson, D.H., Williams, R.B.H., 2019. Annotated bacterial chromosomes from frame-shift-corrected long-read metagenomic data. Microbiome 7, 1–13. https://doi.org/10.1186/s40168-019-0665-y

Bahram, M., Hildebrand, F., Forslund, S.K., Anderson, J.L., Soudzilovskaia, N.A., Bodegom, P.M., Bengtsson-Palme, J., Anslan, S., Coelho, L.P., Harend, H., Huerta-Cepas, J., Medema, M.H., Maltz, M.R., Mundra, S., Olsson, P.A., Pent, M., Põlme, S., Sunagawa, S., Ryberg, M., Tedersoo, L., Bork, P., 2018. Structure and function of the global topsoil microbiome. Nature 560, 233–237. https://doi.org/10.1038/s41586-018-0386-6

Baltz, R.H., 2008. Renaissance in antibacterial discovery from actinomycetes. Current Opinion in Pharmacology, Anti-infectives/New technologies 8, 557–563. https://doi.org/10.1016/j.coph.2008.04.008

Balvočiūtė, M., Huson, D.H., 2017. SILVA, RDP, Greengenes, NCBI and OTT — how do these taxonomies compare? BMC Genomics 18, 114. https://doi.org/10.1186/s12864-017-3501-4

Batani, G., Bayer, K., Böge, J., Hentschel, U., Thomas, T., 2019. Fluorescence in situ hybridization (FISH) and cell sorting of living bacteria. Scientific Reports 9, 18618. https://doi.org/10.1038/s41598-019-55049-2

Blin, K., Shaw, S., Steinke, K., Villebro, R., Ziemert, N., Lee, S.Y., Medema, M.H., Weber, T., 2019. AntiSMASH 5.0: Updates to the secondary metabolite genome mining pipeline. Nucleic Acids Research 47, W81–W87. https://doi.org/10.1093/nar/gkz310

Bodor, A., Bounedjoum, N., Vincze, G.E., Erdeiné Kis, Á., Laczi, K., Bende, G., Szilágyi, Á., Kovács, T., Perei, K., Rákhely, G., 2020. Challenges of unculturable bacteria: Environmental perspectives. Reviews in Environmental Science and Bio/Technology 19, 1–22. https://doi.org/10.1007/s11157-020-09522-4

Bolyen, E., Rideout, J.R., Dillon, M.R., Bokulich, N.A., Abnet, C.C., Al-Ghalith, G.A., Alexander, H., Alm, E.J., Arumugam, M., Asnicar, F., Bai, Y., Bisanz, J.E., Bittinger, K., Brejnrod, A., Brislawn, C.J., Brown, C.T., Callahan, B.J., Caraballo-Rodríguez, A.M., Chase, J., Cope, E.K., Da Silva, R., Diener, C., Dorrestein, P.C., Douglas, G.M., Durall, D.M., Duvallet, C., Edwardson, C.F., Ernst, M., Estaki, M., Fouquier, J., Gauglitz, J.M., Gibbons, S.M., Gibson, D.L., Gonzalez, A., Gorlick, K., Guo, J., Hillmann, B., Holmes, S., Holste, H., Huttenhower, C., Huttley, G.A., Janssen, S., Jarmusch, A.K., Jiang, L., Kaehler, B.D., Kang, K.B., Keefe, C.R., Keim, P., Kelley, S.T., Knights, D., Koester, I., Kosciolek, T., Kreps, J., Langille, M.G.I., Lee, J., Ley, R., Liu, Y.-X., Loftfield, E., Lozupone, C., Maher, M., Marotz, C., Martin, B.D., McDonald, D., McIver, L.J., Melnik, A.V., Metcalf, J.L., Morgan, S.C., Morton, J.T., Naimey, A.T., Navas-Molina, J.A., Nothias, L.F., Orchanian, S.B., Pearson, T., Peoples, S.L., Petras, D., Preuss, M.L., Pruesse, E., Rasmussen, L.B., Rivers, A., Robeson, M.S., Rosenthal, P., Segata, N., Shaffer, M., Shiffer, A., Sinha, R., Song, S.J., Spear, J.R., Swafford, A.D., Thompson, L.R., Torres, P.J., Trinh, P., Tripathi, A., Turnbaugh, P.J., Ul-Hasan, S., Hooft, J.J.J. van der, Vargas, F., Vázquez-Baeza, Y., Vogtmann, E., Hippel, M. von, Walters, W., Wan, Y., Wang, M., Warren, J., Weber, K.C., Williamson, C.H.D., Willis, A.D., Xu, Z.Z., Zaneveld, J.R., Zhang, Y., Zhu, Q., Knight, R., Caporaso, J.G., 2019. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nature Biotechnology 37, 852–857. https://doi.org/10.1038/s41587-019-0209-9

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., 2019. Microbial community drivers of PK/NRP gene diversity in selected global soils. Microbiome 7. https://doi.org/10.1186/s40168-019-0692-8

Boufridi, A., Quinn, R.J., 2018. Harnessing the Properties of Natural Products. Annual Review of Pharmacology and Toxicology 58, 451–470. https://doi.org/10.1146/annurev-pharmtox-010716-105029

Brady, S.F., 2007. Construction of soil environmental DNA cosmid libraries and screening for clones that produce biologically active small molecules. Nature Protocols 2, 1297–1305. https://doi.org/10.1038/nprot.2007.195

Breitwieser, F.P., Salzberg, S.L., 2020. Pavian: Interactive analysis of metagenomics data for microbiome studies and pathogen identification. Bioinformatics 36, 1303–1304. https://doi.org/10.1093/bioinformatics/btz715

Brumfield, K.D., Huq, A., Colwell, R.R., Olds, J.L., Leddy, M.B., 2020. Microbial resolution of whole genome shotgun and 16S amplicon metagenomic sequencing using publicly available NEON data. PLOS ONE 15, e0228899. https://doi.org/10.1371/journal.pone.0228899

Buchfink, B., Xie, C., Huson, D.H., 2014. Fast and sensitive protein alignment using DIAMOND. Nature Methods 12, 59–60. https://doi.org/10.1038/nmeth.3176

Callahan, B.J., McMurdie, P.J., Rosen, M.J., Han, A.W., Johnson, A.J.A., Holmes, S.P., 2016. DADA2: High-resolution sample inference from Illumina amplicon data. Nature methods 13, 581. https://doi.org/10.1038/nmeth.3869

Charlop-Powers, Z., Owen, J.G., Reddy, B.V.B., Ternei, M.A., Guimarães, D.O., Frias, U.A. de, Pupo, M.T., Seepe, P., Feng, Z., Brady, S.F., 2015. Global biogeographic sampling of bacterial secondary metabolism. eLife 4, e05048. https://doi.org/10.7554/eLife.05048

Charlop-Powers, Z., Pregitzer, C.C., Lemetre, C., Ternei, M.A., Maniko, J., Hover, B.M., Calle, P.Y., McGuire, K.L., Garbarino, J., Forgione, H.M., Charlop-Powers, S., Brady, S.F., 2016. Urban park soil microbiomes are a rich reservoir of natural product biosynthetic diversity. Proceedings of the National Academy of Sciences 113, 14811–14816. https://doi.org/10.1073/pnas.1615581113

Chu, J., Koirala, B., Forelli, N., Vila-Farres, X., Ternei, M.A., Ali, T., Colosimo, D.A., Brady, S.F., 2020. Synthetic-Bioinformatic Natural Product Antibiotics with Diverse Modes of Action. Journal of the American Chemical Society 142, 14158–14168. https://doi.org/10.1021/jacs.0c04376

Crits-Christoph, A., Diamond, S., Butterfield, C.N., Thomas, B.C., Banfield, J.F., 2018. Novel soil bacteria possess diverse genes for secondary metabolite biosynthesis. Nature 558, 440–444. https://doi.org/10.1038/s41586-018-0207-y

Crits-Christoph, A., Olm, M.R., Diamond, S., Bouma-Gregson, K., Banfield, J.F., 2020. Soil bacterial populations are shaped by recombination and gene-specific selection across a grassland meadow. ISME Journal 14, 1834–1846. https://doi.org/10.1038/s41396-020-0655-x

Davies, J., Davies, D., 2010. Origins and Evolution of Antibiotic Resistance. Microbiology and Molecular Biology Reviews : MMBR 74, 417–433. https://doi.org/10.1128/MMBR.00016-10

De Maio, N., Shaw, L.P., Hubbard, A., George, S., Sanderson, N.D., Swann, J., Wick, R., Oun, M.A., Stubberfield, E., Hoosdally, S.J., Crook, D.W., Peto, T.E.A., Sheppard, A.E., Bailey, M.J., Read, D.S., Anjum, M.F., Sarah Walker, A., Stoesser, N., 2019. Comparison of long-read sequencing technologies in the hybrid assembly of complex bacterial genomes. Microbial Genomics 5, 1–21. https://doi.org/10.1099/mgen.0.000294

Delgado-Baquerizo, M., Oliverio, A.M., Brewer, T.E., Benavent-González, A., Eldridge, D.J., Bardgett, R.D., Maestre, F.T., Singh, B.K., Fierer, N., 2018. A global atlas of the dominant bacteria found in soil. Science. https://doi.org/10.1126/science.aap9516

Durand, G.A., Raoult, D., Dubourg, G., 2019. Antibiotic discovery: History, methods and perspectives. International Journal of Antimicrobial Agents 53, 371–382. https://doi.org/10.1016/j.ijantimicag.2018.11.010

Eilers, K.G., Debenport, S., Anderson, S., Fierer, N., 2012. Digging deeper to find unique microbial communities: The strong effect of depth on the structure of bacterial and archaeal communities in soil. Soil Biology and Biochemistry 50, 58–65. https://doi.org/10.1016/j.soilbio.2012.03.011

Einsele, G., 1986. The landscape \ "o ecological research project Naturpark Sch \" o nbuch: Water and material balance, bio-, geo-and forestry studies in S \ "u west Germany. VCH-Verlag.

Elfeki, M., Alanjary, M., Green, S.J., Ziemert, N., Murphy, B.T., 2018. Assessing the Efficiency of Cultivation Techniques to Recover Natural Product Biosynthetic Gene Populations from Sediment. ACS Chemical Biology 13, 2074–2081. https://doi.org/10.1021/acschembio.8b00254

FAO and IUSS, 2015. World reference base for soil resources 2014: International soil classification system for naming soils and creating legends for soil maps - Update 2015, World Soil Resources Reports. FAO, Rome, Italy.

Fierer, N., Schimel, J.P., Holden, P.A., 2003. Variations in microbial community composition through two soil depth profiles. Soil Biology and Biochemistry 35, 167–176. https://doi.org/10.1016/S0038-0717(02)00251-1

Ginolhac, A., Jarrin, C., Gillet, B., Robe, P., Pujic, P., Tuphile, K., Bertrand, H., Vogel, T.M., Perrière, G., Simonet, P., Nalin, R., 2004. Phylogenetic analysis of polyketide synthase I domains from soil metagenomic libraries allows selection of promising clones. Applied and Environmental Microbiology 70, 5522–5527. https://doi.org/10.1128/AEM.70.9.5522-5527.2004

Grathwohl, P., Rügner, H., Wöhling, T., Osenbrück, K., Schwientek, M., Gayler, S., Wollschläger, U., Selle, B., Pause, M., Delfs, J.-O., Grzeschik, M., Weller, U., Ivanov, M., Cirpka, O.A., Maier, U., Kuch, B., Nowak, W., Wulfmeyer, V., Warrach-Sagi, K., Streck, T., Attinger, S., Bilke, L., Dietrich, P., Fleckenstein, J.H., Kalbacher, T., Kolditz, O., Rink, K., Samaniego, L., Vogel, H.-J., Werban, U., Teutsch, G., 2013. Catchments as reactors: A comprehensive approach for water fluxes and solute turnover. Environmental Earth Sciences 69, 317–333. https://doi.org/10.1007/s12665-013-2281-7

Gurevich, A., Saveliev, V., Vyahhi, N., Tesler, G., 2013. QUAST: Quality assessment tool for genome assemblies. Bioinformatics 29, 1072–1075. https://doi.org/10.1093/bioinformatics/btt086

Handelsman, J., 2004. Metagenomics: Application of Genomics to Uncultured Microorganisms. Microbiology and Molecular Biology Reviews 68, 669–685. https://doi.org/10.1128/MMBR.68.4.669-685.2004

Hannigan, G.D., Prihoda, D., Palicka, A., Soukup, J., Klempir, O., Rampula, L., Durcak, J., Wurst, M., Kotowski, J., Chang, D., Wang, R., Piizzi, G., Hazuda, D.J., Woelk, C.H., Bitton, D.A., 2018. A Deep Learning Genome-Mining Strategy Improves Biosynthetic Gene Cluster Prediction. bioRxiv 500694. https://doi.org/10.1101/500694

Hao, J., Chai, Y.N., Lopes, L.D., Ordóñez, R.A., Wright, E.E., Archontoulis, S., Schachtman, D.P., 2021. The Effects of Soil Depth on the Structure of Microbial Communities in Agricultural Soils in Iowa (United States). Applied and Environmental Microbiology 87. https://doi.org/10.1128/AEM.02673-20

Jahn, R., Blume, H.P., Asio, V.B., Spaargaren, O., Schad, P., 2006. Guidelines for soil description, 4th edition. FAO, Rome.

Jovel, J., Patterson, J., Wang, W., Hotte, N., O'Keefe, S., Mitchel, T., Perry, T., Kao, D., Mason, A.L., Madsen, K.L., Wong, G.K.-S., 2016. Characterization of the Gut Microbiome Using 16S or Shotgun Metagenomics. Frontiers in Microbiology 7. https://doi.org/10.3389/fmicb.2016.00459

Kautsar, S.A., Blin, K., Shaw, S., Navarro-Muñoz, J.C., Terlouw, B.R., Van Der Hooft, J.J.J., Van Santen, J.A., Tracanna, V., Suarez Duran, H.G., Pascal Andreu, V., Selem-Mojica, N., Alanjary, M., Robinson, S.L., Lund, G., Epstein, S.C., Sisto, A.C., Charkoudian, L.K., Collemare, J., Linington, R.G., Weber, T., Medema, M.H., 2020. MIBiG 2.0: A repository for biosynthetic gene clusters of known function. Nucleic Acids Research 48, D454–D458. https://doi.org/10.1093/nar/gkz882

Klindworth, A., Pruesse, E., Schweer, T., Peplies, J., Quast, C., Horn, M., Glöckner, F.O., 2013. Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. Nucleic Acids Research 41, e1. https://doi.org/10.1093/nar/gks808

Lemetre, C., Maniko, J., Charlop-Powers, Z., Sparrow, B., Lowe, A.J., Brady, S.F., 2017. Bacterial natural product biosynthetic domain composition in soil correlates with changes in latitude on a continent-wide scale. Proceedings of the National Academy of Sciences of the United States of America 114, 11615–11620. https://doi.org/10.1073/pnas.1710262114

Lex, A., Gehlenborg, N., Strobelt, H., Vuillemot, R., Pfister, H., 2014. UpSet: Visualization of Intersecting Sets. IEEE transactions on visualization and computer graphics 20, 1983–1992. https://doi.org/10.1109/TVCG.2014.2346248

Morlon, H., O'Connor, T.K., Bryant, J.A., Charkoudian, L.K., Docherty, K.M., Jones, E., Kembel, S.W., Green, J.L., Bohannan, B.J.M., 2015. The Biogeography of Putative Microbial Antibiotic Production. PLOS ONE 10, e0130659. https://doi.org/10.1371/journal.pone.0130659

Mouncey, N.J., Otani, H., Udwary, D., Yoshikuni, Y., 2019. New voyages to explore the natural product galaxy. Journal of Industrial Microbiology and Biotechnology 46, 273–279. https://doi.org/10.1007/s10295-018-02122-w

Mungan, M.D., Alanjary, M., Blin, K., Weber, T., Medema, M.H., Ziemert, N., 2020. ARTS 2.0: Feature updates and expansion of the Antibiotic Resistant Target Seeker for comparative genome mining. Nucleic Acids Research 48, W546–W552. https://doi.org/10.1093/nar/gkaa374

Navarro-Muñoz, J.C., Selem-Mojica, N., Mullowney, M.W., Kautsar, S.A., Tryon, J.H., Parkinson, E.I., De Los Santos, E.L.C., Yeong, M., Cruz-Morales, P., Abubucker, S., Roeters, A., Lokhorst, W., Fernandez-Guerra, A., Cappelini, L.T.D., Goering, A.W., Thomson, R.J., Metcalf, W.W., Kelleher, N.L., Barona-Gomez, F., Medema, M.H., 2020. A computational framework to explore large-scale biosynthetic diversity. Nature Chemical Biology 16, 60–68. https://doi.org/10.1038/s41589-019-0400-9

Nurk, S., Meleshko, D., Korobeynikov, A., Pevzner, P.A., 2017. MetaSPAdes: A new versatile metagenomic assembler. Genome Research 27, 824–834. https://doi.org/10.1101/gr.213959.116

Pereira, E., 2020. Improvements in natural product biosynthetic gene clusters research and functional trait-based approaches in metagenomics (PhD thesis). Jacobs University Bremen.

Pimentel-Elardo, S.M., Grozdanov, L., Proksch, S., Hentschel, U., 2012. Diversity of Nonribosomal Peptide Synthetase Genes in the Microbial Metagenomes of Marine Sponges. Marine Drugs 10, 1192–1202. https://doi.org/10.3390/md10061192

Price, M.N., Dehal, P.S., Arkin, A.P., 2010. FastTree 2–approximately maximum-likelihood trees for large alignments. PloS one 5, e9490. https://doi.org/10.1371/journal.pone.0009490

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., Glöckner, F.O., 2013. The {SILVA} ribosomal {RNA} gene database project: Improved data processing and web-based tools. Nucleic Acids Research 41, D590–D596. https://doi.org/10.1093/nar/gks1219

Reddy, B.V.B., Kallifidas, D., Kim, J.H., Charlop-Powers, Z., Feng, Z., Brady, S.F., 2012. Natural product biosynthetic gene diversity in geographically distinct soil microbiomes. Applied and Environmental Microbiology 78, 3744–3752. https://doi.org/10.1128/AEM.00102-12

Rodriguez-R, L.M., Gunturu, S., Tiedje, J.M., Cole, J.R., Konstantinidis, K.T., 2018. Nonpareil 3: Fast Estimation of Metagenomic Coverage and Sequence Diversity. mSystems 3. https://doi.org/10.1128/mSystems.00039-18

Sharrar, A., Crits-Christoph, A., Méheust, R., Diamond, S., Starr, E., Banfield, J., 2019. Bacterial secondary metabolite biosynthetic potential in soil varies with phylum, depth, and vegetation type. mBio. https://doi.org/10.1101/818815

Silver, L.L., 2011. Challenges of Antibacterial Discovery. Clinical Microbiology Reviews 24, 71–109. https://doi.org/10.1128/CMR.00030-10

Sorokina, M., Steinbeck, C., 2020. Review on natural products databases: Where to find data in 2020. Journal of Cheminformatics 12, 20. https://doi.org/10.1186/s13321-020-00424-9

Sugimoto, Y., Camacho, F.R., Wang, S., Chankhamjon, P., Odabas, A., Biswas, A., Jeffrey, P.D., Donia, M.S., 2019. A metagenomic strategy for harnessing the chemical repertoire of the human microbiome. Science 366. https://doi.org/10.1126/science.aax9176

Thompson, L.R., Sanders, J.G., McDonald, D., Amir, A., Ladau, J., Locey, K.J., Prill, R.J., Tripathi, A., Gibbons, S.M., Ackermann, G., Navas-Molina, J.A., Janssen, S., Kopylova, E., Vázquez-

Baeza, Y., González, A., Morton, J.T., Mirarab, S., Xu, Z.Z., Jiang, L., Haroon, M.F., Kanbar, J., Zhu, Q., Song, S.J., Kosciolek, T., Bokulich, N.A., Lefler, J., Brislawn, C.J., Humphrey, G., Owens, S.M., Hampton-Marcell, J., Berg-Lyons, D., McKenzie, V., Fierer, N., Fuhrman, J.A., Clauset, A., Stevens, R.L., Shade, A., Pollard, K.S., Goodwin, K.D., Jansson, J.K., Gilbert, J.A., Knight, R., Agosto Rivera, J.L., Al-Moosawi, L., Alverdy, J., Amato, K.R., Andras, J., Angenent, L.T., Antonopoulos, D.A., Apprill, A., Armitage, D., Ballantine, K., Bárta, J., Baum, J.K., Berry, A., Bhatnagar, A., Bhatnagar, M., Biddle, J.F., Bittner, L., Boldgiv, B., Bottos, E., Boyer, D.M., Braun, J., Brazelton, W., Brearley, F.Q., Campbell, A.H., Caporaso, J.G., Cardona, C., Carroll, J.L., Cary, S.C., Casper, B.B., Charles, T.C., Chu, H., Claar, D.C., Clark, R.G., Clayton, J.B., Clemente, J.C., Cochran, A., Coleman, M.L., Collins, G., Colwell, R.R., Contreras, M., Crary, B.B., Creer, S., Cristol, D.A., Crump, B.C., Cui, D., Daly, S.E., Davalos, L., Dawson, R.D., Defazio, J., Delsuc, F., Dionisi, H.M., Dominguez-Bello, M.G., Dowell, R., Dubinsky, E.A., Dunn, P.O., Ercolini, D., Espinoza, R.E., Ezenwa, V., Fenner, N., Findlay, H.S., Fleming, I.D., Fogliano, V., Forsman, A., Freeman, C., Friedman, E.S., Galindo, G., Garcia, L., Garcia-Amado, M.A., Garshelis, D., Gasser, R.B., Gerdts, G., Gibson, M.K., Gifford, I., Gill, R.T., Giray, T., Gittel, A., Golyshin, P., Gong, D., Grossart, H.P., Guyton, K., Haig, S.J., Hale, V., Hall, R.S., Hallam, S.J., Handley, K.M., Hasan, N.A., Haydon, S.R., Hickman, J.E., Hidalgo, G., Hofmockel, K.S., Hooker, J., Hulth, S., Hultman, J., Hyde, E., Ibáñez-Álamo, J.D., Jastrow, J.D., Jex, A.R., Johnson, L.S., Johnston, E.R., Joseph, S., Jurburg, S.D., Jurelevicius, D., Karlsson, A., Karlsson, R., Kauppinen, S., Kellogg, C.T.E., Kennedy, S.J., Kerkhof, L.J., King, G.M., Kling, G.W., Koehler, A.V., Krezalek, M., Kueneman, J., Lamendella, R., Landon, E.M., Lanede Graaf, K., LaRoche, J., Larsen, P., Laverock, B., Lax, S., Lentino, M., Levin, I.I., Liancourt, P., Liang, W., Linz, A.M., Lipson, D.A., Liu, Y., Lladser, M.E., Lozada, M., Spirito, C.M., MacCormack, W.P., MacRae-Crerar, A., Magris, M., Martín-Platero, A.M., Martín-Vivaldi, M., Martínez, L.M., Martínez-Bueno, M., Marzinelli, E.M., Mason, O.U., Mayer, G.D., McDevitt-Irwin, J.M., McDonald, J.E., McGuire, K.L., McMahon, K.D., McMinds, R., Medina, M., Mendelson, J.R., Metcalf, J.L., Meyer, F., Michelangeli, F., Miller, K., Mills, D.A., Minich, J., Mocali, S., Moitinho-Silva, L., Moore, A., Morgan-Kiss, R.M., Munroe, P., Myrold, D., Neufeld, J.D., Ni, Y., Nicol, G.W., Nielsen, S., Nissimov, J.I., Niu, K., Nolan, M.J., Noyce, K., O'Brien, S.L., Okamoto, N., Orlando, L., Castellano, Y.O., Osuolale, O., Oswald, W., Parnell, J., Peralta-Sánchez, J.M., Petraitis, P., Pfister, C., Pilon-Smits, E., Piombino, P., Pointing, S.B., Pollock, F.J., Potter, C., Prithiviraj, B., Quince, C., Rani, A., Ranjan, R., Rao, S., Rees, A.P., Richardson, M., Riebesell, U., Robinson, C., Rockne, K.J., Rodriguezl, S.M., Rohwer, F., Roundstone, W., Safran, R.J., Sangwan, N., Sanz, V., Schrenk, M., Schrenzel, M.D., Scott, N.M., Seger, R.L., Seguinorlando, A., Seldin, L., Seyler, L.M., Shakhsheer, B., Sheets, G.M., Shen, C., Shi, Y., Shin, H., Shogan, B.D., Shutler, D., Siegel, J., Simmons, S., Sjöling, S., Smith, D.P., Soler, J.J., Sperling, M., Steinberg, P.D., Stephens, B., Stevens, M.A., Taghavi, S., Tai, V., Tait, K., Tan, C.L., Taş, N., Taylor, D.L., Thomas, T., Timling, I., Turner, B.L., Urich, T., Ursell, L.K., Van Der Lelie, D., Van Treuren, W., Van Zwieten, L., Vargas-Robles, D., Thurber, R.V., Vitaglione, P., Walker, D.A., Walters, W.A., Wang, S., Wang, T., Weaver, T., Webster, N.S., Wehrle, B., Weisenhorn, P., Weiss, S., Werner, J.J., West, K., Whitehead, A., Whitehead, S.R., Whittingham, L.A., Willerslev, E., Williams, A.E., Wood, S.A., Woodhams, D.C., Yang, Y., Zaneveld, J., Zarraonaindia, I., Zhang, Q., Zhao, H., 2017. A communal catalogue reveals Earth's multiscale microbial diversity. Nature 551, 457–463. https://doi.org/10.1038/nature24621

Tyc, O., Song, C., Dickschat, J.S., Vos, M., Garbeva, P., 2017. The Ecological Role of Volatile and Soluble Secondary Metabolites Produced by Soil Bacteria. Trends in Microbiology 25, 280–292. https://doi.org/10.1016/j.tim.2016.12.002

Verma, S.K., Singh, H., Sharma, P.C., 2017. An improved method suitable for isolation of high-quality metagenomic DNA from diverse soils. 3 Biotech 7, 171. https://doi.org/10.1007/s13205-017-0847-x

Wang, H., Cheng, M., Dsouza, M., Weisenhorn, P., Zheng, T., Gilbert, J.A., 2018. Soil Bacterial Diversity Is Associated with Human Population Density in Urban Greenspaces. Environmental Science and Technology 52, 5115–5124. https://doi.org/10.1021/acs.est.7b06417

Wickham, H., 2011. ggplot2. WIREs Computational Statistics 3, 180–185. https://doi.org/10.1002/wics.147

Will, C., Thürmer, A., Wollherr, A., Nacke, H., Herold, N., Schrumpf, M., Gutknecht, J., Wubet, T., Buscot, F., Daniel, R., 2010. Horizon-Specific Bacterial Community Composition of German Grassland Soils, as Revealed by Pyrosequencing-Based Analysis of 16S rRNA Genes. Applied and Environmental Microbiology 76, 6751–6759. https://doi.org/10.1128/AEM.01063-10

Wilpiszeski, R.L., Aufrecht, J.A., Retterer, S.T., Sullivan, M.B., Graham, D.E., Pierce, E.M., Zablocki, O.D., Palumbo, A.V., Elias, D.A., 2019. Soil Aggregate Microbial Communities: Towards Understanding Microbiome Interactions at Biologically Relevant Scales. Applied and Environmental Microbiology 85. https://doi.org/10.1128/AEM.00324-19

Wohlleben, W., Mast, Y., Stegmann, E., Ziemert, N., 2016. Antibiotic drug discovery. Microbial Biotechnology 9, 541–548. https://doi.org/10.1111/1751-7915.12388

Wood, D.E., Lu, J., Langmead, B., 2019. Improved metagenomic analysis with Kraken 2. Genome Biology. https://doi.org/10.1186/s13059-019-1891-0

Yarza, P., Yilmaz, P., Pruesse, E., Glöckner, F.O., Ludwig, W., Schleifer, K.-H., Whitman, W.B., Euzéby, J., Amann, R., Rosselló-Móra, R., 2014. Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. Nature Reviews Microbiology 12, 635–645. https://doi.org/10.1038/nrmicro3330

Zhang, M.M., Qiao, Y., Ang, E.L., Zhao, H., 2017. Using natural products for drug discovery: The impact of the genomics era. Expert opinion on drug discovery 12, 475–487. https://doi.org/10.1080/17460441.2017.1303478

Ziemert, N., Podell, S., Penn, K., Badger, J.H., Allen, E., Jensen, P.R., 2012. The Natural Product Domain Seeker NaPDoS: A Phylogeny Based Bioinformatic Tool to Classify Secondary Metabolite Gene Diversity. PLOS ONE 7, e34064. https://doi.org/10.1371/journal.pone.0034064

Chapter 6: Metagenomic big-data explorations of natural products diversity in diverse ecosystems

This chapter covers the biosynthesis potential survey of diverse ecosystems and is divided in following parts:

Part I: Manuscript — Evaluating the Distribution of Bacterial Natural Product Biosynthetic Genes Across Lake Huron Sediment

Part II: Dynamics of the human gut secondary metabolome during antibiotic treatment.

Part III: Using linked reads and long reads to recover biosynthetic gene clusters from Tuebingen strain collections.

Part I: Evaluating the Distribution of Bacterial Natural Product Biosynthetic Genes Across Lake Huron Sediment.

Status: Manuscript (*Accepted*)

Citation: *Maryam Elfeki, Shrikant Mantri, Chase M. Clark, Stefan J. Green, Nadine Ziemert, Brian T. Murphy, (2021), Evaluating the Distribution of Bacterial Natural Product Biosynthetic Genes Across Lake Huron Sediment.*

Own contribution: Responsible for scientific ideas (with co-authors)

Responsible for data generation (with co-authors)

Responsible for analysis and interpretation of data. (with co-authors)

Responsible for composing the manuscript (with co-authors)

Abstract

Environmental microorganisms continue to serve as a major source of bioactive natural products (NPs) and as an inspiration for many other scaffolds in the toolbox of modern medicine. Nearly all microbial NP-inspired therapies can be traced to field expeditions to collect samples from the environment. Despite the importance of these expeditions in the search for new drugs, few studies have attempted to document the extent to which NPs or their corresponding production genes are distributed within a given environment. To gain insight into this, the geographic occurrence of NP ketosynthase (KS) and adenylation (A) domains was documented across 53 and 58 surface sediment samples, respectively, covering 59,590 square kilometers of Lake Huron. Overall, no discernable NP geographic distribution patterns were observed for 90,528 NP classes of nonribosomal peptides and polyketides detected in the survey. While each sampling location harbored a similar number of A domain operational biosynthetic units (OBUs), limited overlap of OBU type was observed, suggesting that at the sequencing depth used in this study, no single location served as a NP 'hotspot'. These data support the hypothesis that there is ample variation in NP occurrence between sampling sites and suggests that extensive sample collection efforts are required to fully capture the functional chemical diversity of sediment microbial communities on a regional scale.

6.1 Introduction

The preparation of Pyocyanase in 1899 and the discovery of bioactive natural products (NPs) penicillin and gramicidin in 1928 and 1939, respectively, marked the beginning of modern microbial drug discovery efforts.(Aldrich, 1999; Emmerich and Löw, 1899; Fleming A., 1929; Gause and Brazhnikova, 1944) 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 generated therapies for an array of diseases such as cancer, bacterial infections, immune disorders, and others, as 34% of FDA approved drugs from 2000 to 2014 were NPs or NP-derived.(Newman and

Cragg, 2020) Importantly, nearly all of these microbial NP-inspired therapies resulted from field expeditions to collect samples from the environment. In general, these field expeditions have been guided by the hypothesis that environments in diverse geographic locations contain different ecological pressures, and as a result harbor minimally-overlapping populations of NP biosynthetic pathways.(Cheng et al., 2015; Clardy et al., 2009; Fischbach and Walsh, 2009)

Despite the importance of sample collection expeditions toward the search for new drugs, few studies have attempted to document the extent to which NPs or their corresponding production genes are distributed in any given environment. Charlop-Powers et al. compared the NP biosynthetic potential of soil samples from a diverse array of environmental microbiomes.(Charlop-Powers et al., 2015) Their analyses of 185 soil microbiomes collected from five continents suggested that geographic distance and local environment contributed to biosynthetic diversity differences observed between samples.(Charlop-Powers et al., 2015) Additionally, Lemetre et al. found that changes in latitude correlated with changes in biosynthetic domain composition within soil samples on a continent-wide scale.(Lemetre et al., 2017) Borsetto et al. correlated 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 influence the biosynthetic potential at a given site.(Borsetto et al., 2019) Similarly, Sharrar et al. found that patterns of abundance of BGC types varied by taxonomy in soil bacteria, and that bacteria with higher biosynthetic potential were associated with specific types of soil vegetation.(Sharrar et al., 2020) These studies demonstrate that biosynthetic domain composition can differ with changing geography and/or variables within the soil. Thus, characterizing the geographic distribution of NP-producing BGCs at a finer geographical resolution will inform front-end discovery practices such as sample collection and microbial library generation, which traditionally have a high degree of uncertainty.(Hernandez et al., 2021)

Due to decreasing sequencing costs and availability of online tools, probing microbial-based chemical diversity in nature has become attainable without relying on cultivation techniques. To gain insight into how specific NP classes are distributed in an environment, the occurrence of NP domains was characterized in up to 58 surface sediment samples covering a 59,590 square kilometer region in Lake Huron. Ketosynthase (KS) domains from polyketide synthases (PKS) and adenylation (A) domains from nonribosomal peptide synthetase (NRPS) were examined, as they represent conserved domains within two common classes of NPs that often encode for the production of antibiotics, siderophores, and other bioactive compounds. The current study provides preliminary evidence that there is substantial variation in NP composition between sampling sites on a regional scale and suggests that extensive sample collection efforts will be required to fully capture the BGC diversity that exists in sediment. 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 antibiotic discovery efforts began in the early 20th century.

6.2 Results and Discussion

6.2.1 . 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 (Annexure B, Table S1). To confirm the bacterial diversity present represents populations that commonly occur in freshwater systems, the taxonomic diversity of bacteria at each site was assessed using microbial 16S rRNA gene amplicons (Annexure B, Supp. Experimental Procedures). Results were congruent with those of typical lake bacterial populations (Annexure B, Supp. Table S4).(Newton et al., 2011)To assess the composition of NP domains at each collection site, previously designed degenerate primers were used to amplify the KSα domain for PKS II(Metsä-Ketelä et al., 1999) and the A domain for NRPS genes from genomic DNA (gDNA)

extracted from sediment samples.(Ayuso-Sacido and Genilloud, 2005) The KSα and A domains were selected because they are among the most conserved catalytic domains of the PKS type II and NRPS gene clusters respectively. Furthermore, this sequence conservation has yielded primer sets for PCR amplification(Ayuso-Sacido and Genilloud, 2005; Ginolhac et al., 2004; Metsä-Ketelä et al., 1999) as well as bioinformatic tools and databases to facilitate the annotation and prediction of NPs.(Kautsar et al., 2020; Weber and Kim, 2016)

The selected conserved regions were PCR-amplified from genomic DNA using a two-stage PCR protocol, as described previously.(Naqib et al., 2018) Briefly, 613 bp fragments of KSα (β-ketoacyl synthase) and 700 bp fragments of NRPS A domains were amplified using degenerate oligonucleotides, respectively.(Ayuso-Sacido and Genilloud, 2005; Metsä-Ketelä et al., 1999) All primers were synthesized with a locus-specific sequence as well as a universal 5′ tail.(Naqib et al., 2018) Resulting sequences were filtered using profile hidden Markov models (pHMMs) downloaded from antiSMASH's HMM detection modules to remove non-specific sequences.(Blin et al., 2019) These models are based on known and predicted KSα and A domain architectures.(Adamek et al., 2019) Filtered sequences were then clustered at 85% similarity to approximate compound class designations and to avoid overestimation of chemical diversity in sediment.(Elfeki et al., 2018) Sequences were extracted from the manually curated and annotated BGC database MIBiG(Kautsar et al., 2020), subjected to different clustering thresholds, and 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 and ranged from 80% to 90%. Therefore, analysis proceeded using an 85% similarity threshold. At 85% similarity, the sequence groupings – or operational biosynthetic units (OBUs) – represent an estimation of compound classes. To further scrutinize this clustering method, amplicons from a control *Streptomyces* strain*, Streptomyces coelicolor A3(2),* were subjected to this process (see Methods section 4.5).(Bentley et al., 2002) *S. coelicolor A3(2)* produces two KSα domain-containing compounds (actinorhodin and a spore pigment) and twelve A domain-containing

compounds (CDA1b, CDA2a, CDA2b, CDA3a, CDA3b, CDA4a, CDA4b, coelibactin, coelimycin P1, undecylprodigiosin, SCO-2138, and a putative tris-hydroxamate tetrapeptide iron chelator coelichelin).(Bentley et al., 2002; Lautru et al., 2005) Analysis of *S. coelicolor A3(2)* amplicons at 85% similarity yielded two KSα domain OBUs and fifteen A domain OBUs, and confirmed this as a suitable threshold to organize 300 bp fragments into groups that represent compound classes.

Of the 59 sediment samples, 6 from the KSα dataset and 1 from the A domain dataset did not return sufficient quality data to be included in the analysis. In total, 1,818 KSα OBUs (5,815 total sequences) throughout 53 sediment samples, and 171,527 A domain OBUs (1,730,091 total sequences) throughout 58 sediment samples were observed. This represents approximately 34 KSα and 2,957 A domain OBUs per sediment sample (Table 1). These original numbers were then adjusted to account for suspected overestimation of chemical diversity, as described in the following section. The large disparity in KSα and A domain OBU counts may be attributed to (1) primer biases and accuracy, (2) depth of sequencing, and (3) the size of the family to which these domains belong. A domains belong to a large superfamily of adenylate-forming enzymes,(Schmelz and Naismith, 2009) in contrast to the smaller KSα (α-ketoacyl synthases) domain family, which are known to produce aromatic polyketides and polyenes, and whose primers were designed specifically for strains within the *Streptomyces* genus.(Chen et al., 2018; Du et al., 2018) The number and putative identity of OBUs for each compound class is listed in Supporting Tables S6A-B. As previously reported, the KSα primers are highly degenerate, with substantial off-target amplification.(Liu et al., 2016) Due to this limitation, KSα data, including distribution analysis and maps, can be found in the Supplemental Information.

6.2.2 .Analysis of Characterized NP BGC Distribution in Lake Sediment

In order to assess the occurrence of known NP BGC classes across Lake Huron sediment, the identity of each OBU was verified. Sequence representatives from each OBU were aligned against domain sequences extracted from the MIBiG database using the

DIAMOND alignment tool via its default settings.(Buchfink et al., 2015; Kautsar et al., 2019) MIBiG associates BGCs with known NP structures, allowing prediction of the product of each matching OBU and as a result, estimation of the chemical diversity at each sample site. To ensure that a 300 bp amplicon is sufficient for structural annotation, sequences from control strain *S. coelicolor* A3(2) were amplified, sequenced, and aligned (Supplementary Experimental Procedures).(Bentley et al., 2002) 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.(Scholz et al., 2016) These results were used as a guide to select a list of annotated OBUs to map across lake sediment. Based on empirical tests and comparison to e-values obtained from the *S. coelicolor* A3(2) control, 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.

Once OBU sequence representatives were aligned against sequences from the MIBiG database, the majority of these could not be assigned to known chemical compound classes. In total, of the 1,818 KSα domain OBUs that were observed across 53 samples, 32 (1.7%) were assigned to known compound classes. Similarly, of 171,527 total A domain OBUs observed across 58 samples, 108 (0.06%) were assigned to known compound classes. Of particular note is that some distinct OBU sequence representatives were assigned to the same compound class (for example, five separate OBU sequence representatives aligned to rifamycin), which resulted in an overestimation of compound classes present in sediment. To correct for this, it was necessary to estimate the average number of times a compound class was divided into separate OBUs in the dataset; this average was deemed a “split correction factor” (see Annexure B, Supplementary Table S6 for discussion). The total number of observed OBUs was then divided by that factor, resulting in a more accurate estimation of the compound classes present in sediment: a total

of 1,198 KSα domain OBUs, of which 21 (1.8%) were known compound classes, and a total of 90,528 A domain OBUs, of which 57 (0.06%) were known compound classes. Further details are listed in Supplementary Tables S6A-B.

Table 1

	KSα	A
Total # of OBUs detected	1,818	171,527
Total # of OBUs after adjustment by the split correction factor	1,198	90,528
Average # of OBUs **per sample** after adjustment by the split correction factor	23 (±18)	1,561 (±798)

Table 1. A and KSα domain abundances in sediment.

Figure 1.

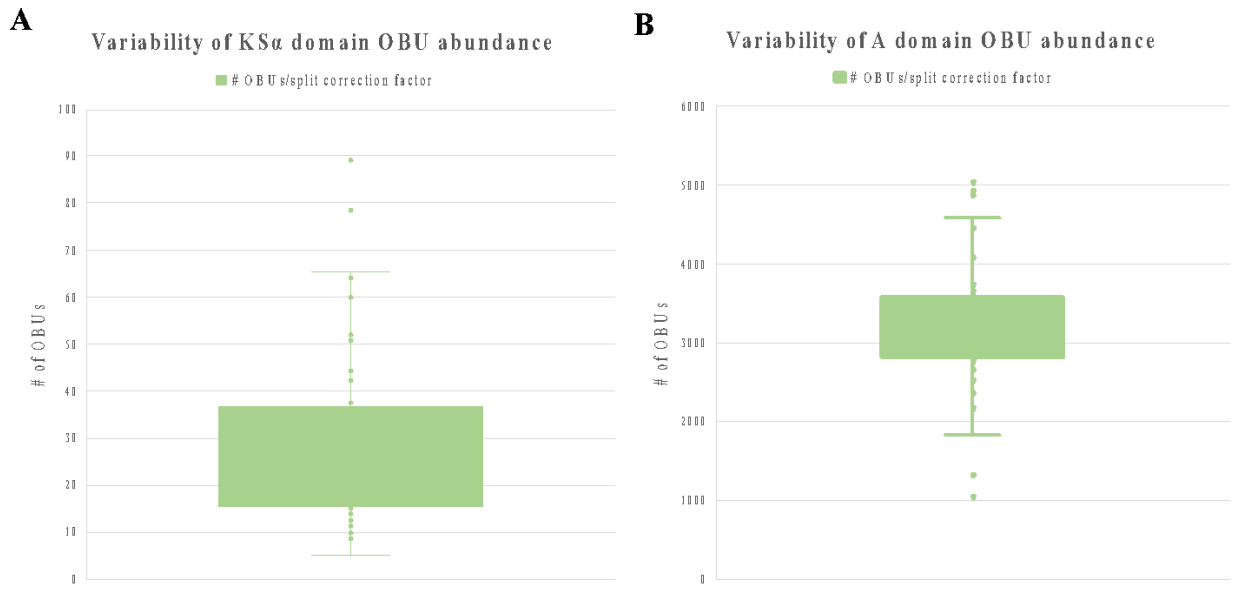


Figure 1. Boxplots depicting the variability of KSα and A domain OBU abundance *in each* sediment sample. A and B represent boxplots of OBU counts after adjustment by the split correction factor, which corrects for overestimations of compound classes present in sediment.

Of the 78 OBUs matched to known classes of PKS (21) and NRPS (57) NPs in MIBiG, distribution maps of compounds that occurred in at least two distinct locations were generated after rarefaction analysis to the lowest sample count (15 sequences for KSα domain OBUs and 3,487 for A domain OBUs). A total of 30 OBUs met these criteria.

These 30 OBUs were further categorized into antibiotics, siderophores, and other bioactive NP classes such as anticancer and antiviral compounds. OBUs from each of the 30 classes were mapped and patterns of occurrence were assessed (representative OBUs per category are shown in Figure 2, while maps for the remaining OBUs are shown in Supplementary Figures S3-5). The size of the colored circles are proportional to the number of sequences detected at each sampling site, after rarefaction. Figures 2A-D show the distribution of cyclomarin, surugamide, pyoverdin, and coelichelin classes. For example, sequence reads for cyclomarin class antibiotics (Figure 2A) were detected in five distinct geographic locations across the lake, while sequence reads for pyoverdin-type siderophores (Figure 2C) were detected in 38 distinct geographic locations across the lake. Four of these locations contained both compounds. Overall, the distribution profiles among the compound classes analyzed were non-overlapping in lake sediment. In general, siderophores were the most frequently detected compound class in lake sediment, exceeding that of antibiotics and other bioactive NPs.

Figure 2

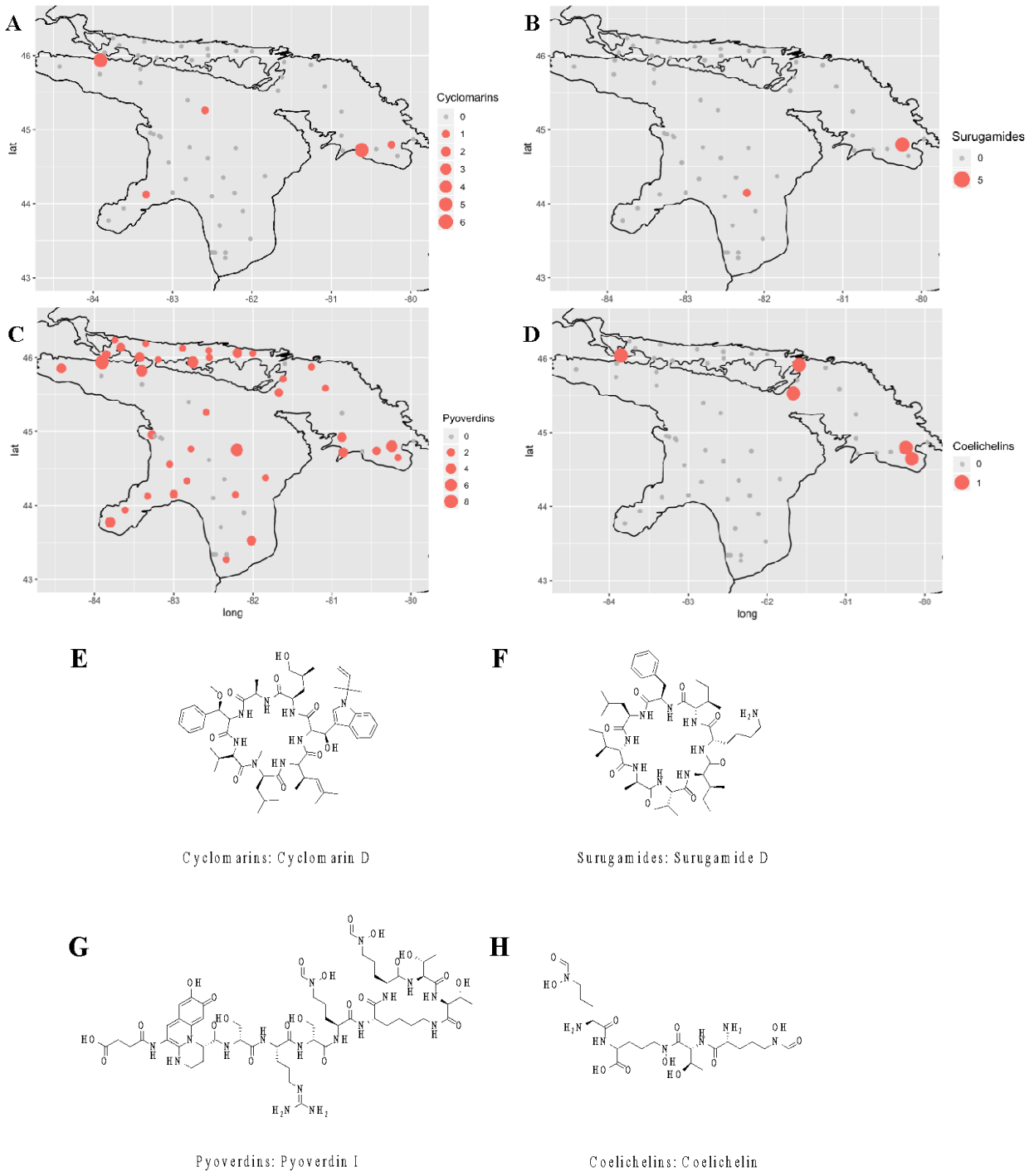


Figure 2. Detection of domain sequences of select NP classes in Lake Huron sediment. Figures 2A-D show the detection and relative read abundance of cyclomarin, surugamide, pyoverdin, and coelichelin classes, respectively. Figures S3-5 depict the distribution of additional NP classes. Different sized circles represent sequence read abundance at a rarefaction depth of 13 sequences per sample for KSα domain sequences and of 3,487 sequences per sample for A domain sequences at each collection site in Lake Huron. Representative structures from each of the four compound classes are shown in Figures 2E-H.

6.2.3 Analysis of Uncharacterized NP BGC Distribution in Lake Sediment

The majority of OBUs detected in Lake Huron sediment were not assigned to known compound classes (98.3% KSα domain OBUs and 99.9% of A domain OBUs, respectively). Instead of constructing maps for all 90,528 uncharacterized A domain OBUs, the number of locations at which a given OBU was detected was plotted (Figure 3). This allowed determination of the frequency of occurrence of OBUs across lake sediment. Figure 3 demonstrates that the vast majority of A domain OBUs (96.5%) occurred in fewer than 10 samples (in varying occurrence patterns, data not shown), across the 58 locations. For example, 40,003 OBUs (83.7%) were detected in only a single sediment location, and 2,524 OBUs (5.3%) were detected in only two locations (in varying occurrence patterns). However, no more than 1,042 OBUs were detected at any single sampling site (Figure 4); thus, the genetic diversity detected is broadly distributed. Figures 3 and 4 together demonstrate that there is little overlap among occurrence patterns of these OBUs, indicating that there are not select NP ‘hotspots’ among our 58 sampling sites and that NP occurrence varies considerably across Lake Huron sediment.

Figure 3

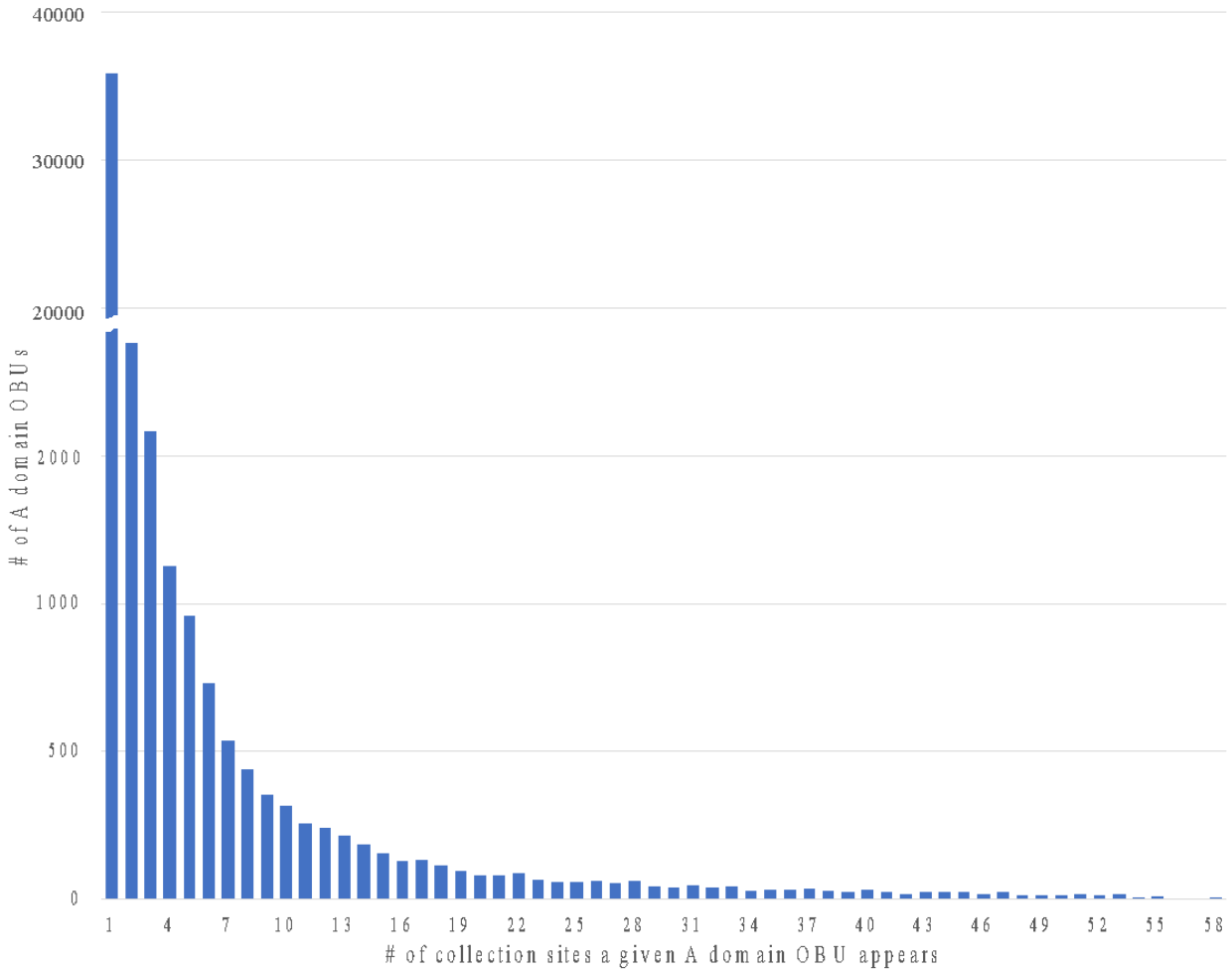


Figure 3. The number of locations from which an A domain OBU occurs. The majority of A domain OBUs occur in fewer than ten locations.

Figure 4

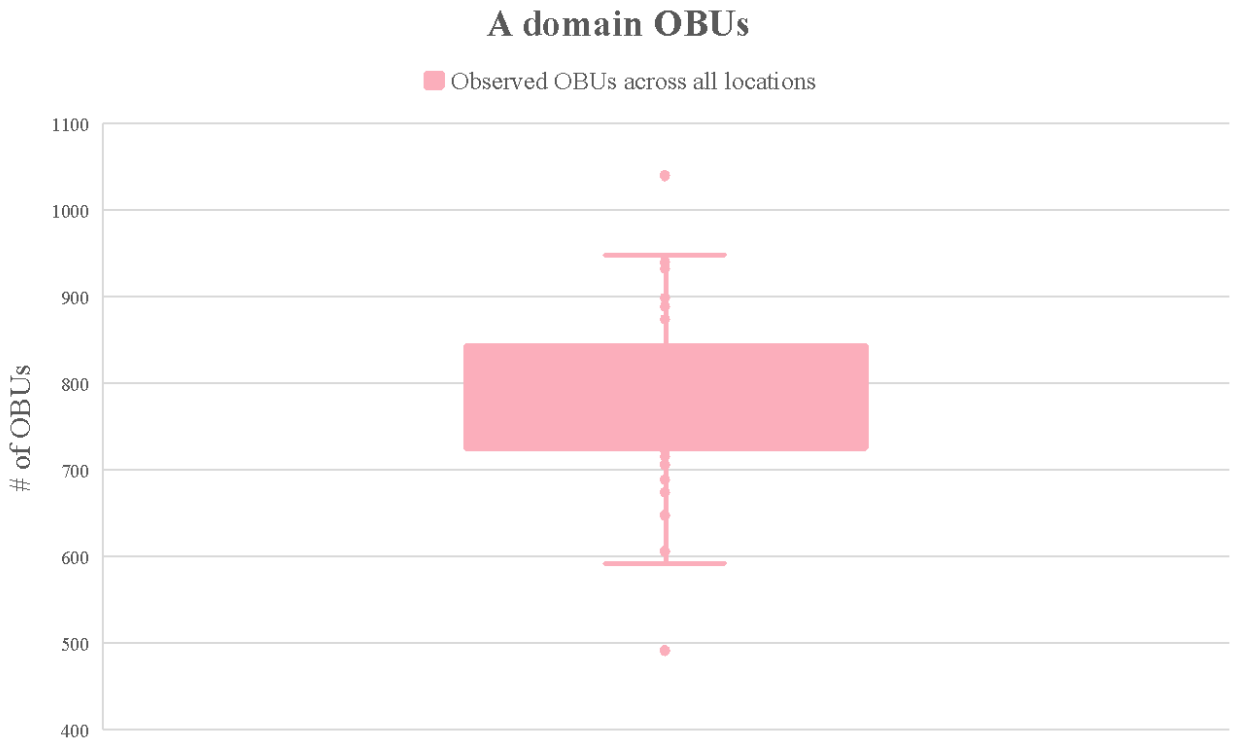


Figure 4. Box plot indicating observed number of A domain OBUs across all locations at a rarefaction depth of 3,487; each point represents the number of A domain OBUs at that location. The number of A domain OBUs that appears in each location ranges from 491 to 1,042.

We sought to determine whether A domain OBUs were likely to co-occur in the environment. Correlation coefficients based on presence/absence and abundance were calculated for each OBU pair for the 1,000 most abundant OBUs from rarified BIOM tables, based on the formula in Supp. Table S7. Among these, 0.16% of OBUs displayed a strong positive correlation with each other (a correlation score of 0.9 or above). This analysis further supports the lack of co-occurrence of dominant OBU classes in these sediments. Similarly, correlation analyses were undertaken to assess whether A domain OBUs correlated with the presence of specific Actinobacteria or Proteobacteria OTUs at each location (see Annexure B, Supplemental Information Table S7). No significant correlations were observed. One possible cause of this may be that the detected OBUs are associated with mobile genetic elements and therefore are associated with multiple taxa.(Penn et al., 2009)Alternatively, primer biases (OBU versus OTU) coupled with insufficient OTU sequencing depth prevented sufficient detection of the necessary sequences needed to observe such correlations. Further experiments using shotgun metagenome sequencing will be required to confirm this result.

This study aimed to generate a preliminary assessment of how NP OBUs are distributed across Lake Huron sediment. As shown in Figure 2 (and Annexure B, Supp. Figures S3-S5), among the select 30 characterized OBUs that were analyzed, no discernable patterns of occurrence in Lake Huron surface sediment were observed. Some NP OBUs exhibited frequent occurrences in sediment across the geographic locations sampled, while others were confined to select sample sites.

This study is one of the few attempts to document the distribution of specific classes of NPs at a regional scale in an environment representative of a collection expedition.(Charlop-Powers et al., 2016) The observed NP distribution profiles lend experimental evidence to a few predictable phenomena, that to the best of our knowledge have seldom been demonstrated on a large scale. First, individual compound profiles, particularly those that represent bioactive NPs (antibiotics, anticancer, etc), exhibit sparse occurrence across Lake Huron sediment. Second, some profiles occur more frequently across the collection sites, such as the pyoverdins and griseorhodins (Figure 2C and Annexure B, Supp. Figure S5H). This may suggest that the NP is highly functional in its environment or is located on a mobile genetic element that is commonly transferred between species, among other possibilities. Regardless, of the greater than 90,000 known and uncharacterized NP OBUs analyzed, there is little evidence for discernable patterns of NP occurrence across Lake Huron sediment. This suggests that robust sampling is required to survey an environment of this magnitude, and that oversampling leading to redundant NP recovery is not a major concern (though cultivation methods will be a significant factor in recovering those NP populations from sediment).[23] Further experiments should be performed to assess whether the OBU distribution trend observed in this study is also detected in the culturable bacterial population, a metric more appropriate to evaluate the efficiency of most microbial drug discovery programs. An attempt to document OBU recovery from culturable bacterial populations was addressed in other complementary studies.(Bech et al., 2020; Elfeki et al., 2018) A similar study that analyzes sequences within an area of higher geographic resolution, at multiple time points, and with consideration toward environmental pressures specific to the benthic lake environment, would provide more detailed information on the available NP chemical space in Lake Huron sediment. Events such as algal blooms or other localized environmental phenomena at the time of collection can influence results in any of the sampled locations.

The need for novel approaches to improve detection of NP BGCs from eDNA

There are a few experimental limitations to the current study (see SI for more detailed explanation). First, the low abundance of sequence reads belonging to NPs can be attributed to undersampling, limited eDNA extracted from sediment, or 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. The design of new primers with a broader detection range can improve discovery of non-traditional BGCs. However, alternative, non-PCR-based approaches such as deep shotgun metagenome sequencing coupled with long-read sequence data (*e.g.* data generated using Oxford Nanopore and Pacific Biosciences sequencing platforms), or enrichment strategies followed by deep sequencing (*e.g.*, Oxford Nanopore selective sequencing,(Edwards et al., 2019) hybridization capture+shotgun metagenome sequencing(Zhou et al., 2015)) will be necessary for further discovery. Finally, the MIBiG database was used to assess compound classes.(Kautsar et al., 2020, p. 2) The number of existing NPs greatly outnumbers the entries in MIBiG, underlining the need for the community to contribute to and expand this valuable resource.

6.3 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, KSα and A domain amplicon sequencing was used to document distribution profiles of NPs across Lake Huron surface sediment. Overall, no discernable NP geographic distribution patterns were observed when comparing OBUs from greater than 90,000 NP classes (NRPS and PKS). We observed that the distribution profiles of the majority of A domain OBUs were non-overlapping across the 58 locations, while each location harbored relatively equal number of OBUs, suggesting that at the sequencing depth used in this study, no single location served as a NP ‘hotspot’. Finally,

analysis of the top 1,000 most abundant OBUs detected in Lake Huron sediment, which belong to unknown/uncharacterized NPs, indicate that co-occurrence patterns are rare, but do exist. This preliminary evidence supports that there is ample variation in NP occurrence between sampling sites and suggests that extensive sample collection efforts will be required to fully capture the diversity that exists in sediment on a regional scale. Overall, investigating BGC distribution patterns and dynamics in Lake Huron has highlighted the need for a more methodical environmental sample collection approach, a great unmet need in NP drug discovery.

6.4 Methods

6.4.1 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 Supp. Table S1. Approximately 1 cm^3 of sediment was homogenized, and an aliquot was placed into a 2 mL cryovial containing 20% glycerol. These were stored in cryogenic vials in a Dewar until transported back to the laboratory where they were stored in a -20°C freezer.

6.4.2 Genomic DNA Isolation from Sediment and Nutrient Agar

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

6.4.3 KSα and A Domain Amplification and Sequencing

KSα and A domain amplicon sequencing was performed using the same two-step PCR strategy described in the Supporting Information. Briefly, a 613 bp fragment of the KSα (β-ketoacyl synthase) was amplified using degenerate primers (5′-TSGCSTGCTTCGAYGCSATC-3′) and (5′-TGGAANCCGCCGAABCCGCT-3′).(Metsä-Ketelä

et al., 1999) 700-bp NRPS A domain gene fragments were amplified using degenerate oligonucleotides A3F (5′-GCSTACSYSATSTACACSTCSGG-3′) and A7R (5′SASGTCVCCSGTSCGGTAS-3′).(Ayuso-Sacido and Genilloud, 2005) All primers were synthesized with 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 at 10 ng/µL, 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 verified by agarose gel electrophoresis and purified using a QIAquick PCR cleanup kit according to the manufacturer's protocol (Qiagen). 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. Sequencing was performed at the Genome Research Core at the University of Illinois at Chicago.

6.4.4 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 trimmed on the ends of the read according to Phred quality scores, then denoised using the DADA2 implemented in Qiime2, and finally chimeras were removed using uchime-denovo as implemented in Qiime2.(Bolyen et al., 2019, p. 2; Callahan et al., 2016, p. 2) The degenerate primer sequences were removed. Filtered and trimmed reads were then 6-frame translated into amino acid sequences using TranslatorX.(Abascal et al., 2010) 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(Johnson et al., 2010) using HMM prebuilt generic detection models downloaded from antiSMASH v5.0.0.(Blin et al., 2019) 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, resulting in a much lower number of sequences per sample (Supp. Table S3). Sequences were then clustered at 80%, 85%, 90%, and 95% using USEARCH v11's UCLUST cluster_fast greedy algorithm via the cluster_fast command. Singletons were kept for the clustering.(Edgar, 2010) A feature-by-sample abundance matrix (a feature table or biological observation matrix, BIOM)(McDonald et al., 2012) file was then created. A representative sequence from each cluster—labeled an OBU—was extracted to a separate file, using the USEARCH v11's makeudb_usearch command, and the file was aligned against the MIBiG database using DIAMOND.(Buchfink et al., 2015) Sequence reads belonging to the same molecular class clustered best at 85%. Therefore, the 85% sequence similarity threshold was used for subsequent analyses. An 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. Singletons were retained during OBU clustering. Since OBU clustering occurred at 85% (as opposed to the single nucleotide/ASV level), changes in a single nucleotide OBU diversity are not expected to change the richness of the sample. In addition, to ensure that singletons were real sequences and not PCR error, 10 singletons from each domain were blasted against the NCBI's protein database, and all singletons mapped to the correct group (A and KSα domains).

6.4.5 Bioinformatic Method Validation Using Reference Strains

Control strain *S. coelicolor* A3(2) was included in wet lab and bioinformatics analysis to ensure clustering methods and compound identities were valid. *S. coelicolor* A3(2) was subjected to the same amplification procedure using the degenerate primers that amplify a fragment of the KSα (β-ketoacyl synthase) and a fragment of the A domain. KSα and A

domain amplicons were sequenced and analyzed using the strategy described in sections 4.3 and 4.4. Resulting sequence data were then filtered using the same HMM prebuild generic detection models described above. Sequences that passed the default e-value thresholds were kept. Sequences were then clustered at 80% 85%, 90%, and 95%. At 85%, KSα amplicons grouped into two OBUs. A representative sequence from each OBU was mapped against MIBIG for compound identification. Indeed, the representative sequence from the first OBU mapped to actinorhodin and the representative sequence from the second OBU mapped to a spore pigment, as expected. Similarly, at 85%, A domain amplicons grouped into 15 OBUs. After mapping against MIBIG, a representative sequence from 10 OBUs mapped to the CDA family of compounds (CDA1b, CDA2a, CDA2b, CDA3a, CDA3b, CDA4a, CDA4b), one representative sequence mapped to coelimycin P1, one representative sequence mapped to coelibactin, and three representative sequences mapped to coelichelin. There were no OBU representative sequences mapped to the remaining A domain containing compounds undecylprodigiosin and SCO-2138.

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. This publication was supported by Vahlteich Scholar research funds, the Illinois–Indiana Sea Grant, the Office of Technology Management at UIC, UIC startup funds, and funding from National Institutes of Health grant R01 GM125943.

6.5 Reference

Abascal, F., Zardoya, R., Telford, M.J., 2010. TranslatorX: Multiple alignment of nucleotide sequences guided by amino acid translations. Nucleic Acids Research 38, 7–13. https://doi.org/10.1093/nar/gkq291

Adamek, M., Alanjary, M., Ziemert, N., 2019. Applied evolution: Phylogeny-based approaches in natural products research. Natural Product Reports 36, 1295–1312. https://doi.org/10.1039/C9NP00027E

Aldrich, S., 1999. Alexander Fleming Discovery and Development of Penicillin - Landmark - American Chemical Society. American Chemical Society International Historic Chemical Landmarks. https://doi.org/10.2307/3561468

Ayuso-Sacido, A., Genilloud, O., 2005. New PCR primers for the screening of NRPS and PKS-I systems in actinomycetes: Detection and distribution of these biosynthetic gene sequences in major taxonomic groups. Microbial Ecology 49, 10–24. https://doi.org/10.1007/s00248-004-0249-6

Bech, P.K., Lysdal, K.L., Gram, L., Bentzon-Tilia, M., Strube, M.L., 2020. Marine Sediments Hold an Untapped Potential for Novel Taxonomic and Bioactive Bacterial Diversity. mSystems 5. https://doi.org/10.1128/mSystems.00782-20

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., Bateman, A., Brown, S., Chandra, G., Chen, C.W., Collins, M., Cronin, A., Fraser, A., Goble, A., Hidalgo, J., Hornsby, T., Howarth, S., Huang, C.H., Kieser, T., Larke, L., Murphy, L., Oliver, K., O'Neil, S., Rabbinowitsch, E., Rajandream, M.A., Rutherford, K., Rutter, S., Seeger, K., Saunders, D., Sharp, S., Squares, R., Squares, S., Taylor, K., Warren, T., Wietzorrek, A., Woodward, J., Barrell, B.G., Parkhill, J., Hopwood, D.A., 2002. Complete genome sequence of the model actinomycete Streptomyces coelicolor A3(2). Nature 417, 141–147. https://doi.org/10.1038/417141a

Bolyen, E., Rideout, J.R., Dillon, M.R., Bokulich, N.A., Abnet, C.C., Al-Ghalith, G.A., Alexander, H., Alm, E.J., Arumugam, M., Asnicar, F., Bai, Y., Bisanz, J.E., Bittinger, K., Brejnrod, A., Brislawn, C.J., Brown, C.T., Callahan, B.J., Caraballo-Rodríguez, A.M., Chase, J., Cope, E.K., Da Silva, R., Diener, C., Dorrestein, P.C., Douglas, G.M., Durall, D.M., Duvallet, C., Edwardson, C.F., Ernst, M., Estaki, M., Fouquier, J., Gauglitz, J.M., Gibbons, S.M., Gibson, D.L., Gonzalez, A., Gorlick, K., Guo, J., Hillmann, B., Holmes, S., Holste, H., Huttenhower, C., Huttley, G.A., Janssen, S., Jarmusch, A.K., Jiang, L., Kaehler, B.D., Kang, K.B., Keefe, C.R., Keim, P., Kelley, S.T., Knights, D., Koester, I., Kosciolek, T., Kreps, J., Langille, M.G.I., Lee, J., Ley, R., Liu, Y.-X., Loftfield, E., Lozupone, C., Maher, M., Marotz, C., Martin, B.D., McDonald, D., McIver, L.J., Melnik, A.V., Metcalf, J.L., Morgan, S.C., Morton, J.T., Naimey, A.T., Navas-Molina, J.A., Nothias, L.F., Orchanian, S.B., Pearson, T., Peoples, S.L., Petras, D., Preuss, M.L., Pruesse, E., Rasmussen, L.B., Rivers, A., Robeson, M.S., Rosenthal, P., Segata, N., Shaffer, M., Shiffer, A., Sinha, R., Song, S.J., Spear, J.R., Swafford, A.D., Thompson, L.R., Torres, P.J., Trinh, P., Tripathi, A., Turnbaugh, P.J., Ul-Hasan, S., van der Hooft, J.J.J., Vargas, F., Vázquez-Baeza, Y., Vogtmann, E., von Hippel, M., Walters, W., Wan, Y., Wang, M., Warren, J., Weber, K.C., Williamson, C.H.D., Willis, A.D., Xu, Z.Z., Zaneveld, J.R., Zhang, Y., Zhu, Q., Knight, R., Caporaso, J.G., 2019. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nat Biotechnol 37, 852–857. https://doi.org/10.1038/s41587-019-0209-9

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., 2019. Microbial community drivers of PK/NRP gene diversity in selected global soils. Microbiome 7, 78. https://doi.org/10.1186/s40168-019-0692-8

Buchfink, B., Xie, C., Huson, D.H., 2015. Fast and sensitive protein alignment using DIAMOND. Nature Methods 12, 59–60. https://doi.org/10.1038/nmeth.3176

Callahan, B.J., McMurdie, P.J., Rosen, M.J., Han, A.W., Johnson, A.J.A., Holmes, S.P., 2016. DADA2: High-resolution sample inference from Illumina amplicon data. Nature Methods 13, 581–583. https://doi.org/10.1038/nmeth.3869

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., 2015. Global biogeographic sampling of bacterial secondary metabolism. eLife 2015, e05048. https://doi.org/10.7554/eLife.05048

Charlop-Powers, Z., Pregitzer, C.C., Lemetre, C., Ternei, M.A., Maniko, J., Hover, B.M., Calle, P.Y., McGuire, K.L., Garbarino, J., Forgione, H.M., Charlop-Powers, S., Brady, S.F., 2016. Urban park soil microbiomes are a rich reservoir of natural product biosynthetic diversity. PNAS 113, 14811–14816. https://doi.org/10.1073/pnas.1615581113

Cheng, K., Rong, X., Pinto-Tomás, A.A., Fernández-Villalobos, M., Murillo-Cruz, C., Huang, Y., 2015. Population genetic analysis of Streptomyces albidoflavus reveals habitat barriers to homologous recombination in the diversification of streptomycetes. Applied and Environmental Microbiology 81, 966–975. https://doi.org/10.1128/AEM.02925-14

Clardy, J., Fischbach, M.A., Currie, C.R., 2009. The natural history of antibiotics. Current Biology 19, R437–R441. https://doi.org/10.1016/j.cub.2009.04.001

Du, D., Katsuyama, Y., Shin-ya, K., Ohnishi, Y., 2018. Reconstitution of a Type II Polyketide Synthase that Catalyzes Polyene Formation. Angewandte Chemie 130, 1972–1975. https://doi.org/10.1002/ange.201709636

Edgar, R.C., 2010. Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26, 2460–2461. https://doi.org/10.1093/bioinformatics/btq461

Edwards, H.S., Krishnakumar, R., Sinha, A., Bird, S.W., Patel, K.D., Bartsch, M.S., 2019. Real-Time Selective Sequencing with RUBRIC: Read Until with Basecall and Reference-Informed Criteria. Scientific Reports 9, 11475. https://doi.org/10.1038/s41598-019-47857-3

Elfeki, M., Alanjary, M., Green, S.J., Ziemert, N., Murphy, B.T., 2018. Assessing the Efficiency of Cultivation Techniques To Recover Natural Product Biosynthetic Gene Populations from Sediment. ACS Chemical Biology 13, 2074–2081. https://doi.org/10.1021/acschembio.8b00254

Emmerich, R., Löw, O., 1899. Bakteriolytische Enzyme als Ursache der erworbenen Immunität und die Heilung von Infectionskrankheiten durch dieselben. Zeitschrift für Hygiene und Infectionskrankheiten 31, 1–65. https://doi.org/10.1007/BF02206499

Fischbach, M.A., Walsh, C.T., 2009. Antibiotics for emerging pathogens. Science 325, 1089–1093. https://doi.org/10.1126/science.1176667

Fleming A., 1929. On the antibacterial action of cultures of a penicillium,with special reference to their use in the isolation of B. influenzae. British journal of experimental pathology 10, 226–236.

Gause, G.F., Brazhnikova, M.G., 1944. Gramicidin S and its use in the treatment of infected wounds. Nature 154, 703. https://doi.org/10.1038/154703a0

Ginolhac, A., Jarrin, C., Gillet, B., Robe, P., Pujic, P., Tuphile, K., Bertrand, H., Vogel, T.M., Perrière, G., Simonet, P., Nalin, R., 2004. Phylogenetic analysis of polyketide synthase I domains from soil metagenomic libraries allows selection of promising clones. Applied and Environmental Microbiology 70, 5522–5527. https://doi.org/10.1128/AEM.70.9.5522-5527.2004

Hernandez, A., T. Nguyen, L., Dhakal, R., T. Murphy, B., 2021. The need to innovate sample collection and library generation in microbial drug discovery: a focus on academia. Natural Product Reports. https://doi.org/10.1039/D0NP00029A

Johnson, L.S., Eddy, S.R., Portugaly, E., 2010. Hidden Markov model speed heuristic and iterative HMM search procedure. BMC Bioinformatics 11, 1471–2105. https://doi.org/10.1186/1471-2105-11-431

Kautsar, S.A., Blin, K., Shaw, S., Navarro-Muñoz, J.C., Terlouw, B.R., van der Hooft, J.J.J., van Santen, J.A., Tracanna, V., Suarez Duran, H.G., Pascal Andreu, V., Selem-Mojica, N., Alanjary, M., Robinson, S.L., Lund, G., Epstein, S.C., Sisto, A.C., Charkoudian, L.K., Collemare, J., Linington, R.G., Weber, T., Medema, M.H., 2020. MIBiG 2.0: a repository for biosynthetic gene clusters of known function. Nucleic Acids Res 48, D454–D458. https://doi.org/10.1093/nar/gkz882

Lemetre, C., Maniko, J., Charlop-Powers, Z., Sparrow, B., Lowe, A.J., Brady, S.F., 2017. Bacterial natural product biosynthetic domain composition in soil correlates with changes in latitude on a continent-wide scale. Proceedings of the National Academy of Sciences of the United States of America 114, 11615–11620. https://doi.org/10.1073/pnas.1710262114

Liu, L., Salam, N., Jiao, J.Y., Jiang, H.C., Zhou, E.M., Yin, Y.R., Ming, H., Li, W.J., 2016. Diversity of Culturable Thermophilic Actinobacteria in Hot Springs in Tengchong, China and Studies of their Biosynthetic Gene Profiles. Microbial Ecology. https://doi.org/10.1007/s00248-016-0756-2

McDonald, D., Clemente, J.C., Kuczynski, J., Rideout, J.R., Stombaugh, J., Wendel, D., Wilke, A., Huse, S., Hufnagle, J., Meyer, F., Knight, R., Caporaso, J.G., 2012. The Biological Observation Matrix (BIOM) format or: How I learned to stop worrying and love the ome-ome. GigaScience 464, 1–6. https://doi.org/10.1186/2047-217X-1-7

Metsä-Ketelä, M., Salo, V., Halo, L., Hautala, A., Hakala, J., Mäntsälä, P., Ylihonko, K., 1999. An efficient approach for screening minimal PKS genes from *Streptomyces*. FEMS Microbiology Letters 180, 1–6. https://doi.org/10.1016/S0378-1097(99)00453-X

Naqib, A., Poggi, S., Wang, W., Hyde, M., Kunstman, K., Green, S.J., 2018. Making and sequencing heavily multiplexed, high-throughput 16S ribosomal RNA gene amplicon libraries using a flexible, two-stage PCR protocol. Methods in molecular biology (Clifton, N.J.) 1783, 149–169. https://doi.org/10.1007/978-1-4939-7834-2_7

Newman, D.J., Cragg, G.M., 2020. Natural Products as Sources of New Drugs over the Nearly Four Decades from 01/1981 to 09/2019. Journal of Natural Products 83, 770–803. https://doi.org/10.1021/acs.jnatprod.9b01285

Newton, R.J., Jones, S.E., Eiler, A., McMahon, K.D., Bertilsson, S., 2011. A Guide to the Natural History of Freshwater Lake Bacteria. Microbiology and Molecular Biology Reviews 75, 14–49. https://doi.org/10.1128/MMBR.00028-10

Penn, K., Jenkins, C., Nett, M., Udwary, D.W., Gontang, E.A., McGlinchey, R.P., Foster, B., Lapidus, A., Podell, S., Allen, E.E., Moore, B.S., Jensen, P.R., 2009. Genomic islands link secondary metabolism to functional adaptation in marine Actinobacteria. The ISME Journal 3, 1193–1203. https://doi.org/10.1038/ismej.2009.58

Scholz, M., Ward, D. V., Pasolli, E., Tolio, T., Zolfo, M., Asnicar, F., Truong, D.T., Tett, A., Morrow, A.L., Segata, N., 2016. Strain-level microbial epidemiology and population genomics from shotgun metagenomics. Nature Methods 13, 435–438. https://doi.org/10.1038/nmeth.3802

Sharrar, A.M., Crits-Christoph, A., Méheust, R., Diamond, S., Starr, E.P., Banfield, J.F., 2020. Bacterial Secondary Metabolite Biosynthetic Potential in Soil Varies with Phylum, Depth, and Vegetation Type. mBio 11. https://doi.org/10.1128/mBio.00416-20

Weber, T., Kim, H.U., 2016. The secondary metabolite bioinformatics portal: Computational tools to facilitate synthetic biology of secondary metabolite production. Synthetic and Systems Biotechnology 1, 69–79. https://doi.org/10.1016/j.synbio.2015.12.002

Zhou, J., He, Z., Yang, Y., Deng, Y., Tringe, S.G., Alvarez-Cohen, L., 2015. High-Throughput Metagenomic Technologies for Complex Microbial Community Analysis: Open and Closed Formats. mBio 6. https://doi.org/10.1128/mBio.02288-14

Part II: Dynamics of the human gut secondary metabolome during antibiotic treatment.

Abstract

Antibiotic-mediated perturbation of the human gut microbiome increases the risk of a variety of diseases, including infections (Willing, Russell, and Finlay 2011). The increased infection risk is the result of unrestrained proliferation of opportunistic microbes that may occupy ecological niches previously unavailable to them (De La Cochetière et al. 2008). Bacterial secondary metabolites are known to play crucial roles in microbe-microbe and microbe-host interactions (Ziemert, Alanjary, and Weber 2016). However, not much is known about their antimicrobial role within the human gut. The aim of this project was to monitor changes within the secondary metabolome potential of the gut microbiome in the course of antibiotic treatment to identify bacterial biosynthetic gene clusters and metabolites with a potential role in community stabilization and host defense. In this exploratory study we have determined patterns, identified crucial pathways, and built solid hypotheses for testing in wet lab experiments.

6.6 Introduction

Secondary metabolites (SM) are fundamental units with which microbes sense and respond to their environment. SM key functions include microbial communication, defense, nutrient acquisition and development (O'Brien and Wright 2011). In the past, these molecules have been mainly studied as natural products for their use as antibiotics or chemotherapeutics. However, the availability of improved genomic data as well as of more sensitive detection methods and increasing insights into biological systems recently uncovered a role for SM in a wide range of symbiotic relationships not only among microbes but also in connection to their multicellular hosts (Donia and Fischbach 2015). In this respect, multiple genome mining efforts revealed the presence of a multitude of secondary metabolite biosynthetic gene clusters (BGCs) within the human microbiome (Donia and Fischbach 2015). Furthermore,

various bioactive metabolites have been isolated from human-associated microbes, leading to question their role within the human microbiome and their relevance in microbe-microbe and microbe-host interactions (Sugimoto et al. 2019; Zipperer et al. 2016). As antibiotic treatment alters the composition of the human gut microbiota and increases the risk of infections, for example with Clostridium difficile (De La Cochetière et al. 2008), in this study we wanted to understand how the use of antibiotics affects the secondary metabolome of human gut commensal bacteria and uncover a potential role for bacterial SM in stabilizing microbial communities and preventing infections.

6.7 Methods

The group of Matthias Willmann and Silke Peter analyzed two clinical cohorts in Tuebingen and Cologne and gained shotgun metagenomic data of 41 hematological patients before, during, and after prophylactic treatment with ciprofloxacin and cotrimoxazole. This clinical study - Amplification and Selection of Antimicrobial Resistance in the Intestine (ASARI)- showed that antibiotic treatment had major effects on the diversity of the human microbiome and resistome and provides the basis for the proposed project (Willmann et al. 2019). During the last years one of the main research areas in the Ziemert lab has been the determination of SM biosynthetic potential in environmental bacteria (Ziemert et al. 2012; Elfeki et al. 2018). For that purpose, we developed the pipeline MBEZ, which performs a standardized analysis of the distribution and diversity of secondary metabolite gene clusters in various kinds of metagenomic data, including shotgun metagenomes and amplicon sequences. MBEZ implements programs such as QIIME II, BiG-MEx, antiSMASH, and BiGSCAPE, and allows a fast and reproducible screen of metagenomic data for secondary metabolite diversity, bacterial taxonomic diversity and correlations between the two. Within the framework of this project, MBEZ was employed to analyze secondary metabolite patterns within human microbiome data in order to unravel BGC abundance dynamics during the course of antibiotic treatment. Briefly, the metagenomic contigs were analysed with antiSMASH for annotating BGCs. The detected BGCs were clustered with known MIBiG

BGCs using BiGSCAPE. The BGC abundance plots were developed using the ggplot2 package.

6.8 Results and Discussion

A preliminary analysis of the ASARI data revealed the distinctive presence of biosynthetic gene clusters (BGCs) in each patient sample (Table 1). Interestingly, many of these BGCs have so far remained uncharacterized as only a few BGCs clustered with the known BGCs in MIBiG database (Table 2). In the ciprofloxacin treatment cohort, BGC abundance across treatment stages showed a decreasing trend (Figure 1). BGC abundance across cotrimoxazole treatment stages was comparatively constant (Figure 2). At treatment stage T0 and T2 the BGC abundance across cotrimoxazole and ciprofloxacin cohorts showed differential abundance (Figure 3). Comparative view of BGC biosynthesis class abundance across treatment stages in cotrimoxazole and ciprofloxacin cohorts is shown in Figure 4.

Furthermore, a pattern analysis across the different time points of ciprofloxacin treatment revealed a major decrease in the abundance of sactipeptide gene clusters within the microbial communities (Figure 5). Magnitude of decrease in abundance of sactipeptide gene clusters was not that prominent in the metagenomic dataset of the cotrimoxazole treated cohort as compared to the ciprofloxacin treated cohort (Figure 6).

Sactipeptides are ribosomally assembled and posttranslationally modified natural product peptides that currently consist of five members (Flühe and Marahiel 2013). Interestingly, some members of this class show a narrow antimicrobial activity against Clostridium difficile (Flühe and Marahiel 2013), thus providing a possible explanation for the increased susceptibility to this bacterial pathogen after antibiotic treatment.

	RIPPs	Others	NRPS	PKS-NRPS-Hybrids	PKS-others	Terpenes	PKSI
Number of families	795	363	766	18	48	67	22
Average number of BGCs per family	5	5	3	3	2	4	2
Max number of BGCs in a family	67	76	46	15	15	32	7
Families with MIBiG Reference BGCs	8	2	5	3	0	0	1

Table1 : BGC class annotation and abundance based on BiG-SCAPE clustering

Table 2: Know BGC clusters detected in ASARI metagenome

MiBIG ID	**BGC Name**
BGC0000624.1:	Salivaricin CRL1328 alpha peptide / salivaricin CRL1328 beta peptide biosynthetic gene cluster
BGC0000619.1:	Gassericin T biosynthetic gene cluster
BGC0001602.1:	Gassericin-T biosynthetic gene cluster
BGC0001388.1:	Gassericin E biosynthetic gene cluster
BGC0000526.1:	Macedocin biosynthetic gene cluster
BGC0000534.1:	Mutacin K8 biosynthetic gene cluster
BGC0001788.1:	Suicin 65 biosynthetic gene cluster
BGC0000485.1:	Acidocin B biosynthetic gene cluster
BGC0000491.1:	Gassericin A biosynthetic gene cluster
BGC0001222.1:	Acidocin B biosynthetic gene cluster
BGC0000547.1:	Salivaricin 9 biosynthetic gene cluster
BGC0000545.1:	Ruminococcin A biosynthetic gene cluster
BGC0001701.1:	Nisin O biosynthetic gene cluster
BGC0001575.1:	Dipeptide aldehydes biosynthetic gene cluster
BGC0001055.1:	Yersiniabactin biosynthetic gene cluster
BGC0000972.1:	Colibactin biosynthetic gene cluster
BGC0000467.1:	Yersiniabactin biosynthetic gene cluster
BGC0001686.1:	N-octanoyl-Met-Phe-H biosynthetic gene cluster
BGC0001055.1:	Yersiniabactin biosynthetic gene cluste
BGC0001499.1:	Aerobactin biosynthetic gene cluster
BGC0001555.1:	Colicin V biosynthetic gene cluster
BGC0000836.1:	APE Ec biosynthetic gene cluster

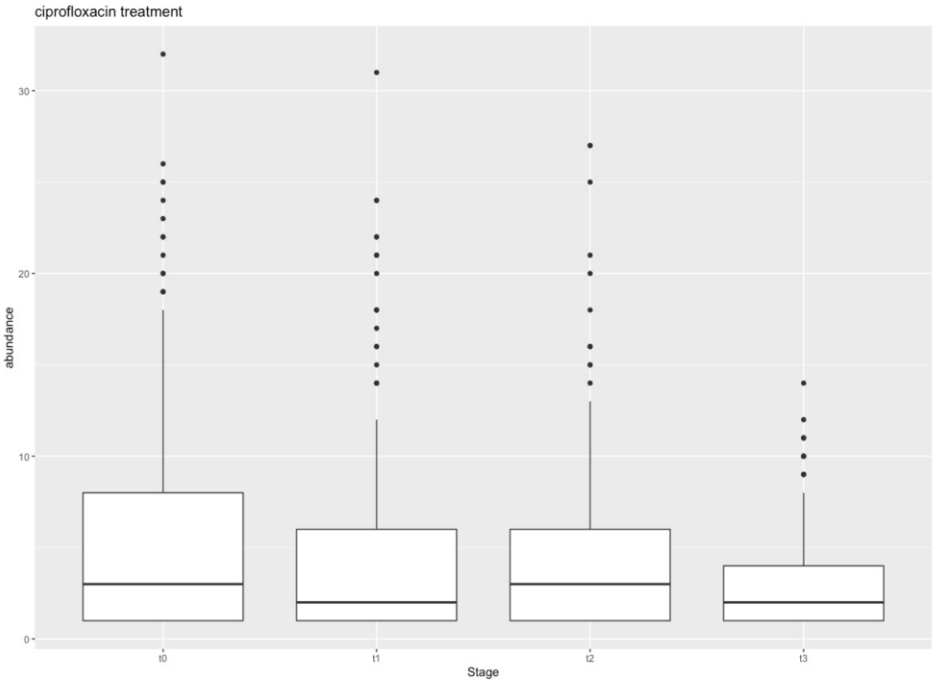


Figure 1: BGC abundance across ciprofloxacin treatment stages (Sampling stages: T0, within a maximum of 3 days before the start of antibiotic prophylaxis; T1, 1 day after initiation of prophylaxis; T2, after 3 days of prophylaxis; T3, at the end of the observation period). Individual patient assembled metagenomic contigs were annotated (BGC annotations) using antiSMASH

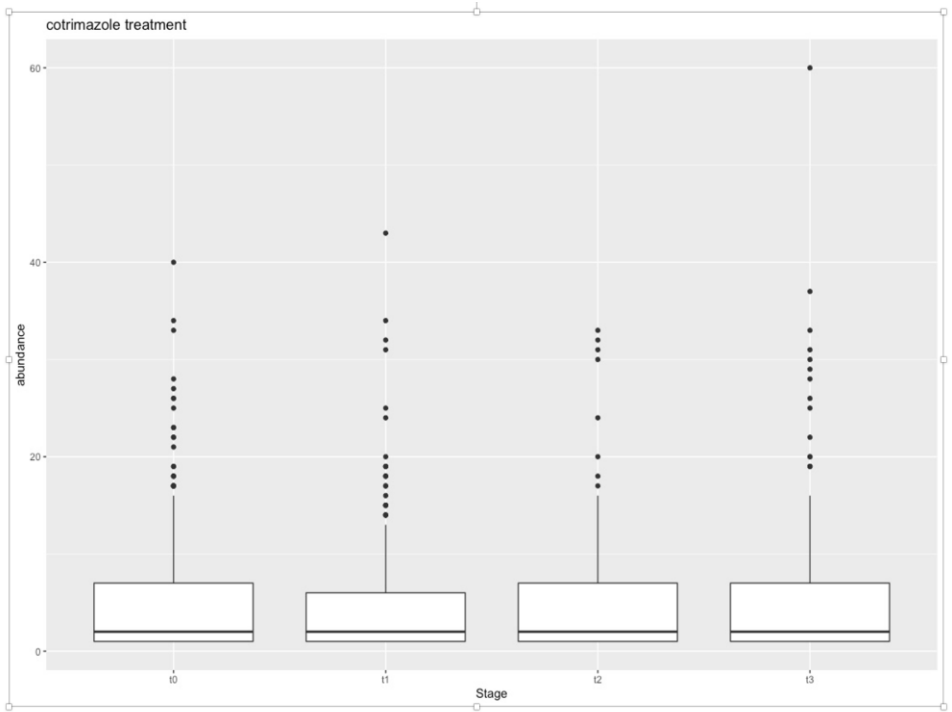


Figure 2: BGC abundance across cotrimoxazole treatment stages (Sampling stages: T0, within a maximum of 3 days before the start of antibiotic prophylaxis; T1, 1 day after initiation of prophylaxis; T2, after 3 days of prophylaxis; T3, at the end of the observation period).

Individual patient assembled metagenomic contigs were annotated (BGC annotations) using antiSMASH

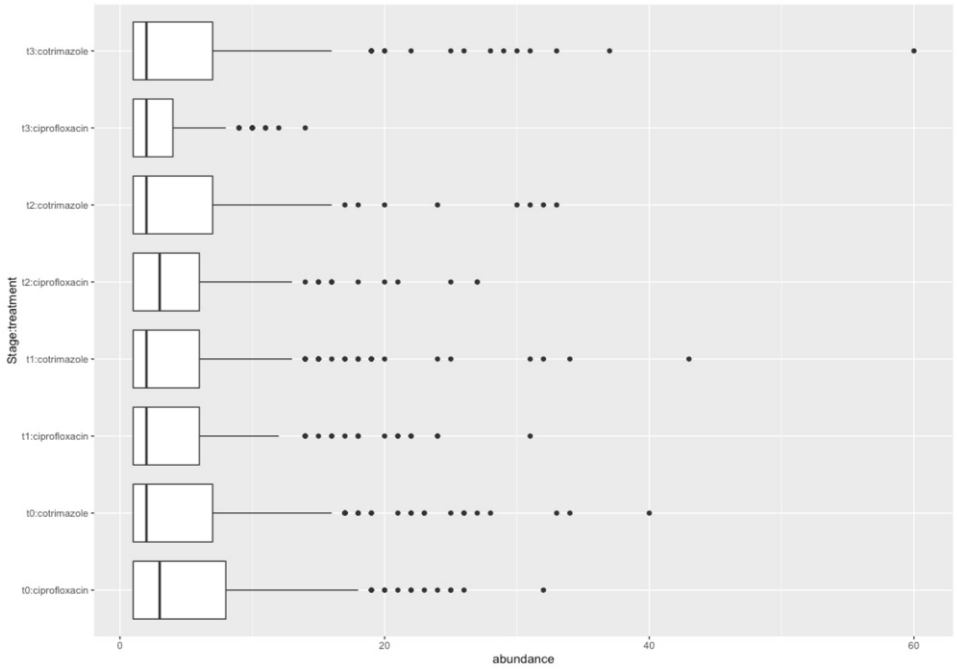


Figure 3: Comparative view of BGC abundance across cotrimoxazole and ciprofloxacin treatment stages (Sampling stages: T0, within a maximum of 3 days before the start of antibiotic prophylaxis; T1, 1 day after initiation of prophylaxis; T2, after 3 days of prophylaxis; T3, at the end of the observation period).

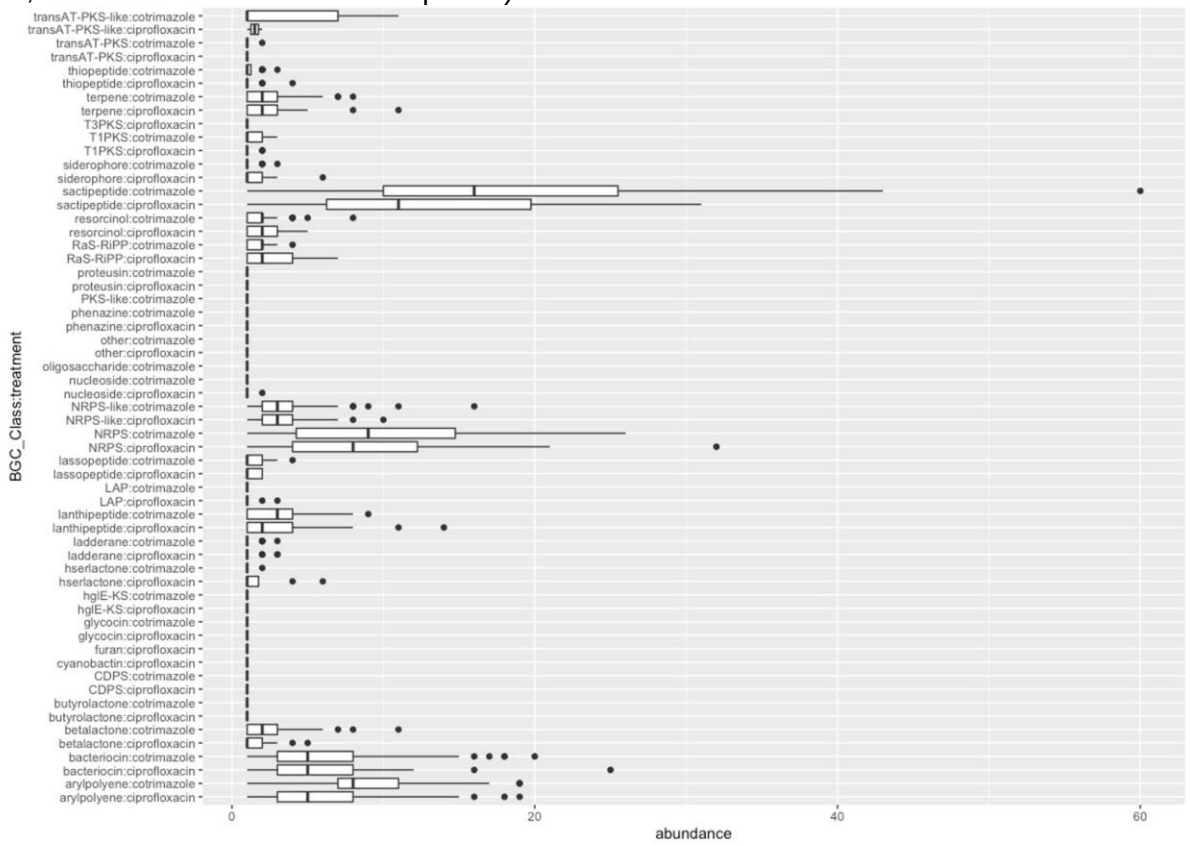


Figure 4: Comparative view of BGC biosynthesis class abundance across cotrimoxazole and ciprofloxacin treatment stages (Sampling stages: T0, within a maximum of 3 days before the

start of antibiotic prophylaxis; T1, 1 day after initiation of prophylaxis; T2, after 3 days of prophylaxis; T3, at the end of the observation period).

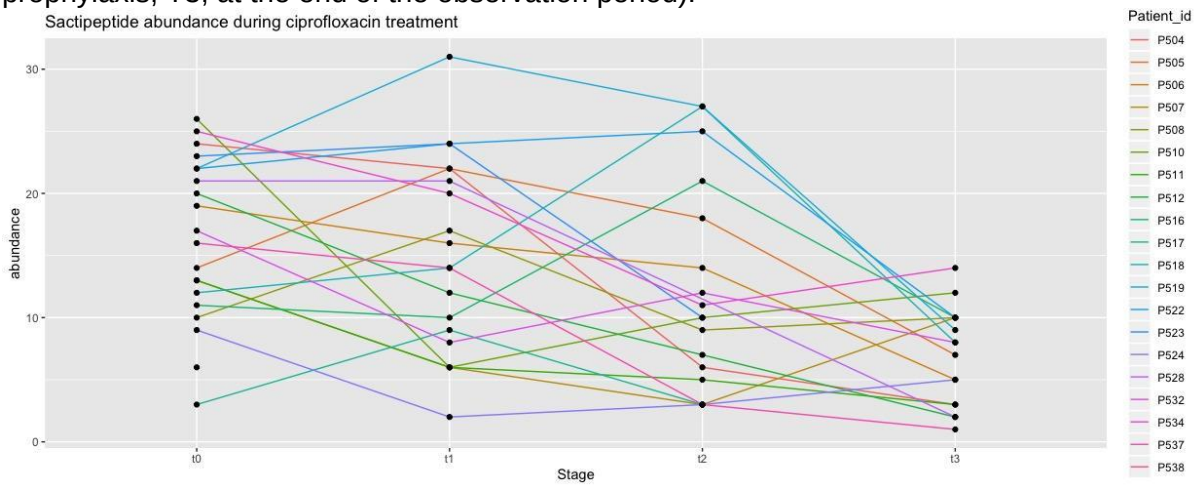


Figure 5: Sactipeptide BGC abundance in metagenomes of individual patients before, during and after ciprofloxacin treatment. (Sampling stages: T0, within a maximum of 3 days before the start of antibiotic prophylaxis; T1, 1 day after initiation of prophylaxis; T2, after 3 days of prophylaxis; T3, at the end of the observation period).

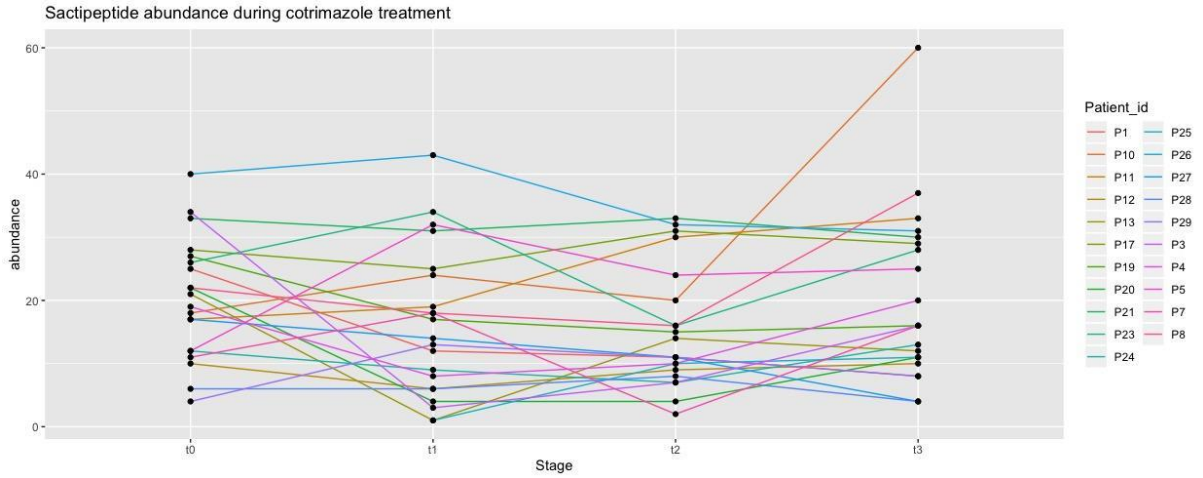


Figure 6: Sactipeptide BGC abundance in metagenomes of individual patients before, during and after cotrimoxazole treatment. (Sampling stages: T0, within a maximum of 3 days before the start of antibiotic prophylaxis; T1, 1 day after initiation of prophylaxis; T2, after 3 days of prophylaxis; T3, at the end of the observation period).

In our search for novel BGCs and NP, we have undertaken this exploration of metagenomes from human gut samples. Apart from finding many novel BGCs, as the dataset contained data from sampling and sequencing of metagenomic DNA during the course of antibiotics treatment, it became possible to observe the dynamic changes that happens over the course of antibiotics treatment. While the unique sactipeptide pattern observed in our exploration is helpful in generating new hypothesis, a follow up exploration and validation of such patterns

in a similar repeat trial or reanalysis of publicly available datasets having analogous trial design, would be necessary.

Further explorations in gut microbiome, will not only uncover the BGC diversity, but also has the huge potential for designing novel strategies for controlling microbes to fight infections.

6.9 References

De La Cochetière, M. F., T. Durand, V. Lalande, J. C. Petit, G. Potel, and L. Beaugerie. 2008. "Effect of Antibiotic Therapy on Human Fecal Microbiota and the Relation to the Development of Clostridium Difficile." *Microbial Ecology* 56 (3): 395–402. https://doi.org/10.1007/s00248-007-9356-5.

Donia, Mohamed S., and Michael A. Fischbach. 2015. "HUMAN MICROBIOTA. Small Molecules from the Human Microbiota." *Science (New York, N.Y.)* 349 (6246): 1254766. https://doi.org/10.1126/science.1254766.

Elfeki, Maryam, Mohammad Alanjary, Stefan J Green, Nadine Ziemert, and Brian T Murphy. 2018. "Assessing the Efficiency of Cultivation Techniques to Recover Natural Product Biosynthetic Gene Populations from Sediment." *ACS Chemical Biology* 13 (8): 2074–81. https://doi.org/10.1021/acschembio.8b00254.

Flühe, Leif, and Mohamed A. Marahiel. 2013. "Radical S-Adenosylmethionine Enzyme Catalyzed Thioether Bond Formation in Sactipeptide Biosynthesis." *Current Opinion in Chemical Biology* 17 (4): 605–12. https://doi.org/10.1016/j.cbpa.2013.06.031.

O'Brien, Jonathan, and Gerard D. Wright. 2011. "An Ecological Perspective of Microbial Secondary Metabolism." *Current Opinion in Biotechnology* 22 (4): 552–58. https://doi.org/10.1016/j.copbio.2011.03.010.

Sugimoto, Yuki, Francine R Camacho, Shuo Wang, Pranatchareeya Chankhamjon, Arman Odabas, Abhishek Biswas, Philip D Jeffrey, and Mohamed S Donia. 2019. "A Metagenomic Strategy for Harnessing the Chemical Repertoire of the Human Microbiome." *Science* 366 (6471): 1–17. https://doi.org/10.1126/science.aax9176.

Willing, Benjamin P., Shannon L. Russell, and B. Brett Finlay. 2011. "Shifting the Balance: Antibiotic Effects on Host-Microbiota Mutualism." *Nature Reviews. Microbiology* 9 (4): 233–43. https://doi.org/10.1038/nrmicro2536.

Willmann, Matthias, Maria J G T Vehreschild, Lena M Biehl, Wichard Vogel, Daniela Dörfel, Axel Hamprecht, Harald Seifert, Ingo B Autenrieth, and Silke Peter. 2019. "Distinct Impact of Antibiotics on the Gut Microbiome and Resistome : A Longitudinal Multicenter Cohort Study," 1–18.

Ziemert, Nadine, Mohammad Alanjary, and Tilmann Weber. 2016. "The Evolution of Genome Mining in Microbes-a Review." *Natural Product Reports* 33 (8): 988–1005. https://doi.org/10.1039/c6np00025h.

Ziemert, Nadine, Sheila Podell, Kevin Penn, Jonathan H Badger, Eric Allen, and Paul R Jensen. 2012. "The Natural Product Domain Seeker NaPDoS: A Phylogeny Based Bioinformatic Tool to Classify Secondary Metabolite Gene Diversity." *PLoS ONE* 7 (3): 1–9. https://doi.org/10.1371/journal.pone.0034064.

Zipperer, Alexander, Martin C Konnerth, Claudia Laux, Anne Berscheid, Daniela Janek, Christopher Weidenmaier, Marc Burian, et al. 2016. "Human Commensals Producing a Novel Antibiotic Impair Pathogen Colonization." *Nature* 535 (7613): 511–16. https://doi.org/10.1038/nature18634.

Part III : Using linked reads and long reads to recover biosynthetic gene clusters from Tuebingen actinomycetes strain collections.

Abstract

Current capability of next generation sequencing makes this technology suitable for deciphering the complete genomes of thousands of bacterial species in a single sequencing run. The reconstruction of bacterial genomes can be further optimized to get efficient assemblies at economical costs using hybrid data of both long and short reads. Linked short reads technologies such as Transposase Enzyme Linked Long-read Sequencing (TELL-Seq™) and 10X genomics can further improve the approach and can be used to sequence the complete strain collections. Using TELL-seq and Nanopore data generated for one such pool consisting of 115 streptomyces species from our Tuebingen Strain Collection resulted in single contigs with more than 4 megabases. Biosynthetic gene cluster (BGC) annotation of such large contigs can lead to discovery of complete sequences of novel BGCs. In this project we have sequenced and analysed 10 such pools consisting of 110 streptomyces species in each pool. Tracing of strains and BGCs of interest has become easier using this data and this has also led to accelerated novel natural product discovery using this rare and unique strain collection.

6.10 Overview and motivation

Microorganisms derived natural products and their analogs have been to date our most important source and inspiration for medically used antibiotics. The aim of the German Center for infectious research, Thematic Translation Unit: DZIF TTU9 -Novel Antibiotics, is to facilitate the discovery of new natural products as new antibiotics through innovative and effective methods including genome mining and synthetic biology. Two important prerequisites for the application of these methods are whole genome sequencing of available strain collections as a source for novel natural products, and the development of molecular tools in available strains. At the Tübingen DZIF Partner site, a unique proprietary

strain collection of more than 2000 actinomycetes producing natural compounds is available, which have been isolated over a span of 50 years by Prof. Zähner and Prof. Fiedler from a diversity of locations worldwide. Over 100 novel natural products with novel structures have been already isolated from these strains in a multitude of publications (Hennrich et al. 2020; Ortlieb 2019). So far, only about 150 strains have been sequenced, all of which show high genetic potential to produce many more natural products than have been isolated so far. Facilitating sequencing of the entire Tübingen strain collection would allow to uncover the full genetic potential of all strains, and significantly expand the possibilities for genome mining and host engineering within the DZIF TTU9. Current capability of next generation sequencing makes this technology suitable for deciphering the complete genomes of thousands of bacterial species in a single sequencing run, however, growing and DNA isolation of 2000 strains is a time- and cost intensive effort. To reduce cost and labour we developed a metagenomic approach to sequence the strain collection in pools, each consisting of 115 strains, and combining long- and short-read technologies. Linked short-reads technologies such as Transposase Enzyme Linked Long-read Sequencing (TELL-Seq™) can further improve the approach and can be applied to sequence the complete strain collection (Chen et al. 2020).

6.11 Methods

6.11.1 Library preparation and NGS sequencing

1100 strains from the Tübinger strain collection have been grown and harvested on plates and pooled into 10 different pools containing cell material from 110 strains each. For each pool DNA has been isolated and tested if amount, size, and purity is suitable for long-read sequencing methods. The pools are stored in the freezer and are all ready for sequencing. Our pilot study including only one pool has shown that a 30x coverage of Nanopore data and 60x coverage of Illumina Novaseq data is sufficient to gain very long contigs of almost 4 Mb suitable for effective genome mining approaches. All pools were sent to the NGS

Competence Center Tübingen (NCCT) for Illumina TELL-Seq and Nanopore library preparation. The libraries were barcoded, and sequenced on Nanopore flowcells and S2 Novaseq Illumina FlowCell.

6.11.2 Bioinformatics Analysis:

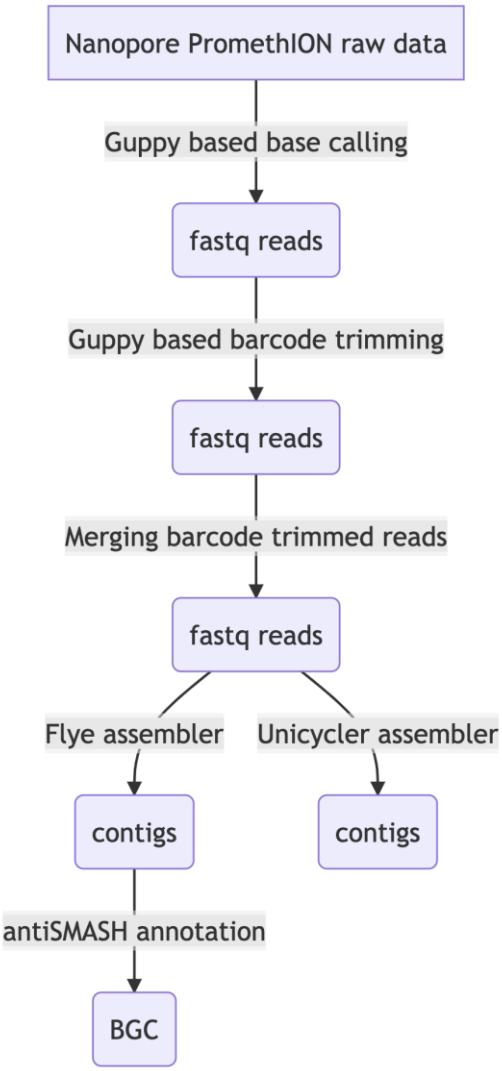


Figure 1: Analysis workflow for Nanopore data.

Nanopore PromethION raw data was processed with Guppy 4.0.14 (Figure 1). The base called fastq reads were then barcode trimmed and merged. Flye 2.8.1-b1676 and Unicycler v0.4.8 were used separately to assemble the nanopore only reads of each sample pool (Wick et al. 2017; Kolmogorov et al. 2020).

6.11.3 TELL-Seq analysis

The TELL-seq Illumina raw data was processed with a tell-read pipeline to generate linked reads. Using tell-link de novo pipeline the linked reads for each of the sample pools was assembled to generate contigs (Chen et al. 2020). The assembled contigs were annotated using antiSMASH (Blin et al. 2019).

6.12 Results and Discussion

On PromethION around 9 million long reads were sequenced and a total of more than 13 Gigabases sequence data was generated for 10 sample pools. The average read length was nearly 1.5 kb. On Illumina Novaseq, 1030 million paired end reads (2 X 150 bp) were generated for the same 10 sample pools. Raw barcode statistics and tell-read pipeline barcode processing statistics are shown in Table 1.

Raw Barcode Statistics

sample	BHVYYYDRXX-results-lane1_T502	BHVYYYDRXX-results-lane1_T503	BHVYYYDRXX-results-lane1_T504	BHVYYYDRXX-results-lane1_T505	BHVYYYDRXX-results-lane1_T506
total_reads	96,341,666	105,409,364	107,821,693	104,363,337	95,716,696
reads_with_barcode_all_Gs	1,776,390	920,718	1,521,628	1,734,331	1,230,417
reads_with_correct_barcode	92,169,286	102,232,667	104,072,921	100,309,066	92,155,544
reads_with_error_barcode	4,172,380	3,176,697	3,748,772	4,054,271	3,561,152
%reads_with_correct_barcode	95.7%	97.0%	96.5%	96.1%	96.3%
%reads_with_error_barcode	4.3%	3.0%	3.5%	3.9%	3.7%
unique_barcode	6,699,373	6,573,284	6,462,665	6,670,438	6,289,259
unique_correct_raw_barcode	5,424,590	5,444,577	5,349,383	5,481,938	5,015,258
mean_#reads/correct_barcode	17.0	18.8	19.5	18.3	18.4

Barcode Processing Statistics

correction	T502	T503	T504	T505	T506
barcode_with_single_read	4,063,549	3,976,054	3,981,153	4,097,860	4,112,732
barcode_with_more_than_3_reads	1,838,317	1,720,749	1,724,389	1,748,191	1,490,046
reads_related_to_barcode_with_more_than_3_reads	90,429,097	99,393,630	102,083,276	98,351,333	90,015,518
1mismatch_barcode_corrected	1,324,743	1,317,677	1,327,047	1,335,681	1,255,892
error_barcode_number	937,191	824,331	819,156	875,004	964,676
final_correct_barcode_number	4,437,439	4,431,276	4,316,462	4,459,753	4,068,691
final_reads_number	92,261,482	102,334,055	104,176,623	100,408,753	92,245,364

Raw Barcode Statistics

sample	BHVYYYDRXX-results-lane2_T501	BHVYYYDRXX-results-lane2_T502	BHVYYYDRXX-results-lane2_T503	BHVYYYDRXX-results-lane2_T505	BHVYYYDRXX-results-lane2_T507	BHVYYYDRXX-results-lane2_T508
total_reads	110,765,075	83,328,488	7,537,348	135,629,052	118,067,006	40,113,859
reads_with_barcode_all_Gs	3,647,243	585,337	936,425	6,216,092	1,284,181	1,357,018
reads_with_correct_barcode	104,075,705	81,140,162	6,382,646	126,572,685	114,488,206	37,754,571
reads_with_error_barcode	6,689,370	2,188,326	1,154,702	9,056,367	3,578,800	2,359,288
%reads_with_correct_barcode	94.0%	97.4%	84.7%	93.3%	97.0%	94.1%
%reads_with_error_barcode	6.0%	2.6%	15.3%	6.7%	3.0%	5.9%
unique_barcode	7,137,174	6,871,713	1,372,739	8,727,352	6,561,084	3,769,864
unique_correct_raw_barcode	5,569,486	5,980,140	1,235,262	7,226,778	5,466,624	3,197,180
mean_#reads/correct_barcode	18.7	13.6	5.2	17.5	20.9	11.8

Barcode Processing Statistics

correction	T501	T502	T503	T505	T507	T508
barcode_with_single_read	4,763,038	3,898,254	645,030	5,474,447	4,092,452	2,077,596
barcode_with_more_than_3_reads	1,648,234	2,068,995	420,078	2,022,747	1,721,529	1,196,946
reads_related_to_barcode_with_more_than_3_reads	104,339,015	77,326,232	6,160,368	127,300,074	112,244,243	36,873,265
1mismatch_barcode_corrected	1,603,645	1,150,398	98,898	1,866,500	1,401,801	553,401
error_barcode_number	1,163,115	689,171	117,463	1,123,095	818,229	457,303
final_correct_barcode_number	4,370,414	5,032,144	1,156,378	5,737,757	4,341,054	2,759,160
final_reads_number	104,157,852	81,210,399	6,387,624	126,679,694	114,579,985	37,785,729

Table 1: TELL-Seq raw barcode statistics and barcode processing statistics for 10 Sample pools run on 2 lanes of Illumina flowcell.

Statistics without reference	scaffold.full
# contigs	78 849
# contigs (>= 0 bp)	78 849
# contigs (>= 1000 bp)	18 670
# contigs (>= 5000 bp)	126
# contigs (>= 10000 bp)	94
# contigs (>= 25000 bp)	47
# contigs (>= 50000 bp)	29
Largest contig	4 342 632
Total length	82 970 907
Total length (>= 0 bp)	82 970 907
Total length (>= 1000 bp)	42 659 143
Total length (>= 5000 bp)	14 266 213
Total length (>= 10000 bp)	14 033 714
Total length (>= 25000 bp)	13 319 207
Total length (>= 50000 bp)	12 666 110
N50	1027
N75	679
L50	17 512
L75	42 843
GC (%)	66.9
Mismatches	
# N's	60 727
# N's per 100 kbp	73.19

Table 2: TELL-Seq assembly statistics (using the tell-link de novo assembly pipeline) of a representative sample pool. Largest contig length of more than 4 megabases was produced.

Using TELL-seq and Nanopore data generated for one such pool, consisting of 115 streptomyces species from our Tuebingen Strain Collection, resulted in single contigs with more than 4 megabases (Table 2). Annotating these large contigs and applying genome mining tools led to the identification of novel biosynthetic gene clusters (BGCs). Easy backtracking of strains from our rare and unique strain collection is feasible. Using this set-up, the discovery of novel natural products has been drastically accelerated. This project is currently an ongoing project in ZiemertLab and the sequencing strategy, analysis pipelines and robust comparative methods are under development.

6.13 Conclusion and Outlook

TELL-Seq method is being continuously updated, improved and optimised. Currently, only limited samples data was processed (tell-link based de novo assembly) without facing any software or data related issues. Nevertheless the limited results obtained so far were sufficient for appreciating the power of this novel method. Algorithms and software tools for hybrid assembly of linked reads along with long reads might be developed in future. Using these will further improve the assembly of pooled metagenome data. Recovery of near complete metagenome assembled genomes would be then possible and the true potential of linked reads and long reads can be then realised for discovering biosynthetic gene clusters from the strain collections.

6.14 References

Blin, Kai, Simon Shaw, Katharina Steinke, Rasmus Villebro, Nadine Ziemert, Sang Yup Lee, Marnix H Medema, and Tilmann Weber. 2019. “AntiSMASH 5.0: Updates to the Secondary Metabolite Genome Mining Pipeline.” *Nucleic Acids Research* 47 (W1): W81–87. https://doi.org/10.1093/nar/gkz310.

Chen, Zhoutao, Long Pham, Tsai-Chin Wu, Guoya Mo, Yu Xia, Peter L. Chang, Devin Porter, et al. 2020. “Ultralow-Input Single-Tube Linked-Read Library Method Enables Short-Read Second-Generation Sequencing Systems to Routinely Generate Highly Accurate and Economical Long-Range Sequencing Information.” *Genome Research* 30 (6): 898–909. https://doi.org/10.1101/gr.260380.119.

Hennrich, Oliver, Franziska Handel, Regina Ort-Winklbauer, and Yvonne Mast. 2020. “Genome Sequences of Two Putative Streptogramin Producers, Streptomyces Sp. Strains TÜ 2975 and TÜ 3180, from the Tübingen Strain Collection.” *Microbiology Resource Announcements* 9

(21): e01582-19. https://doi.org/10.1128/MRA.01582-19.
Kolmogorov, Mikhail, Derek M. Bickhart, Bahar Behsaz, Alexey Gurevich, Mikhail Rayko, Sung Bong Shin, Kristen Kuhn, et al. 2020. “MetaFlye: Scalable Long-Read Metagenome Assembly Using Repeat Graphs.” *Nature Methods* 17 (11): 1103–10. https://doi.org/10.1038/s41592-020-00971-x.
Ortlieb, Nico. 2019. “Characterization of Natural Products from Actinobacteria of the Tübingen Strain Collection – Screening, Isolation & Structure Elucidation.” Dissertation, Universität Tübingen. https://doi.org/10.15496/publikation-31678.
Wick, Ryan R., Louise M. Judd, Claire L. Gorrie, and Kathryn E. Holt. 2017. “Unicycler: Resolving Bacterial Genome Assemblies from Short and Long Sequencing Reads.” *PLOS Computational Biology* 13 (6): e1005595. https://doi.org/10.1371/journal.pcbi.1005595.

Chapter 7: General Conclusion

7.1 Concluding remarks on results

In this thesis, I have presented a metagenomic approaches to explore novel regions of natural products chemical space. An easy to use metagenome data mining pipeline for exploring natural products diversity was presented in Chapter 4. We used this pipeline to explore the biosynthesis potential present in the different horizons of soil sampled from three sites in close vicinity from the Schoenbuch forest. Amplicon sequencing and analysis of 16S rRNA gene and BGC domains were highlighted in the manuscript included in Chapter 5. Shotgun metagenome sequencing using Illumina Short reads and Nanopore based long reads further helped in recovering the BGC clusters.

Further in Chapter 6, that covers metagenomic big-data explorations in diverse ecosystems, in part 1 I have included our manuscript in which we report the distribution of bacterial natural product biosynthetic genes across lake Huron sediment. While this exploration helped in appreciating the tremendous NP diversity present in the lake sediments, a need for wider and extensive sampling was experienced to fully capture the functional chemical diversity.

Part 2 of this chapter covers the aspects of how the dynamics of the human gut secondary metabolome changes during antibiotic treatment. The sactipeptide BGC abundance pattern observed in the metagenomic samples collected during the course of ciprofloxacin treatment helped in generating new hypotheses that can be further tested in future.

In part 3 a novel approach is presented that helps to discover the BGC sequences from the pooled bacterial strains from the strain collections. Linked reads technology improves the BGC discovery, and could be widely adopted by the research community until the long read sequencing becomes economical and error rate is reduced.

Using the approaches mentioned in this thesis, it will be possible to map the NP chemical space at different scales, based on the availability of resources, funds and rationale. To map the NP chemical space of the entire earth, even at kilometer resolution, megaprojects - worth billions of dollars- would be required. Such megaprojects can change the entire landscape of the field of NP discovery. It might lead to discovering many new NP diversity hotspots, novel phyla and gifted microbial strains. Fruits of understanding the NP chemical space, would be massive, and have the potential to revolutionize science. Novel NP chemical classes, truly unknown (BGC) unknown (chemical structure), can guide development of new survey rationale and discovery algorithm development. The massive amount of metagenomic data generated in such projects will also improve the quality of the taxonomy database. This will improve our understanding of how microbial diversity community structure gets shaped, how new strains evolve, how nature innovates to produce novel NP chemical classes, and what ecological roles these molecules and microbes play. Using the new NP knowledge, designing novel and more potent chemical probes capable of fishing out unique chemical compounds directly from environmental samples, would become possible. In this context, the future of chemical and synthetic biology would certainly be brighter than even what we can imagine.

There are already many orphan BGCs - those for which we don't know the structure of natural products they encode - available from public databases, and it is difficult to predict their chemical structure. Recent breakthrough (AlphFold 2) that improved prediction of protein structure by many folds, uses artificial intelligence and was developed by Google’s DeepMind, has started to positively impact all biomedical studies that require structural insights for solving numerous challenges. Aided by such powerful methods and ever increasing metagenomic big-data, future machine learning based predictions has the potential to fill the so-called “Genes to molecules” gap.

7.2 Impact and applications of developed approaches and methods

Acceleration of discovering novel genes, domains and clusters involved in biosynthesis of novel natural products would have a direct positive impact in reviving the antibiotic discovery pipeline in industry and academia. While we all are in the current Covid-19 pandemic, the need for searching or synthesizing novel antibiotics and antiviral drugs has tremendously increased. Specifically in context to the increase in hospitalization experienced during the past several waves of Covid-19, this would be driving the havoc of antimicrobial resistance and we need to be geared up to combat this future storm.

Amplicon based explorations have the potential to quickly and inexpensively uncover the microbial diversity and biosynthetic gene/domain diversity. Discovery of novel biosynthetic domains can drive the rational prioritization of the environmental samples for deeper metagenome sequencing via shotgun approaches for the characterization of full length BGCs and subsequent characterization of biosynthetic pathways. MBEZ pipeline will assist in such analysis for identifying patterns and correlations.

Biosynthetic cluster diversity comparisons and resulting patterns would make it possible to infer how the BGCs shape the microbial community structure. Ecological and evolutionary dynamics that govern the distribution of specialised metabolites could be then hypothesized using metagenomic big-data. Temporal and longitudinal metagenomic dataset analysis might even make it possible to study the BGC evolutionary history and mechanisms involved in creation of chemical diversity in nature. The long-standing outstanding questions of the natural products research field, such as, can one rationally choose the best natural ecosystem to survey metagenomes and discover novel antibiotics/natural products? Has the world profiled enough metagenomes for such NP discovery or still serendipity is best bet yet? — some of these questions have been previously articulated by Prof. Paul R. Jensen — such as “At what rate are BGCs created and lost? How often do new chemical scaffolds evolve?”(Jensen, 2016). We can hope to answer these questions in future.

7.3 Future challenges

Current metagenomic sequence data sizes demand sophisticated and bigger computing resources such as high memory workstations, computing cloud and high performance computing clusters. Novel algorithms and software tools that can accelerate metagenome mining and make the analysis possible on smaller personal computers and laptops are needed for democratising this powerful approach to discovering novel natural products.

Subsequent bottleneck after discovering novel BGCs is to heterologously express them in suitable hosts for getting the metabolic products that these BGCs encode. Currently due to availability of limited hosts, expressing BGCs from rare phylum is a big challenge.

Large-insert metagenomic library creating methods along-with high throughput cloning and functional characterization novel methods would be needed to cope up with the high rates of discovery achieved via metagenome sequencing methods.

Due to high rates of microbial species extinctions that we are experiencing, those that were partly fueled by climate change, it is difficult to fathom the magnitude of natural products chemical space that we are continuously losing forever. Novel metagenome mining methods, and mega diversity expeditions will be needed to map and uncover the entire natural products chemical space that earth currently holds. Apart from academic interests, the survival and flourishing of humanity is dependent on these novel natural product discoveries.

7.4 References

Jensen, P.R., 2016. Natural Products and the Gene Cluster Revolution. Trends in Microbiology 24, 968–977. https://doi.org/10.1016/j.tim.2016.07.006

Annexure A:

Publication: The confluence of big data and evolutionary genome mining for the discovery of natural products

Natural Product Reports

ROYAL SOCIETY OF CHEMISTRY

REVIEW

View Article Online
View Journal

Cite this: DOI: 10.1039/d1np00013f

The confluence of big data and evolutionary genome mining for the discovery of natural products

Marc G. Chevrette, †[a] Athina Gavrilidou, †[bc] Shrikant Mantri, †[bcd] Nelly Selem-Mojica, ‡*[e] Nadine Ziemert *[bc] and Francisco Barona-Gómez *[e]

Received 2nd March 2021

DOI: 10.1039/d1np00013f

rsc.li/npr

This review covers literature between 2003–2021

The development and application of genome mining tools has given rise to ever-growing genetic and chemical databases and propelled natural products research into the modern age of Big Data. Likewise, an explosion of evolutionary studies has unveiled genetic patterns of natural products biosynthesis and function that support Darwin's theory of natural selection and other theories of adaptation and diversification. In this review, we aim to highlight how Big Data and evolutionary thinking converge in the study of natural products, and how this has led to an emerging sub-discipline of evolutionary genome mining of natural products. First, we outline general principles to best utilize Big Data in natural products research, addressing key considerations needed to provide evolutionary context. We then highlight successful examples where Big Data and evolutionary analyses have been combined to provide bioinformatic resources and tools for the discovery of novel natural products and their biosynthetic enzymes. Rather than an exhaustive list of evolution-driven discoveries, we highlight examples where Big Data and evolutionary thinking have been embraced for the evolutionary genome mining of natural products. After reviewing the nascent history of this sub-discipline, we discuss the challenges and opportunities of genomic and metabolomic tools with evolutionary foundations and/or implications and provide a future outlook for this emerging and exciting field of natural product research.

1. Introduction
1.1 Origins of evolutionary genome mining of natural products
2. Big data and evolutionary genome mining of natural products: from key concepts to databases and algorithms
2.1. Key big data concepts in natural products research
2.2. Key evolutionary concepts in natural products research
2.3 Natural products databases available for evolutionary genome mining
2.4 Big data and natural products evolutionary genome mining algorithms
3. Genomic and enzymatic evolution of natural products
3.1 Evolution of the genome of NP-producing organisms
3.2. BGC and multidomain enzyme evolution
4. What lies ahead? Needs and opportunities for evolutionary genome mining of NPs
5. Conflicts of interest
6. Acknowledgments
7. References

[a]Wisconsin Institute for Discovery, Department of Plant Pathology, University of Wisconsin-Madison, Madison, WI, USA

[b]Interfaculty Institute of Microbiology and Infection Medicine Tübingen (IMIT), Interfaculty Institute for Biomedical Informatics (IBMI), University of Tübingen, Germany

[c]German Centre for Infection Research (DZIF), Partner Site Tübingen, Germany. E-mail: nadine.ziemert@uni-tuebingen.de

[d]Computational Biology Laboratory, National Agri-Food Biotechnology Institute (NABI), Mohali, Punjab, India

[e]Laboratorio de Evolución de la Diversidad Metabólica, Unidad de Genómica Avanzada (Langebio), Cinvestav-IPN, Irapuato, Guanajuato, Mexico. E-mail: francisco.barona@cinvestav.mx

† Authors contributed equally (MGC, AG, SM).

‡ Current address: Centro de Ciencias Matemáticas (CCM), UNAM, Morelia, Mexico.

1. Introduction

Evolution is a process; therefore, evolutionary theory seeks to describe the series of events that have allowed life to appear, develop, and diversify. Natural selection, postulated by Charles Darwin more than one hundred and fifty years ago, is perhaps the most recognized of these theories, linking the natural histories of all living forms to their reproductive fitness.[1] In the years since Darwin, we have come to appreciate that evolutionary processes display enormous complexity and act through both selective and neutral forces of varying physicochemical, ecological, temporal, and population-level constraints.[2] Neutral,

non-adaptive evolution was once thought to be discordant with Darwinian evolution; now we appreciate that evolutionary histories provide evidence of both selective pressures and neutral events.[3,4] Founder effects, genetic drift, gene flow, and many other neutral mechanisms shape the genetic variation within populations upon which natural selection operates.[5] The enzymes of natural product (NP) biosynthesis are encoded in genomic information, and as such do not escape these forces of evolution.[6,7] This distinction is as important to recognize as it is easy to neglect: NPs with antagonistic functions, like antibiotics or other biocides, are typically assumed to be under positive selection to maintain the interactions with their molecular target(s) necessary to retain function. Paradoxically, the historical use of the term 'secondary metabolism', synonymous with trivial or unimportant metabolism, at the same time suggests neutral evolution, free to drift from one structure to the next. This conundrum highlights the importance of better defining evolutionary principles during chemical and biological investigation of natural products.

In this review, we aim at providing basic evolutionary principles as they have been embraced by genome miners interested in natural products-based drug discovery and the development of bioinformatics tools useful for this purpose. We discuss the origins of this sub-discipline (Sub-section 1.1), as well as working definitions and core evolutionary and Big Data principles, both generally and specifically regarding evolution-driven genome mining approaches (Sub-sections 2.1 and 2.2). We distinguish and highlight selected examples in which the confluence of Big Data and evolutionary genome mining for the discovery of natural products is more evident; and provide information to better understand and efficiently use these tools, but also to prompt newcomers and pave the way for the development of tools embracing the predictive power of the theory of evolution and the wealth of Big Data. Both databases and algorithms with relevant evolutionary features are presented in Sub-sections 2.3 and 2.4. Selected examples of NPs research

Marc G. Chevrette received a B.Sc. in Molecular Biology & Bioinformatics from Rensselaer Polytechnic Institute, Master's degrees in Bioengineering and Genetics from Harvard University Extension and the University of Wisconsin-Madison, respectively, and a PhD in Genetics from the University of Wisconsin-Madison. Marc was the Head of Experimental Genomics at Warp Drive Bio and an Associate at the Broad Institute of MIT & Harvard. He is currently a Postdoctoral Associate at the Wisconsin Institute of Discovery focused on the genomics and evolution of secondary metabolism, microbial chemical diversity, and interspecies interactions.

Athina Gavriilidou studied Biology at the Aristotle University of Thessaloniki, Greece. During her undergraduate studies she became interested in the application of informatics tools that further biological goals. She completed a Masters Degree in Bioinformatics and she currently conducts her PhD research in Bioinformatics at the University of Tübingen, Germany, with a focus on genome mining for Natural Products.

Shrikant Mantri received his B. Pharmacy from University of Pune and M.Tech in Bioinformatics from Indian Institute of Information Technology Allahabad. He leads the computational biology lab at National Agri-Food Biotechnology Institute, Mohali, India. His interests in interdisciplinary genomics research and bioinformatics led him to work on his Ph.D in Bioinformatics from University of Tuebingen, Germany.

Nelly is a Mexican mathematician and bioinformatician interested in genome evolution in prokaryotes. She has developed genome mining tools for which she earned the Mexican L'oréal award for women in science 2021. She also likes to develop software and software education resources such as lessons for "The Carpentries" and Wikipedia content. After a Ph.D. and a Post-doctorate in Integrative Biology at Evolution of Metabolic Diversity lab at Langebio-Cinvestav she is starting her research group at Centro de Ciencias Matematicas UNAM.

embracing evolutionary thinking – from enzymes to whole microbiomes – are provided in Sub-sections 3.1 and 3.2. The selected cases highlight evolutionary thinking and include the few examples that involve tools of what we call evolutionary genome mining of natural products. The final Sub-section 4 provides future directions for the development of this emerging sub-discipline as an important area of research to better understand NPs as whole and direct their biotechnological exploitation.

1.1 Origins of evolutionary genome mining of natural products

Advances in DNA sequencing have allowed for the study of allelic variation and how it relates to different phenotypes and evolutionary pressures.[8] These genetic investigations have developed into entire fields of molecular and genome evolution research, most notably advancing the areas of population genetics and phylogenetics. Population genetics investigates the frequencies and dynamics of genetic differences in and across populations, aiming to understand how some gene variants are more or less frequent than others.[5] In contrast, phylogenetics seeks to relate gene variants to each other by inferring an evolutionary history that explains differences between both genes and species.[9] Indeed, one might argue that phylogenetics was the first molecular biology Big Data method used broadly in biology, and remains so, as it aims to unveil hidden patterns otherwise ambiguous using empirical knowledge alone.[10] These inferences can be used to predict evolutionary histories through building networks of relatedness (*e.g.* phylogenetic trees) and reconstructing ancestral states, and therefore, in order to adopt evolutionary theory properly, these frameworks should be considered when approaching the evolution of NPs, especially when mining large datasets.

While evolutionary frameworks increasingly appear in the study of NPs, the extreme interdisciplinarity of NP research has led to adoption of evolutionary principles at different rates in different subdisciplines, depending on scientific goals and availability of data and the technologies used for their generation and analysis. For example, NP chemists often focus on empirical and mechanistic data to direct future investigations, and by doing so, they reinforce working models of biosynthetic logic in well-studied enzymes, for instance, nonribosomal peptide synthetases (NRPS)[11] and polyketide synthases (PKS).[12] In contrast, phylogenetics, whether at the species, gene, or genome level, aims to unveil broader patterns and place them into evolutionary context. This is increasingly done for bacterial,[13–15] fungal[16,17] and plant[18,19] NP biosynthetic enzymes, and even across different taxonomic lineages that produce similar NPs.[20,21] Phylogenetic insights may have limited mechanistic value, but they can assist in posing novel mechanistic hypotheses that can be experimentally tested. The combination of both approaches is embraced by Dean and Thornton's functional synthesis, which proposes that sequence analyses should be coupled with empirical, molecular experiments to retrace the evolutionary histories of biochemical processes and their phenotypes.[22]

In recent years, these two apparently disparate schools of thoughts have converged, yielding new protein evolution theory[23,24] and NP genome-mining applications.[25–27] Indeed, the marriage of phylogenies and mechanistic insights, implicit in early protein evolution-rate studies,[28] is the essence of evolutionary genome mining of NPs. The genes involved in NP biosynthesis and function, a subset of which have been validated through mechanistic studies, can be used to reconstruct large-scale phylogenies of multiple genes and their proteins. The genetic patterns uncovered by this Big Data approach can then feed back into more mechanistic predictions, providing hypotheses to further validate *via* new empirical, mechanistic studies. As these patterns can be affected by both evolutionary forces and the genetic mechanisms underlying them (in bacteria,[6,7] fungi[29–31] and plants[32,33] alike, yet each with their own intricacies) it is of utmost importance that these are clearly

Nadine Ziemert received her Diploma and PhD degrees from the Humboldt University in Berlin, followed by a postdoc and project scientist position at the Scripps Institution of Oceanography in La Jolla, California. Since 2015, she is a Professor at the University of Tübingen, where she leads an interdisciplinary research group focusing on genome mining approaches and the evolution of secondary metabolites in bacteria and their diverse functions.

Paco is a Mexican chemist and microbiologist interested in deciphering the multi-scale evolutionary mechanisms underlying the evolution of metabolism during bacterial adaptation. In addition to the roles played by natural products in mediating microbial interactions, he is also interested in deciphering novel biosynthetic logics and discovering chemical scaffolds, together with the development of bioinformatics tools for genome mining of natural products. After a PhD (Biology) and Postdoctoral research position (Chemistry) at Warwick University, UK, he founded the Evolution of Metabolic Diversity Laboratory at Cinvestav, Mexico, where he sustains a Newton Advanced Fellowship of the Royal Society, UK.

defined and appreciated by the natural products community when describing NP evolution.

2. Big data and evolutionary genome mining of natural products: from key concepts to databases and algorithms

Genomic assemblies from DNA sequencing data and a strain's associated phenotypic and/or meta information are the source of Big Data needed for the development of NP evolutionary genome mining databases (training sets) and algorithms (tools). This stems from the fact that the interactions between the chemical products of natural product biosynthesis and their molecular targets are shaped by evolutionary processes that control chemical structure, regulation, and/or availability.[6] Thus, the enzymes that assemble natural products are subject to these evolutionary pressures as well.[6,51] Biosynthesis of natural products is typically a series of incorporating building blocks into a larger structure and the stepwise addition of chemical modifications. Precursors may be sourced from other parts of metabolism, the environment, or synthesized within the biosynthetic gene cluster itself.[6,27] Some biosynthetic enzymes are large macromolecular machines, like NRPSs[11] or PKSs,[12] while others are single domain enzymes.[51] BGCs can be as simple as a few genes or as complex as many dozens of genes whose encoded enzymes work in concert to produce the final product(s). The enzymes at work within natural product biosynthesis are as diverse and varied as the chemical structures they biosynthesize, the molecular targets with which they engage, and the interactions within and between species that they mediate. Taking this context into account, we next define evolutionary and Big Data key concepts as the foundations of evolutionary genome mining of natural products databases and algorithms.

2.1. Key big data concepts in natural products research

Big Data refers to datasets that fit four major criteria: volume, velocity, variety, and validation. First, volume: Big Data must be big.[34] This typically refers to having many different entries or examples or replicates, depending on your data type. The distinction between "normal" datasets and Big Data is an ever-changing definition: what is considered Big Data today will likely not be Big Data in the future. This is mainly due to scientific breakthroughs leading to technological improvements and data generation. Second, velocity: Big Data grows quickly, which is mainly prompted by technological advances. A useful example of volume and velocity is shown in Fig. 1, highlighting the growth (velocity) of the number (volume) of genomes in NCBI over time. Third, variety: Big Data typically has several layers of information, which will be discussed below specifically for NP research. Finally, validation: a Big Data approach is only as good as its training data, so ensuring that training information is verified in some way is necessary for confidence in making forward predictions and identifying patterns. While validation is not strictly required for a dataset to be considered "Big", applications will have limited value if they are based on unverified information. This may sound fairly obvious yet is something that needs to be explicitly stated. Gene annotations are a common example where validation becomes very important: comparing your gene of interest to a validated dataset (*e.g.* UniProt, SwissProt) yields classifications that are much higher confidence than if you were to compare to unvalidated datasets (*e.g.* NCBI-NR) where the annotations of the dataset itself are unvalidated and errors can compound.[35]

As datasets grow bigger (volume) at faster rates (velocity), an unvalidated dataset made up only of predictions may have misannotations. These errors can lead to many more subsequent misannotations, which themselves can further exacerbate these errors.[36] Thus, understanding the level of validation for your dataset is necessary to properly interpret your results. Together, these four Vs present analysis challenges, as Big Data is often too large or complex such that non-traditional or parallel computing tools are needed for analysis with *ad hoc* algorithms.[37,38] In general, for a natural products researcher in the early 2020s, data becomes 'Big Data' when it is too large or too complex to do simple statistics in spreadsheet-based software (*e.g.* Microsoft Excel). These data, moreover, are hard to process and visualize with available tools within tolerable computing times.

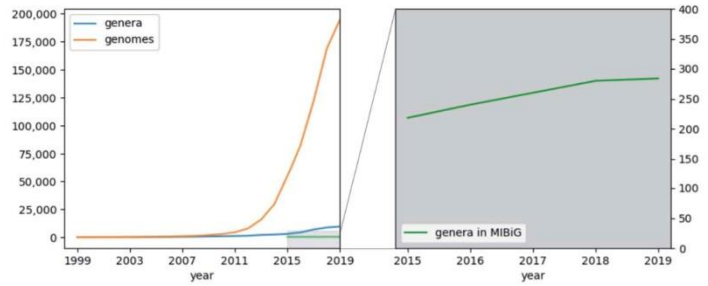


Fig. 1 Growth of the number of NCBI Genomes (bacteria and archaea) and Genera per year from 1999 to 2019. Data from GTDB (release 95). Inset: number of Genera represented by data in MIBiG.

Table 1 Genomic databases to explore natural products diversity and evolution

Database name[a]	Parameter name	Parameter value	Current version (date)
MIBiG[65]	BGCs	1923	2.0 (2019)
IMG-ABC[69]	BGCs	410 683	5.0
antiSMASH-db[67]	BGCs	147 517	3.0
BiG-FAM[70]	BGCs	1 225 071	1.0
NCBI genome	Bacteria spp.	278 820	November 2020
	Archaea spp.	5625	November 2020
	Eukaryote spp.	14 486	November 2020
MGnify[71]	Metagenomes	32 746	November 2020
IMG/M[72]	MAGs	52 515	November 2020
	BGCs	104 211	November 2020
CARD[73]	Alleles	213 809	February 2021
	Reference sequences	3146	February 2021
SRA (bacteria)	Datasets	1 466 494	November 2020
SRA (archaea)	Datasets	38 592	November 2020
NCBI WGS (bacteria)	Projects	941 266	December 2020
NCBI WGS (archaea)	Projects	6225	December 2020
Resfinder 4.0 [74]	Resistance genes	2690	December 2020
MG-RAST 4.0.3 [75]	Metagenome	447 497	January 2021

[a] Most of the listed databases in Tables 1 and 2 arguably satisfy the Big Data characteristics of volume and variety. Since there have been only few periodic releases for some of these databases, the velocity characteristics of Big Data can be appreciated for only a few of these. The month and year (date) of each database in Tables 1 and 2, when last accessed, are provided. Exact dates for current versions are not provided as are not available.

Standard genome mining approaches to uncover NP biosynthesis have been used to explore a wide range of taxa and environments, identifying "microbial dark matter" as a promising source of hidden chemical treasures. In evolutionary genome mining of NPs this becomes an essential consideration with potentially confounding factors. As shown in Fig. 1, the first two 'Vs', volume and velocity, are currently covered by the sequence data in large databases. In NP research, however, data is not limited to genetics, but it has many other layers, including chemical, gene expression, ecological, and evolutionary data. For instance, the MIBiG[39] data repository is a good example of 'variety', in that it includes multifaceted chemical and genetic data. It also has a high standard of validation, as the level of validation is listed for each entry. These advantages come at the cost of volume and velocity: keeping the standards of variety and validation high mean that this repository grows at slower rates than for example the NCBI genome database. Important to evolutionary genome mining, MIBiG and other repositories tend to be biased towards a limited number of taxa that have been investigated in great detail, like species of the genus *Aspergillus* in fungi[16,31] or *Streptomyces*[40–44] within the Actinobacteria. While a bias towards bacterial genera clearly exists, this issue is slowly decreasing with other Genera such as *Nocardia*,[45] *Amycolatopsis*,[15] *Salinispora*,[46] *Micromonospora*,[47] *Pseudonocardia*,[48] *Rhodococcus*,[49,50] *etc.* emerging as promising NP producers. Yet, bias in sampling remains a critical consideration in evolutionary studies as they can confound results and sometimes lead to erroneous conclusions, as argued recently in the case of *Aspergillus*.[31]

In summary, Big Data available for evolutionary studies and genome mining of natural products come from several sources, including both broad and specialized chemical and genetic databases (see Tables 1 and 2). As an example, NCBI database contains over 1.4 million bacterial and over 38 thousand

Table 2 Chemical databases to explore natural products diversity and evolution

Database name[a]	Parameter name	Parameter value	Current version (date)
MACADAM[154]	Metabolites	7921	1
PubChem[77]	Compounds	111 456 896	November 2020
GNPS[64]	NP compounds	18 163	1
	Spectra	221 083	
NP Atlas[78]	Compounds	24 594	v 2020_06
COCONUT[155]	Compounds	406 747	March 2021
StreptomeDB[156]	Compounds	4000	2
PoDP[145]	Paired (meta)genomes and metabolomes	4853	2021 GitHub v0.9.2
Siderophore DB	Compounds	262	June 2021
LOTUS[157]	NP compounds	276 518	February 2021

[a] Refer to table notes in Table 1.

archaeal samples at the writing of this manuscript, with data existing as either genomes, transcriptomes, or metagenomes. These data however are far from being informative into NP research unless they are organized and/or translated into other forms or layers of information and analyzed with suitable tools. Based on our own experience, Big Data for natural products research today implies algorithms fast enough to conveniently analyze the genomes and/or metabolomes of over 30 thousand strains or samples. These numbers will rapidly multiply in the future, and thus it is critical to continually reassess "natural classifications" seen in evolutionary relationships, keeping in mind that sampling bias of training data remains a fundamental, yet often overlooked, issue. Scalability of tools is also a consideration. For example, multiple sequence alignments and phylogenies of hundreds or thousands of genes was once considered Big Data, and remains so, yet now we can perform phylogenomic comparisons across entire kingdoms of life on an inexpensive laptop computer or free public web server.[25,27] This scalability of datasets and analysis tools can provide the genetic context necessary to perform evolutionary genome mining.

2.2. Key evolutionary concepts in natural products research

Evolutionary pressures that drive the appearance and that overall shape the physicochemical and biomolecular features of natural products biosynthesis, can be incredibly dynamic and complex. Nevertheless, overarching principles of evolution of NP enzymes and/or pathways emerge. Just as biochemical principles (*e.g.* adenylation (A) domain specificity of NRPSs or chain elongation during PKS-catalyzed synthesis) are mechanistically fundamental for the understanding of NP biosynthesis, the following broad evolutionary principles, with a mechanistic bearing, can be considered:

(i) Enzyme promiscuity drives pathway evolution through genetic expansion-and-recruitment events, providing the building blocks to assemble, shuffle, and combine NP biosynthetic pathways.[52–54]

(ii) Once enzymes (or domains) are recruited into NP biosynthesis, they tend to cluster together as multidomain megasynthases and/or biosynthetic gene clusters (BGC).[6,7,29]

These two corollaries are valid across bacteria[6,27,40,55] fungi[16,31] and plants[32,56–58] within their unique physiological, morphological, and chromosomal peculiarities. They also hold across different taxonomic lineages that share homologous NP biosynthetic enzymes.[59,60] It is starting to be widely appreciated that the phenomena from which these corollaries derive can occur under strong positive selection, but growing evidence and theory suggests a key role for negative selection and neutral forces on BGC dynamics.[6] Once recombination events cluster enzymes together, either as multidomain enzymes or BGCs, the resulting pathways can recruit other auxiliary elements, such as regulators, domain–domain interactors, transporters, and importantly, resistance genes.[51] As these principles were comprehensively demonstrated in the last decade or so, they were exploited by researchers for the development of the four main evolutionary genome mining tools that the NP community has used to identify and investigate novel pathways: (i) EvoMining,[26,27] (ii) ARTS[25,61](iii) BiG-SCAPE[40] and (iv) CORASON.[40] These tools are placed in the context of Big Data and discussed in further detail in Sub-section 2.4.

Using phylogenetics to unveil the evolutionary patterns of NPs follows two main approaches. On the one hand, gene trees can be used to infer a gene's evolutionary history and provide evidence for past events that have led to present-day data (*i.e.* branches or leaves of the tree). For evolutionary genome mining, gene trees can be useful in identifying expansions (*e.g.* duplications) and subsequent diversification of biosynthetic genes of interest. On the other hand, species trees describe the reconstructed evolutionary history of a set of species or individuals, and thus are useful for identifying larger-scale evolutionary events.[62] Critically assessing how the topologies of genes and species agree and disagree can shed light on important evolutionary events, such as horizontal transfers.[63] While NP research is focused on BGCs (a collection of genes), much can be learned from studying single-gene and species trees. Understanding the distribution and evolution of NPs within taxa, for example, is a prerequisite for effective sampling and bioprospecting strategies.

For those interested in evolutionary genome mining of NPs, it is important to note that the above-mentioned approaches are the result of properly embracing phylogenetics and evolutionary principles, often implementing concepts and principles not typically studied by NP chemists. Fig. 2 shows the main concepts that those interested in the use and development of these tools should take into account. As mentioned, the main two evolutionary mechanisms driving the appearance of novel NP biosynthetic pathways are diversification (enzyme promiscuity and BGC dynamics) and selection (positive, negative, and neutral). However, it is only when these forces combine and impact the fitness of the NP-producing organism that pathways are assembled and reassembled during the course of evolution.[51] The main genetic mechanisms driving these evolutionary events have been identified and have been used in the development of NP evolutionary genome-mining tools (thicker arrows, Fig. 2). However, much remains to be deciphered regarding the evolution of NPs, especially in terms of their expression and function in the real environmental settings of their producing organisms, where fitness operates. Study cases are available (see Sub-section 3), but their scarcity makes them anecdotal and thus more data is needed to develop mining tools based on Big Data principles to investigate this layer of complexity (thinner and/or dashed arrows, Fig. 2).

2.3 Natural products databases available for evolutionary genome mining

As mentioned, data available for investigating natural products in the Big Data era comes from several sources. However, this information only becomes useful when organized on databases (training sets) that can be coupled with metadata of the organisms themselves, but also with information about the technology and methods used to generate the data. Examples of

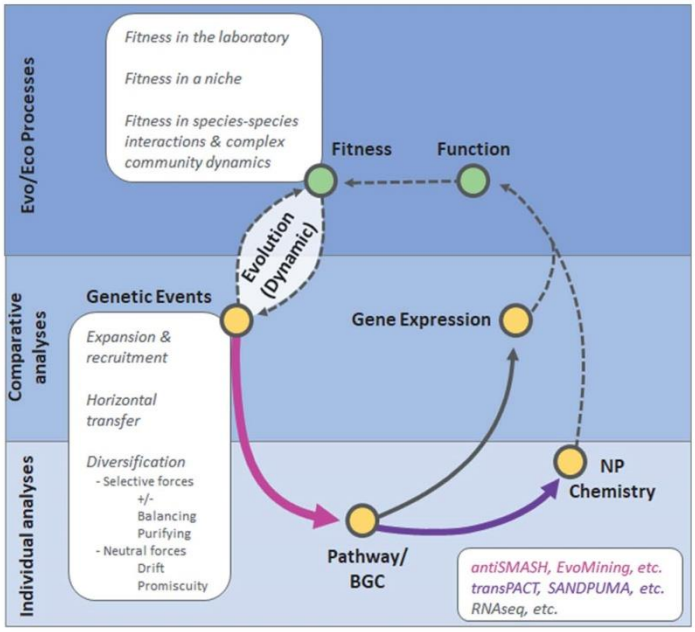


Fig. 2 Evolutionary genome mining of natural products in a concept-driven framework. Studies on the evolutionary histories of NPs, their biosynthetic genes, and their producing organisms are driven by analyses at different levels of organization. Individual analyses (bottom) focus on a pathway/BGC and their molecular product(s) or chemistry. Examples of tools that predict NP chemistry from BGCs are shown in purple. These individual data can then be contextualized with comparative analyses (middle) across many conditions or strains/species, with an emphasis in the genetic events underlying the evolution of NPs BGCs. One example is Gene Expression studies (gray, RNAseq) where comparisons of transcriptional patterns can place genes in a broader biological context. Analyses at the level of ecological and/or evolutionary processes (top) are the most challenging, and as a field we have only just begun to understand how Gene Expression, BGC, NP chemistry, and other "lower-level" data contribute to molecular function, and in turn how function contributes to an organism's fitness (linked by dotted lines to highlight that there are not yet standardized methods, but there is opportunity to develop them integrating Big Data). This remains a major challenge, as fitness is often a function of the environment. Evolution occurs as a dynamic process in which the fitness impact of a BGC's product influences the BGCs genetic components (*e.g.* diversification, selection, and other processes; see box). These in turn can feed back into fitness. Previously characterized genes and/or patterns of genetic events can then be used to identify and characterize BGCs *de novo* from genomic data (pink), either through rules-based or evolutionary methods.

well-executed databases include the GNPS mass spectra public database,[64] the MIBiG repository with experimentally validated datasets,[39,65] and the bioinformatically predicted BGCs of the antiSMASH-db[66,67] (Tables 1 and 2). Recently, the first evolutionary database, *i.e.* ActDES, which is specific for the Actinobacteria, has been reported.[68] All of these databases, despite complying with the four 'Vs' in one way or another, including variety, are useful in comparative or evolutionary studies, but not sufficient as none of them provide a comprehensive multi-layer database including or embracing evolution. In turn, at this stage, it is the responsibility of the evolutionary genome miner to select and integrate the most suitable and relevant DBs from those provided in Tables 1 and 2, within a phylogenomics framework. Selected DBs are highlighted throughout this review with the aim of emphasising their value in relation to the four 'Vs'.

2.4 Big data and natural products evolutionary genome mining algorithms

Communication between evolutionary biologists, computer scientists and mathematicians has historically led to biological insight, including the developments of population genetics theory and the transition matrices that are key to common genomic search algorithms like BLAST.[76] These disciplines have successfully converged again in recent years for the development of sophisticated NP genome-mining algorithms and platforms (Table 3). In this subsection, we list and explain major evolutionary genome mining of NPs approaches available to date with a focus on those that directly or indirectly rely on the use of the theory of evolution in any of its forms, either within the algorithms themselves or in their visualizations. The availability of genomic data (*e.g.* MIBiG, CARD, antiSMASH-db, Table 1) is fundamental, but probably more often will also be

Table 3 Big Data algorithms for exploring natural products diversity and evolution

Algorithm	Validation dataset	Type of data	Method	Date
ARTS 2.0 (ref. 61)	Bacterial kingdom genomes and metagenomes	Genomes	Duplication and BGC proximity, phylogeny and resistance screen	May 2020
BiG-SCAPE[40]	Clusters from ~3000 genomes	BGCs	Jaccard index plus maximum likelihood FastTree	November 2019
EvoMining 2.0 [27]	~100 conserved families from ~1000 genomes	Biosynthetic genes	Duplication and gene proximity to MIBiG, phylogeny	December 2019
BiG-SLICE[98]	BiG-FAM (1 225 071)	BGCs	Balanced iterative reducing and clustering using hierarchies	August 2020
CORASON[40]	~3000	Genomes or BGCs (visualization)	Blast plus FastTree	November 2019
Clinker[95]	NA	BGCs (visualization)	Hierarchical clustering	January 2021
FlaGs[96]	324	BGCs (visualization)	BGC's hidden Markov model	September 2020
TREND[97]	NA	BGCs (visualization)	Hierarchical clustering	April 2020
MicroReact[88]	NA	Trees with metadata (visualization)	Libraries: Chart.js, leaflet, phylocanvas, react, Sigma	November 2016
Anvi'o[93]	NA	Pangenomes (visualization)	Hidden Markov models	October 2015

inputs from purely chemical DBs (Table 2), *e.g.* GNPS, Paired Omics Data Platform (PODP), which can also serve as training data in supervised algorithms. Notably, some of these genomic-based algorithms already include input from chemical databases.[64,77,78] Thus, the integration of data types, as in MIBiG or PODP, may provide training datasets with valuable links between genomic and chemical data, further embracing variety. This integration holds great promise and value to the field, but

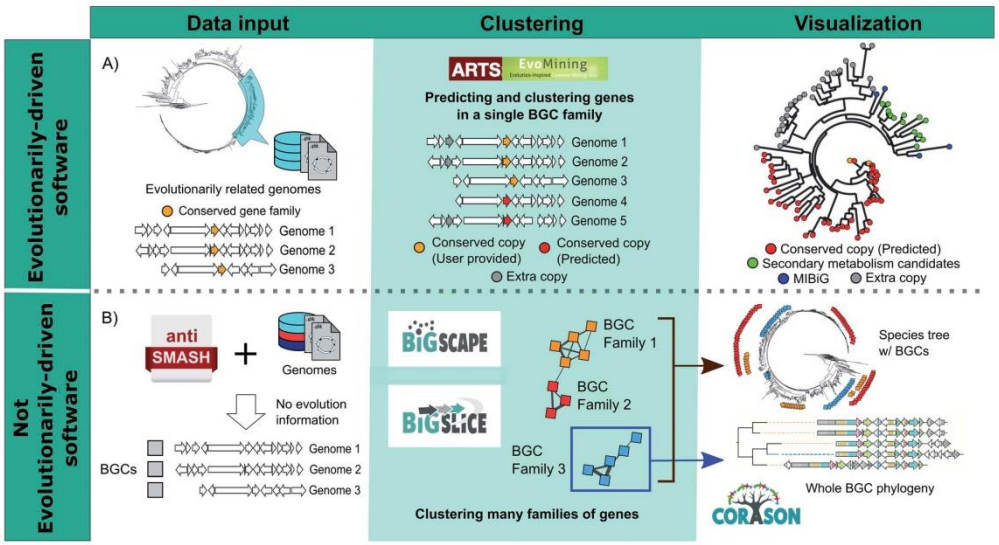


Fig. 3 Evolution-driven genome mining tools. (A) Evolutionary algorithms need as inputs genomes from taxonomically related lineages, where conserved protein families (orange) are selected for further exploration (ARTS/EvoMining). Conserved (orange and red) and extra (gray) copies of these families are identified and compared by a phylogenetic distance against proteins from NP databases (blue). Finally, the tree used in the phylogenetic distance is provided as a visualization, where predictions are included (green). (B) Algorithms with an evolutionary visualization but without evolutionary driven distances does not restrict their input genomes to be phylogenetically related. Gene clusters obtained from these algorithms are gathered in gene cluster families (GCF) by classification methods. Finally, evolutionary visualizations can be provided, either as a whole-BGC network of phylogenetic tree (BiG-SCAPE/CORASON) or as the occurrence of each GCF throughout a species tree (BiG-SLICE).

since it is only beginning to occur, it remains to be seen how regularly chemical data will be embraced by evolution-driven genome mining efforts.

Currently, evolutionary genome-mining for the discovery of novel NPs[79] aims to provide answers to two main questions, and by doing so, generate predictions: (i) which genes and/or BGCs produce metabolites not typically associated with central metabolism? and (ii) which genes or domains specific to a lineage represent innovation and diversification compared to ancestral states? As mentioned, several specialty databases[81] (Tables 1 and 2) are available and are used by the main evolutionary genome mining tools that the NP community has used to identify and investigate novel pathways: (i) EvoMining,[26,27] (ii) ARTS[25,61] (iii) BiG-SCAPE[40] and (iv) CORASON.[40] Following a similar rationale, a conceptual framework for mining siderophore BGCs based on their transporters has recently been reported.[80] Importantly, available tools can be used independently or in combination, and go in hand with species-level phylogenetic analyses which directly integrate NP biosynthesis (*e.g.* AutoMLST[91]) or analyses that are part of more generalized phylogenetic pipelines.[82] The combination of the latter, *i.e.* a species tree, with large-scale BGC prediction and their taxonomic distribution, is BiG-SLiCE output[98]

Supervised algorithms make use of the DBs mentioned in the previous sub-section in the form of training sets with validated labels about what is an NP BGC and what is not.[36] Here, the "correct" classifications are known for training data and used to make predictions about new data. These methods typically require heavy (and often manual) curation of training sets, and thus the importance of the fourth V, validation. So far, most of NP research adopting genome mining approaches employs supervised algorithms, mainly used in classification problems that require prior knowledge.[83] Unsupervised algorithms, instead aim to extract patterns and trends from unlabeled data,[84] similar to phylogenies. These can be helpful to identify data features (*e.g.* genes and domains) that are important for categorization, but since no "true" answer is known false-positive errors may be more frequent. Clustering or other grouping methods used in unsupervised methods attempt to give some structure to a dataset. Typically, supervised and unsupervised strategies are complementary, as it is the case in NP evolutionary genome-mining (Fig. 3).

Within NP research, supervised problems are used to identify and classify domains, genes, and BGCs. ClusterFinder[85] was one of the first algorithms that attempted to classify regions of the genome as NP BGC (or not) by calculating a moving average of a "biosynthetic score", calculated based on domain- and gene-level agreement with profile Hidden Markov Models of biosynthetic enzymes. Although ClusterFinder[85] does not directly leverage evolutionary theory in its algorithm, it is indirectly inferring the evolutionary processes that shaped BGC regions throughout the genome. Many of these algorithms have been trained primarily (or exclusively) on bacterial data, and thus accurate and reliable identification of fungal BGCs remains a challenge. Fortunately, recent work has begun to take fungal-specific genes and genetic structure into account to address this issue.[86–88] A similar scenario in plants[89] has now been encountered since the realization that BGCs actually exist in this large and prominent group of NP producing organisms.

Identifying shared and novel features within and between taxonomic lineages is attempted by unsupervised algorithms, such as BiG-SCAPE, BiG-SLICE and CORASON. For example, BiG-SCAPE, and more recently BiG-SLICE, clusters BGCs into gene cluster families (GCFs) without requiring prior knowledge of these families. This is done after calculating distance scores between BGCs on the basis of shared protein families and BGC organization. After clustering, it can be useful to sort and/or connect these GCFs with each other into bigger "clans", that are related but more distantly so than members of the same GCF. This broader context can be used to track evolutionary events of related BGCs and investigate how these events are distributed across gene and/or strain phylogenies. An alternative-yet-complementary approach employed by CORASON involves phylogenetic trees of shared enzymatic features, including in some instances whole-BGCs phylogenies. Importantly, these processes use *supervised* classifications of genes and domains to perform *unsupervised* clustering into GCFs, so they too require high quality (*i.e.* validated, or at least carefully curated) genomic and chemical databases.

In contrast, EvoMining and ARTS, represent the first (and to our knowledge, thus far the only) heuristic algorithms that incorporate evolutionary thinking as part of the supervised approach itself, attempting to infer what is central metabolism and what may be secondary metabolism, with a certain degree of diversification hinting towards the appearance of an specialized pathway. Evolution is inferred as a distance metric, which can be seen as similar to a support vector machine algorithm,[90–92] but implemented using a tree to determine appropriate groupings (and thus classifications) for biosynthetic enzymes. Put in another way, it seeks to identify which query enzymes cluster more closely with central metabolism and which cluster more closely with secondary or specialized metabolism. Extra gene copies are assessed by EvoMining as potential recruitments into NP biosyntheses, and these gene families may differ from one taxonomic lineage to another (Fig. 3A).

After classification into BGC families (*e.g.* with BiG-SLICE and/or BiG-SCAPE), further evolutionary context can be added in the visualization stage with CORASON according to the phylogenetic history of genes within the BGC or the strain-level phylogeny of the producing organism itself. In turn, CORASON identifies gene clusters in a genomes database and sorts them according to their evolutionary relationships. Tools such as MicroReact[88] can also allow for visual exploration of large phylogenetic trees annotated with metadata. EvoMining and ARTS both start with labeled sets (genes that are either the primary copy or specialized metabolism copies that belong to other databases, *e.g.* CARD/MIBiG) and employ supervised methods where evolutionary distance is used to classify putative BGCs. As a consequence, their predictions are intuitively displayed phylogenetically. Other software suites that perform pangenomic visualization (*e.g.* Anvio[93]) are also useful in that they allow identification of families with potential gene expansion and/or recruitment events. Many recent tools aim to sort

and visualize relations between BGCs: for example, MultiGeneBlast[94] (implemented in antiSMASH), finds gene homologs in BGC comparisons. Given otherwise identified BGCs (*e.g.* by antiSMASH or other tools), BiG-SCAPE[40] can classify them into BGC families and other visualization tools such as clinker,[95] FlaGs[96] and TREND[97] allow for interactive visualizations (Fig. 3B).

3. Genomic and enzymatic evolution of natural products

3.1 Evolution of the genome of NP-producing organisms

Multiple studies have been conducted on the evolution of NP producers, providing useful indications for targeted bioprospecting. Biosynthetic potential and diversity appear to be related to the ecological niche of the producers, as was confirmed in multiple instances.[99–109] In some cases, though, phylogeny is more important, as observed in microbial taxa where secondary metabolism is most similar in closely related organisms rather than those isolated from the same source.[105,109] Such investigations showcase possible promising targets for NP research, be they specific known[14,109] or understudied taxa[14,49,105] or different environments/niches.[100,102–104,108] As such, it is clear therefore that evolution can be applied for the discovery of novel natural products, which can powerful if properly embraced.

Comparative genomic analyses have shown that most bacterial taxa harbor only a few BGCs while some dedicate a large proportion of their genomes to specialized or secondary metabolism[82,99–101,104–106,110–112]. The quantity and diversity of BGC content differs among the taxa, with extreme cases reported.[46,102] How disperse the phylogenetic distribution of a BGC is, can allude to the various effects selection has had on its related pathways.[113] Most notably, horizontal gene transfer (HGT) is a relatively frequent phenomenon in BGCs, which is one likely explanation for their extended distribution across distant taxa and their observed diversity.[6,15,99,101,109,110,114–117] While HGT is observed frequently in BGCs compared to other genetic elements, it is important to note that the evolutionary timescales involved are still quite large[6,99,118] and depend on both population structure and genetic identity of donor and recipient.[6,99,118] Vertical inheritance of BGCs within the same lineage is the dominant means through which biosynthetic information is transferred.[6,119] This is a key distinction that should be made when studying the evolution of BGCs, as the more subtle vertical evolutionary dynamics happen from generation to generation, while HGT events are typically observed at timescales closer to thousands, millions, or billions of years.

Thus far, all analyses mentioned in this subsection were not conducted on a Big Data scale. Indeed, the information discovered so far is being confirmed by multiple independent inquiries, yet still issues of small taxonomic coverage and sampling biases remain. In 2014, three articles were published that followed a more global approach to NP producer genomics. Cimermancic[85] and co-authors analyzed more than 1000 genomes from across the bacterial kingdom and created a "global map" of biosynthesis, encompassing ~33 000 predicted BGCs. Doroghazi[44] and co-authors focused on one phylum and, using different metrics and methods than Cimermancic, reached a similar conclusion by collecting information on the producers capacity and potential. At the same time, Medema[116] and co-authors examined a large number of known BGCs and proved that the rates of evolutionary events within such units are much higher than in clusters of primary metabolism. Since these studies were first published, the available data has multiplied and so too have the methods for processing them; more universal-scope analyses will soon follow and give the answers to questions that remain open, including how and when biosynthetic diversity evolved[112] or the capacity of nature to keep providing us with new compounds.[120]

The above-mentioned studies have focused on microbes that have been cultured under laboratory conditions. However, the number of unculturable organisms is vast and metagenomic analyses have begun to unravel their hidden biosynthetic potential, indicating promising new sources for NP bioprospecting (see next paragraph). Furthermore, investigating evolutionary patterns based on environmental samples can shed light on the functions of the NPs found in nature as well as their raison d'etre within their microcosm.[121] This is important as NP evolution occurs at the population level, as highlighted by recent examples where population genomics frameworks have been adopted to mine NPs in genomic data, both in fungi and bacteria.[31,102,122–125] Such approaches have even proven valuable at the bacterial colony-level of a domesticated model laboratory strain, *i.e. Streptomyces coelicolor*.[126,127]

Soil metagenomic surveys in urban greenspaces, grassland meadows, and areas covering up to continent-wide scale have reported microbial diversity patterns.[128–131,158] These patterns are drastically affected by the environment and massive sequencing efforts are required to comprehensively capture their diversity, even at kilometer scale. High throughput functional studies involving creation of large-insert metagenomic libraries provides a novel approach to examine the functional and phylogenetic diversity of sampled ecosystems.[132–134] Economically attractive approaches using amplicon sequencing have been used to prove the domain-level diversity of environmental NPs. Such approaches have provided clues to answer the long standing question of which sites should be surveyed to maximize the discovery of novel natural products.[64,104,135–138,158] Massive amounts of shotgun metagenomic data are already easily available from public repositories. Analyzing these Big Data to infer significant NP patterns has now become the next bottleneck and faster algorithms and easy to use tools are badly required to mine the potential resource. Additionally, detailed documentation, standardized sampling procedures, and still more metadata are required to be incorporated into public databases in order to exploit patterns and extract useful information.

3.2. BGC and multidomain enzyme evolution

The evolutionary history of BGCs can be studied by building separate and/or concatenated trees of their genes and protein products. These can have very different topologies than the

species trees of the NP producers themselves, suggesting unconventional sequence transmission events, such as Horizontal Gene Transfer (see previous section), gene conversion, intra-genomic recombination,[116] and others. Together, these trees and functional information of NP genes can be used as a foundation to predict the activity of yet-unknown compounds and suggest potential links between fitness and the evolutionary forces at work.

Natural products exhibit extremely diverse chemistry. Their evolutionary complexity is no less complex. Domains evolve in the context of genes, genes in the context of BGCs, and BGCs in the context of their the producers' genomes.[6,139] Further, how these metabolites contribute to the fitness of their producing organisms depends largely on their environmental niche, which is often completely unknown or has poorly-understood factors and boundaries.[140] Because of this interdependence between multiple levels of organization, evolution does not affect clusters uniformly.[116] Indicatively, trans-acyltransferase (trans-AT) AT domains have evolved independently from cis-AT AT domains: the latter cluster into NP-specific clades and are known to be acquired vertically, while the prior are present in many different phyla and appear to be transferred horizontally.[141] Based on the clades formed in trans-AT AT and KS trees, it appears their evolution is strongly linked to their elongation substrate specificities.[99,116,141,142] Indeed, computational pipelines such as transPACT[143] place KS sequence information into a phylogenetic framework to predict substrate specificity for unknown sequences. Cis-AT and trans-AT PKS variants can produce similar metabolites even though they have distinct evolutionary histories. This case of evolution may be influenced by the modularity of Type I PKS clusters that can be more plastic due to intragenic recombinations and may allow for adaptability in a wide range of ecological niches.[141]

Although much of NP evolution is thought of at the level of BGCs or genes, important evolutionary changes can also happen at even smaller scales. Substrate specificity of different NP enzymes is often dictated by the three-dimensional organization of their active sites and/or protein–protein interaction surfaces, so subtle changes to the protein sequence of these areas can steer specificity (and promiscuity) in multiple evolutionary directions. In some cases, these changes correlate with phylogeny, so knowledge of the evolutionary mechanisms behind BGCs can allow for collecting reliable information from domain phylogeny. NRPS domains also show evolutionary patterns linking phylogeny and chemistry.[141] Similar to the trans-AT KS domains of the PKS clusters, A-domains of NRPSs cluster into clades according to substrate specificity, while C-domains are highly conserved and follow a BGC-specific pattern.[21,99,116] Computational methods such as SANDPUMA[144] and others have used this phylogenetic signal to reliably predict the substrate specificity of A-domains. Recently, "substrate level" evolutionary signals, like in trans-AT KS and NRPS A-domains, can be used to predict substrate specificity, while "pathway level" evolutionary signals, like in NRPS C-domains can be used to predict BGC-level patterns of similar molecules.[46]

4. What lies ahead? Needs and opportunities for evolutionary genome mining of NPs

Evolutionary genome mining of natural products in the Big Data era has inherited the tradition of phylogenetics, in the sense that natural history coupled with genetic and chemical observations can provide mechanistic insight. With this heritage comes the promise of discovering "the known unknowns, unknown knowns, and unknown unknowns of secondary metabolism", which has important implications in gene expression and the distinctions between "cryptic" and "silent" BGCs.[79] Although genomic and metabolomic speciality databases have made considerable progress, we envisage an ever-growing need for novel speciality datasets merging different layers of information. A promising current endeavor is the assemblage of metabologenomics databases, where genetic information and predictions are merged with chemical data (*e.g.* Paired Omics database[145]). Nevertheless, the systematic inclusion of other data types, including evolutionary relationships, remains a challenge. One notable evolutionary database has been recently released for Actinobacteria,[68] but those with larger scale and broader taxonomic coverage are much needed. These high-variety databases promise new insights in the NP field as a whole. Similarly, the accompanying algorithms needed to efficiently compute high volume datasets will allow us to perform these analyses at scale and keep pace with the technological advances that generate data at high velocity. In the near future we expect these data to go beyond only genomes, metabolomes, and metagenomes and begin to encompass ecological and functional metadata.[146]

Biosynthetic enzyme domains are the focus of current, and likely future, algorithms. This presents unique challenges for enzyme families whose classifications are problematic and/or understudied in the community. For instance, chemists have provided insights into why sequence-based phylogenies are insufficient for certain enzymes: transition-state intermediaries can be highly reactive and plastic, and therefore sequence space is less constrained than in enzymes with well-defined active sites.[147] Examples of this include the terpene cyclases, cytochrome P450s, hydrolases and type III polyketide synthases, amongst others. In these examples, analyses could benefit from alternative methods to establish relationships useful to provide classification and dataset structure. In turn, this may provide more informative training sets within well-structured databases, increasing the quality of predictions surrounding these important classes of natural products biosynthetic enzymes. It should be noted that classification of some of these enzymes within abovementioned DBs, such as antiSMASH-db, does not necessarily mean that this problem has been sorted out (see validation; previous sections). Pangenomic analyses[93,148] to identify expanded enzyme families within lineages may provide an interesting possibility to classify enzyme families on evolutionary grounds.

Here, by reviewing the nascent history of evolutionary genome mining of natural products as a sub-discipline, it has

become apparent that a prerequisite for the development of successful algorithms is the realization and characterization of genetic events driving the evolution of biosynthetic enzymes in their genomic context (*e.g.* BGCs). As such, we highlight the following evolutionary concepts with the promise to link evolution to genetic and chemical mechanisms. It has become clearer that "natural" evolution of natural products can be governed by dynamic processes that result in functional replacements. For example, in convergent evolution of chemically related scaffolds with diverse biomolecular activities,[149] whose biosynthesis is directed by non-related BGCs that produce functionally similar molecules. It has also become clearer that biosynthetic pathways can be encoded by physically unrelated loci (in contrast to BGCs), which may consist of subclusters,[150] and that the same BGC can produce diverse natural products with different biological functions in response to the environmental conditions.[151] This intragenomic cross-talk might be seen as a simplified version of the metabolic exchange between different organisms within a microbiome, for which evolutionary experimental and conceptual frameworks have been developed.[107,152,153] Both levels of metabolic cross-talk represent an immanent Big Data challenge: to genomically mine large datasets to correlate physically unlinked loci and propose metabolic relationships[72,104] How to best embrace evolutionary processes, many of which we are only beginning to understand, in Big Data genome mining for natural products remains an exciting yet challenging endeavor; one that will surely provide many possibilities for the future of this emerging sub-discipline.

5. Conflicts of interest

There are no conflicts to declare.

6. Acknowledgments

We are grateful to Jorge Navarro-Muñoz for useful discussions and Erika V. Cruz for help with figures. Support for M. G. C. provided by grant 2020-67012-31772 (accession 1022881) from the USDA National Institute of Food and Agriculture. F. B. G. and N. S. M. are supported by Conacyt, Mexico (grant No. 285746) and the Royal Society of the United Kingdom, Newton Advanced Fellowship (NAF\R2\180631) to F. B. G. A. G. is grateful for the support of the Deutsche Forschungsgemeinschaft (DFG; Project ID # 398967434-TRR 261). S. M. is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2124 – 390838134. N. Z. is funded by the German Center for Infection Research (TTU09.716).

7. References

1 A. Sugden, C. Ash, B. Hanson and L. Zahn, Happy Birthday, Mr. Darwin, *Science*, 2009, **323**, 727.
2 A. D. Goldman and D. A. Liberles, The Journal of Molecular Evolution Turns 50, *J. Mol. Evol.*, 2021, **89**, 119–121.
3 M. Lynch, *et al.*, Genetic drift, selection and the evolution of the mutation rate, *Nat. Rev. Genet.*, 2016, **17**, 704–714.
4 J. G. Wideman, A. Novick, S. A. Muñoz-Gómez and W. F. Doolittle, Neutral evolution of cellular phenotypes, *Curr. Opin. Genet. Dev.*, 2019, **58–59**, 87–94.
5 M. B. Hamilton, *Population Genetics*, 2nd edn, Wiley, 2021.
6 M. G. Chevrette, *et al.*, Evolutionary dynamics of natural product biosynthesis in bacteria, *Nat. Prod. Rep.*, 2020, **37**, 566–599.
7 P. R. Jensen, Natural Products and the Gene Cluster Revolution, *Trends Microbiol.*, 2016, **24**, 968–977.
8 K. H. Wolfe and W.-H. Li, Molecular evolution meets the genomics revolution, *Nat. Genet.*, 2003, **33**, 255–265.
9 M. Nei and S. Kumar, *Molecular Evolution and Phylogenetics*, Oxford University Press, 2000.
10 C. R. Woese, O. Kandler and M. L. Wheelis, Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya, *Proc. Natl. Acad. Sci. U. S. A.*, 1990, **87**, 4576–4579.
11 R. D. Süssmuth and A. Mainz, Nonribosomal Peptide Synthesis—Principles and Prospects, *Angew. Chem., Int. Ed.*, 2017, **56**, 3770–3821.
12 A. Nivina, K. P. Yuet, J. Hsu and C. Khosla, Evolution and Diversity of Assembly-Line Polyketide Synthases: Focus Review, *Chem. Rev.*, 2019, **119**, 12524–12547.
13 J. S. Larsen, L. A. Pearson and B. A. Neilan, Genome Mining and Evolutionary Analysis Reveal Diverse Type III Polyketide Synthase Pathways in Cyanobacteria, *Genome Biol. Evol.*, 2021, **13**, 1–15.
14 K. Gutiérrez-García, *et al.*, Phylogenomics of 2,4-Diacetylphloroglucinol-Producing Pseudomonas and Novel Antiglycation Endophytes from Piper auritum, *J. Nat. Prod.*, 2017, **80**, 1955–1963.
15 M. Adamek, *et al.*, Comparative genomics reveals phylogenetic distribution patterns of secondary metabolites in Amycolatopsis species, *BMC Genomics*, 2018, **19**, 426.
16 A. L. Lind, *et al.*, Drivers of genetic diversity in secondary metabolic gene clusters within a fungal species, *PLoS Biol.*, 2017, **15**, e2003583.
17 K. E. Bushley and B. G. Turgeon, Phylogenomics reveals subfamilies of fungal nonribosomal peptide synthetases and their evolutionary relationships, *BMC Evol. Biol.*, 2010, **10**, 26.
18 B. T. Piatkowski, *et al.*, Phylogenomics reveals convergent evolution of red-violet coloration in land plants and the origins of the anthocyanin biosynthetic pathway, *Mol. Phylogenet. Evol.*, 2020, **151**, 106904.
19 A. E. Wilson and L. Tian, Phylogenomic analysis of UDP-dependent glycosyltransferases provides insights into the evolutionary landscape of glycosylation in plant metabolism, *Plant J.*, 2019, **100**, 1273–1288.
20 Y. Shimizu, H. Ogata and S. Goto, Type III Polyketide Synthases: Functional Classification and Phylogenomics, *ChemBioChem*, 2017, **18**, 50–65.

21 H. Jenke-Kodama, A. Sandmann, R. Müller and E. Dittmann, Evolutionary Implications of Bacterial Polyketide Synthases, *Mol. Biol. Evol.*, 2005, **22**, 2027–2039.
22 A. M. Dean and J. W. Thornton, Mechanistic approaches to the study of evolution, *Nat. Rev. Genet.*, 2007, **8**, 675–688.
23 M. A. DePristo, D. M. Weinreich and D. L. Hartl, Missense meanderings in sequence space: a biophysical view of protein evolution, *Nat. Rev. Genet.*, 2005, **6**, 678–687.
24 C. Pál, B. Papp and M. J. Lercher, An integrated view of protein evolution, *Nat. Rev. Genet.*, 2006, **7**, 337–348.
25 M. Alanjary, *et al.*, The Antibiotic Resistant Target Seeker (ARTS), an exploration engine for antibiotic cluster prioritization and novel drug target discovery, *Nucleic Acids Res.*, 2017, **45**, W42–W48.
26 P. Cruz-Morales, *et al.*, Phylogenomic Analysis of Natural Products Biosynthetic Gene Clusters Allows Discovery of Arseno-Organic Metabolites in Model Streptomycetes, *Genome Biol. Evol.*, 2016, **8**, 1906–1916.
27 N. Sélem-Mojica, C. Aguilar, K. Gutiérrez-García, C. E. Martínez-Guerrero and F. Barona-Gómez, EvoMining reveals the origin and fate of natural product biosynthetic enzymes, *Microb. Genomics*, 2019, **5**(12), e000260.
28 D. Alvarez-Ponce, Richard Dickerson, Molecular Clocks, and Rates of Protein Evolution, *J. Mol. Evol.*, 2021, **89**, 122–126.
29 A. Rokas, J. H. Wisecaver and A. L. Lind, The birth, evolution and death of metabolic gene clusters in fungi, *Nat. Rev. Microbiol.*, 2018, **16**, 731–744.
30 A. Rokas, M. E. Mead, J. L. Steenwyk, H. A. Raja and N. H. Oberlies, Biosynthetic gene clusters and the evolution of fungal chemodiversity, *Nat. Prod. Rep.*, 2020, **37**, 868–878.
31 M. T. Drott, *et al.*, Microevolution in the pansecondary metabolome of Aspergillus flavus and its potential macroevolutionary implications for filamentous fungi, *Proc. Natl. Acad. Sci. U. S. A.*, 2021, **118**(21), 1–10.
32 J.-K. Weng, The evolutionary paths towards complexity: a metabolic perspective, *New Phytol.*, 2014, **201**, 1141–1149.
33 G. D. Moghe and R. L. Last, Something Old, Something New: Conserved Enzymes and the Evolution of Novelty in Plant Specialized Metabolism, *Plant Physiol.*, 2015, **169**, 1512–1523.
34 F. M. Megahed and L. A. Jones-Farmer, Statistical Perspectives on "Big Data", in *Frontiers in Statistical Quality Control 11*, ed. S. Knoth and W. Schmid, Springer International Publishing, 2015, pp. 29–47, DOI: 10.1007/978-3-319-12355-4_3.
35 F. Barona-Gómez, Re-annotation of the sequence > annotation: opportunities for the functional microbiologist, *Microb. Biotechnol.*, 2015, **8**, 2–4.
36 E. M. Cahan, T. Hernandez-Boussard, S. Thadaney-Israni and D. L. Rubin, Putting the data before the algorithm in big data addressing personalized healthcare, *npj Digit. Med.*, 2019, **2**, 1–6.
37 V. Marx, The big challenges of big data, *Nature*, 2013, **498**, 255–260.
38 X. Jin, B. W. Wah, X. Cheng and Y. Wang, Significance and Challenges of Big Data Research, *Big Data Res.*, 2015, **2**, 59–64.
39 M. H. Medema, *et al.*, Minimum Information about a Biosynthetic Gene cluster, *Nat. Chem. Biol.*, 2015, **11**, 625–631.
40 J. C. Navarro-Muñoz, *et al.*, A computational framework to explore large-scale biosynthetic diversity, *Nat. Chem. Biol.*, 2020, **16**, 60–68.
41 K. C. Belknap, C. J. Park, B. M. Barth and C. P. Andam, Genome mining of biosynthetic and chemotherapeutic gene clusters in Streptomyces bacteria, *Sci. Rep.*, 2020, **10**, 2003.
42 E. A. Barka, *et al.*, Taxonomy, Physiology, and Natural Products of Actinobacteria, *Microbiol. Mol. Biol. Rev.*, 2016, **80**, 1–43.
43 N. F. AbuSara, *et al.*, Comparative Genomics and Metabolomics Analyses of Clavulanic Acid-Producing Streptomyces Species Provides Insight Into Specialized Metabolism, *Front. Microbiol.*, 2019, **10**, 1–17.
44 J. R. Doroghazi and W. W. Metcalf, Comparative genomics of actinomycetes with a focus on natural product biosynthetic genes, *BMC Genomics*, 2013, **14**, 611.
45 D. Männle, *et al.*, Comparative Genomics and Metabolomics in the Genus Nocardia, *mSystems*, 2020, **5**, e00125-20.
46 N. Ziemert, *et al.*, Diversity and evolution of secondary metabolism in the marine actinomycete genus Salinispora, *Proc. Natl. Acad. Sci. U. S. A.*, 2014, **111**, E1130–E1139.
47 M. S. Hifnawy, *et al.*, The genus Micromonospora as a model microorganism for bioactive natural product discovery, *RSC Adv.*, 2020, **10**, 20939–20959.
48 S. L. Goldstein and J. L. Klassen, Pseudonocardia Symbionts of Fungus-Growing Ants and the Evolution of Defensive Secondary Metabolism, *Front. Microbiol.*, 2020, **11**, 621041.
49 M. A. Schorn, *et al.*, Sequencing rare marine actinomycete genomes reveals high density of unique natural product biosynthetic gene clusters, *Microbiology*, 2016, **162**, 2075–2086.
50 A. Undabarrena, *et al.*, Rhodococcus comparative genomics reveals a phylogenomic-dependent non-ribosomal peptide synthetase distribution: insights into biosynthetic gene cluster connection to an orphan metabolite, *Microb. Genomics*, 2021, **7**(7), 1–17.
51 M. G. Chevrette, P. A. Hoskisson and F. Barona-Gómez, Enzyme Evolution in Secondary Metabolism, in *Comprehensive Natural Products III*, Elsevier, 2020, pp. 90–112, DOI: 10.1016/B978-0-12-409547-2.14712-2.
52 O. Khersonsky and D. S. Tawfik, Enzyme promiscuity: a mechanistic and evolutionary perspective, *Annu. Rev. Biochem.*, 2010, **79**, 471–505.
53 L. Noda-Garcia, W. Liebermeister and D. S. Tawfik, Metabolite–Enzyme Coevolution: From Single Enzymes to Metabolic Pathways and Networks, *Annu. Rev. Biochem.*, 2018, **87**, 187–216.

54 L. Noda-Garcia and D. S. Tawfik, Enzyme evolution in natural products biosynthesis: target- or diversity-oriented?, *Curr. Opin. Chem. Biol.*, 2020, **59**, 147–154.
55 E. Dittmann, M. Gugger, K. Sivonen and D. P. Fewer, Natural Product Biosynthetic Diversity and Comparative Genomics of the Cyanobacteria, *Trends Microbiol.*, 2015, **23**, 642–652.
56 Z. Liu, *et al.*, Formation and diversification of a paradigm biosynthetic gene cluster in plants, *Nat. Commun.*, 2020, **11**, 5354.
57 P. Fan, *et al.*, Evolution of a plant gene cluster in Solanaceae and emergence of metabolic diversity, *eLife*, 2020, **9**, e56717.
58 Z. Liu, *et al.*, Drivers of metabolic diversification: how dynamic genomic neighbourhoods generate new biosynthetic pathways in the Brassicaceae, *New Phytol.*, 2020, **227**, 1109–1123.
59 M.-C. Tang, Y. Zou, K. Watanabe, C. T. Walsh and Y. Tang, Oxidative Cyclization in Natural Product Biosynthesis, *Chem. Rev.*, 2017, **117**, 5226–5333.
60 M. Montalbán-López, *et al.*, New developments in RiPP discovery, enzymology and engineering, *Nat. Prod. Rep.*, 2021, **38**, 130–239.
61 M. D. Mungan, *et al.*, ARTS 2.0: feature updates and expansion of the Antibiotic Resistant Target Seeker for comparative genome mining, *Nucleic Acids Res.*, 2020, **48**, W546–W552.
62 L. Nakhleh, Evolutionary Trees, in *Brenner's Encyclopedia of Genetics*, Elsevier, 2013, pp. 549–550, DOI: 10.1016/B978-0-12-374984-0.00504-0.
63 E. Avni and S. Snir, A New Phylogenomic Approach For Quantifying Horizontal Gene Transfer Trends in Prokaryotes, *Sci. Rep.*, 2020, **10**, 12425.
64 M. Wang, *et al.*, Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking, *Nat. Biotechnol.*, 2016, **34**, 828–837.
65 S. A. Kautsar, *et al.*, MIBiG 2.0: a repository for biosynthetic gene clusters of known function, *Nucleic Acids Res.*, 2019, gkz882, DOI: 10.1093/nar/gkz882.
66 K. Blin, M. H. Medema, R. Kottmann, S. Y. Lee and T. Weber, The antiSMASH database, a comprehensive database of microbial secondary metabolite biosynthetic gene clusters, *Nucleic Acids Res.*, 2017, **45**, D555–D559.
67 K. Blin, S. Shaw, S. A. Kautsar, M. H. Medema and T. Weber, The antiSMASH database version 3: increased taxonomic coverage and new query features for modular enzymes, *Nucleic Acids Res.*, 2021, **49**, D639–D643.
68 J. K. Schniete, *et al.*, ActDES – a curated Actinobacterial Database for Evolutionary Studies, *Microb. Genomics*, 2021, **7**(1), 000498.
69 K. Palaniappan, *et al.*, IMG-ABC v.5.0: an update to the IMG/Atlas of Biosynthetic Gene Clusters Knowledgebase, *Nucleic Acids Res.*, 2019, gkz932, DOI: 10.1093/nar/gkz932.
70 S. A. Kautsar, K. Blin, S. Shaw, T. Weber and M. H. Medema, BiG-FAM: the biosynthetic gene cluster families database, *Nucleic Acids Res.*, 2021, **49**, D490–D497.
71 A. L. Mitchell, *et al.*, MGnify: the microbiome analysis resource in 2020, *Nucleic Acids Res.*, 2020, **48**, D570–D578.
72 S. Nayfach, *et al.*, A genomic catalog of Earth's microbiomes, *Nat. Biotechnol.*, 2020, 1–11, DOI: 10.1038/s41587-020-0718-6.
73 B. P. Alcock, *et al.*, CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database, *Nucleic Acids Res.*, 2020, **48**, D517–D525.
74 V. Bortolaia, *et al.*, ResFinder 4.0 for predictions of phenotypes from genotypes, *J. Antimicrob. Chemother.*, 2020, **75**, 3491–3500.
75 F. Meyer, *et al.*, The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes, *BMC Bioinf.*, 2008, **9**, 386.
76 S. F. Altschul, W. Gish, W. Miller, E. W. Myers and D. J. Lipman, Basic local alignment search tool, *J. Mol. Biol.*, 1990, **215**, 403–410.
77 S. Kim, *et al.*, PubChem 2019 update: improved access to chemical data, *Nucleic Acids Res.*, 2019, **47**, D1102–D1109.
78 J. A. van Santen, *et al.*, The Natural Products Atlas: An Open Access Knowledge Base for Microbial Natural Products Discovery, *ACS Cent. Sci.*, 2019, **5**, 1824–1833.
79 P. A. Hoskisson and R. F. Seipke, Cryptic or Silent? The Known Unknowns, Unknown Knowns, and Unknown Unknowns of Secondary Metabolism, *mBio*, 2020, **11**(5), e02642–20.
80 A. Crits-Christoph, N. Bhattacharya, M. R. Olm, Y. S. Song and J. F. Banfield, Transporter genes in biosynthetic gene clusters predict metabolite characteristics and siderophore activity, *Genome Res.*, 2020, **31**(2), 239–250.
81 M. Alanjary, K. Steinke and N. Ziemert, AutoMLST: an automated web server for generating multi-locus species trees highlighting natural product potential, *Nucleic Acids Res.*, 2019, **47**, W276–W282.
82 M. Adamek, M. Alanjary and N. Ziemert, Applied evolution: phylogeny-based approaches in natural products research, *Nat. Prod. Rep.*, 2019, **36**, 1295–1312.
83 D. Bzdok, M. Krzywinski and N. Altman, Machine learning: supervised methods, *Nat. Methods*, 2018, **15**, 5–6.
84 J. Y. Yang and O. K. Ersoy, *Combined Supervised and Unsupervised Learning in Genomic Data Mining*, 2003, p. 143.
85 P. Cimermancic, *et al.*, Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters, *Cell*, 2014, **158**, 412–421.
86 T. A. J. van der Lee and M. H. Medema, Computational strategies for genome-based natural product discovery and engineering in fungi, *Fungal Genet. Biol.*, 2016, **89**, 29–36.
87 T. Wolf, V. Shelest, N. Nath and E. Shelest, CASSIS and SMIPS: promoter-based prediction of secondary metabolite gene clusters in eukaryotic genomes, *Bioinformatics*, 2016, **32**, 1138–1143.
88 S. Argimón, *et al.*, Microreact: visualizing and sharing data for genomic epidemiology and phylogeography, *Microb. Genomics*, 2016, **2**(11), e000093.
89 S. A. Kautsar, H. G. Suarez Duran, K. Blin, A. Osbourn and M. H. Medema, plantiSMASH: automated identification,

annotation and expression analysis of plant biosynthetic gene clusters, *Nucleic Acids Res.*, 2017, **45**, W55–W63.
90 L. Krause, *et al.*, GISMO—gene identification using a support vector machine for ORF classification, *Nucleic Acids Res.*, 2007, **35**, 540–549.
91 A. S. Walker and J. Clardy, A Machine Learning Bioinformatics Method to Predict Biological Activity from Biosynthetic Gene Clusters, *J. Chem. Inf. Model.*, 2021, **61**(6), 2560–2571.
92 A. M. Kloosterman, *et al.*, Expansion of RiPP biosynthetic space through integration of pan-genomics and machine learning uncovers a novel class of lanthipeptides, *PLoS Biol.*, 2020, **18**, e3001026.
93 A. M. Eren, *et al.*, Anvi'o: an advanced analysis and visualization platform for 'omics data, *PeerJ*, 2015, **3**, e1319.
94 M. H. Medema, E. Takano and R. Breitling, Detecting Sequence Homology at the Gene Cluster Level with MultiGeneBlast, *Mol. Biol. Evol.*, 2013, **30**, 1218–1223.
95 C. L. M. Gilchrist and Y.-H. Chooi, clinker & clustermap.js: automatic generation of gene cluster comparison figures, *Bioinformatics*, 2021, btab007.
96 C. K. Saha, R. Sanches Pires, H. Brolin, M. Delannoy and G. C. Atkinson, FlaGs and webFlaGs: discovering novel biology through the analysis of gene neighbourhood conservation, *Bioinformatics*, 2020, **37**(9), 1312.
97 V. M. Gumerov and I. B. Zhulin, TREND: a platform for exploring protein function in prokaryotes based on phylogenetic, domain architecture and gene neighborhood analyses, *Nucleic Acids Res.*, 2020, **48**, W72–W76.
98 S. A. Kautsar, J. J. J. van der Hooft, D. de Ridder and M. H. Medema, BiG-SLiCE: A highly scalable tool maps the diversity of 1.2 million biosynthetic gene clusters, *GigaScience*, 2021, **10**, giaa154.
99 M. G. Chevrette and C. R. Currie, Emerging evolutionary paradigms in antibiotic discovery, *J. Ind. Microbiol. Biotechnol.*, 2019, **46**, 257–271.
100 M. G. Chevrette, *et al.*, The antimicrobial potential of Streptomyces from insect microbiomes, *Nat. Commun.*, 2019, **10**, 516.
101 I. J. Miller, M. G. Chevrette and J. C. Kwan, Interpreting Microbial Biosynthesis in the Genomic Age: Biological and Practical Considerations, *Mar. Drugs*, 2017, **15**, 165.
102 E. J. Caldera, M. G. Chevrette, B. R. McDonald and C. R. Currie, Local Adaptation of Bacterial Symbionts within a Geographic Mosaic of Antibiotic Coevolution, *Appl. Environ. Microbiol.*, 2019, **85**(24), e01580-19.
103 A. Iglesias, A. Latorre-Pérez, J. E. M. Stach, M. Porcar and J. Pascual, Out of the Abyss: Genome and Metagenome Mining Reveals Unexpected Environmental Distribution of Abyssomicins, *Front. Microbiol.*, 2020, **11**.
104 A. M. Sharrar, *et al.*, Bacterial Secondary Metabolite Biosynthetic Potential in Soil Varies with Phylum, Depth, and Vegetation Type, *mBio*, 2020, **11**(3), e00416–20.
105 S. G. Silva, J. Blom, T. Keller-Costa and R. Costa, Comparative genomics reveals complex natural product biosynthesis capacities and carbon metabolism across host-associated and free-living Aquimarina (Bacteroidetes, Flavobacteriaceae) species, *Environ. Microbiol.*, 2019, **21**, 4002–4019.
106 Y. Yang, *et al.*, Genomic characteristics and comparative genomics analysis of the endophytic fungus Sarocladium brachiariae, *BMC Genomics*, 2019, **20**, 782.
107 K. Gutiérrez-García, *et al.*, Cycad Coralloid Roots Contain Bacterial Communities Including Cyanobacteria and Caulobacter spp. That Encode Niche-Specific Biosynthetic Gene Clusters, *Genome Biol. Evol.*, 2019, **11**, 319–334.
108 R. M. Stubbendieck, *et al.*, Competition among Nasal Bacteria Suggests a Role for Siderophore-Mediated Interactions in Shaping the Human Nasal Microbiota, *Appl. Environ. Microbiol.*, 2019, **85**(10), e02406-18.
109 M. G. Chevrette, *et al.*, Taxonomic and Metabolic Incongruence in the Ancient Genus Streptomyces, *Front. Microbiol.*, 2019, **10**, 2170.
110 Â. Brito, *et al.*, Comparative Genomics Discloses the Uniqueness and the Biosynthetic Potential of the Marine Cyanobacterium Hyella patelloides, *Front. Microbiol.*, 2020, **11**(1527), 1–15.
111 J. R. Doroghazi, *et al.*, A roadmap for natural product discovery based on large-scale genomics and metabolomics, *Nat. Chem. Biol.*, 2014, **10**, 963–968.
112 T. Hoffmann, *et al.*, Correlating chemical diversity with taxonomic distance for discovery of natural products in myxobacteria, *Nat. Commun.*, 2018, **9**, 803.
113 E. Gluck-Thaler, *et al.*, The Architecture of Metabolism Maximizes Biosynthetic Diversity in the Largest Class of Fungi, *Mol. Biol. Evol.*, 2020, **37**, 2838–2856.
114 F. Baldeweg, D. Hoffmeister and M. Nett, A genomics perspective on natural product biosynthesis in plant pathogenic bacteria, *Nat. Prod. Rep.*, 2019, **36**, 307–325.
115 E. V. Koonin, Archaeal ancestors of eukaryotes: not so elusive any more, *BMC Biol.*, 2015, **13**(84), 1–7.
116 M. H. Medema, P. Cimermancic, A. Sali, E. Takano and M. A. Fischbach, A Systematic Computational Analysis of Biosynthetic Gene Cluster Evolution: Lessons for Engineering Biosynthesis, *PLoS Comput. Biol.*, 2014, **10**, e1004016.
117 N. M. Vior, *et al.*, Discovery and Biosynthesis of the Antibiotic Bicyclomycin in Distantly Related Bacterial Classes, *Appl. Environ. Microbiol.*, 2018, **84**(9), e02828-17.
118 B. R. McDonald and C. R. Currie, Lateral Gene Transfer Dynamics in the Ancient Bacterial Genus Streptomyces, *mBio*, 2017, **8**(3), e00644-17.
119 A. B. Chase, D. Sweeney, M. N. Muskat, D. Guillén-Matus and P. R. Jensen, Vertical inheritance governs biosynthetic gene cluster evolution and chemical diversification, *bioRxiv*, 2020, 12.19.423547, DOI: 10.1101/2020.12.19.423547.
120 J. Bérdy, Bioactive Microbial Metabolites, *J. Antibiot.*, 2005, **58**, 1–26.
121 M. F. Traxler and R. Kolter, Natural products in soil microbe interactions and evolution, *Nat. Prod. Rep.*, 2015, **32**, 956–970.

122 C. P. Andam, M. J. Choudoir, A. Vinh Nguyen, H. Sol Park and D. H. Buckley, Contributions of ancestral interspecies recombination to the genetic diversity of extant Streptomyces lineages, *ISME J.*, 2016, **10**, 1731–1741.
123 Y. Li, *et al.*, Population Genomics Insights into Adaptive Evolution and Ecological Differentiation in Streptomycetes, *Appl. Environ. Microbiol.*, 2019, **85**, e02555-18.
124 A.-R. Tidjani, *et al.*, Massive Gene Flux Drives Genome Diversity between Sympatric Streptomyces Conspecifics, *mBio*, 2019, **10**, e01533–19.
125 B. R. McDonald, *et al.*, Biogeography and Microscale Diversity Shape the Biosynthetic Potential of Fungus-growing Ant-associated Pseudonocardia, *bioRxiv*, 2019, 545640, DOI: 10.1101/545640.
126 V. M. Zacharia, *et al.*, Genetic Network Architecture and Environmental Cues Drive Spatial Organization of Phenotypic Division of Labor in Streptomyces coelicolor, *mBio*, 2021, e00794–21.
127 Z. Zhang, *et al.*, Antibiotic production in Streptomyces is organized by a division of labor through terminal genomic differentiation, *Sci. Adv.*, 2020, **6**, eaay5781.
128 M. Bahram, *et al.*, Structure and function of the global topsoil microbiome, *Nature*, 2018, **560**, 233–237.
129 M. Delgado-Baquerizo, *et al.*, A global atlas of the dominant bacteria found in soil, *Science*, 2018, **359**, 320–325.
130 L. R. Thompson, *et al.*, A communal catalogue reveals Earth's multiscale microbial diversity, *Nature*, 2017, **551**, 457–463.
131 H. Wang, *et al.*, Soil Bacterial Diversity Is Associated with Human Population Density in Urban Greenspaces, *Environ. Sci. Technol.*, 2018, **52**, 5115–5124.
132 J. Handelsman, M. R. Rondon, S. F. Brady, J. Clardy and R. M. Goodman, Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products, *Chem. Biol.*, 1998, **5**, R245–R249.
133 S. Nasrin, *et al.*, Chloramphenicol Derivatives with Antibacterial Activity Identified by Functional Metagenomics, *J. Nat. Prod.*, 2018, **81**, 1321–1332.
134 A. L. R. Santana-Pereira, *et al.*, Discovery of Novel Biosynthetic Gene Cluster Diversity From a Soil Metagenomic Library, *Front. Microbiol.*, 2020, **11**(585398), 1–17.
135 B. Dror, Z. Wang, S. F. Brady, E. Jurkevitch and E. Cytryn, Elucidating the Diversity and Potential Function of Nonribosomal Peptide and Polyketide Biosynthetic Gene Clusters in the Root Microbiome, *mSystems*, 2020, **5**(6), e00866–20.
136 M. Elfeki, M. Alanjary, S. J. Green, N. Ziemert and B. T. Murphy, Assessing the Efficiency of Cultivation Techniques To Recover Natural Product Biosynthetic Gene Populations from Sediment, *ACS Chem. Biol.*, 2018, **13**, 2074–2081.
137 C. Lemetre, *et al.*, 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**, 11615–11620.
138 B. V. B. Reddy, *et al.*, Natural Product Biosynthetic Gene Diversity in Geographically Distinct Soil Microbiomes, *Appl. Environ. Microbiol.*, 2012, **78**, 3744–3752.
139 N. Waglechner, A. G. McArthur and G. D. Wright, Phylogenetic reconciliation reveals the natural history of glycopeptide antibiotic biosynthesis and resistance, *Nat. Microbiol.*, 2019, **4**, 1862–1871.
140 R. D. Firn and C. G. Jones, Natural products? a simple model to explain chemical diversity, *Nat. Prod. Rep.*, 2003, **20**, 382.
141 T. Nguyen, *et al.*, Exploiting the mosaic structure of trans-acyltransferase polyketide synthases for natural product discovery and pathway dissection, *Nat. Biotechnol.*, 2008, **26**, 225–233.
142 J. Masschelein, M. Jenner and G. L. Challis, Antibiotics from Gram-negative bacteria: a comprehensive overview and selected biosynthetic highlights, *Nat. Prod. Rep.*, 2017, **34**, 712–783.
143 E. J. N. Helfrich, R. Ueoka, M. G. Chevrette, *et al.* Evolution of combinatorial diversity in trans-acyltransferase polyketide synthase assembly lines across bacteria., *Nat. Commun.*, 2021, **12**(1), 1422.
144 M. G. Chevrette, F. Aicheler, O. Kohlbacher, C. R. Currie and M. H. Medema, SANDPUMA: ensemble predictions of nonribosomal peptide chemistry reveal biosynthetic diversity across Actinobacteria, *Bioinformatics*, 2017, **33**, 3202–3210.
145 M. A. Schorn, *et al.*, A community resource for paired genomic and metabolomic data mining, *Nat. Chem. Biol.*, 2021, 1–6, DOI: 10.1038/s41589-020-00724-z.
146 V. Tracanna, *et al.*, Dissecting Disease-Suppressive Rhizosphere Microbiomes by Functional Amplicon Sequencing and 10× Metagenomics, *mSystems*, 2021, **6**(3), e01116-20.
147 M. B. Austin, P. E. O'Maille and J. P. Noel, Evolving biosynthetic tangos negotiate mechanistic landscapes, *Nat. Chem. Biol.*, 2008, **4**, 217–222.
148 W. Ding, F. Baumdicker and R. A. Neher, panX: pan-genome analysis and exploration, *Nucleic Acids Res.*, 2018, **46**, e5.
149 N. L. Grenade, G. W. Howe and A. C. Ross, The convergence of bacterial natural products from evolutionarily distinct pathways, *Curr. Opin. Biotechnol.*, 2021, **69**, 17–25.
150 F. Del Carratore, *et al.*, Computational identification of co-evolving multi-gene modules in microbial biosynthetic gene clusters, *Commun. Biol.*, 2019, **2**, 1–10.
151 L. Martinet, *et al.*, A Single Biosynthetic Gene Cluster Is Responsible for the Production of Bagremycin Antibiotics and Ferroverdin Iron Chelators, *mBio*, 2019, **10**, e01230–19.
152 S. Wiegand, *et al.*, Cultivation and functional characterization of 79 planctomycetes uncovers their unique biology, *Nat. Microbiol.*, 2020, **5**, 126–140.
153 A. Cibrián-Jaramillo, Increasing Metagenomic Resolution of Microbiome Interactions Through Functional Phylogenomics and Bacterial Sub-Communities, *Front. Genet.*, 2016, **7**, 1–8.

154 M. Le Boulch, P. Déhais, S. Combes and G. Pascal, The MACADAM Database: A MetAboliC PAthways DAtabase for Microbial Taxonomic Groups for Mining Potential Metabolic Capacities of Archaeal and Bacterial Taxonomic Groups, *Database*, 2019, 1–14.

155 M. Sorokina, P. Merseburger, K. Rajan, M. A. Yiriki and C. Steinbeck, COCONUT online: Collection of Open Natural Products database, *J. Cheminf.*, 2021, **13**(1), 2.

156 D. Klementz, K. Döring, X. Lucas, K. K. Telukunta and D. Deubel, StreptomeDB 2.0—an extended resource of natural products produced by streptomycetes, *Nucleic Acids Res.*, 2016, **44**(D1), D509–D514.

157 A. Rutz, M. Sorokina, J. Galgonek, *et al.* The LOTUS Initiative for Open Natural Products Research: Knowledge Management through Wikidata., *BioRxiv*, 2021, DOI: 10.1101/2021.02.28.433265.

158 A. Crits-Christoph, M. R. Olm, S. Diamond, K. Bouma-Gregson and J. F. Banfield, Soil Bacterial Populations Are Shaped by Recombination and Gene-Specific Selection across a Grassland Meadow, *ISME J.*, 2020, **14**(7), 1834–1846.

Annexure B

Supplemental Information

Evaluating Distribution of Bacterial Natural Product Biosynthetic Genes Across Lake Huron Sediment

Supplementary Experimental Procedures.

16S rRNA Gene Amplification and Sequencing

The V4 region of microbial small subunit rRNA genes (16S rRNA) was PCR-amplified from genomic DNA using a two-stage PCR protocol, as described previously.[1] 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).[2] Each 25 μL PCR reaction mixture consisted of 0.5 μL 10 ng/μL gDNA, 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 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[3] 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 using the DADA2 option within the software package Qiime2[4] for sequence quality control and feature table construction. The resulting analysis generated 141,078 amplicon sequence variants (ASVs).[4] A sequence representative of each ASV was classified using the Silva_128 database.[5] A taxon-by-sample abundance matrix (a feature table or biological observation matrix, BIOM)[6] file was then created.

Estimated amount of DNA in KSα and A Domain Amplification

To demonstrate that the amount of gDNA used in a PCR reaction was representative of a full biome: 1 PCR reaction contained ~10 ng DNA. 10 ng of gDNA is roughly equivalent to gDNA from ~500,000 to 20,000,000 soil bacterial cells (a single gram of healthy soil contains 11.0×10^{10} cells in 1 g dry weight at the sediment surface, and a bacterial cell contains approximately 0.5–20 fg DNA).

Accession Codes.

Sample SRA data can be accessed using Accession code PRJNA690811.

Table S1. Sediment sample collection data for Lake Huron expedition.

Sample name	Longitude	Latitude
GB01	-80.8563933	44.71783667
GB03	-80.61701	44.72527
GB04	-80.1672617	44.64574667
GB05	-80.2431117	44.796915
GB06	-80.435975	44.73815167
GB09	-79.9675	44.87164167
GB12	-80.8747533	44.92021167
GB17	-80.8742217	45.24485
GB29	-81.0829917	45.58357
GB35	-81.670485	45.52572833
GB36	-81.620125	45.70816833
GB39	-81.2583983	45.87294667
GB42	-81.5954067	45.91245667
H001	-83.6141917	43.937425
H002	-83.3324467	44.12494167
H006	-82.0184967	43.52649333
H012	-82.1130467	43.900655
H027	-82.5024567	44.09988833
H032	-82.3596233	44.35418333
H037	-82.7836283	44.76185333
H038	-82.2023783	44.75069333
H048	-82.5911867	45.26139333
H054	-83.402845	45.63384
H061	-83.9164083	45.74978833
H096	-82.83258	44.33275
H101	-82.3348767	43.26900667
H102	-82.403855	43.70586833
H103	-82.2209217	44.14485833
H104	-81.83796	44.37196167
H107	-82.554065	44.61541667
H108	-83.05021	44.557415
H109	-83.000015	44.150185
H110	-83.8036883	43.77230833
H118	-83.165955	44.91682333
H119	-82.8106817	45.39766833
H121	-83.403945	45.81889667
H123	-83.90591	45.93646167
H124	-84.4215683	45.85121

HTXD	-82.33345	43.33989
HTXM	-82.46681	43.33977
HTXS	-82.4991167	43.33974333
HXSG	-82.4991167	43.33974333
NC68	-83.8536033	46.04127
NC70	-83.671975	46.13648
NC71	-83.74624	46.23346833
NC73	-83.3551783	46.18685167
NC76	-83.4329117	46.00034
NC77	-83.1977083	45.97041667
NC79	-82.886655	46.12299667
NC82	-82.7588	45.93686333
NC83	-82.5497	45.99998167
NC84	-82.5564417	46.09173833
NC87	-82.197085	46.06112167
NC88	-81.999815	46.05529667
NC89	45.91649	-82.1617117
TB01	-83.1496367	44.89958667
TB02	-83.240505	44.93872833
TB03	-83.277	44.95524667
TB04	-83.0352944	44.15244444

Table S2. Sediment sample gDNA quantification using Nanodrop.

Sample	ng/μL	Sample	ng/μL
GB01-A	17.5	GB01-B	N/A
GB03-A	16.5	GB03-B	22
GB04-A	14.4	GB04-B	15.9
GB05-A	20	GB05-B	14.9
GB06-A	13	GB06-B	N/A
GB09-A	26.3	GB09-B	21.2
GB12-A	N/A	GB12-B	15
GB17-A	8.2	GB17-B	8.8
GB29-A	17.6	GB29-B	16
GB35-A	17.9	GB35-B	N/A
GB36-A	10.9	GB36-B	19
GB39-A	13.8	GB39-B	17.5
GB42-A	15.3	GB42-B	20.2
H001-A	25.1	H001-B	20.1
H002-A	18.1	H002-B	18.2
H006-A	26.5	H006-B	24.4
H012-A	13.8	H012-B	14.8
H027-A	23.3	H027-B	30.7
H032-A	13.9	H032-B	17.9
H037-A	16.5	H037-B	12.1
H038-A	10.6	H038-B	22.2
H048-A	30.6	H048-B	26.4
H054-A	13	H054-B	16.7
H061-A	15.3	H061-B	11.4
H096-A	17.8	H096-B	14.2
H101-A	18.4	H101-B	17.6
H102-A	24.4	H102-B	22.5
H103-A	18.1	H103-B	15.2
H104-A	7.4	H104-B	9.8
H107-A	9.8	H107-B	15.9
H108-A	18.1	H108-B	20.2
H109-A	20.1	H109-B	24.3
H110-A	30.8	H110-B	32.3
H118-A	6.2	H118-B	5.9
H119-A	29.1	H119-B	25
H121-A	14.3	H121-B	13.4
H123-A	24	H123-B	25.1
H124-A	27.4	H124-B	17
HTXD-A	9.3	HTXD-B	13.5
HTXM-A	6.8	HTXM-B	6.4
HTXS-A	4	HTXS-B	5
HXSG-A	3.1	HXSG-B	3.8
NC68-A	11.4	NC68-B	19.9
NC70-A	23.8	NC70-B	19.1
NC71-A	4.9	NC71-B	19.5
NC73-A	22.5	NC73-B	13.3
NC76-A	19.3	NC76-B	20.1
NC77-A	13.1	NC77-B	13.2
NC79-A	23.1	NC79-B	32.5
NC82-A	21.7	NC82-B	19.3

NC83-A	15.5	NC83-B	17.4
NC84-A	13	NC84-B	13
NC87-A	21.1	NC87-B	10.6
NC88-A	24.4	NC88-B	6.7
NC89-A	14.9	NC89-B	14.4
NE01-A	6.8	NE01-B	3.9
NE02-A	5.3	NE02-B	7.4
NE03-A	9.9	NE03-B	9.4
NE04-A	5.1	NE04-B	1.8
TB01-A	17.5	TB01-B	15.7
TB02-A	6.6	TB02-B	4.3
TB03-A	6.1	TB03-B	8.4
TB04-A	6.9	TB04-B	8.2

Table S3. Sequence reads per sample before and after filtering.

S3A. A domain sequence reads per sample before and after filtering

Sample	*A domain sequence count*	
	before filtering	*after filtering*
GB01	100068	29538
GB03	87304	33514
GB04	81992	30589
GB05	79378	30822
GB06	94934	23731
GB09	16605	3487
GB12	84785	20090
GB17	63581	19046
GB29	93624	30221
GB35	104613	24546
GB36	120819	19103
GB39	90825	29671
GB42	66836	26490
H001	47575	22825
H002	46549	24705
H006	58045	26476
H012	52019	27589
H027	54526	28092
H032	55399	29310
H037	43365	11407
H038	74084	28311
H048	42879	25550
H054	57279	30289
H054B	47330	27082
H061	27332	19480
H096	32191	19325

H101	49232	31717
H102	38635	24216
H103	49020	28022
H104	43140	23638
H107	46000	26520
H108	56986	30330
H109	50588	25213
H110	61123	31547
H118	54969	29164
H119	60357	32965
H121	42440	24175
H123	50282	29375
H124	64667	39958
HTXD	44915	25960
HTXM	61883	31140
HTXS	64642	34153
HXSG	65939	36061
NC68	93915	44407
NC68B	86165	27941
NC70	85876	35580
NC71	118527	33900
NC73	78861	29530
NC76	67029	32849
NC77	76025	29920
NC79	63627	14229
NC82	80280	30078
NC83	77778	29099
NC84	77518	29234
NC87	79403	25970
NC88	74994	32434
NC89	53	2
TB01	91511	33540
TB02	109450	31474
TB03	122427	14537
TB04	91202	31209

S3B. KSα domain sequence reads per sample before and after filtering

Sample	*KSα domain sequence count*	
	before filtering	*after filtering*
GB01	142872	40
GB03	116849	42
GB04	117531	43

GB05	52214	29
GB06	106274	45
GB09	94289	33
GB12	122097	65
GB17	93897	15
GB29	103481	49
GB35	131773	35
GB36	134277	47
GB39	110956	41
GB42	145182	40
H001	115109	32
H002	106612	43
H006	113030	33
H012	200784	33
H027	103413	39
H032	146624	31
H037	115310	109
H038	116346	26
H048	8575	108
H054	129969	929
H061	115873	13
H096	123566	89
H101	126539	29
H102	78024	70
H103	64881	63
H104	94503	10
H107	66572	19
H108	116249	30
H109	1929	13
H110	257816	313
H118	76047	212
H119	201467	50
H121	123562	71
H123	121709	101
H124	112397	63
HTBXM	100854	54
HTBXD	522	47
HTBXS	196054	60
HXSGB	182584	50
NC79	11704	84
NC82	130853	101
NC83	125930	106

NC84	112485	64
NC87	126142	60
NC88	15451	58
NC89	146028	35
TB01	159206	81
TB02	131451	66
TB03	143707	64
TB04	73951	21

Table S4. Most abundant phyla in sediment

Kingdom	**Phylum**	**average (%)**
Bacteria	Proteobacteria	43.94
Bacteria	Acidobacteria	7.07
Bacteria	Bacteroidetes	6.21
Bacteria	Nitrospirae	5.73
Bacteria	Actinobacteria	5.60
Bacteria	Planctomycetes	5.22
Bacteria	Verrucomicrobia	5.16
Unassigned	Other	4.54
Bacteria	Cyanobacteria	4.07
Bacteria	Chloroflexi	2.50

Table S5. Number of sequences in most abundant OBUs

S5A. A domain OBUs

Location	*% sequences of most abundant A domain OBU*
GB01	5.7
GB03	1.9
GB04	3.2
GB05	3.8
GB06	8.5
GB09	4.0
GB12	5.4
GB17	4.9
GB29	2.1
GB35	2.0
GB36	6.3
GB39	3.2
GB42	2.2
H001	3.9
H002	2.9
H006	3.2
H012	5.6

H027	4.5
H032	7.2
H037	23.0
H038	2.5
H048	2.1
H054	4.1
H061	5.2
H096	3.5
H101	2.4
H102	3.9
H103	5.9
H104	2.7
H107	3.7
H108	4.2
H109	2.5
H110	3.4
H118	1.8
H119	2.8
H121	2.8
H123	3.0
H124	4.1
HTXD	2.5
HTXM	1.9
HTXS	1.5
HXSG	0.8
NC68	3.9
NC70	2.7
NC71	3.6
NC73	5.3
NC76	5.1
NC77	5.3
NC79	3.8
NC82	3.6
NC83	2.7
NC84	2.5
NC87	2.7
NC88	3.2
NC89	1.9
TB01	2.2
TB02	1.7
TB03	2.6
TB04	4.0

S5B. KSα domain OBUs

Location	*% sequences of most abundant* KSα *domain OBU*
GB01	10.0
GB03	4.8
GB04	27.9
GB05	10.0
GB06	15.6
GB09	4.5
GB12	15.4
GB17	7.7
GB29	11.1
GB35	20.0
GB36	7.4
GB39	15.4
GB42	13.0
H001	15.6
H002	17.6
H006	13.0
H012	30.3
H027	17.9
H032	12.9
H037	27.5
H038	15.4
H048	15.7
H054	16.7
H061	25.0
H096	29.2
H101	10.3
H102	17.1
H103	52.4
H104	20.0
H107	28.6
H108	20.0
H109	38.5
H110	54.3
H118	32.4
H119	26.0
H121	22.5
H123	5.9
H124	10.4

HTXD	12.8
HTXM	11.1
HTXS	15.0
HXSG	4.3
NC79	10.7
NC82	7.2
NC83	9.4
NC84	15.6
NC87	11.7
NC88	6.9
NC89	8.8
TB01	12.2
TB02	4.5
TB03	18.8
TB04	33.3

Table S6. List of molecular classes that KSα and A domain sequences aligned to in the MIBiG 2.0 database.[7] The entry under molecular class denotes the molecular product for which the biosynthetic gene cluster (BGC) was annotated in MIBiG. If a KSα or an A domain sequence aligned against a molecular product of a given BGC, that sequence was classified as a sequence producing a compound from the molecular class corresponding to that molecular product. 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.

Split correction factor estimation for estimating OBU counts.
To assign OBU sequence representatives to chemical compound classes, the former were aligned against sequences in the MIBiG database using the DIAMOND alignment tool.[8] MIBiG associates biosynthetic gene clusters (BGCs) with known natural product (NP) structures, allowing prediction of the product of each matching OBU and as a result, estimation of the chemical diversity at each sample site. Some distinct OBU sequence representatives were assigned to the same compound class (for example, five separate OBU sequence representatives aligned to rifamycin), which resulted in an overestimation of compound classes present in sediment. To correct for this, the number of times the same chemical compound class was represented by different OBU sequence representatives was computed for each chemical compound class. This number was then averaged for all observed chemical compounds classes and called the “split correction factor” (i.e. a residual error). To avoid overestimating chemical compound classes present in the sediment, the total number of observed OBUs was divided by that factor, resulting in less biased estimation of the chemical compound classes variance present in sediment.
Split correction factor for KSα: 1.517241379
Split correction factor for A: 1.894739749

S6A. List of identified A domain hits after rarefaction and stringent filtering.

Molecular class	Molecular class detected in x samples	# of sequences belonging to molecular class
Pyoverdin	39	89
Scabichelin	13	26
Salinichelin	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
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
Oxalomycin B	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

S6B. List of identified KSα domain hits after rarefaction and stringent filtering.

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

Table S7. 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:[9]

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

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.

S7A. Correlation coefficients between KSα and A domain OBUs.

	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 any single KSα or A domain OBU with a 16S OTU within the phyla Actinobacteria or Proteobacteria. These data suggest that of the NP biosynthetic pathways detected, very few co-occur in the environment. As discussed in the main article, the lack of correlation may be that either the detected OBUs are associated with mobile genetic elements and therefore are associated with multiple taxa, or that differential primer biases (OBU versus OTU) prevented sufficient detection of the necessary sequences needed to observe said correlation. Further experiments are required to confirm this.

S7B. Correlation coefficients between KSα domain OBUs.

To test for co-occurrence patterns, correlation coefficients were calculated for the twenty 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.[10] This yielded an identity of 68.93%. This suggests 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	-0.04	-0.03	-0.04	-0.03	-0.03	-0.05
KSα_2			0.13	-0.08	0.18	-0.07	-0.1	-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	-0.1	-0.05	-0.07	0.02	-0.04	0	-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	-0.1	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	-0.07	0.2
KSα_7								-0.04	0.07	-0.06	0.08	0.4	-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
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	-0.1	0.05	0.14	0.1
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	-0.09	0	-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																				

S7C. Correlation coefficients between all A domain OBUs.

Similarly, co-occurrence patterns were examined by calculating correlation coefficients for the twenty 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 identical, the sequence representative for these OBUs were aligned against each other using BLAST.[10] 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
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																				

S7D. Correlation coefficients between observed A domain OBUs and Shannon indices.

The correlation coefficient between observed A domain OBUs and Shannon indices was calculated using the formula in Table S7. The resulting coefficient was 0.69, indicating a small positive correlation between abundance and diversity in this dataset.

S7E. Correlation coefficients between observed A domain sequence abundance and OBU occurrence.

The correlation coefficient between observed A domain sequence abundance and OBU occurrence (the number of times a given OBU appeared at a given location) was calculated using the formula in Table S7. The resulting coefficient was 0.37, indicating no correlation between sequence abundance and OBU occurrence per site in this dataset.

Table S8. 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.[11] 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 S8A-B.

S8A. 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	5.3	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	4.1	23	3.46	12
H048	4.09	36	3.14	10
H054	**1.52**	**28**	**1.56**	**5**
H061	3.5	13	3.46	12
H096	3.79	30	3.46	12
H101	4.83	35	3.64	13
H102	4.1	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
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	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	6.5	97	3.91	15
TB03	5.8	77	3.64	13
TB04	4.18	21	3.77	14

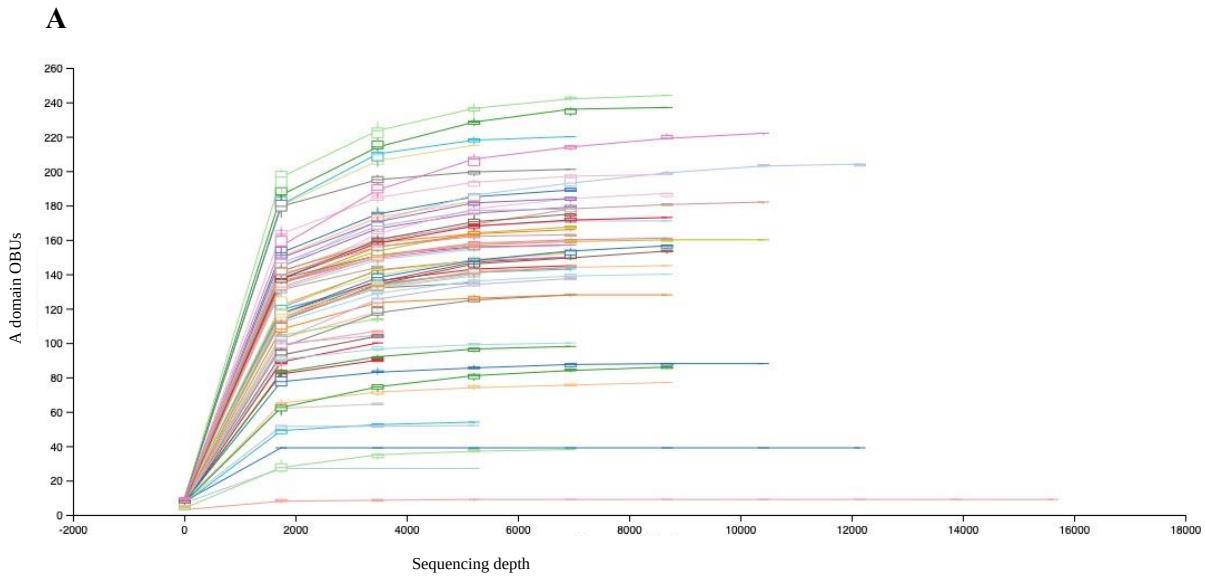
S8B. Shannon index for A domain OBUs before and after rarefaction.

Before rarefaction			After rarefaction	
Sample	**Shannon**	**OBU count**	**Shannon**	**OBU count**
GB01	8.95	1326	9.76	5956
GB03	9.47	1478	10.4	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	9.3	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	9.3	1576	10.12	5936
H002	9.22	1431	9.98	5677
H006	9.25	1465	10.1	6030
H012	9.06	1382	9.8	5580
H027	9.36	1576	10.13	6407
H032	8.96	1399	9.88	6165
H037	6.13	919	6.55	2508
H038	9.38	1494	10.29	6472
H048	9.21	1404	9.91	5052
H054	8.89	1175	9.5	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	9.6	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
NC68	9.76	1728	10.77	9532
NC70	9.65	1721	10.82	9611
NC71	9.8	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	10.8	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	9.7	4152
TB04	8.97	1156	9.49	4105

Figure S1. Rarefaction curves to estimate A and KSα domain OBU diversity

S1A. Rarefaction curve for A domain OBUs



S1B. Rarefaction curve for KSα domain OBUs

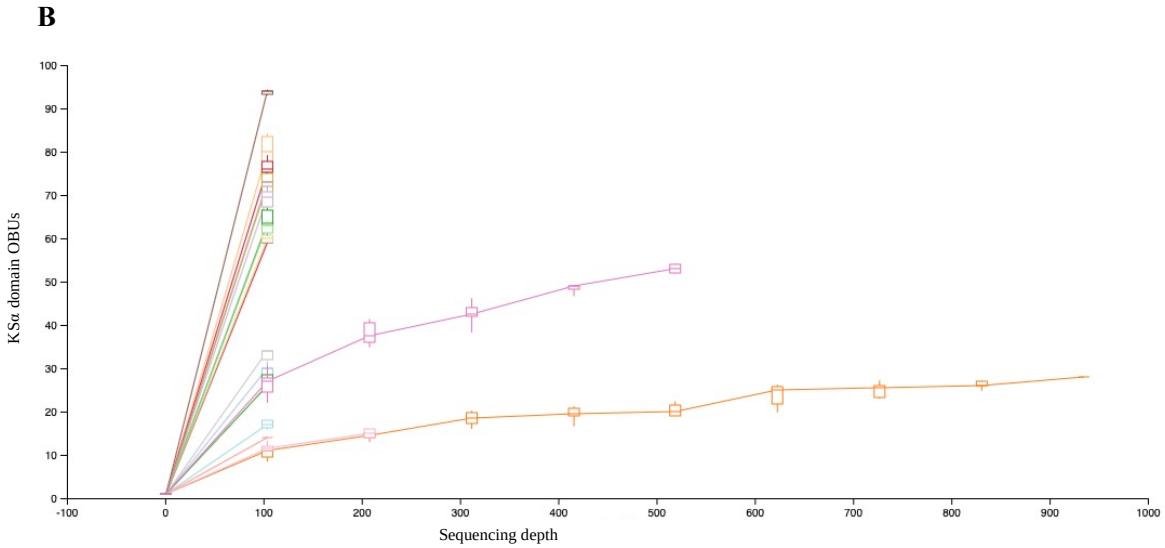


Figure S2. A and KSα domain OBU and sequence abundances.

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. Sequences were rarified at 13 sequences per sample for KSα domain sequences and at a rarefaction of 3,487 sequences per sample for A domain sequences to map known sequences.

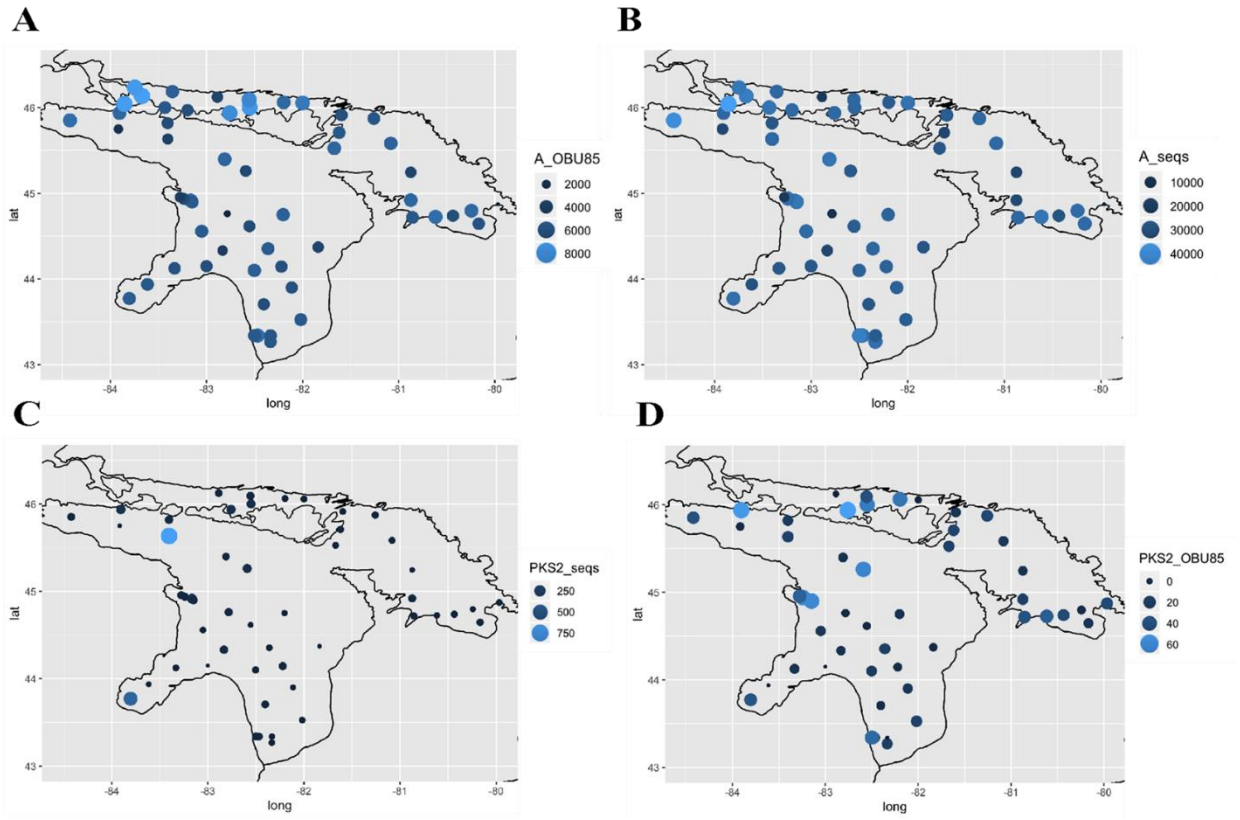
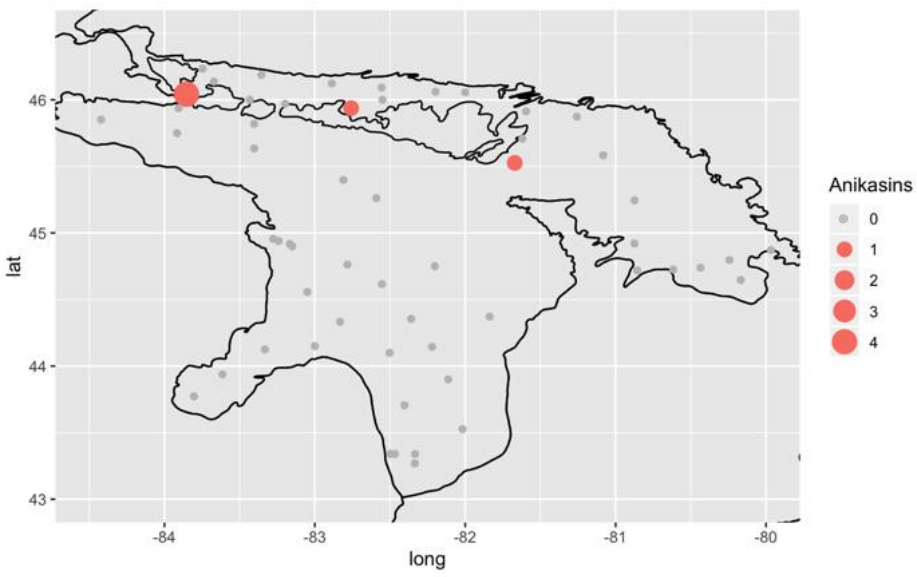
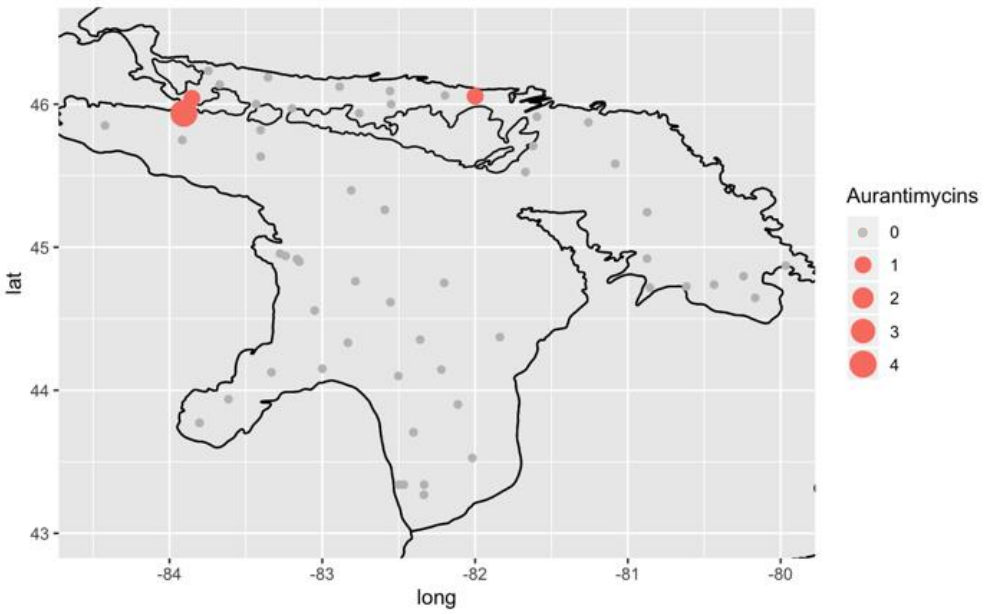


Figure S3. Occurrence of known antibiotics in Lake Huron sediment. Sequences were rarified at 13 sequences per sample for KSα domain sequences and at a rarefaction of 3,487 sequences per sample for A domain sequences before mapping.

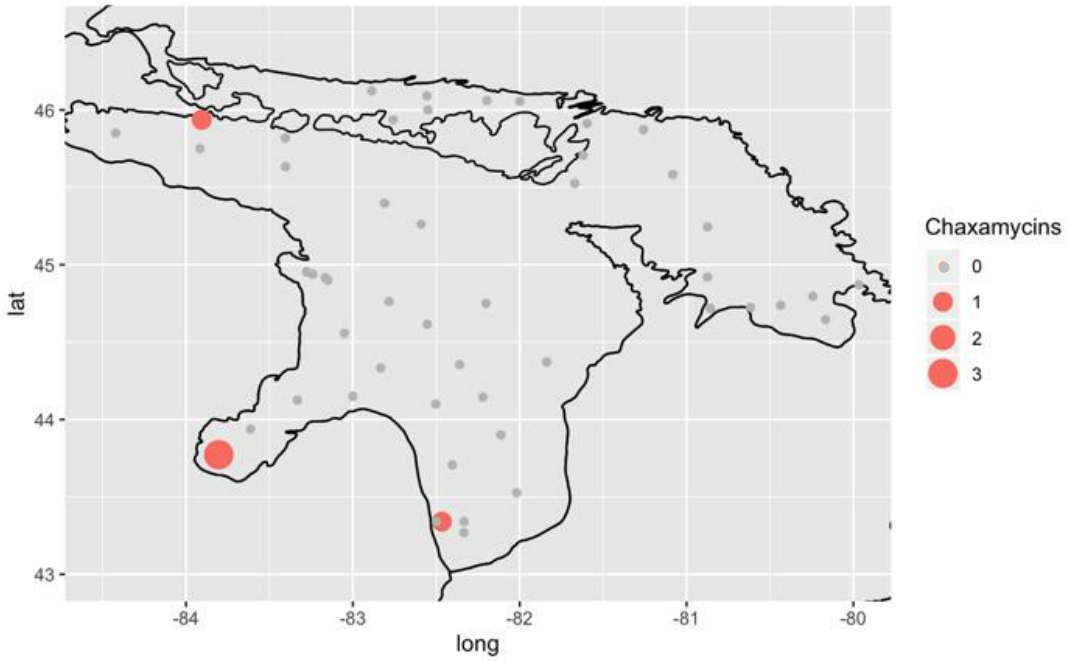
S3A. Occurrence of anikasin-like molecules in Lake Huron sediment.



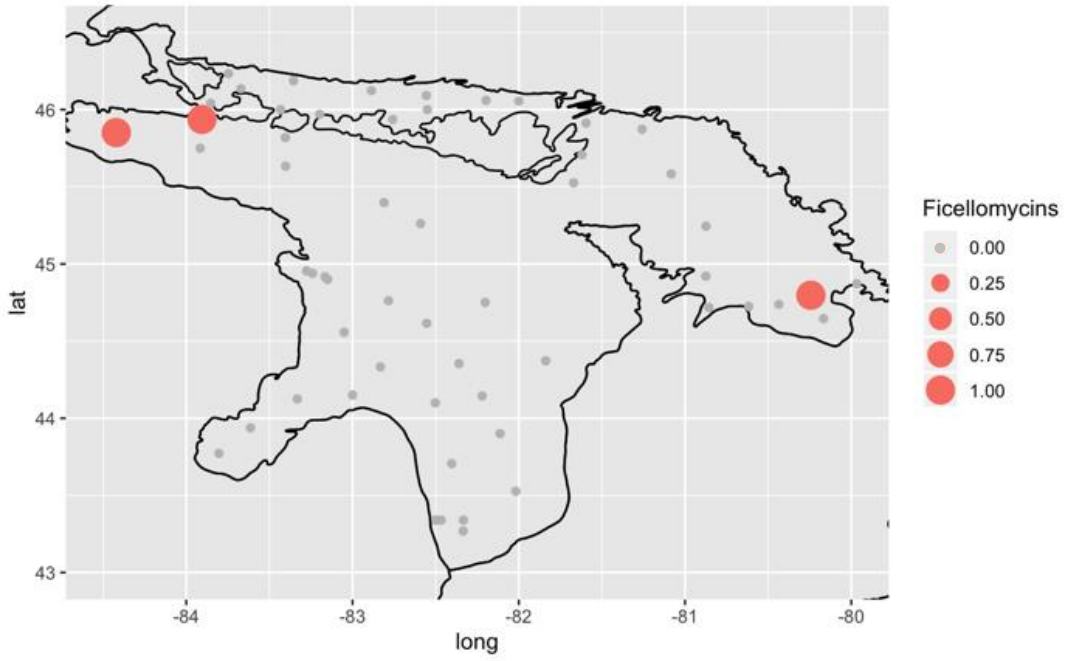
S3B. Occurrence of aurantimycin-like molecules in Lake Huron sediment.



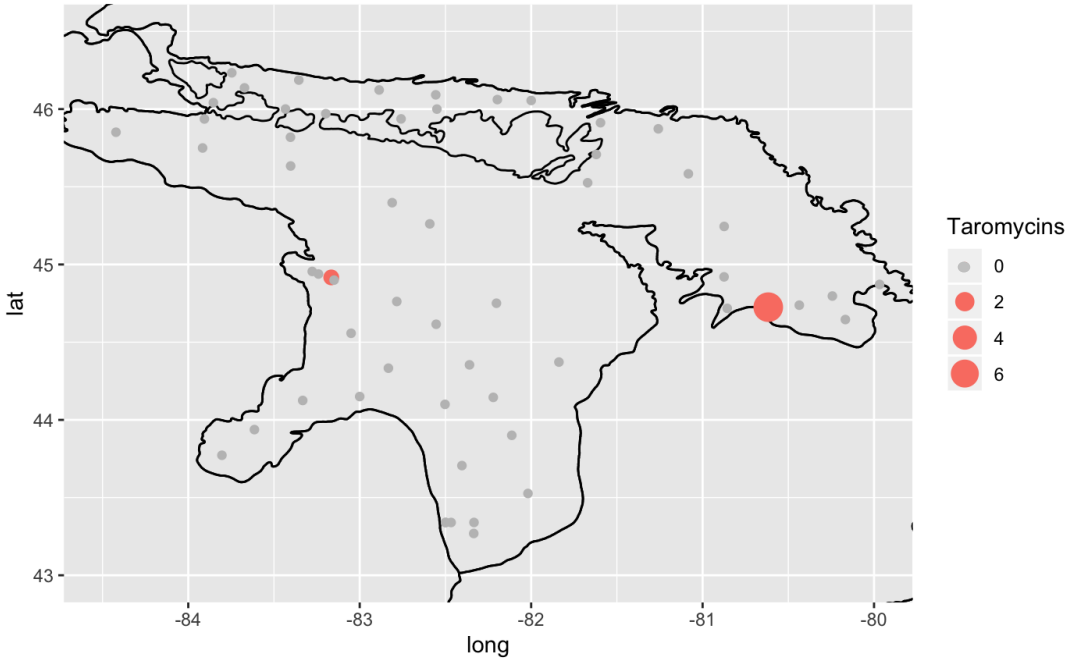
S3C. Occurrence of chaxamycin-like molecules in Lake Huron sediment.



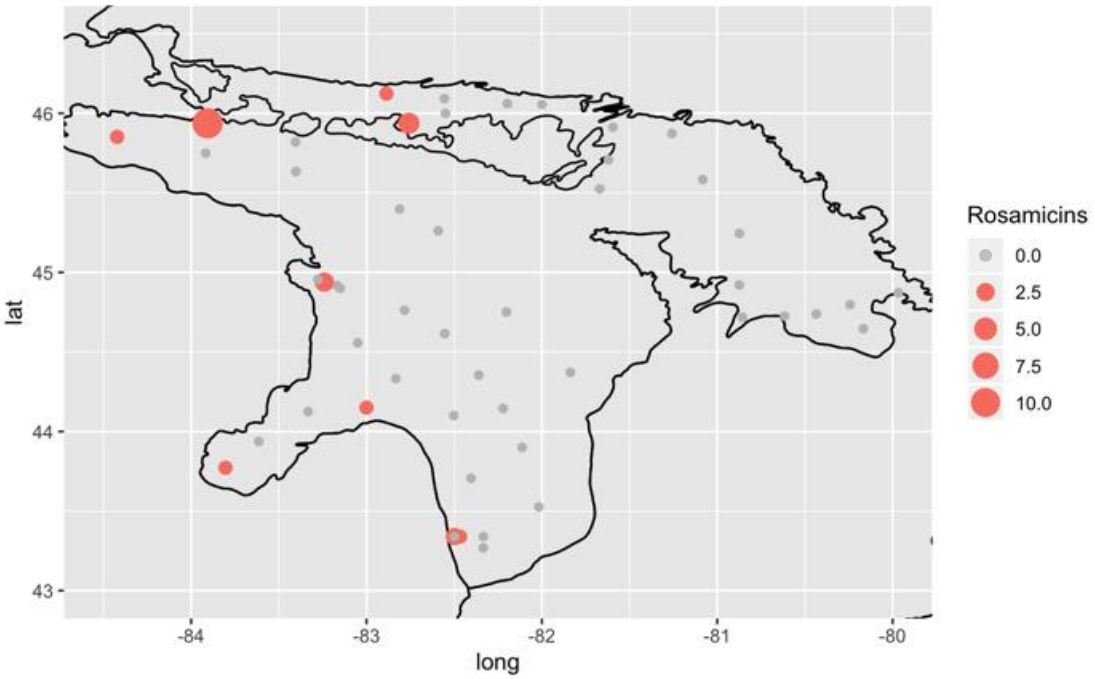
S3D. Occurrence of ficellomycin-like molecules in Lake Huron sediment.

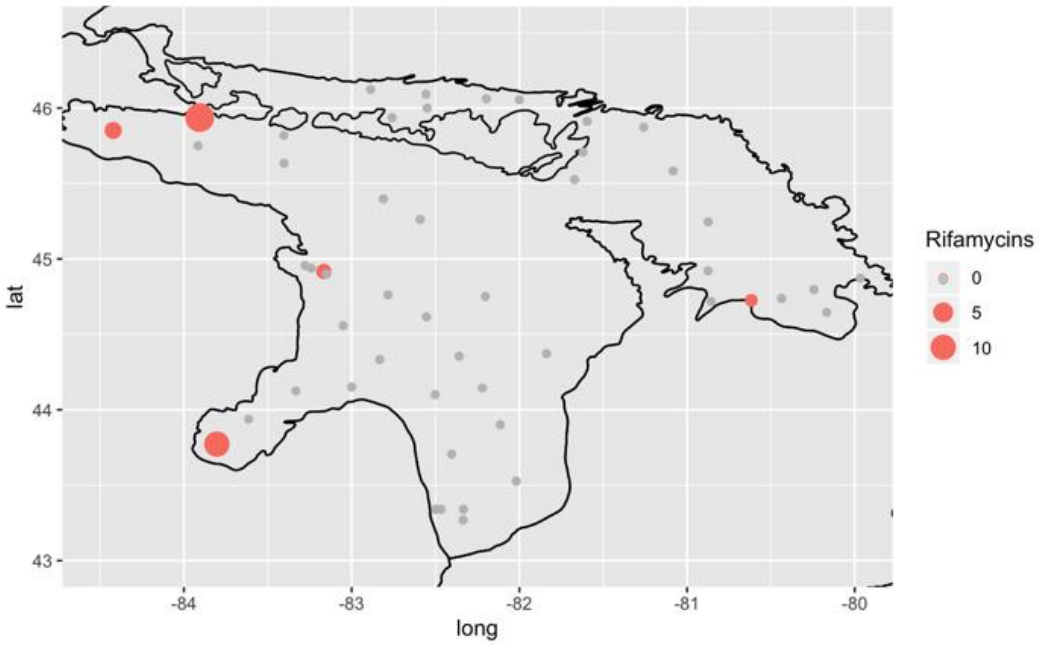


S3E. Occurrence of taromycin-like molecules in Lake Huron sediment.



S3F. Occurrence of rosamycin-like molecules in Lake Huron sediment.

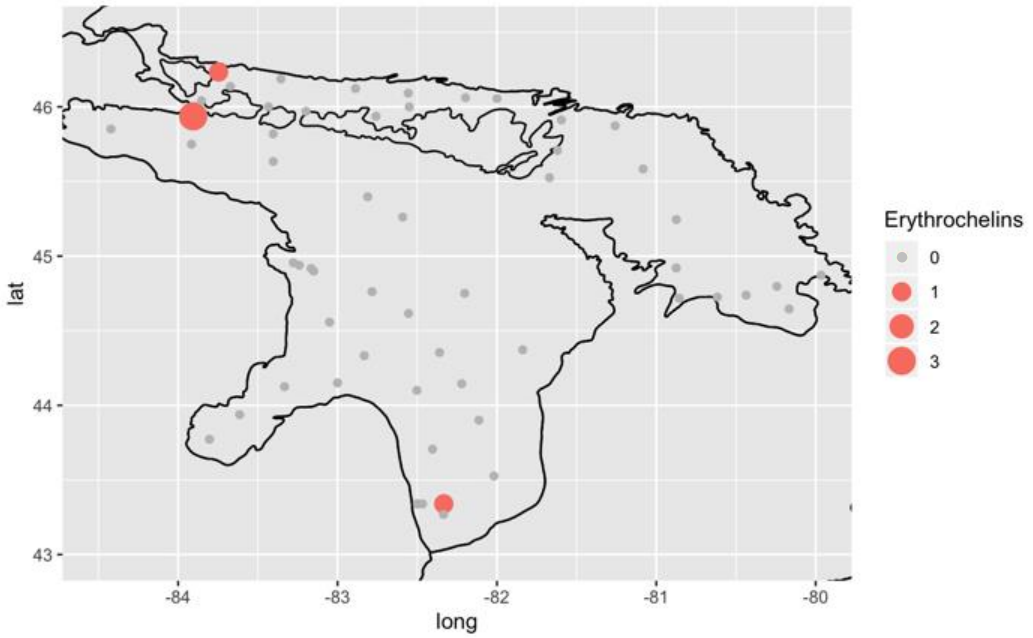




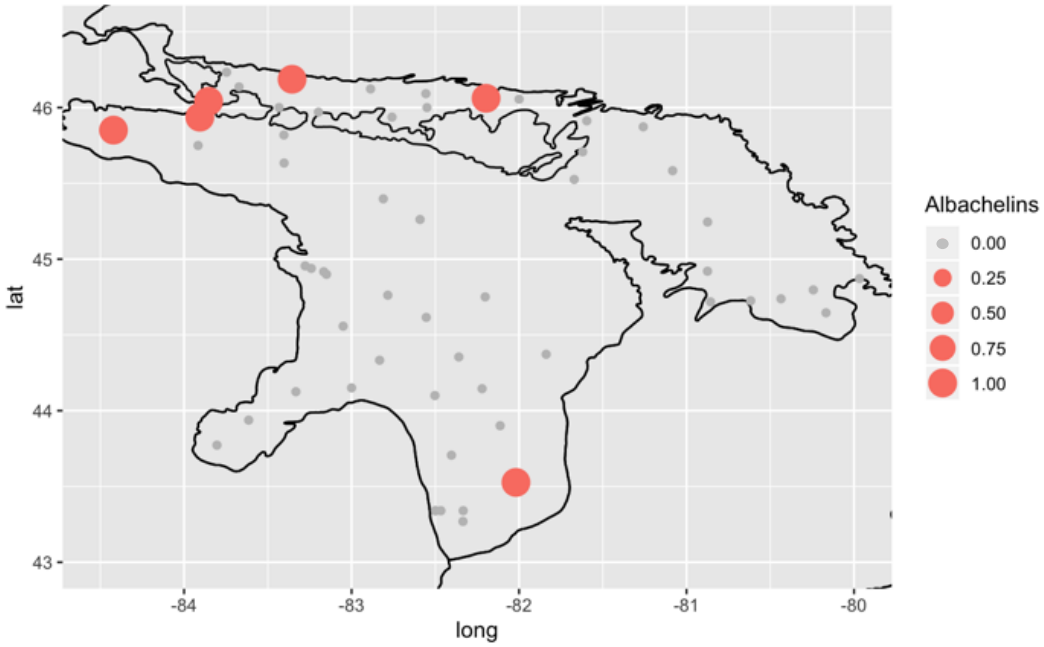
S3G. Occurrence of rifamycin-like molecules in Lake Huron sediment.

Figure S4. Occurrence of known siderophores in Lake Huron sediment.

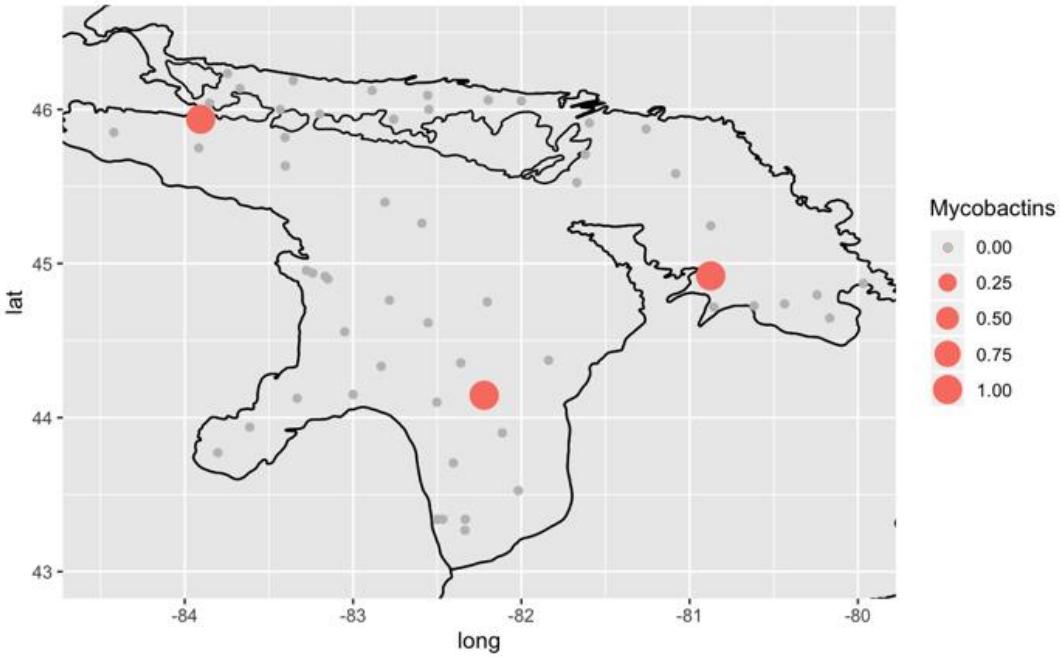
S4A. Occurrence of erythrochelin-like molecules in Lake Huron sediment.



S4B. Occurrence of albachelin-like molecules in Lake Huron sediment.



S4C. Occurrence of mycobactin-like molecules in Lake Huron sediment.



S4D. Occurrence of coelibactin-like molecules in Lake Huron sediment.

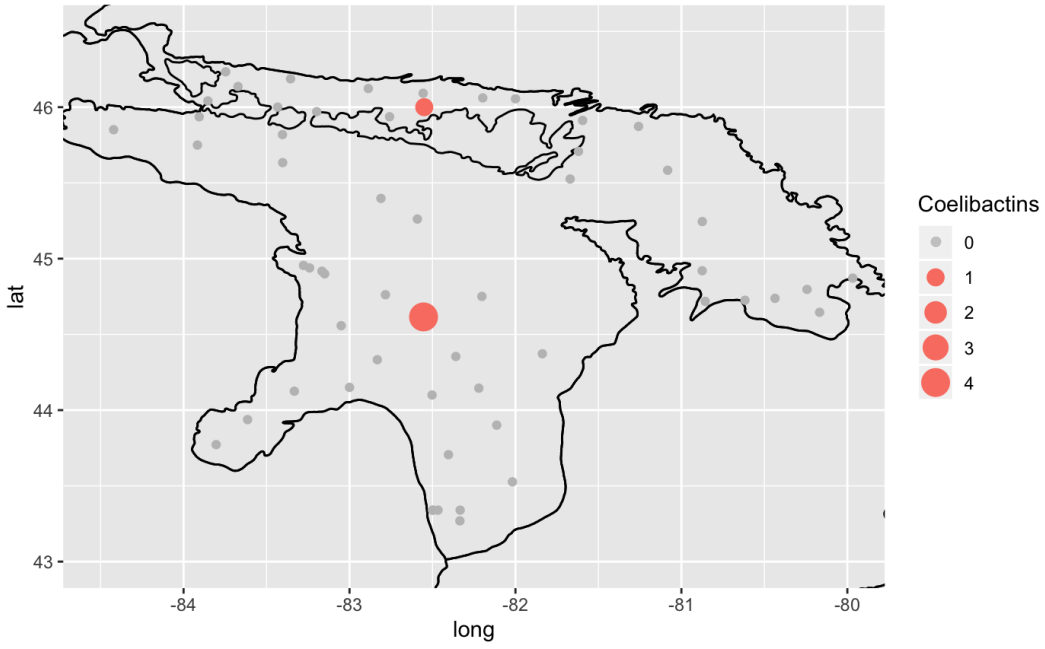
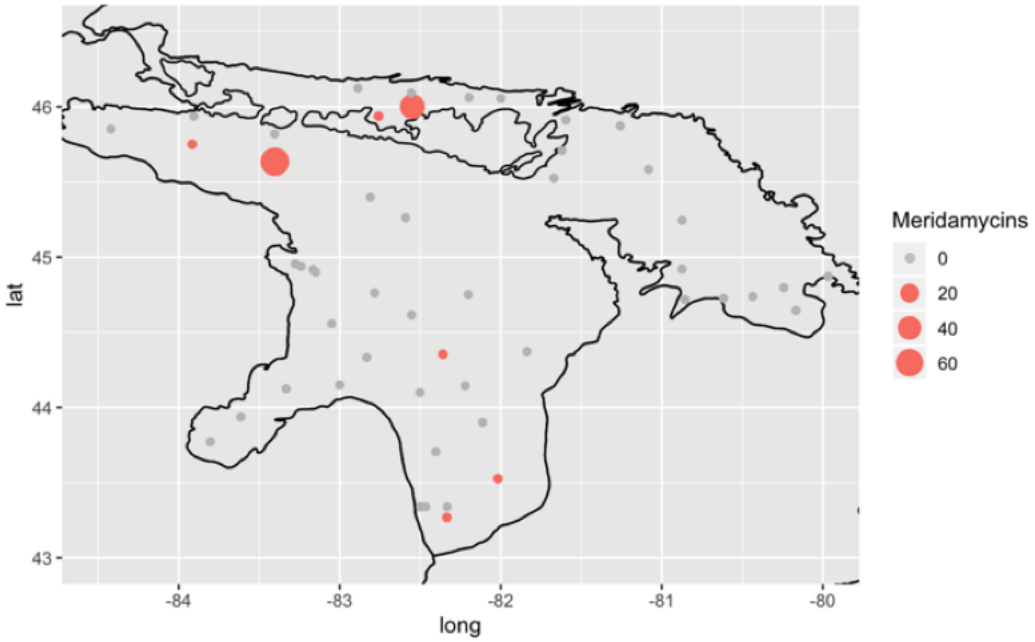
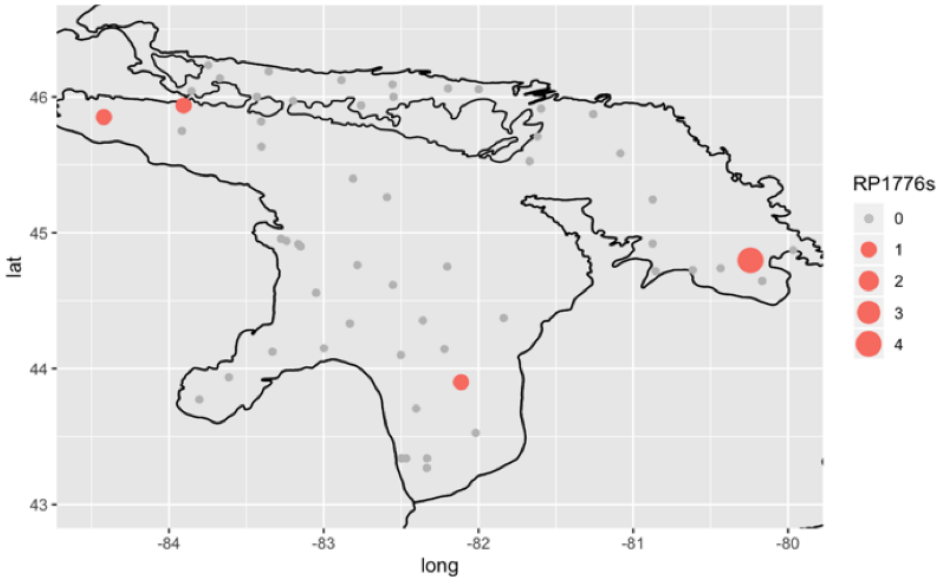


Figure S5. Occurrence of all other known, detected bioactive natural products in Lake Huron sediment. Sequences were rarified at 13 sequences per sample for KSα domain sequences and at a rarefaction of 3,487 sequences per sample for A domain sequences before mapping.

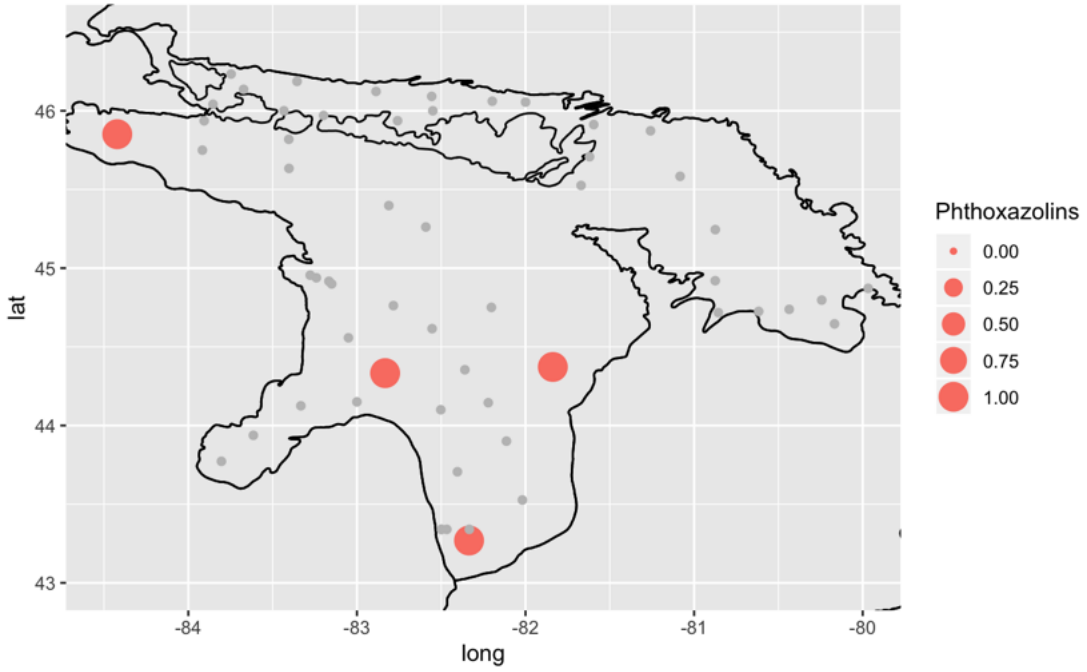
S5A. Occurrence of meridamycin-like molecules in Lake Huron sediment.



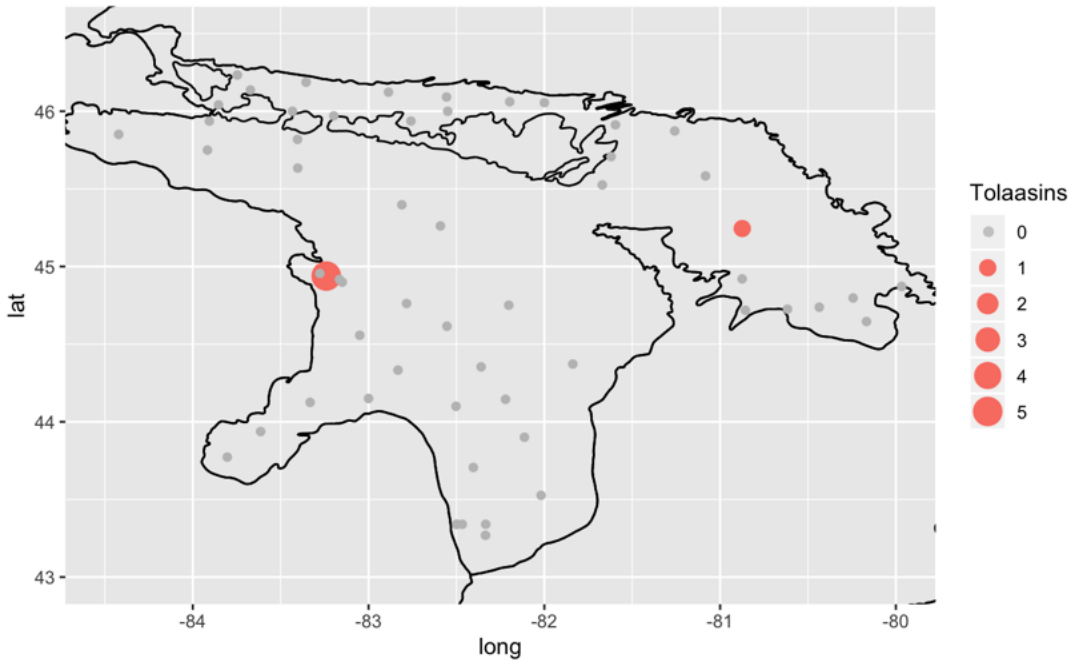
S5B. Occurrence of RP1776-like compounds in Lake Huron sediment.



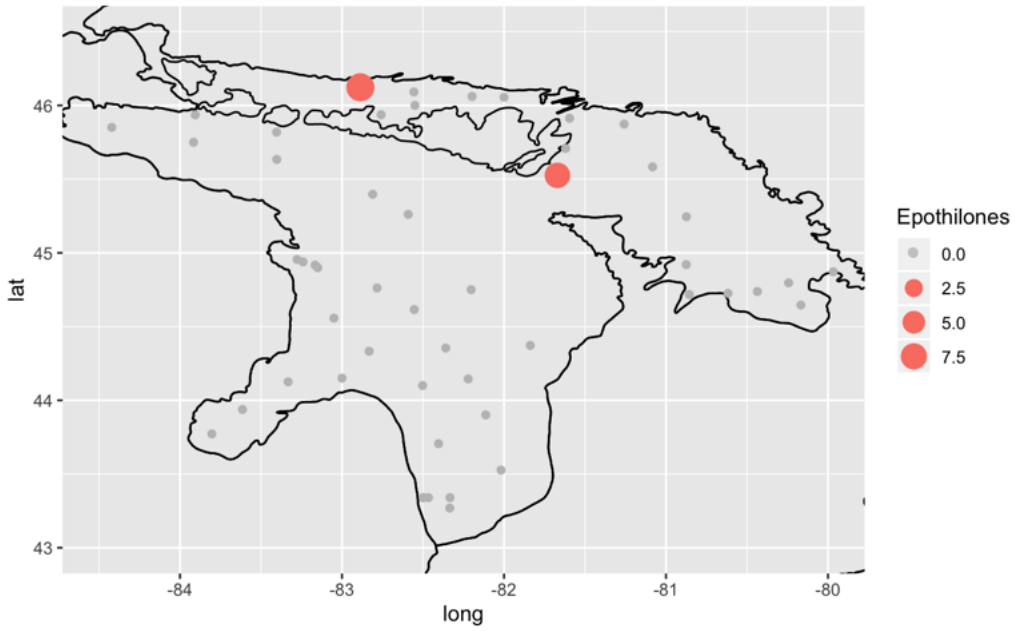
S5C. Occurrence of phthoxazolin-like molecules in Lake Huron sediment.



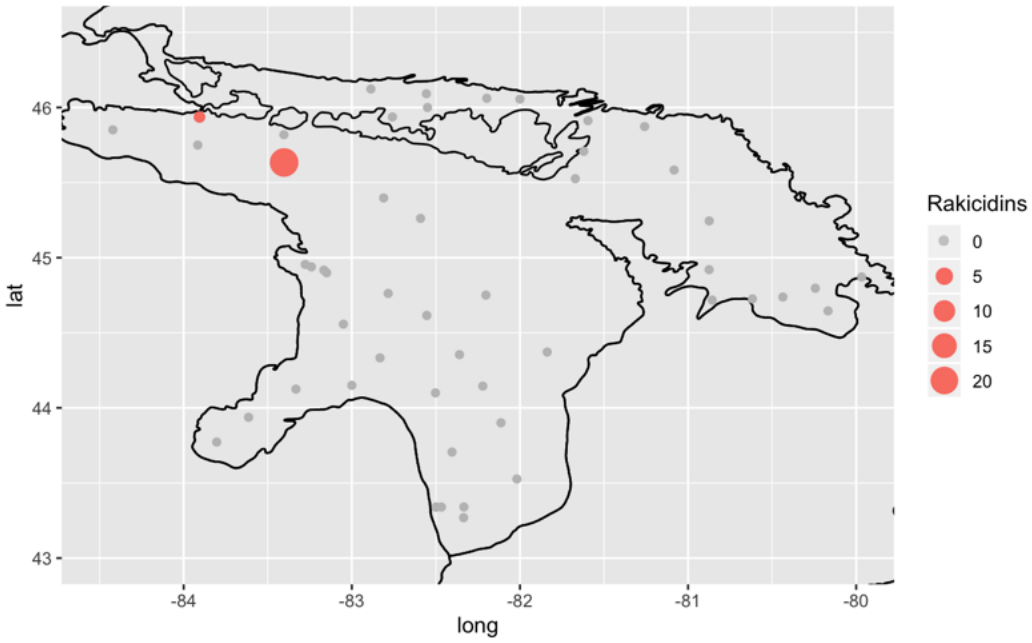
S5D. Occurrence of tolaasin-like molecules in Lake Huron sediment.



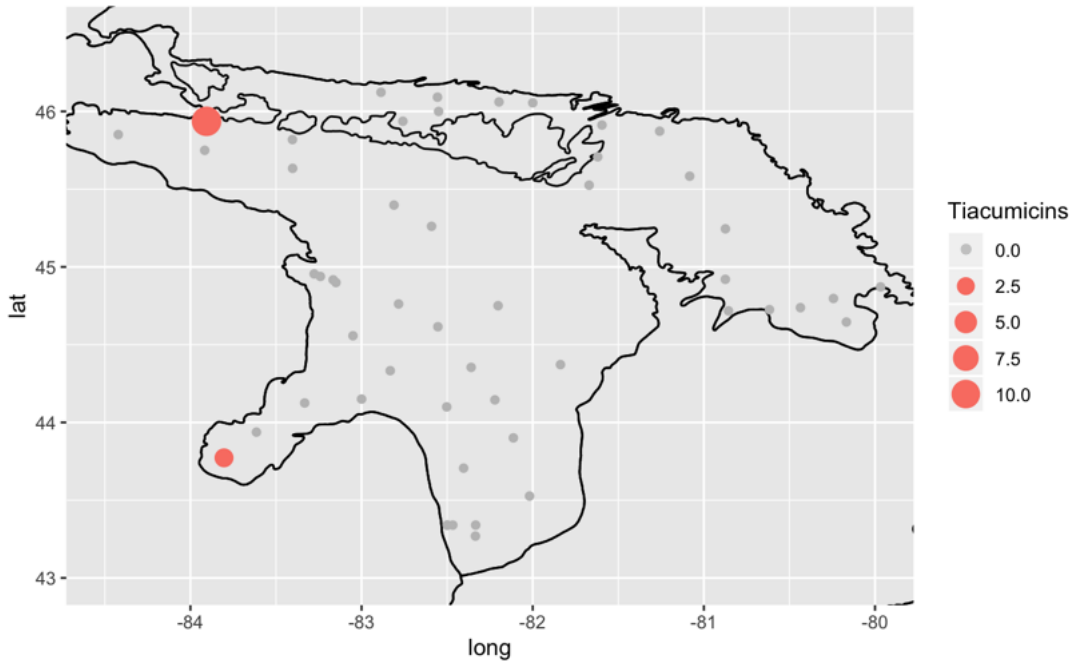
S5E. Occurrence of epothilone-like molecules in Lake Huron sediment.


46
45
44
43
lat
-84
-83
-82
-81
-80
long
Epothilones
0.0
2.5
5.0
7.5

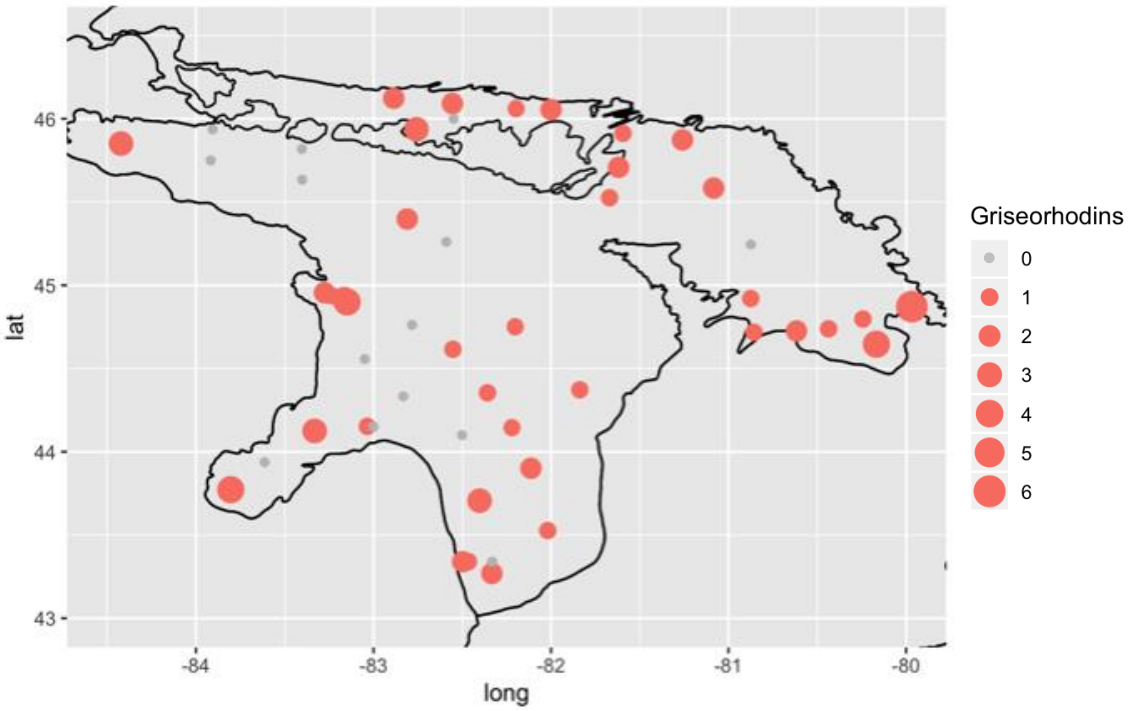
S5F. Occurrence of rakicidin-like molecules in Lake Huron sediment.



S5G. Occurrence of tiacumicin-like molecules in Lake Huron sediment.



S5H. Occurrence of griseorhodin-like molecules in Lake Huron sediment**.**



Supplementary Discussion.

Of the known identified OBUs, those 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 that bacteria more commonly use siderophores in the environment. Siderophores are essential for a microbe's survival: they chelate essential metals and thereby make them available for use in processes such as oxygen metabolism, and DNA and RNA syntheses.[12,13] Strains that are relevant for the field of NP drug discovery are present in undetectable amounts in sediment.[14] This might be the reason a large proportion of the OBUs (98.3% KSα domain OBUs and 99.9% of A domain OBUS, respectively) failed to match any of the compounds available in the MIBiG database. It is also worth noting there was no observed correlation between OBU presence/abundance and OTU presence/abundance (Supp. Table S7).

Expanded experimental limitations discussion.

The need for novel approaches to improve detection of NP BGCs from eDNA.

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 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 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.[7] The number of existing NPs greatly outnumbers the number of entries in MIBiG, underlining the need for the community to contribute to this and similar existing databases to identify NPs.

References:

1. Naqib, A. *et al.* 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.)* vol. 1783 149–169 (2018).

2. Caporaso, J. G. *et al.* Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proceedings of the National Academy of Sciences of the United States of America* **108 Suppl 1**, 4516–22 (2011).

3. Caporaso, J. G. *et al.* QIIME allows analysis of high-throughput community sequencing data. *Nature methods* **7**, 335–6 (2010).

4. Callahan, B. J. *et al.* DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods* **13**, 581–583 (2016).

5. Quast, C. *et al.* The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. *Nucleic Acids Research* **41**, 590–596 (2013).

6. McDonald, D. *et al.* The Biological Observation Matrix (BIOM) format or: How I learned to stop worrying and love the ome-ome. *GigaScience* **464**, 1–6 (2012).

7. Kautsar, S. A. *et al.* MIBiG 2.0: a repository for biosynthetic gene clusters of known function. *Nucleic Acids Research* (2019) doi:10.1093/nar/gkz882.

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

9. 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. *Biotechnology Letters* **26**, 509–515 (2004).

10. Camacho, C. *et al.* BLAST+: Architecture and applications. *BMC Bioinformatics* **10**, 1–9 (2009).

11. Caporaso, J. G. *et al.* QIIME allows analysis of high- throughput community sequencing data intensity normalization improves color calling in SOLiD sequencing. *Nature Publishing Group* **7**, 335–336 (2010).

12. Ahmed, E. & Holmström, S. J. M. Siderophores in environmental research: roles and applications. *Microbial Biotechnology* **7**, 196–208 (2014).

13. Cox, C. D. & Adams, P. Siderophore activity of pyoverdin for Pseudomonas aeruginosa. *Infection and Immunity* **48**, 130–138 (1985).

14. 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* **13**, 2074–2081 (2018).

Annexure C: List of publications and manuscripts

Review Articles:

1. Chevrette, Marc G., Athina Gavrilidou, **Shrikant Mantri**, Nelly Selem-Mojica, Nadine Ziemert, and Francisco Barona-Gómez. "The confluence of big data and evolutionary genome mining for the discovery of natural products." Natural Product Reports (2021).

   Referred in text as Publication 1

Research Articles:

1. Männle, Daniel, Shaun MK McKinnie, **Shrikant Mantri**, Katharina Steinke, Zeyin Lu, Bradley S. Moore, Nadine Ziemert, and Leonard Kaysser. "Comparative genomics and metabolomics in the genus Nocardia." Msystems 5, no. 3 (2020): e00125-20.

   Referred in text as Publication 2

2. Maryam Elfeki, **Shrikant Mantri**, Chase M. Clark, Stefan J. Green, Nadine Ziemert, Brian T. Murphy, (2021), Evaluating the Distribution of Bacterial Natural Product Biosynthetic Genes Across Lake Huron Sediment. ACS ChemBio. (Accepted)

   Referred in text as Publication 3

Manuscripts:

1. **Shrikant Mantri**, Nadine Ziemert (2021): MBEZ: Easy biosynthetic potential exploration metagenome mining pipeline.

   Referred in text as Manuscript 1

2. **Shrikant Mantri**, Timo Negri, Helena Sales-Ortells, Angel Angelov, Silke Peter, Harald Neidhardt, Yvonne Oelmann, Nadine Ziemert, (2021), Metagenomic sequencing of multiple soil horizons and sites in close vicinity revealed novel secondary metabolite diversity.

   Referred in text as Manuscript 2

Annexure D: Declaration of Contributions

Chapters 1, 2 and 3 are original works of Shrikant Mantri. Chapter 1 covers the introduction of the dissertation, describes the research problem and project objectives. The biological concepts and topics crucial for understanding the basics involved in the area of natural products genome mining and metagenome mining have been described in Chapter 2. Briefly, this covers topics about microbial diversity, natural products chemical space, biosynthesis pathways of natural products, next generation sequencing technologies and metagenome mining. Chapter 3 covers technical background. The algorithms and databases involved in genome and metagenome mining have been surveyed and reviewed in this chapter. Microbial community diversity profiling methods, natural products domain exploration methods, natural products biosynthetic cluster exploration methods, metagenome assembly and biosynthesis potential exploration method, and easy to use tools and techniques have been covered.

Chapter 4

Manuscript 1: MBEZ: Easy biosynthetic potential exploration metagenome mining pipeline.

Shrikant Mantri
: Study Conception; Implemented the pipeline and scripts; Analysed and interpreted the data; Manuscript writing; Manuscript editing.

Nadine Ziemert
: Study Conception; Manuscript writing; Manuscript editing; Funding acquisition.

Chapter 5

Manuscript 2: Metagenomic sequencing of multiple soil horizons and sites in close vicinity revealed novel secondary metabolite diversity.

Shrikant Mantri
: Study Conception; Soil Sampling; Bioinformatics Analysis; Manuscript writing; Manuscript editing.

Timo Negri
: Study Conception; Soil Sampling; DNA isolation, library preparation and sequencing; Manuscript writing; Manuscript editing.

Helena Sales-Ortells
: Soil Sampling.

Angel Angelov
: DNA isolation, library preparation and sequencing; Manuscript editing.

Silke Peter
: DNA isolation, library preparation and sequencing; Manuscript editing.

Harald Neidhardt
: Soil Analysis; Manuscript editing.

Yvonne Oelmann

Soil Analysis; Manuscript editing.

Nadine Ziemert
Study Conception; Manuscript writing; Manuscript editing; Funding acquisition.

Chapter 6.1

Publication 3: Evaluating the Distribution of Bacterial Natural Product Biosynthetic Genes Across Lake Huron Sediment

Maryam Elfeki
Study conception and scientific ideas; Data generation; Analysis and interpretation of data; Manuscript writing; Manuscript editing.

Shrikant Mantri
Study conception and scientific ideas; Data generation; Analysis and interpretation of data; Manuscript writing; Manuscript editing.

Chase Clark
Data generation.

Stefan Green
Data generation.

Nadine Ziemert
Study conception and scientific ideas; Manuscript writing; Manuscript editing.

Brian Murphy
Study conception and scientific ideas; Manuscript writing; Manuscript editing; Funding acquisition.

Chapters 6.2. and 6.3 are original works of Shrikant Mantri.

Chapter 6.2 and 6.3 covers the biosynthesis potential survey of diverse ecosystems, specifically gut microbiomes, and Tuebingen strain collection.

Chapter 7 is the original works of Shrikant Mantri.

In the final Chapter 7 overall conclusion, expected future impact of the developed methods and approaches, and future challenges in the field are discussed.

Annexure E: List of Abbreviations

16S rRNA	16S ribosomal ribonucleic acid (rRNA), where S (Svedberg) is a unit of measurement of sedimentation rate
AMR	antimicrobial resistance
antiSMASH	antibiotics & Secondary Metabolite AnalysisShell
ARTS	Antibiotic Resistant Target Seeker
AT	acyl transferase
ATP	Adenosine Triphosphate
BGC	Biosynthetic Gene Cluster
BinAC	Bioinformatics and Astrophysics Cluster
BLAST	Basic Local Alignment Search Tool
bp	basepair
CMFI	Controlling Microorganisms to Fight Infections
DEBS	6-Deoxyerythronolide B Synthase
deNBI	German Network for Bioinformatics Infrastructure
DFG	German Research Foundation (Deutsche Forschungsgemeinschaft)
DNA	Deoxyribonucleic acid
DZIF	Deutsche Zentrum für Infektionsforschung
GCF	Gene Cluster Family
HMM	Hidden Markov Model
IBMI	Interfaculty Institute for Biomedical Informatics
IMIT	Interfaculty Institute for Microbiology and Infection Medicine Tübingen
kb	kilobase
KR	ketoreductase
KS	ketosynthase
MAG	Metagenome Assembled Genome
MBEZ	Metagenome biosynthesis potential exploration easy tool
MDR	Multi-Drug Resistant
MiBIG	Minimum Information about a Biosynthetic Gene cluster
NABI	National Agri Food Biotechnology Institute
NaPDoS	Natural Product Domain Seeker
NCBI	National Center for Biotechnology Information
NCCT	NGS Competence Center Tübingen
NGS	Next Generation Sequencing
NIH	National Institute of Health
NP	Natural Product
NRPS	Non-Ribosomal Peptide Synthetase
PacBio	Pacific Biosystems
PCP	Peptide Carrier Protein
PCR	Polymerase Chain Reaction
Pfam	Protein family
pHMM	profile Hidden Markov Model
PKS	Polyketide Synthases
RiPP	ribosomally synthesized and post-translationally modified peptides
RNA	Ribonucleic acid
SM	Secondary Metabolite
TE	thioesterase
TELL-seq	Transposase Enzyme Linked Long-read Sequencing