Informatic strategies for the discovery and characterization of peptidic natural products

Informatic strategies for the discovery and characterization of peptidic natural products

By Nishanth Merwin, B.H.Sc.

A Thesis Submitted to the School of Graduate Studies in the Partial Fulfillment of the Requirements for the Degree MSc.

McMaster University May 17, 2018

McMaster University

MSc. (2018)

Hamilton, Ontario (Biochemistry and Biomedical Sciences)

TITLE: Informatic strategies for the discovery and characterization of peptidic natural products

AUTHOR: Nishanth Merwin B.H.Sc (McMaster University)

SUPERVISOR: Dr. Nathan Magarvey

NUMBER OF PAGES: x, 116

Lay Abstract

Biochemistry is the study in which life is built upon a series of diverse chemistry and their interactions. Some of these chemicals are not essential for the maintaining basic metabolism, but are instead tailored for alternative functions best suited to their environment. Often, these molecules mediate biological warfare, allowing organisms to compete and establish dominance amongst their neighbours. Understanding this, several of these molecules have been exploited in our modern pharmaceutical regimen as effective antibiotics. Due to the ever rising reality of antibiotic resistance, we are in dire need of novel antibiotics. With this goal, I have developed several software tools that can both identify these molecules encoded within bacterial genomes, but also predict their effects on neighbouring bacteria. Through these computational tools, I provide an updated strategy for the discovery and characterization of these biologically derived chemicals.

Abstract

Microbial natural products have served a key role in the development of clinically relevant drugs. Despite significant interest, traditional strategies in their characterization have lead to diminishing returns, leaving this field stagnant. Recently developed technologies such as low-cost, high-throughput genome sequencing and high-resolution mass spectrometry allow for a much richer experimental strategy, allowing us to gather data at an unprecedented scale. Naive efforts in analyzing genomic data have already revealed the wealth of natural products encoded within diverse bacterial phylogenies. Herein, I leverage these technologies through the development of specialized computational platforms cognizant of existing natural products and their biosynthesis in order to reinvigorate our drug discovery protocols.

As a first, I present a strategy for the targeted isolation of novel and structurally divergent ribosomally synthesized and post-translationally modified peptides (RiPPs). Specifically, this software platform is able to directly compare genomically encoded RiPPs to previously characterized chemical scaffolds, allowing for the identification of bacterial strains producing these specialized, and previously unstudied metabolites. Further, using metabolomics data, I have developed a strategy that facilitates direct identification and targeted isolation of these uncharacterized RiPPs. Through these set of tools, we were able to successfully isolate a structurally unique lasso peptide from a previously unexplored *Streptomyces* isolate.

With the technological rise of genomic sequencing, it is now possible to survey polymicrobial environments with remarkable detail. Through the use of metagenomics, we can survey the presence and abundances of bacteria, and further metatranscriptomics is able to reveal the expression of their biosynthetic pathways. Here, I developed a platform which is able to identify microbial peptides exclusively found within the human microbiome, and further characterize their putative antimicrobial properties. Through this endeavour, we identified a bacterially encoded peptide that can effectively protect against pathogenic *Clostridium difficile* infections.

With the wealth of publicly available multi-omics datasets, these works in conjunction demonstrate the potential of informatics strategies in the advancement of natural product discovery.

Acknowledgements

First, and foremost, I would like to thank my supervisor Nathan Magarvey. For making science exciting, and challenging me to learn more than I could have ever imagined, I truly appreciate your guidance. I must also thank my committee members Paul McNicholas and Michael Surette, for the insightful meetings and rich discussions.

To my amazing colleagues, the interdisciplinary work we do would not be possible without the level of collaboration we were able to achieve. To Walaa, I hope to continue our scientific journey. To Keshav, Chris and Mathusan, who understand the struggles of writing software, I hope we continue to bring the database down with our shenanigans and always share the lessons we've learned.

To my family, your unconditional love from across the world has made all of this possible. To Natasha, you were my foundation throughout my graduate studies, and I'm so lucky to have your love and support.

Contents

List of Figures

Acronyms

A adenylation

AMP antimicrobial peptide

BGC biosynthetic gene cluster

C condensation

iHMP integrated human microbiome project

LCMS liquid chromatography coupled mass spectrometry

MIC minimum inhibitory concentration

MS^2 tandem mass spectrometry

NMR nuclear magnetic resonance

NP natural product

NRP nonribosomal peptide

NRPS nonribosomal peptide synthase

PTM post translational modification

RiPP ribosomally synthesized and post-translationally modified peptide

T thiolation

TE thioesterase

Declaration of Authorship

I, Nishanth Merwin, declare that this thesis titled, "Informatic strategies for the discovery and characterization of peptidic natural products" and the work presented in it are my own. The work presented here is a compilation of multiple prepared journal articles, and author contributions are prepended to each body of work.

Chapter 1

Introduction

1.1 Thesis context

Biochemistry is, in many ways, the insight into how life is built upon a series of non-living molecular components. Apart from the core set of molecules required to facilitate living, many organisms have developed a set of so-called secondary metabolites. Since these molecules are non-essential for basic metabolism, they are often tailored for very specific adaptations to niche environments. In many cases, these molecules mediate biological competition through inducing deleterious effects on neighbouring organisms [6]. It was this property that intrigued many scientists in the early to mid 1900s to investigate these natural products (NPs).

Biological extracts have been used medically predating modern history. From the widespread usage of the opium extract [4] for pain relief, to the *Cinchoa* bark extract used to treat malaria [1], it was evident that these NPs played a significant role in historical medicine. It was not until the isolation of morphine in 1803 [4] and the first synthetic production of a natural product in 1828 [29] that we began to understand the underlying chemical components responsible for manifesting these important therapeutic effects. Koch's postulates published 1884 demonstrated a clear link between disease and microbiological organisms [12]. With the combination of these understandings, we began to hunt aggressively for molecular components that can selectively antagonize biological agents responsible for disease. In this pursuit, it was Paul Erlich in the early 1900s that first attempted to systematically test molecules that could selectively treat syphilis growth without harming human tissue. Through this work, one of the first antibiotics, salvarasan, was discovered [28]. However, as an arsenic containing compound, salvarasan was both highly unstable, and often caused severe adverse effects inspiring the search for further antibiotics.

Through the early 1900s, the works by Selman Waksman and Alexander Fleming showed proof that bacteria and fungi can also be rich sources of antibacterial agents with the discovery of streptomycin and penicillin, respectively [25, 9]. These highly successful examples spurred both academic and industrial research into further microbial sources of antibiotics. Using bioactivity assays, microbes with demonstrable antibacterial activity were collected and investigated at scale. Through successive fractionations of biological extracts, scientists were able to trace the antibiotic phenotype to specific purified compounds. This strategy was highly successful, and led to a wealth of antibiotics developed during the 1950s-1970s [15], which has since been termed the "Golden Age" of antibiotic discovery. As nuclear magnetic resonance (NMR) technologies advanced during this era, scientists were further able to characterize these discovered NPs. Further, isotope labeling studies were able to demonstrate that these antibiotic agents were often built up from monomers already present in primary metabolites. Looking back, its clear that many of the antibiotic agents can be grouped into families, according to their monomer units. These NPs were often composed of sugars (aminoglycosides), isoprene units (terpenes), amino acids (peptides), acetate units (polyketides and fatty acids) and nucleic acids (nucleosides).

Despite the remarkable success of NP discovery in the mid-1900's, scientists and industrial leaders noticed that these large scale bioactivity guided searches were waning [15]. Towards the late 1900s, screening programs for synthetic molecules began using strategies such as combinatorial chemistry to generate large libraries of candidate molecules. However, despite keen interest and significant investment, these libraries have not produced nearly the number or diversity of antibiotic agents expected. Other than the kinase inhibitors and G-protein coupled receptor antagonists developed [7], these platforms have only able to generate a single family of new antibacterials (oxazolidinones [26]). This can be largely attributed to the decreased diversity of compounds observed in combinatorial libraries, where molecules in general lack as many chiral centers and complex ring systems [8]. Further, while many candidate compounds are discovered to effectively antagonize specific bacterial targets, a larger problem that is often overlooked is bacterial cell wall and membrane permeability [22]. Looking back, it is clear that NPs and their derivatives were our best source for antimicrobial agents. In fact, almost all new drugs clinically approved have been either NPs or NP derived / inspired. Only 4% of drugs approved since 1981 have been developed entirely from synthetic chemistry based approaches [20].

Peptides are of particular interest in this body of work, in part due to their immense functionality within biological systems. Even in primary metabolism, the inherent modularity of amino acid backbones fuel a wealth of molecular diversity. Using just the twenty base amino acids, and a few minor post translational modifications (PTMs) (e.g. disulfide bridges, phosphorylation, etc.), all forms of life create functional proteins that act as catalytic, structural and signaling elements. However, in many cases, these peptidic scaffolds have evolved niche roles in microbial antagonism. These antimicrobial peptides can be largely classified according to their size and relative decree of PTMs. Small unmodified antimicrobial peptides (AMPs) are found across all phylogenetic orders of life, and are believed to primarily target cell membranes. Since bacterial outer membranes are chemically and structurally distinct from eukaryotic membranes, and are often varied between bacterial taxonomies, AMPs such as these are able to exert specific antibacterial effects. While the specific mechanisms of action are believed to vary depending on concentration, peptide, target organism and membrane composition, it is generally believed that these AMPs all derive their action through instilling membrane pore formation and instability. Similarly, large unmodified peptides, such as the pyocins produced by various species of the genus *Pseudomonas* [18] and the colicins produced by *E. coli* [14], are also known for biological antagonism through a myriad of highly specialized mechanistic interactions with specific cellular membrane targets and intracellular processes.

With decreasing size, peptides often employ a wider set of modifications to achieve the structural and functional diversity required to mediate wide functionalities. These often use a much larger set of amino acids and extensive PTMs to achieve greater structural diversity [2, 10]. While early structural elucidation techniques were able to reveal the peptidic nature of these antibiotics discovered in the golden era, very little was known about their complex biosynthesis.

As molecular biology techniques were developed in the 1980s, along with DNA sequencing techniques, scientists were able to identify the genes responsible for NP biosynthesis. In many cases, these genes were clustered within the genome in regions now known as biosynthetic gene clusters (BGCs), and often transcriptionally linked within operons. We can now classify these heavily modified peptides into two main families, as directed by their biosynthetic pathways: (1) ribosomally synthesized and post-translationally modified peptides (RiPPs) and (2) nonribosomal peptides (NRPs).

RiPPs, as nomenclature suggests, are biosynthesized through ribosomal translation, followed by extensive post-translational processing by a variety of tailoring enzymes to form mature products of distinct classes. Due to these extensive PTMs, the biological activity of RiPPs is also highly variant. From the quorum sensing mediator ComX [21], morphogenic sapB [13], antifungal pinensins [19] to the antibacterial and recently discovered anticancer property of thiopeptides [3, 11], the chemical diversity in this class of compounds is demonstrably wide enough to accomplish these various biological functions. While at first indistinguishable from NRPs, this class of molecules is distinct because the core chemical backbone begins as a standard peptide, comprised of the twenty proteinogenic amino acids. This precursor peptide is then modified with the use of various co-localized genes. Prior to excretion, cleavage is guided by various proteases along specific motifs to release a mature peptide. To date, over 21 subfamilies of RiPPs have been characterized [27], each denoted with unique tailoring motifs and intriguing biosynthetic machinery capable of assembling a diverse pool of chemical scaffolds.

Contrary to the prior peptidic NPs, NRPs are assembled via multimodular protein complexes called nonribosomal peptide synthases (NRPSs). Unlike the ribosomal peptide assembly systems that are limited to the set of amino acids covalently bonded to tRNA molecules, these factories are much more flexible in their design [17]. Within the genome, these assembly lines are encoded into modules, where each successive amino acid is encoded by a distinct group of genetic domains. A minimal module is comprised of three domains: adenylation (A) domains selectively activate amino acids with ATP which is then transfered onto neighbouring thiolation (T) domains where they are temporarily covalently bonded to a long, flexible phosphopantethenine extension. At this point, the adjacent condensation (C) domain is able to create the amide bond between the activated amino acid and the growing peptide chain. Through this process, each module is able to append a wide range of amino acids, with as many as five hundred possibilities recognized to date [5]. These growing chains are then released by a thioesterase (TE) domain which uses water or a in-product nucleophile to release linear or macrocyclized product respectively. Further diversity can be introduced with the addition of various starter units and incorporation of polyketide modules. This complex, and highly substitutable modular machinery allows for the creation of extremely diverse chemical structures, tailored towards to biological activity at hand.

In the past, these products were identified through largely through bioactivity guided screening platforms. While successful at the time, it is widely accepted that this methodology has aged. Originally, bacterial and fungal strains were gathered and cultured *en masse* for subsequent biological assays. While this was exceptionally successful for compounds relevant to the assay of interest, it was only possible to effectively identify compounds produced in abundance. However, recent genomic efforts have demonstrated that many actinomycetes, responsible for over two thirds of current antibiotics [16], may be hiding several so-called cryptic gene clusters [24]. These are BGCs that can be detected clearly through genomic scans, but are missed through bioactivity guided screens likely due to their low abundance or unique regulation pathways. These BGCs represent a wealth of untapped NPs that have not been investigated for antibacterial activity. The work here aims to develop a comprehensive pipeline to selectively target these NPs to reveal novel peptides with strong antibiotic activity.

Recently, several new technologies have emerged to better broadly survey biological extracts, microbial genomes and their behaviour in ecological niches. Metabolomics, powered by liquid chromatography coupled mass spectrometry (LCMS), is able to identify a wealth of molecules in crude biological samples. Due to the high resolution nature of of this technique, we can reveal exact masses of the thousands of metabolites produced by a single microbe. Further, selected metabolites can be further examined through tandem mass spectrometry (MS^2) where these metabolites are broken down into fragments revealing key structural information. Techniques such as this are crucial in revealing the metabolites of these cryptic BGCs, as we are able to detect metabolites far below the effective minimum inhibitory concentration (MIC) of potential antibiotics.

Genomics, or more specifically, the significant decrease in sequencing cost, has lead to an explosion of publicly available genomic sequence data. More than one hundred thousand bacterial isolates have been sequenced, assembled and uploaded to NCBI [23]. Since bacterial metabolites are directly encoded within their genomes, this data can be used to estimate the antibiotic potential of a given bacterium. Further, if used correctly, this data can infer key structural properties of the NPs produced. This data can effectively then be used to assess the novelty and relationship to known antibiotics. Further, these structural cues can be used to relate the putative products of BGCs to their actual product measured in metabolomic data.

Studying microbial communities and their produced metabolites *in situ* can reveal key insights about the function of specific metabolites on With the rise in large scale DNA sequencing techniques, it is now possible to conduct even more untargeted experiments to survey microbial interactions within their ecological niche. Specifically, metatranscriptomic experiments can illuminate the expression level of various genes responsible for NP biosynthesis, and thus estimate levels of a given metabolite within sample. Further, metagenomic data can estimate the overall abundance of various microbes within a given sample. While these techniques are still in their infancy, the integrated human microbiome project (iHMP) has recently conducted an in depth survey of the lower intestinal tract, studying multiple healthy and diseased patients over time using these techniques. Although the goals of the iHMP are multidisciplinary, we can leverage this data to find antibiotics that are already native to the human ecosystem.

Although natural products discovery and characterization is a crucial source of new bioactive chemical scaffolds, pharmacophores and lead molecules, the methods used in the past have since lead to diminishing returns. With an ever increasing crisis and antibiotic resistance, it is crucial to accelerate this discovery pipeline using technologies recently developed. In particular, these data rich experimental platforms such as genomics, metagenomics, metabolomics and metatranscriptomics have made significant impacts into our understanding of biology and biochemistry as a whole. Unlike technologies in the past that allowed for reductionist, highly controlled environments, these next generation platforms generate massive amounts of data that are not apprehensible without the use of modern day compute power. In the context of drug discovery, this body of work aims to develop scientific software that can better enable researchers to appreciate these data in the endeavour of targeted drug discovery. **I propose that through a stronger machine understanding of NP biosynthesis, we can leverage this data to effectively accelerate the search for novel antibiotics.**

1.2 Scope and nature of this work

The collection of works presented here represents a curated sample of the projects undertaken through my graduate studies that best demonstrate my endeavours in developing scientific software targeted towards accelerating NP discovery. First, I begin in **Chapter 2** with an in depth exploration of genomically encoded RiPPs among the totality of publicly available genome sequences. This aims to provide a complete platform for NP discovery and characterization. Through the development of software that can accurate classify and compare genomically encoded RiPPs, I demonstrate the wealth of scaffolds yet to be characterized while providing key metrics upon the various bacterial clades housing the most divergent examples. Further, I present here a clear protocol for the isolation and characterization of these products through targeted metabolomics.

Following this, I present a second pipeline to leverage *in situ* microbial dynamics in the characterization of antimicrobial peptides (**Chapter 3**). Here, we look into the secondary metabolites produced exclusively by microbial entities found on and within the human body. Our body is a host to a diverse and complex polymicrobial environment that is in constant flux. While it is known that certain microbial flora may provide a protective role against foreign pathogens, relatively little has been explored into the chemicals mediating these interactions. Through the integrated use of genomic, metagenomic and metatranscriptomics, I present a platform to visualize the rich network of antimicrobial interactions mediated by secondary metabolites. We glean from this analysis a molecule responsible for inhibiting the infection of *Clostridium difficile*, a notorious infectious agent of the GI tract.

Taken together, these software platforms provide a novel way to perform experimentation in the modern information era. Using informatics, and more specifically, software systems understanding of NP biosynthesis customized for this research field, I have demonstrated here the recent successes of this platform. These selection of works that I have completed show a small glimpse into the potential of informatics strategies in this field, and towards the overall goal of furthering our understanding of microbial natural products.

Bibliography

[1] Achan, J. et al. Quinine, an old anti-malarial drug in a modern world: Role in the treatment of malaria. *Malaria Journal*, 10:144, May 2011. ISSN 1475-2875. doi: 10.1186/1475-2875-10-144.

[2] Arnison, P.G. et al. Ribosomally synthesized and post-translationally modified peptide natural products: Overview and recommendations for a universal nomenclature. *Natural product reports*, 30(1):108–160, January 2013. ISSN 0265-0568. doi: 10.1039/c2np20085f.

[3] Bagley, M.C. et al. Thiopeptide Antibiotics. *Chemical Reviews*, 105(2):685–714, February 2005. ISSN 0009-2665. doi: 10.1021/cr0300441.

[4] Brownstein, M.J. A brief history of opiates, opioid peptides, and opioid receptors. *Proceedings of the National Academy of Sciences*, 90(12):5391–5393, June 1993. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.90.12.5391.

[5] Dejong, C.A. et al. Polyketide and nonribosomal peptide retro-biosynthesis and global gene cluster matching. *Nature Chemical Biology*, 12(12):1007–1014, December 2016. ISSN 1552-4469. doi: 10.1038/nchembio.2188.

[6] Demain, A.L. Pharmaceutically active secondary metabolites of microorganisms. *Applied Microbiology and Biotechnology*, 52(4):455–463, October 1999. ISSN 0175-7598, 1432-0614. doi: 10.1007/s002530051546.

[7] Drews, J. Drug Discovery: A Historical Perspective. *Science*, 287(5460):1960–1964, March 2000. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.287.5460.1960.

[8] Feher, M. and Schmidt, J.M. Property Distributions: Differences between Drugs, Natural Products, and Molecules from Combinatorial Chemistry. *Journal of Chemical Information*

and Computer Sciences, 43(1):218–227, January 2003. ISSN 0095-2338. doi: 10.1021/ci0200467.

[9] Fleming, A. On the Antibacterial Action of Cultures of a Penicillium, with Special Reference to their Use in the Isolation of B. influenzæ. *British journal of experimental pathology*, 10 (3):226–236, June 1929. ISSN 0007-1021.

[10] Heel, V. et al. BAGEL3: Automated identification of genes encoding bacteriocins and (non-)bactericidal posttranslationally modified peptides. *Nucleic Acids Research*, 41(W1): W448–W453, July 2013. ISSN 0305-1048. doi: 10.1093/nar/gkt391.

[11] Hegde, N.S. et al. The transcription factor FOXM1 is a cellular target of the natural product thiostrepton. *Nature Chemistry*, 3(9):725–731, September 2011. ISSN 1755-4349. doi: 10.1038/nchem.1114.

[12] Kaufmann, S.H.E. and Schaible, U.E. 100th anniversary of Robert Koch's Nobel Prize for the discovery of the tubercle bacillus. *Trends in Microbiology*, 13(10):469–475, January 2005. ISSN 0966-842X. doi: 10.1016/j.tim.2005.08.003.

[13] Kodani, S. et al. The SapB morphogen is a lantibiotic-like peptide derived from the product of the developmental gene ramS in Streptomyces coelicolor. *Proceedings of the National Academy of Sciences of the United States of America*, 101(31):11448–11453, August 2004. ISSN 0027-8424. doi: 10.1073/pnas.0404220101.

[14] Lazdunski, C.J. Pore-forming colicins: Synthesis, extracellular release, mode of action, immunity. *Biochimie*, 70(9):1291–1296, September 1988. ISSN 0300-9084.

[15] Lewis, K. Antibiotics: Recover the lost art of drug discovery. *Nature*, 485(7399):439–440, May 2012. ISSN 1476-4687. doi: 10.1038/485439a.

[16] Lucas, X. et al. StreptomeDB: A resource for natural compounds isolated from Streptomyces species. *Nucleic Acids Research*, 41(Database issue):D1130–D1136, January 2013. ISSN 0305-1048. doi: 10.1093/nar/gks1253.

[17] Marahiel, M.A., Stachelhaus, T. and Mootz, H.D. Modular Peptide Synthetases Involved in Nonribosomal Peptide Synthesis. *Chemical Reviews*, 97(7):2651–2674, November 1997. ISSN 0009-2665. doi: 10.1021/cr960029e.

[18] Michel-Briand, Y. and Baysse, C. The pyocins of Pseudomonas aeruginosa. *Biochimie*, 84 (5-6):499–510, 2002 May-Jun. ISSN 0300-9084.

[19] Mohr, K.I. et al. Pinensins: The first antifungal lantibiotics. *Angewandte Chemie (International Ed. in English)*, 54(38):11254–11258, September 2015. ISSN 1521-3773. doi: 10.1002/anie.201500927.

[20] Newman, D.J. and Cragg, G.M. Natural Products as Sources of New Drugs from 1981 to 2014. *Journal of Natural Products*, 79(3):629–661, March 2016. ISSN 0163-3864. doi: 10.1021/acs.jnatprod.5b01055.

[21] Okada, M. et al. Structure of the Bacillus subtilis quorum-sensing peptide pheromone ComX. *Nature Chemical Biology*, 1(1):23–24, June 2005. ISSN 1552-4469. doi: 10.1038/nchembio709.

[22] Overbye, K.M. and Barrett, J.F. Antibiotics: Where did we go wrong? *Drug Discovery Today*, 10(1):45–52, January 2005. ISSN 1359-6446. doi: 10.1016/S1359-6446(04)03285-4.

[23] Pruitt, K.D. et al. NCBI Reference Sequences (RefSeq): Current status, new features and genome annotation policy. *Nucleic Acids Research*, 40(D1):D130–D135, January 2012. ISSN 0305-1048. doi: 10.1093/nar/gkr1079.

[24] Rutledge, P.J. and Challis, G.L. Discovery of microbial natural products by activation of silent biosynthetic gene clusters. *Nature Reviews Microbiology*, 13(8):509–523, August 2015. ISSN 1740-1534. doi: 10.1038/nrmicro3496.

[25] Schatz, A., Bugle, E. and Waksman, S.A. Streptomycin, a Substance Exhibiting Antibiotic Activity Against Gram-Positive and Gram-Negative Bacteria. *Proceedings of the Society for Experimental Biology and Medicine*, 55(1):66–69, January 1944. ISSN 0037-9727. doi: 10.3181/00379727-55-14461.

[26] Shaw Karen Joy and Barbachyn Michael R. The oxazolidinones: Past, present, and future. *Annals of the New York Academy of Sciences*, 1241(1):48–70, December 2011. ISSN 0077-8923. doi: 10.1111/j.1749-6632.2011.06330.x.

[27] Skinnider, M.A. et al. Genomic charting of ribosomally synthesized natural product chemical space facilitates targeted mining. *Proceedings of the National Academy of Sciences*, 113(42): E6343–E6351, October 2016. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1609014113.

[28] Williams, K. The introduction of 'chemotherapy' using arsphenamine – the first magic bullet. *Journal of the Royal Society of Medicine*, 102(8):343–348, August 2009. ISSN 0141-0768. doi: 10.1258/jrsm.2009.09k036.

[29] Wohler, F. On the artificial formation of urea. *Trends in Biochemical Sciences*, 3(1):17–18, 1978.

Chapter 2

Machine learning platform for the targeted discovery of divergent ribosomal natural products

2.1 Preface

While ribosomally synthesized and post-translationally modified peptides (RiPPs) have been extensively studied, there have been no formal attempts to develop metrics for comparing and assessing compound similarity specifically for this biosynthetic class as a whole. In the pursuit of advancing our software capabilities in deciphering large scale genomic, chemical and metabolomic databases, we have developed a tool kit that aims to accelerate the discovery of highly divergent and novel RiPPs. Specifically, the software here aims to create several metrics that can identify genomically encoded RiPPs, to facilitate targeted analysis. In doing so, we conduct a wide genomic scan on the entirety of publicly available bacterial genome sequences available through NCBI, identifying taxonomic clades strongly enriched in novel encoded RiPPs. Further, here we present a targeted isolation strategy that is able to specifically identify metabolites associated with specific RiPP gene clusters of interest. This was accomplished by first using genomic data to identify a collection of bacterial strains naturally lacking the specific encoded RiPP of interest using BARLEY. Through the development of CLAMS, an algorithm for the identification and dereplication of metabolites in large scale metabolomics, this collection of strains can be used to drastically reduce the number of candidate metabolites in a given metabolomics experiment. Of the remaining metabolites, we present a further strategy to reduce candidate metabolites using the predicted scaffold library generated by PRISM to both identify metabolites with similar fragmentation patterns and exact masses. Through this platform, we have identified a novel lasso peptide from a previously uncharacterized *Streptomyces*, and have demonstrated its targeted isolation. The software tools and protocols developed through this work are hoped to accelerate the discovery of novel and divergent RiPPs.

The following chapter is a manuscript prepared for submission for which I will be the lead author. For this work, I developed and validated BARLEY, an extension of the previous works developed by Christopher Dejong developed for NRPs and PKs. I conducted a large scale genomic scan, and used BARLEY in combination with CLAMS to develop a targeted isolation strategy leading to the discovery of streptopeptin. I also contributed to overall study design and wrote the manuscript. Walaa Mousa curated data, cultured bacteria, isolated streptopeptin, contributed to study design and wrote the manuscript. Michael Cannon and Christopher A. Dejong developed CLAMS, Michael Skinnider developed PRISM, extended LEMONS for assessing BARLEY, and edited the manuscript. Professor Nathan Magarvey contributed to study design and wrote the manuscript.

2.2 Abstract

Bacteria are historically recognized as an invaluable source of natural products with diverse chemical scaffolds. Noted for their unique biosynthesis, ribosmally synthesized post-translationally modified peptides (RiPPs) are a class of natural products with interesting structural diversity and bioactivities. Recent years have witnessed a revolutionary shift in in natural products discovery from bioactivity-guided to bioinformatics genome-based protocols aided by the exposed wealth of genomic data and the vast advances in bioinformatics algorithms. However, we still lack an automated tool to connect genomic predictions to metabolomic data. Here, we present

BARLEY/CLAMs pipeline, a machine-learning algorithm which expands on genome prediction tools to discover encoded novelty and selectively identify the putative cluster product in crude extract mixtures. BARELY is a local alignment tool that employs a unique multiple scoring matrix to assign a divergence score to predicted clusters relative to all known entities. CLAMS is metabolomics peak detection tool and can take biosynthetic predictions to link candidate ions in mass spectrometry data. BARLEY/CLAMS, as an integrated platform, facilitates the discovery of products from their native hosts and as such, subverts the laborious processes of cloning and heterologous expression to detect products, enabling targeted natural products discovery. As a proof of concept, we purified a new RiPP, we named it streptopeptin from an unexplored *Streptomyeces*. Streptopeptin is fully structurally divergent from all known RIPPs discovered to date.

2.3 Introduction

Natural products, in particular those of microbial origins, are appreciated for their structural diversity. Much of the strategies designed to enable natural products discovery rely on finding products from microbial extracts or metabolomes. Creating an efficient workflow from genomic startpoints, with an emphasis on diversity-staged discovery, would produce the changes that are necessary in a genomic era and would fix the issues of high rediscovery of known molecules facing natural product discovery [24]. Emergence of large-scale microbial genome sequence data has suggested that high numbers of natural products remain undiscovered [3]. However, few such tools are designed explicitly to define genomic loci according to their encoding of novel agents and further, carry this information forward to facilitate targeted isolation. Combined bio- and cheminformatic analyses have demonstrated that novel chemical scaffolds are more likely to affect divergent or previously unknown targets [10]. Consequently, a method capable of selectively targeting structurally novel agents for genome-guided discovery would be a highly desirable complement.

Ribosomally synthesized and post-translationally modified peptides (RiPPs) are noted for their structural diversity with RiPP biosynthesis proceeding from gene translation, post-translational tailoring to form mature products of distinct chemical classes. Like other microbial secondary products, gene clustering exists, facilitating computational identification within genomic sequences. In previous work, we introduced a computational pipeline, RiPP-PRISM, designed to identify and predict the structures of RiPPs of 21 structural and biosynthetic families [25]. A shortcoming was an inability to explicitly define the genetic clusters encoding novel RiPPs as compared with the known RiPPs. Moreover, to close the cycle of genomes to natural products and define novel low-abundance RiPPs, an explicit pipeline with genomic and metabolomics data integration needs formulation.

In the present study, we introduce an algorithm, Basic Alignment of RibosomaL Encoded products locallY (BARLEY), that permits direct alignment of biosynthetic information encoded in microbial genomes to a database of known RiPP chemical structures. To ensure the functionality of this suite, GRAPE was extended to recognize and comprehensively retrobiosynthesize RiPP scaffolds. Integration of BARLEY, with a metabolomics matching package, Computational Library for the Analysis of Mass Spectral data (CLAMS), pairs genomic and metabolomic data to facilitate selective isolation of novel RiPPs within complex microbial extracts. In doing so, we formalize an automated genome to natural products pipeline to permit the directed discovery of unknown RiPPs.

2.4 Results

The aim of this work is to survey the genomic landscape of encoded RiPPs and to connect the genomic prediction to real metabolomics data for targeted identification of new RiPPs in their native host producer. Previously, we attempted to map the encoded RiPPs using chemoinformatic methods based on the predicted chemical scaffolds identified by PRISM. Here, we extend on RiPP-PRISM capability with an integrated platform that dereplicates known RiPPs, assign divergence distance to known entities, and directly link the predicted gene cluster to the corresponding metabolite in the crude extracts. Here, we present BARLEY-CLAMS pipeline. Basic Alignment of Ribosomally Encoded products locally, BARLEY generates a divergence score of each predicted gene clusters to all known RiPPs using three unique scoring metrics. Comparative aLignment Algorithm of Mass Spectral, CLAMS connects predicted masses, identified by RiPP-PRISM, of each cluster to the corresponding putative metabolite in the crude extract. Specifically, CLAMS is able to accurately identify and categorize MS1 ions in crude metabolomic extracts. Using a pool of samples, CLAMS facilitates comparative analysis strategies to identify metabolites uniquely found in relation to specific samples of interest through subtracting metabolites observed in non-candidate bacteria and natural product databases. In combination with BARLEY, CLAMS identifies metabolites uniquely associated with specific RiPP cluster.

Comprehensive cataloguing of all known RiPPs

As a first step, we curated all known landscape of RiPPs (Figure 2.1) and created a database with their structures (Supplementary table 2). Each of the known RiPP classes has defined features that lead to their definition. Many of these are highly differentiated from other non-ribosomal peptides, NRPs. All RiPPs are biosynthesized by direct translation of their propeptide by the ribosome, the propeptide is then undergoes various structural modifications and cleavage to yield the mature product. Post-translational modifications that tailor the core peptide can be extensive when considering the complete catalog of known Ripps. We have summarized the existing known modifications within a simulated core peptide backbone for illustrative summary of the chemical functionalities (Figure 2.1). Examples of these structural alterations include; 1. simple head-to-tail macrocyclization, as observed in cyclic bacteriocins, 2. lanthionine and labionin bonds as in the lantibiotics, 3. multiple thiazole and oxozole ring formations as in cyanobactins, thiopeptides, YMs, and linear azol(in)e peptides (LAPs) 4. dehydro-amino acids, 5. Secondary amide bonds and characteristic knot-like topology observed in lasso, 6. O and S -glycosylations in gylcosins, 7. thioamide bonds in thioviridamides, 8. D-amino acids, β-hydroxylation and methylations in proteusins, 9. Varying prenylations in cyanobactins and ComX, 10. Pyridines, hydroxy pyridines and piperidines in thiopeptides among many more.

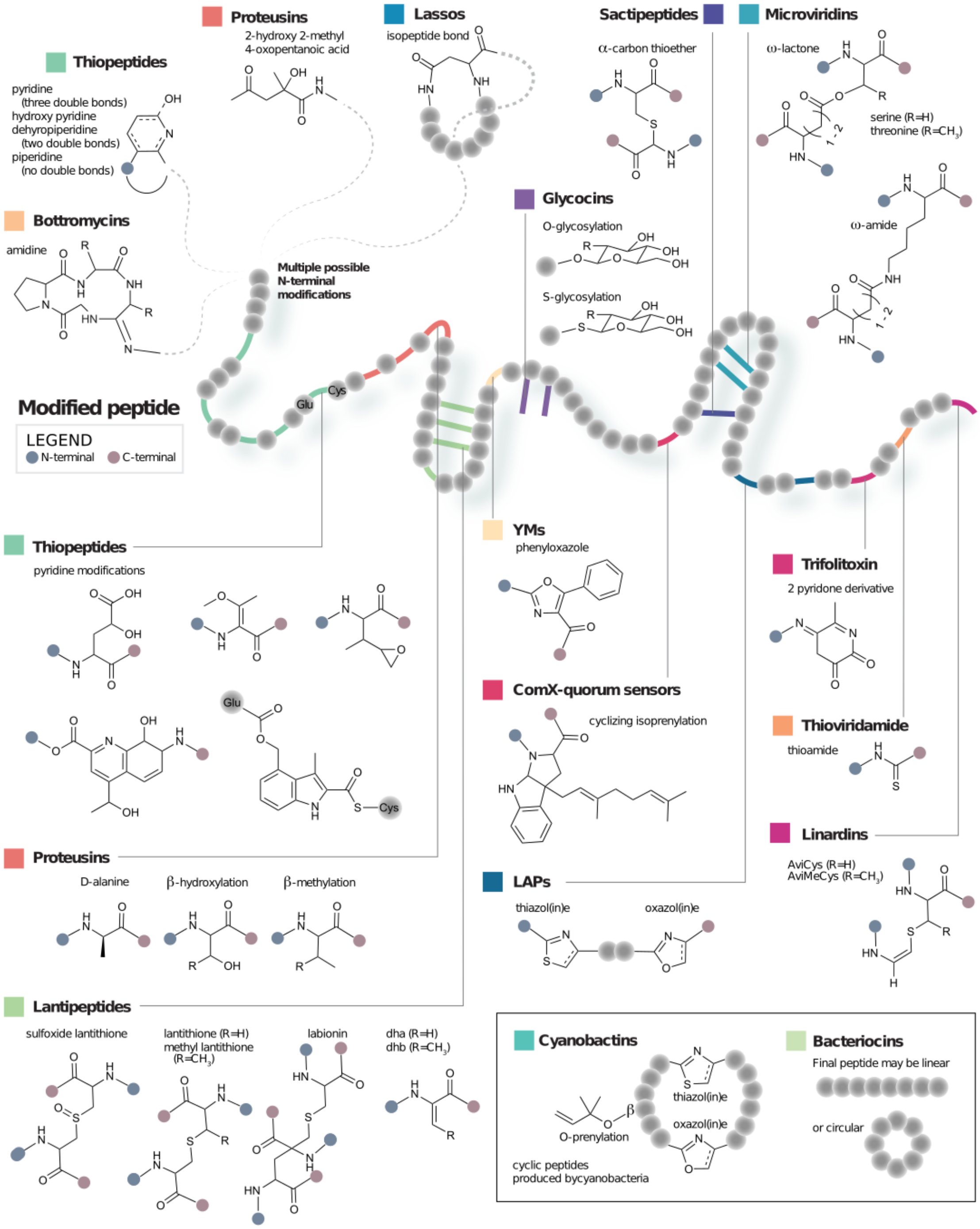

FIGURE 2.1: **Structural diversity of known RiPPs.** Shown is a summary of known posttranslational tailoring that defines subclasses within RiPPs. A simulated core peptide backbone was used for the illustrative summary.

The BARLEY/CLAMS workflow

Using PRISM, genomically encoded RiPPs of bacteria can be identified. Using this information, in conjunction with the totality of characterized RiPP scaffolds to date, BARLEY is able to identify novel RiPPs and dereplicate these across multiple bacterial strains. This information is translated to CLAMS to facilitate identification of metabolites that are uniquely present in strains carrying an encoded RiPP product of interest (Figure 2.2). In total, we are able to comparatively analyze metabolites in extracts of 463 strains, each with a multitude of media and growth conditions and genomic data, through BARLEY. Media constituents are also eliminated through the discarding of metabolite signatures present in any of 118 diverse blank media extractions. Further, we are able to dereplicate metabolites against a database of 50,317 bacterial and fungal natural products. In total, the combination of these software platforms allows for a much more targeted analysis of candidate metabolites, drastically reducing the amount of noise present in metabolomic experiments.

BARLEY implements multiple unique scoring metrics

BARLEY is an alignment algorithm designed for accurate scoring of genome-predicted RiPPs to all known entities. Dereplication is an essential preliminary step in the modern discovery pipeline to help diminish the rediscovery of natural products. BARLEY uses three scoring metrics to measure similarity between RiPP scaffolds, capture genomically encoded diversity of RiPPs and identify the novelty of encoded RiPPs. Through these metrics, BARLEY aims to create a method that accurately leads to the identification of novel and divergent encoded chemistry.

Measuring chemical similarity

Previously, we have demonstrated that chemoinformatic methods used for measuring chemical similarity have varying degrees of efficacies [26], and are highly dependent on the dataset [9]. In particular, due to the large size, and inherent modularity of peptidic scaffolds, typical hashing fingerprints may not be the best representative for plotting chemical similarity. Thus, GRAPE

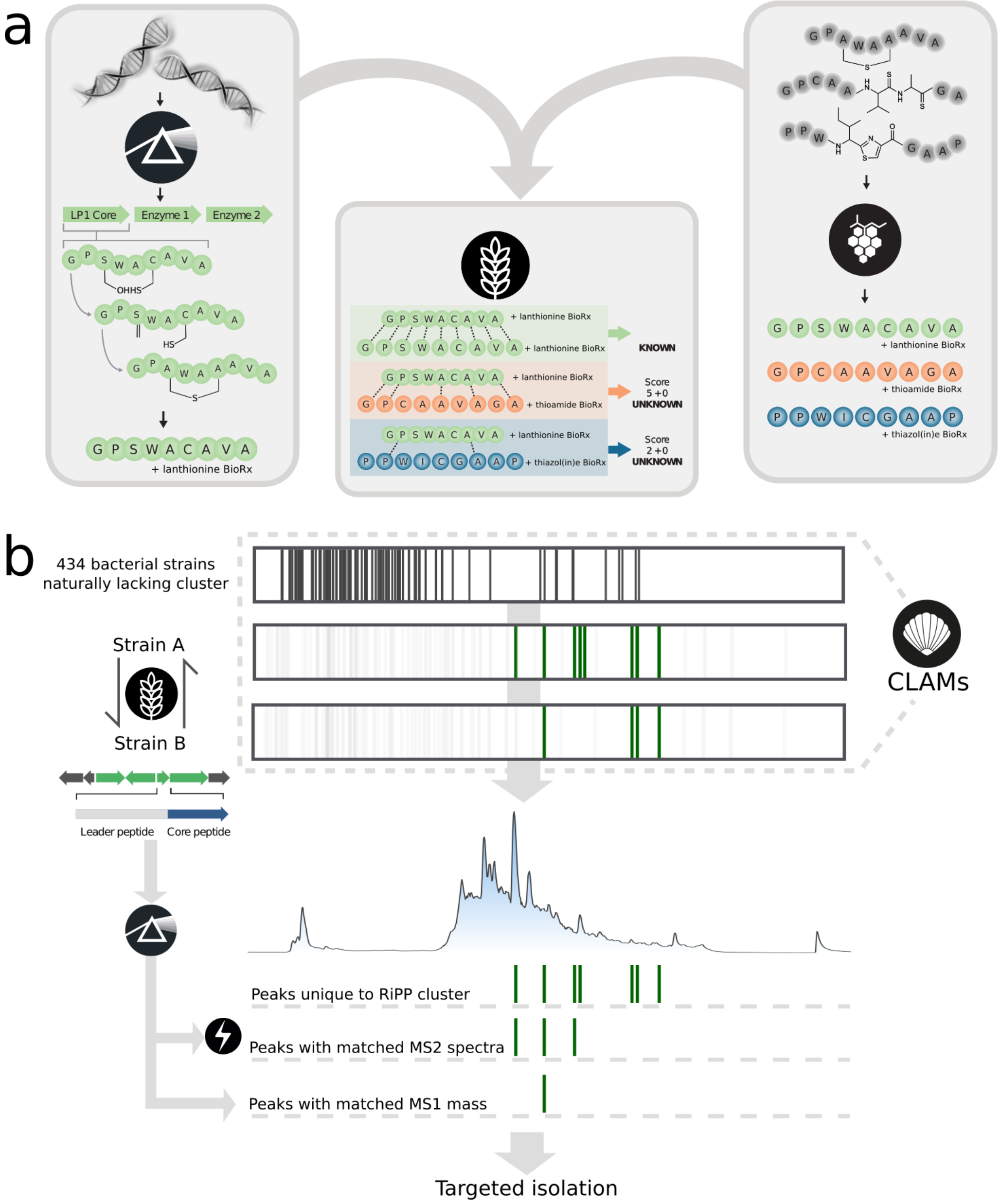

FIGURE 2.2: **Workflow for BARLEY/CLAMS pipeline.** PRISM parses genomic data to identify RiPPs as a core amino acid sequence alongside subsequent tarilorings. GRAPE retrobiosynthetically processes RiPP chemical scaffolds to identify the proteinogenic core amino acid sequence, and denotes any associated tailoring reactions. The input to BARLEY is these two streams of data. BARLEY performs local alignment between the amino acid sequences while scoring the number of tailorings that are matched. These scores are then classified to identify an encoded RiPP as new or previously characterized. Using a BARLEY genome guided scan, we can identify bacterial strains that are known to carry a specific encoded RiPP of interest, while identifying the numerous strains that naturally lack the product. CLAMs enables the identification of metabolites present uniquely in strains carrying the specific encoded RiPP. From the encoded RiPP cluster, PRISM generates a library of predicted chemical scaffolds. Using *in silico* fragmentation, this library is then compared to the experimental fragmentation patterns of metabolites, retaining only those with significant overlaps. Finally, candidate peaks may be selected by matching the exact mass as predicted by PRISM.

was extended to perform retrobiosynthesis on RiPP scaffolds. In particular, this algorithm is now capable of recognizing 60 post-translational modifications (PTMs) and is able to convert these resulting residues into the likely standard set of amino acids from which they are derived (Supplementary table 2). To determine a chemical similarity score, these derived amino acids are then compared using local alignment, while PTMs are scored independently. Using the naive scoring parameters, detailed in the methods section, we performed two stages of validation. Firstly, we examined the efficacy and accuracy of BARLEY to determine the class of a particular RiPP cluster. Specifically, we looked to validate how accurate BARLEY chemical similarity score can be used to assign RiPPs to a specific class. For comparison, a Tanimoto coefficient using the extended connectivity fingerprint with a radius of three (ECFP6-Tc) [23] was used as it was previously determined as the likely optimal choice for peptidic natural products [26]. In comparison to the ECFP6-Tc method, BARLEY scores are a better classifier of inter and intra family RiPP chemical relationships (Figure A1.1-b).

As a second stage of validation, we sought to compare more detailed chemical patterns within a single class. In particular, we use the case of class I lantipeptides and examine how chemoinformatic similarity tools work with peptides of increasing divergence. To conduct this experiment, we used LEMONS, a tool to generate hypothetical scaffolds and random substitutions given an initial product template [26]. LEMONS was extended to perform the LanB and LanC catalyzed reactions in silico, introducing dehyrdated amino acids and lanthionine bridges to scaffolds where possible. Thus LEMONS was used to generate a library of scaffolds using the initial proteinogenic sequence for nisin, where divergence was manually increased through random substitutions of individual amino acids. The results of this analysis demonstrate that while most metrics demonstrably have a negative correlation with increasing divergence, BARLEY is the most consistent similarity metric that can be correlated to this divergence, with the least degree of variance (Figure A1.1-b). Through these stages of validation, we deem that model based chemoinformatic solutions such as BARLEY are a better estimator of chemical divergence within RiPPs. In Figure 2.3, we have demonstrated the effectiveness of this novel metric in grouping RiPP families.

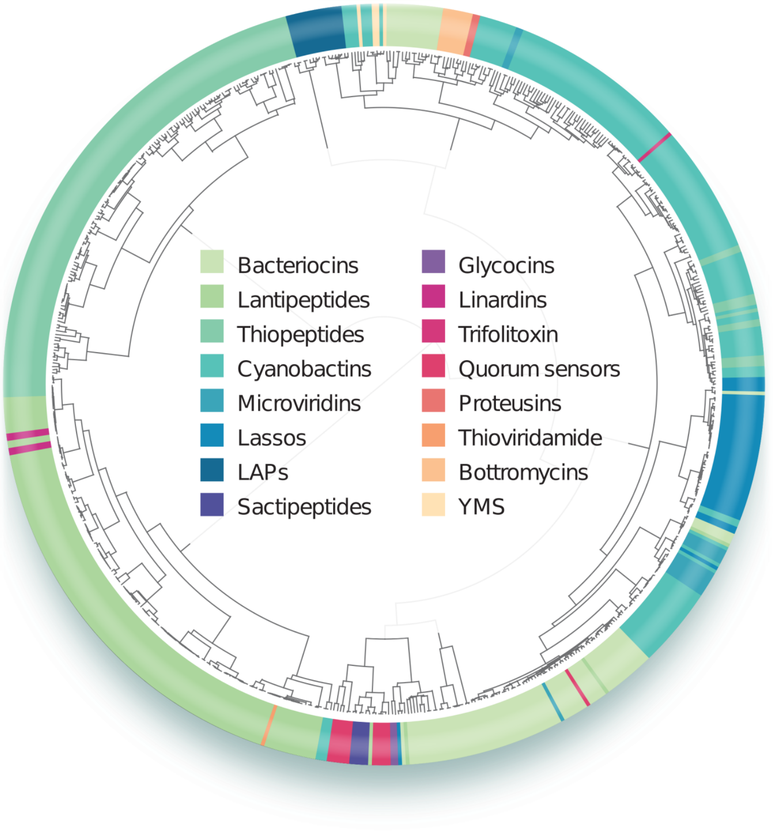

Figure 2.3: **Dendrogram of known RiPP chemical scaffolds as grouped by BARLEY.** Hierarchical clustering was performed using the Ward linkage method, using a distance matrix comprised of all pairwise scores generated by BARLEY between 640 unique RiPP scaffolds.

Measuring genomically encoded RiPP similarity

The main goal of measuring similarity between genomically encoded RiPPs, is to use this data as a proxy for real chemical structural information. Previously, we have shown that while RiPP-PRISM is highly accurate in translating encoded-BGCs to their respective scaffolds, it does generate an entire library of scaffolds to reach this level of accuracy. While these chemical scaffolds are essential for downstream targeted mass spectral identification, we demonstrate here that these predicted chemical libraries are not the most effective tool in charting and identifying diversity of genomically encoded RiPPs. As we have demonstrated that BARLEY is likely an effective tool for measuring chemical diversity, we measure how well genomic comparisons by BARLEY correlate to the same true chemical comparisons.

In total, our dataset is comprised of 136 gene clusters encoding for 161 products (Supplementary table 3). In cases where a single gene cluster might encode for multiple products due to multiple encoded precursor peptides as predicted by PRISM, BARLEY is able to consider these independently and assumes the same tailoring reactions occur for both products. However, for this analysis, we consider the maximum similarity between both structural products, to best ensure that we are capturing the event when two BGC clusters may produce the same product. Each cluster was compared by three metrics: (1) BARLEY, (2) median tanimoto similarity of predicted RiPPs, (3) BiGScape, a BGC comparison tool used to determine relative similarity between clusters. The characterized products of these BGCs were then respectively compared according to BARLEY and the ECFP-Tc (Figure 2.4-b). When assessing the full spectrum of similarity scores and their correlation, the highest correlation is seen between the Tc of PRISM's predicted structures and the Tc of their characterized products. However, it is important to note that the majority of their similarities occur at relatively low Tc values ¡ 0.5. Further, a major concern is that in several instances, clusters that encode the same product are often not distinctly recognized through PRISM's predicted structures. These cases were more pronounced in RiPP families with large combinatorial space such as thiopeptides and lantipeptides. To enrich for these cases, we examined cases where the respective encoded products were similar (i.e. Tc / BARLEY ¿ 0.5). With this, we found a stark reversal, where BARLEY genomic similarity scores are more consistently mapped to both chemical similarity scores both measured by BARLEY and

ECFP6-Tc. In all cases, BigScape, while demonstrating some correlation, was significantly less useful as a proxy for chemical similarity (Figure 2.4-b). Overall, this analysis reveals that when comparing related genomic clusters within the same family, BARLEY is more refined to capture the true underlying chemical diversity than existing tools. Further, BARLEY is able to more accurately depict duplicate clusters than through examination of PRISM structure predictions.

Measuring novelty among genetically encoded RiPPs

A final, but crucial functionality of BARLEY is its ability to determine the novelty of genomically encoded RiPPs. This metric is essential in focusing future research towards putative RiPPs that are yet to be characterized, and promotes the discovery of diverse chemical scaffolds. Unlike the metrics defined above, this metric was specifically trained using example data to better classify pairwise relationships between genetically encoded RiPPs and characterized RiPPs to fall in these three ordinal categories: unknown, within family, or exact match. In this attempt, the dataset was split into a training and test dataset (75 : 25 % split) of BGC-structure pairs, and stratified according to family such that any comparisons made in the test set were between structures and BGCs that were never trained on.

While both streams of data from PRISM and GRAPE contain a similar data format (core amino acids, identified modifications), this data is further annotated by BARLEY to describe 5 features; two features describe the strength of the local alignment, and three represent the similarity of PTMs. This training set was further split for validation and model selection, where two models were chosen and evaluated using 10-fold cross validation. In particular, it was found that a random forest performed the best to predict structural novelty. Three parameters were further tuned using this same validation set to generate a final random forest model with 400 base tree estimators, a minimum of 5 randomly selected features for each estimator, and a minimum terminal node size of 1 (Supplementary figures A1.2, A1.3, A1.4). Finally, this model was compared to the structure prediction engine within PRISM using the test dataset, in the context of classifying a candidate RiPP-BGC as either representing a novel, or previously characterized product. In this task, BARLEY outperforms the RiPP structure prediction engine

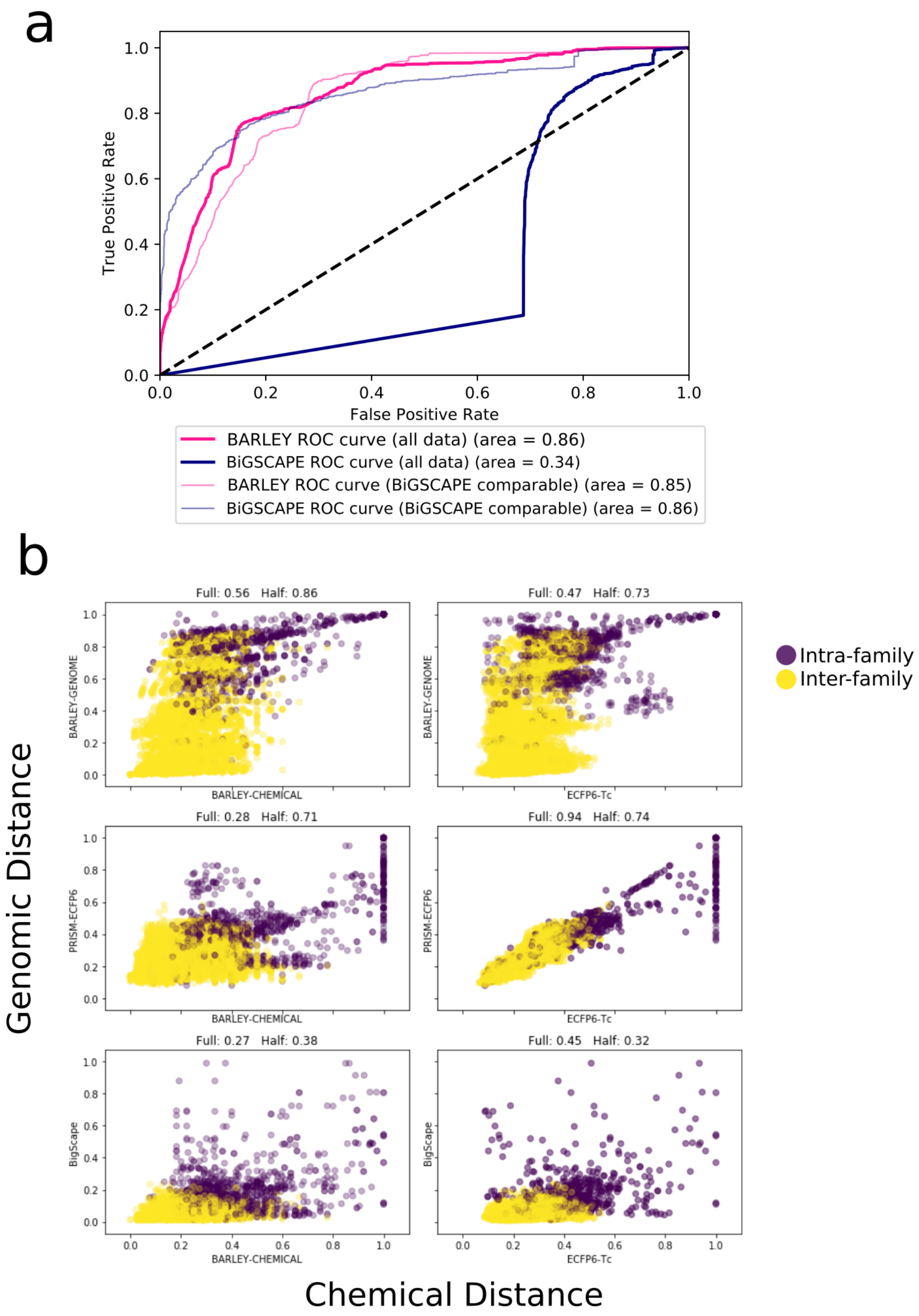

Figure 2.4: **Evaluating genomic similarities predicted by BARLEY.** (a) Comparing BigScape to BARLEY in determining encoded RiPP pairs as belonging to the same family. (b) Comparing genomic RiPP distances measured by BARLEY, median ECFP6-Tc distance of PRISM predicted scaffold library, and BigScape on y-axis to corresponding chemical distances measured by BARLEY and ECFP6-Tc on x-axis. Above each chart is the corresponding Spearman correlation between variables measure on both all data (full) and all data points above ECFP6-Tc 0.5 (half). Colours represent comparisons between RiPPs of same family (purple), or different family (yellow).

when considering the maximum Tc index between predicted scaffolds and the candidate comparison (Supplementary figure A1.5). BARLEY can comparatively analyze BGCs to candidate compounds with an accuracy of 99.7% when using a cutoff of 0.2.

Refined genomic charting

Using these key metrics defined, we set forward to reassess the genomic landscape of RiPPs. In particular, we set to define the diversity potential of genera according to each of RiPP subfamilies. Of total of 65,421 prokaryotic genomes analysed through PRISM, a total of 19,113 contained at least one BGC. A total of 27,393 products were predicted, using BARLEY, the total number of uniquely encoded products can be predicted as 21,849, with 5,062 predicted as being known (Figure 2.5). As BARLEY considers multiple RiPPs encoded within the same BGC as separate entities, the total number of novel unique compounds to be potentially isolated given the current set of genomes available is 16,787.

Among each of the RiPP families, we can use BARLEY to estimate the total encoded diversities. Here, we look at the mean BARLEY genomic distances between all encoded RiPPs as a measure for the total diversity. This effectively normalizes for the number of instances we observe encoded RiPPs, where certain rare families such as proteusins (8 encoded RiPPs) can be directly compared to more frequently encoded AIPs (8365 encoded RiPPs). This diversity metric effectively demonstrates the total space encoded within each biosynthetic family. Interestingly, a common trend seen here is that many of the rare RiPP families such as the YMs, cyanobactins, proteusins and prochlorosins in general encode for much more diversity than very common families. Surprisingly, class III/IV lantipeptides are the least diverse class of RiPPs, despite a total of 4154 encoded products detected through this analysis. Among these, these, sactipeptides and bottromycins are the two families which are mostly comprised of previously characterized RiPPs (Figure 2.6-a).

With this in mind, we set to evaluate the encoded chemical diversity across genera to evaluate these bacterial clades for targeted mining. This is particularly difficult due to the large divergence in sample size per genera, with the Streptococcus being the most sampled genera

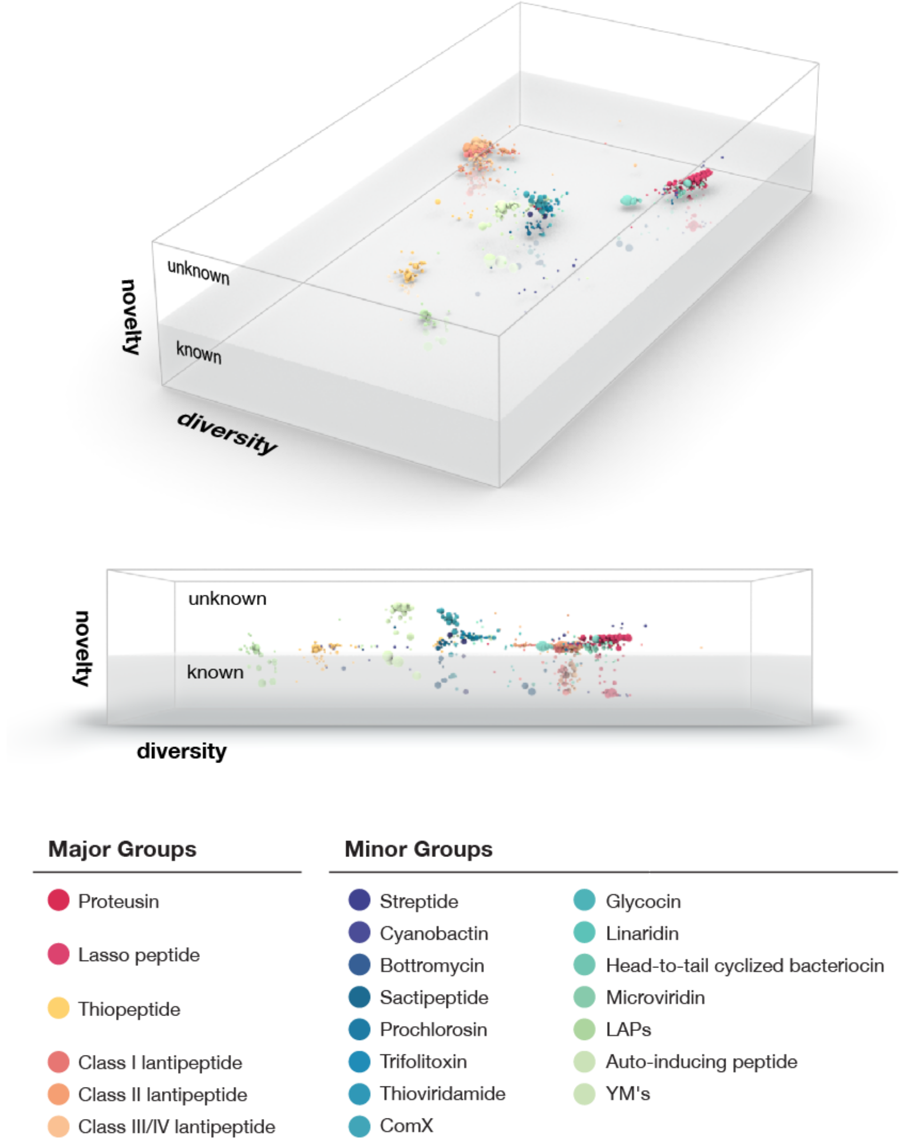

FIGURE 2.5: **Topological distribution of 21,849 encoded RiPPs according to diversity and novelty as determined by BARLEY.** Diversity is plotted on two axes, shown at two angles for appreciation. Diversity was plotted using classical multidimensional scaling using a distance matrix of BARLEY pairwise distance scores. Novelty is presented on the y-axis as labeled and all encoded RiPPs predicted as known fall within the grey-filled space. Colours represent the diverse RiPP families, and the size of the points represent the number of occurrences a unique RiPP scaffold is encoded in multiple genomes.

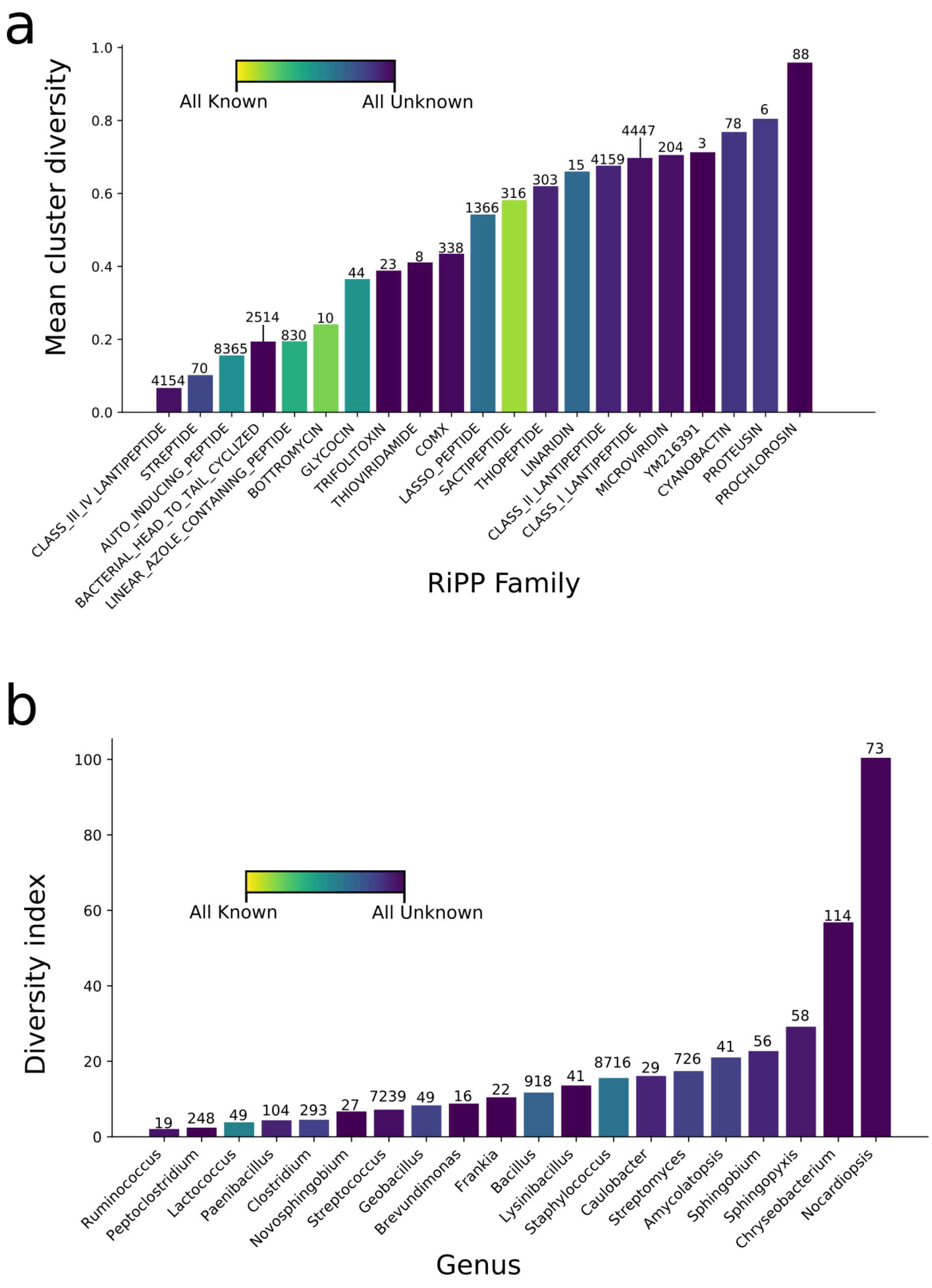

Figure 2.6: **Encoded RiPP diversity and novelty across chemical subfamilies and genera.** (a) Mean diversity of encoded RiPPs across chemical subfamilies, numbers above bars represent the number of encoded RiPPs in each class. (b) Top 20 most diverse RiPP encoding genera. Numbers above bars represent the number of unique RiPPs encoded in each genera observed, only genera with at least 20 sequenced members were included here. In both charts, colour represents the percentage of uniquely encoded RiPPs that were predicted as known.

with over 9934 deposited sequences, and 703 genera with only a single sequenced member. Using an iterative sampling strategy and calculating total diversity as the sum of BARLEY defined distance measurements between the detected RiPPs per sample, we generate a slope for each genera representing the average amount of RiPP diversity increased per new genome sequenced. In this context, we evaluate each genera according to this diversification index in the context of all RiPPs, followed by a specific evaluation according to each RiPP family. This index (Di, Diversity index), represents how frequently we can expect to find a highly divergent RiPP through sampling more members of a specific genera. To ensure accurate sampling, only genera with at least twenty sequenced members were included, leaving 163 genera for analysis (Figure 2.6-b).

When concerning all RiPP families, the genus with the largest projected diversity of RiPPs is Nocardiopsis (Figure 2.6-b). In this genus, despite only having 28 sequenced members, these strains carry a total of 75 unique products of which none are predicted as characterized. In particular, this genera is rich in diverse thiopeptides, linear azole containing peptides (LAPs), lasso peptides and all classes of lantipeptides (Figure 2.6-b). The total Di for this genus is 100.4.

Within the class I lantipeptides, Staphylococcus represents the genera with the most diverse encoded scaffolds (Di=2.4), with a total 2699 unique predicted products from a total of 7235 genomes analysed (Figure 2.6-b). Following this is Paenibacillus (Di=0.7), Lacotoccus (Di=0.44) and Geobacillus (Di=0.33) where despite containing diverse encoded RiPPs, these genera are also abundant in previously characterized products. Of the class II lantipeptides, Bacillus (Di=1.55) and Carnbobacterium (Di=1.24) carry the most diversity, with 23% and 66% of these products representing characterized products. Balancing these two factors, Streptococcus seems to hold the most divergent class II lantipeptides (Di=1.13), while only 1% of these encoded products are predicted as known (Figure 2.6-b). For class III/IV lantipeptides, Streptomyces are the carriers of the most diverse products (Di=0.73), while 36% of its products are predicted as characterized. Following this genus are Nocardiopsis (Di=0.58) and Cellulomonas (Di=0.25), where none of their predicted products are predicted to be known.

For thiopeptides, Nocardiopsis represents the genus with the most diverse encoded products (Di=8.76), with only 3% of its products marked as known. Following this are the Deinococcus (Di=0.71), and Streptomyces (Di=0.4), with 0% and 9% of its clusters marked as known. In general, apart from the limited diversity carried by Lactobacillus (Di=0.0008, 75% known) and Bacillus (Di=0.01, 31% known), most genera carry a majority uncharacterized thiopeptides, making this class of RiPPs a tractable target for isolation of novel scaffolds (Figure 2.6-b).

ComX, previously characterized in Bacillus [20], was detected here in three genera, Lysinibacillus, Bacillus and Anoxybacillus. Of these, only Bacillus carried ComX peptides that are previously characterized, but even this only represents a small portion (1%) of their total predicted diversity (Figure 2.6-b).

Autoinducing peptides (AIPs) represent the largest family of products predicted. AIPs were first characterized from Staphylococcus [16], and this trend is apparent here where 54% of the diversity encoded by this genus (Di=1.26) represents previously characterized products. Interestingly, Peptoclostridium carries the most diversity (Di=2.01), and is completely uncharacterized. Apart from these, diverse AIPs are found in Clostridium (Di=0.96), Ruminococcus (Di=0.1) and Lactobacillus (Di=0.001) that are all uncharacterized (Figure 2.6-b).

Linear azole containing peptides (LAPs) are encoded by eight genera, with Nocardiopsis (Di=0.004), Streptomyces (Di=0.003) and Streptococcus (Di=0.0027) contain the most diversity. containing the most predicted diversity (Di=0.004). Of these LAP encoding genera, Clostridium (Di=0.0004, 94% known) and Bacillus (Di=0.0003, 84% known) carry a number of unique products, however the majority of these are closely related and predicted as characterized, leading to low diversity indices.

Lasso peptides are found among a large diversity of genera. Of these, Sphingopyxis (Di=31.49), Sphingobium (Di=19.95) and Caulobacter (Di=14.8) represent the most enriched genera with diverse lasso peptides. Of the remaining genera, all contain a fairly high density of divergent RiPPs, however it is interesting to note that of all the 436 unique products encoded by the genus Burkholderia, the majority (91%) share significant similarity to characterized products (Figure 2.6-b).

Microviridins are very abundantly encoded within the Chryseobacterium genus, with 104 products found in just 61 genomes. While also found in Alteromonas and Flavobacterium, these products are generally quite rare outside Chryseobacterium (Figure 2.6-b).

Several families of RiPPs were denoted as much rare, appearing in only select genera. Of these, prochlorosins were found in two genera, Prochlorococcus (Di=0.47) and Butyrivibrio (Di=0.02), where both genera carried completely uncharacterized products (Figure 6b). Linaridins were only found in Streptomyces, with 12 encoded products, of which 59% are predicted to be novel. Likewise, trifolitoxins were only found within the Rhizobium genera, for a total of three gene clusters. Bottromycins were only found in select Streptomyces where 10 unique products were encoded, however 8 of these were very similar to the currently characterized member of this family. Glycocins are only found in Bacillus, however, are fairly diverse, with only 10% of these marked as previously characterized. Thioviridamides were only found in Streptomyces, of which the majority are uncharacterized. Sactipeptides were only found in Bacillus, of which almost all (94%) were denoted as previously isolated.

Mass spectral strategies for genomic RiPP identification

In order to facilitate targeted mining, we have developed an explicit workflow for the identification of RiPP scaffolds in native strains using metabolomic data. The goal of this work was to develop an analytical workflow that can drastically reduce the number of LC-MS signatures for targeted isolation. Through a subtractive analysis, where BARLEY is used to identify strains sharing identical products, thus when searching for mass spectral signatures corresponding to a specific RiPP gene cluster, extracts from other bacterial strains can be used as a negative control to eliminate these redundant signatures. Herein, we present an example where this protocol is able to identify a novel lasso peptide from a previously unexplored Streptomyces isolate (Figure 2.7).

A lasso peptide cluster that was predicted as novel was found encoded in *Streptomyces* sp. BTA 0171 was sought after in this manner. This strain was grown in four media conditions, resulting a total of 20,007 MS1 ions. An initial filter for ions present in over fifty strains and dereplication of ions between media conditions identified a total of 611 MS1 ions relevant to this pursuit. After filtering against a bacterial metabolite database a total of 142 ions were eliminated as previously characterized compounds. A further 421 ions were eliminated due to very low abundance. BARLEY identified that no other strain within our internal collection contained this exact encoded product, and thus a further 29 ions were removed, resulting in a remaining

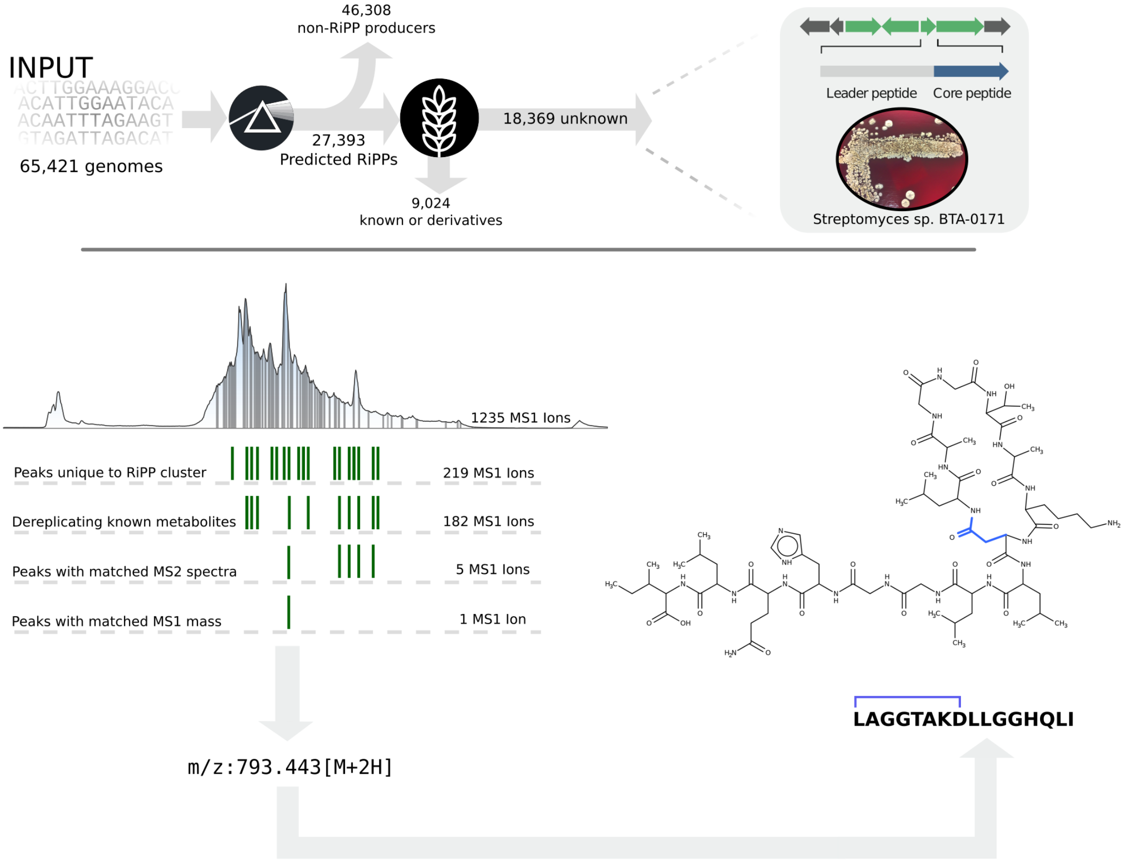

FIGURE 2.7: **Genomic and metobolomic guided isolation of novel lasso peptide.** All publicly available prokaryotic genomes were analysed through PRISM to identify 27,393 predicted RiPPs. Of these, RiPP-BARELY identified 18,369 of these as previously uncharacterized, with 844 of these being lasso peptides. Within these, *Streptomyces* sp. BTA-0171 was chosen as a candidate. This strain was grown in four media conditions, resulting in a total of 20,007 ions. This was dereplicated to 611 MS1 ions. After filtering for an intensity of sixty thousand, the remaining metabolites were then dereplicated against a bacterial metabolite database, removing a further 421 low intensity compounds and 142 known compounds. In total, the metabolomic database currently contains 459 other strains with both genomic and metabolomic data for the process of metabolite dereplication. This metabolite was not observed in any other strains according to BARLEY, and so 29 metabolites were removed for being observed in other bacterial cultures. Of the remaining 19 metabolites, 5 had some fragmentation patterns in common with in-silico fragmentation of the predicted lasso peptide, but one in particular had both strong fragmentation similarities and also shared the exact predicted mass of the lasso peptide. This metabolite was then targeted for isolation through large-scale fermentation and its structure was determined to be exactly as predicted.

19 metabolites. PRISM was used to generate a combinatorial structure library for this encoded product, revealing 17 unique masses, and an *in silico* generated library of 1000 fragment ions. Of these remaining ions, 5 were revealed to share some fragmentation similarity to the encoded product, while one was found to both share significant fragmentation similarity and match a mass within PRISMs structural library. The resulting ion was then selected for targeted isolation from a large scale fermentation in KE medium, given this ion was most intensely witnessed in GGYM and KE media. NMR spectroscopy data revealed the structure to be exactly as predicted by PRISM, further validating this pipeline (Figure 2.7, supplementary figures: A1.11 – A1.18). Currently no biological activity was seen against gram positive (*S. aureus* Newman) and gram negative (*P. aeruginosa* PAO1) bacteria or fungi (*C. albicans* ATCC 90028), however a much wider array of bioactivity assays may be needed to assess the total functional spectrum of this novel scaffold considering the diverse targets of existing lasso peptides [29, 11, 21].

Identifying noncanonical precursor peptides

Traditional genome mining approaches for RiPPs is dependent on the identification of precursor peptides adjacent to modifying enzymes. In certain scenarios, it may be that precursor peptides can leverage enzymes located in distant regions of the genome to facilitate post-translational modifications. In many cases, PRISM is not able to identify a definitive precursor peptide within gene clusters. Of the 30,261 BGCs identified by PRISM, 5,459 were denoted to either not have a precursor peptide at all, or to only have one on the basis of a heuristic rule determined by PRISM. With this idea, we acknowledge this as a limitation of our current platform, and the encoded diversities as described above may be a conservative estimate. As seen in Figure A1.7, all subclasses of lantipeptides are frequently observed with this phenomenon. To investigate this, a candidate bacterium (*Flavobacterium ginsengiterrae* JCM 17337) was chosen with an abundance of class I lantipeptides gene clusters lacking canonical precursor peptides. Specifically, this bacterium was denoted with 5 class I lantipeptide gene clusters, 3 of which were found without a definitive precursor peptide. Due to the inherent failure of homology based models (HMMs and motifs) in recognizing the precursor peptides in this scenario, we set to develop a new set of heuristic rules for the size and possible cleavage sites of precursor peptides using the totality of identified precursor peptides from our genomic analysis. As seen in figure A1.8, precursor ORFs in class I lantipeptides are typically between 40 and 80 amino acids in length, while cleavage sites within these peptides are typically located on the relative position 0.4 (Figure A1.9), but can be variable. Using this information, a heuristic rule was developed such that putative class I lantipeptide ORFs must be between 40 and 80 amino acids in length, and must contain at least two cysteines and two of either serine or threonines in the C-terminal 60% region of the ORF. Further, it was gleaned from this data, that the median size of mature products within class I lantipeptides was 22 AA, but was variable with a range from 8 to 53. Thus, a combinatorial cleavage and structure generation strategy was implemented to create putative products at position 22 ± 10 from the C-terminal to use for subsequent structure prediction using PRISM. Using this strategy, Prodigal was used to extract protein coding ORFs within this genome of *F. ginsengiterrae* JCM 17337, and a total of 70 ORFs were found to meet this set of heuristics. These ORFs were used to generate a total of 1,400 cleavage possibilities. In total 165,040 predicted structures across 3,232 unique masses were generated. To evaluate this approach, this bacteria was fermented, and its biological extracts were analyzed using LCMS/CLAMS to reveal a total of XX metabolites. Of these, 20 were identified to match a mass from this set of predictions where two products (ginsebactin and ginsecidin) were isolated. In both cases, the precursor peptides of these ORFs were encoded on large contigs (flavopeptin I: 466 kbp contig, ORF at 377 kbp; II: 500 kbp contig, ORF at 60 kbp) containing no genes detected by PRISM to share homology with any lantipeptide modifying enzymes. Although seemingly unlikely, NMR spectroscopy revealed these two products to match exactly to the structures as predicted by this analysis (seen in Figures A1.19 and A1.27. This process shows that these products exist and are not produced through canonical pathways, while this software strategy is able to identify peptides outside of the canonical rules for RiPP genomic encoding.

2.5 Discussion

Presented here is a comprehensive platform in the identification and targeted isolation of novel genomically encoded RiPP scaffolds. Due to their direct genomic translations, genome mining of RiPPs has garnered significant interest [1]. However, the main focus of most current algorithms is biased towards detection of BGCs [2, 28], without making any attempt to link this information to metabolomic datasets. Further, there are no current platforms which aim to targetedly dereplicate genomically encoded RiPPs to known scaffolds. Previous works have used sequence similarity metrics to determine novelty within a single subfamily of RiPPs [28], however in moving towards a generalist platform, BARLEY integrates the various PTMs associated with RiPPs to generate a more comprehensive novelty index.

Recently, due to advancements in molecular biology techniques, heterologous expression of RiPP BGCs has been increasingly used to isolate natural products following genome identification [15, 30, 5]. Currently, only a limited number of hosts are optimized for this process, including *S. coelicolor* [32], *S. avermitilis* [12], *E. coli* [31] and *S. cerivesiae* [14]. However, this analysis has demonstrated the wealth of novel and divergent RiPPs encoded across almost all bacterial phyla. Developing a strategy for each to be expressed in the same biosynthetic capability as such divergent hosts is not a simple endeavour. First, researchers need to consider the differential codon bias and regulatory elements across bacteria to optimize expression [6], and second, heterologous biosynthesis may drastically alter the original metabolic balance of the hosts, leading to inadequate or incomplete biosynthetic intermediates [8]. Further, many RiPPs act as antibacterial agents and may induce significant toxicity to heterologous hosts [4]. For these reasons, it is improbable that simple identification of encoded RiPPs and heterologous expression technologies will allow for efficient characterization of the total diversity of encoded RiPPs.

Unlike heterologous expression platforms, identifying the precise metabolite produced by a RiPP BGC of interest is challenging due to the large number of metabolites seen in a typical crude bacterial extract. One previous approach has demonstrated that mass spectral fragmentation patterns can be used to link lantipeptide metabolites to specific BGCs in native hosts [18, 17]. Other work has demonstrated the viability of mass spectral fragment residue identification and genomic matching of metabolites, which is demonstrably successful across a variety of peptidic natural products [19]. Issues with fragmentation of more complex RiPPs confounds this approach and lacks an ability to provide resolution among complex peptidic media extracts. While fragmentation patterns are used to guide metabolite selection in this platform, the iSNAP algorithm used here does not make strong assumptions or limitations in the possible residues detected [7]. Instead a library of candidate fragment masses is generated through successive in silico bond cleavages of the predicted scaffold library by PRISM, allowing for a targeted analysis across 21 RiPP classes. Further, we describe CLAMS, a novel subtractive strategy to effectively leverage genomic data from multiple bacterial strains to effectively discard commonly produced metabolites and shed light on strain-unique products. While other peak identification tools have been published [13, 22, 27], CLAMS allows, and was built specifically for, large scale metabolomic analysis, facilitating the dereplication of MS1 ions across thousands of individual experiments. Thus, we can leverage wide scale metabolomic and genomic data across a large diversity of bacterial strains to effectively target specific and novel genomically encoded RiPPs.

Overall, our analyses not only provide clear directions for future efforts in targeted discovery of RiPP scaffolds, we provide here the tools and protocols necessary to expedite this process. To exemplify the success of our platform, we have selected a lasso peptide highly divergent from all currently characterized and encoded RiPPs as directed by BARLEY. Further, we have demonstrated that our mass spectral strategy can drastically reduce the number of candidate metabolites in crude bacterial extracts to identify a single ion corresponding to this lasso peptide. The genomic analysis presented here represents an exploration into the genomically encoded chemical space at an unparalleled fine resolution. Overall, this platform is built as a flexible toolkit to guide the discovery of novel RiPP scaffolds, as demonstrated through the discovery of two lantipeptides biosynthetically encoded via a noncanonical pathway. Although the adjacency of precursor peptides and their modifying enzymes is a highly prevelant pattern within the genomic encoding of RiPPs, this analysis has revealed that it is not universal. It is our hope that the directions and strategies provided here will accelerate and facilitate efficient genome-guided searches for novel RiPP scaffolds.

2.6 Methods

Genomic and chemical datasets

A total of 138 gene clusters stored in FASTA format, mapped to 161 chemical scaffolds, stored in SMILES format, were used to validate the genomic distance analysis and to train BARLEY's genome to chemical novelty index (Supplementary table 3). A total of 640 chemical scaffolds with family level annotation, but without mapped clusters were used to validate BARLEY's chemical distances (Supplementary table 1).

Construction of GRAPE

As described before [3], GRAPE is used to retro-biosynthetically process chemical structures in SMILES format to their corresponding amino acids and a list of chemical reactions detected in doing so. Specifically, for RiPPs, GRAPE was extended to annotate 60 specific PTMs (Supplementary table 2).

BARLEY - Chemical distance

For a comparison between a query chemical scaffold to a subject, a Smith-Waterman alignment is calculated between the query and the subject using an identity matrix, scoring 1 for exact matches, and a gap opening and extension penalty of -2. From this alignment, two scores are denoted for the local alignment, the total number of amino acids in the query that were exactly matched and mismatched to the subject, which are weighted 1 and -1 respectively. From the resulting PTMs identified, three metrics are derived with respect to the query: the number of PTMs observed in both the query and the subject, the number of PTMs observed in either the query or the subject but not in both, and the number of PTMs between query and subject that were marked as similar. A table of similar PTMs are denoted (Supplementary table 6). These three scores (PTM match, PTM mismatch and PTM similar) are weighted 5, -5 and 5 respectively. The sum of these five weighted scores are used to determine a total score. To generate a relative score between 0 and 1, a self-score is generated between the query and itself. The total score is divided by the self score to determine the relative similarity of two RiPP BGCs.

BARLEY - Genomic distance

BARLEY uses the genes identified by PRISM within RiPP gene clusters to build a model of the propeptide core amino acid sequence, and the total PTMs possibly encoded. A total of 112 genes are identified in PRISM as performing PTMs (Supplementary table 6). Each reaction is encoded into BARLEY with three main parameters: a list of genes required to perform the reaction, a list of precursors required, and the resulting PTM. Since certain PTMs are required precursors for other PTMs, such as the dehydration of serine and threonine residues by LanB prior to thioether formation by LanC, all possible reactions are performed iteratively until the total number of PTMs converges upon a maximum, guaranteeing that all possible PTMs encoded by a gene cluster are available. Precursors for a PTM include both amino acids and other PTMs, and are not consumed during the execution of a reaction, thus allowing for all combinatorial possibilities.

For a comparison between a query gene cluster to a subject, all potential PTMs as described above are gathered for both query and subject. The corresponding core amino acid sequence for each BGC are then identified using the predictions by PRISM. In cases where multiple precursor peptides are identified within a RiPP gene cluster, each is considered as a unique entity, and all PTMs are generated for each independently. The query and subject are then scored in the exact same manner as described above for chemical distance.

BARLEY - Novelty index

To generate a novelty index, BARLEY compares encoded RiPPs from PRISM to the database of all RiPP chemical scaffolds processed by GRAPE.As described above, BARLEY is able to parse PRISM and GRAPE data to model RiPPs in the same manner, a sequence of proteinogenic amino acids and a set of PTMs. The same five scores are generated between the PRISM query and GRAPE subject to generate a list of five scores per comparison. These five scores are used as features in a random forest regression model to sort comparison types into three categories.

Here, an outputted score of -1 represents comparisons from different RiPP families, 0 represents a same family comparisons, and 1 represents an exact comparison between a RiPP BGC and its corresponding product. To train and tune this model, the set of 138 gene clusters and their corresponding matched small molecules and family annotations were used. 25% of this data was saved for final testing, while the remaining was used to tune the model using 10-fold cross validation across three parameters: number of base tree estimators, number of randomly selected features for each estimator, and the minimum terminal node size. Using these results (Supplementary figures A1.2-A1.4), a final model was constructed using 400 base estimators, 5 randomly selected features, and a minimum terminal node size of 1.

Genomic analysis

This analysis was performed on the same dataset as published previously by Skinnider et. al, 2016 [25]. A total, 65,421 genomes were run through PRISM, and revealed 24,756 BGCs. The JSON output of PRISM was parsed through BARLEY, and all pairwise scores were stored in an n x n distance matrix where n represents the total number of identified and cleaved precursor peptides identified by PRISM. Since BARLEY scores are directionally dependent, the maximum score of each side is considered for subsequent analysis. Each encoded product was also compared to library of 641 microbial peptide chemical scaffolds using BARLEY. If the highest scoring chemical scaffold to an encoded product was above a cutoff of 0.2, it was determined to be previously characterized.

Diversity Index

To score genera according to their encoded RiPP diversity, all genera with at least twenty sequenced strains were collected and scored according a diversity index. This index is calculated from an iterative random sampling strategy of genomes. The distance described below is 1 - BARLEY similarity score.

A sample here is defined as a selection of genomes. To generate the diversity score for a sample, all unique RiPP gene clusters within a sample were identified, and the sum of the pairwise distances between each was used as the total diversity within a sample. To generate the diversity index, a randomly selected sample of genomes, of size n, were collected and scored for diversity. This was repeated 10 times for each integer increment between 1 and 20. From this, a linear regression was fit to this data, where the slope was used to estimate the average amount of diversity increase per new genome. This analysis was repeated for every genera with more than 20 sequenced members, across all RiPP families.

For estimating the total encoded diversity according to specific RiPP families across all genomes, the mean of pairwise distance scores between all encoded products was generated as it better reflected an intermediate value between modes in the multimodal distributions witnessed (Figure A1.6).

Metabolomic mass spectral analysis

For analytical separation and to record high resolution LC-MS/MS spectra, a SciEX 5600+ TripleTOF mass spectrometer (ABSciEX) with an electrospray ionization (ESI) source was used. The system operates using CID with helium for fragmentation, coupled with an Agilent 1100 series HPLC system using an luna C18 column (150 mm × 2.1 mm, Phenomenex). For preparative separation we used Dionex UltiMate 3000 HPLC system, coupled with a Luna C18 column (250 mm × 15 mm, Phenomenex). For both analytical and preparative separation, the mobile phase consists of gradient mixture of double distilled H2O with 0.1% formic acid and acetonitrile. 0.1% formic acid was used as buffer for both solvents.

Mass spectrometry data was analysed using CLAMs to format MS1 ions as individual entities, mapping to each their relative isotopic distribution, monoisotopic *m/z*, retention time, charge and intensity. Precise values were obtained for each MS1 ion at their maximal intensity. Where observed, MS2 spectra containing relative intensity and *m/z* of each ion were associated with each MS1 ion. To generate a profile of MS1 ions per strain, all detected MS1 ions above a baseline intensity of 10,000 from each experimental analysis across multiple media conditions were compared. Ions within 5 PPM and a 30 second retention time window were considered the same metabolite. Of these overlapping ions across media conditions, a single candidate ion was chosen based on its relative intensity as a representative for subsequent analysis. All MS1 ions in this representative set were then compared to a dataset of 118 blank media extractions to remove any compounds not associated to bacterial metabolism using the same PPM and retention time tolerance described above. The remaining MS1 ions were then compared to all analytical experiments from 463 strains with associated genomic data. BARLEY was used to determine strains with identical encoded RiPPs, any MS1 ions from the candidate strain that were overlapped with non-RiPP carrying strains were eliminated using the same PPM and retention time tolerances described above. Remaining MS1 ions were then evaluated for similarity to a library of PRISM generated scaffolds for a genomically encoded RiPP product of interest. As described before [7], a library of fragment masses were generated *in silico* from a randomly chosen set of 100 predicted scaffolds. The top 1000 most frequently observed fragment masses were then compared to the MS2 spectra for each candidate peak using a 20 PPM tolerance, where a score was generated representing the fraction of MS2 ions that were matched to an in *silico* prediction. For each MS1 ion, all adducts were considered to generate a set of predicted exact masses. MS1 ion mass matches were considered with a tolerance of 20 PPM.

Detection of noncanonical precursor peptides

To identify all putative precursor peptides within the genome of *Flavobacterium ginsengiterrae* JCM 17337, all ORFs were gathered from its genome using Prodigal v2.6.2. These ORFs were filtered according to size (between 40 and 80 AAs in size), and were further filtered according to containing at least two cysteines and two of either serines or threonines in the C-terminal 60% region of the ORF, as these residues are required for lanthionine bridging. For each of these ORFs, a combinatorial set of cleavage predictions was generated at position 22 ± 10 from the C-terminal. These cleaved ORFs were then analysed through the structure prediction engine of PRISM using the LanB and LanC catalysed reactions to generate a total set of predicted structures. *Flavobacterium ginsengiterrae* JCM 17337 was cultured (as described below), and its metabolites were anlaysed using CLAMS to generate a total set of metabolites. For each MS1 ion, all adducts were considered to generate a set of predicted exact masses. The total set of predictions were compared to these MS1 ions to generate mass matches with a tolerance of 20 PPM.

General Experimental Procedures

For analytical separation and to record high resolution LC-MS/MS spectra, a SciEX 5600+ TripleTOF mass spectrometer (ABSciEX) with an electrospray ionization (ESI) source was used. The system is coupled with an Shimadzu Nexera XR HPLC system (Mandel Scientific Company) using an luna C18 column (150 mm × 2.1 mm, Phenomenex). For preparative separation we used Dionex UltiMate 3000 HPLC system, coupled with a Luna C18 column (250 mm × 10 mm, Phenomenex). For both analytical and preparative separation, the mobile phase consists of gradient mixture of double distilled H_2O with 0.1% formic acid and acetonitrile. 0.1% formic acid was used as buffer for both solvents.

To record nuclear magnetic resonance (NMR) spectra of streptapeptin we used Bruker AVIII 700 MHz. Recorded spectra included, 1D (^{1}H and DEPTq), 2D (^{1}H-^{1}H) COSY, TOCSY, ROESY, and NOESY, and 2D (^{1}H-^{13}C) HSQC, and HMBC.

Microbial Strains and Culturing

Streptomyces sp. BTA 0171 was obtained from Pfizer culture collection and maintained on ISP3 agar, or KE or GGYM broth with shaking at 200 RPM, at 28 °C. ISP3 medium consists of 4 g/L yeast extract, 10 g/L malt extract, 4 g/L dextrose, and 15 g/L agar. KE medium consists of 1 g/L glucose, 10 g/L potato dextrin, 5 g/L NZ-amine, 5 g/L yeast extract, 3 g/L beef extract, 0.5 g/L CaCO3, 0.05 g/L MgSO4.7H2O, 2 mL/L filter-sterilized phosphate buffer, added after autoclaving (consists of 91 g/L potassium phosphate monobasic and 95 g/L potassium phosphate dibasic at pH 7). GGYM medium consists of 4 g/L glucose, 4 g/L yeast extract, 10 g/L malt extract, and 5 g/L glycine. *Flavobacterium ginsengiterrae* JCM 17337 was obtained from Japan Collection of Microorganisms nad maintained on Nutrient agar, Nutrient broth or CY broth with shaking at 200 RPM, at 28 °C. CY medium consists of casitone 5 g/L and yeast extract 1 g/L. For antimicrobial assay, *P. aeruginosa* PAO1, *S. aureus* Newman, and *C. albicans* ATCC

90028 were maintained and cultured on cation-adjusted Mueller Hinton, CAMH (4 g/L, Difco^{TM}, USA), tryptic soy broth, TSB (24 g/L, Difco^{TM}, USA), and potato dextrose broth, PDB (30g/L, Sigma-Aldrich, USA), respectively

Production and purification of streptopeptin

Streptomyces sp. BTA 0171 was cultured in small 250 mL Erlenmeyer flasks to produce a seed culture. A 48 h single colony of strain BTA 0171 grown on ISP3 agar was used to inoculate 50 ml GGYM medium. The seed cultures were incubated for 48 h at 28 °C with continuous shaking at 200 rpm. Thereafter, 10 ml of the seed cultures were aseptically transferred to 2.8-L Fernbach flask containing one liter KE broth and incubated for 5 days at 28 °C with continuous shaking at 200 RPM. A total of 18 liters KE media was used for large scale production of streptopeptin. Cells were harvested by centrifugation at 4000 rpm for 20 min at 4 °C then pellets were extracted with one liter of methanol for 4 h. Simultaneously, a resin mixture of 1:1 HP20 and XAD7 were added to the supernatant at ratio of 3:100 W/V, shacked at 100 RPM for 3 h at room temperature, filtered under vacuum, washed with water. The resin was then extracted three times with methanol (1:4 W/V) followed by a final wash in acetone (1:4 W/V). The methanol extract from the mycelial pellets and resin was combined and dried under vacuum using a rotary evaporator followed by nitrogen air for complete drying to yield 10 g of crude extract. The dried extract was then suspended in water and subjected to liquid-liquid partition between 1:1 n-butanol and water. The n-butanol fraction was dried under vacuum using rotary evaporator then nitrogen gas to yield 600 mg dry residue. The residue was suspended in 2 ml methanol and applied on the top of size exclusion column (Sephadex LH20, 1.6 × 80 cm), methanol was used as the mobile phase with flow rate of 1 ml/min. Fractions containing streptopeptin were pooled and dried under nitrogen to yield 120 mg dry mass which was dissolved in 5 ml methanol and subjected to a semi-preparative reversed-phase HPLC with a luna 5 µm C18 column (Phenomenex, 250 mm × 15 mm) using water and acetonitrile with 0.1% formic acid as mobile phase employing a linear gradient of 5% to 80% acetonitrile over 30 min followed by a wash of 100% acetonitrile for 10 min. Fractions were collected at 5 sec interval with streptopeptin being eluted at 18 min. HPLC fractions contain streptopeptin were pooled and dried under nitrogen line to yield 10 mg of pure compound which is then dissolved in methanol-d3 and subjected to NMR analysis. Detailed NMR spectra and assignment of chemical shifts are detailed in supplementary Figures A1.10-A1.18.

Production and purification of ginsebactin and ginsecidin

Flavobacterium ginsengiterrae JCM 17337 was allowed to grow in 250 mL Erlenmeyer flasks to produce seed culture. A 48 h single colony of strain JCM 17337 grown on nutrient agar was used to inoculate 50 ml nutrient medium. The seed cultures were incubated for 48 h at 28 °C with continuous shaking at 200 rpm. Thereafter, 10 ml of the seed cultures were aseptically transferred to 2.8-L Fernbach flask containing one liter CY broth and incubated for 5 days at 28 °C with continuous shaking at 200 RPM. A total of 18 liters CY media was used for large scale production of ginsebactin and ginsecidin. A resin mixture of 1:1 HP20 and XAD7 were added to the fermented broth at ratio of 3:100 W/V, shacked at 100 RPM for 3 h at room temperature, filtered under vacuum, washed with water. The resin was then extracted three times with methanol (1:4 W/V) followed by a final wash in acetone (1:4 W/V). The mixture of methanol and acetone was dried under vacuum to yield 7.9 g of crude extract, which was dissolved in water (500 mL) and partitioned with EtOAc (3 × 500 mL) to yield an EtOAc fraction of 4.2 g. The EtOAc fraction (4.2 g) was then subjected to a flash column chromatography (Teledyne) with a 30 g SNAP Ultra C18 column (Biotage) using water and acetonitrile as mobile phase at 35 ml/min employing a linear gradient of 10% to 100% acetonitrile over 18 mins followed by 5 mins of 100% acetonitrile wash. Fractions containing ginsebactin and ginsecidin were pooled to result a dry mass of 230 and 160 mg, respectively. ginsebactin containing fraction (230 mg) was then subjected to a semi-preparative reverse phase HPLC with a Luna 5 µm C18 column (Phenomenex, 250 mm × 10 mm) using water and acetonitrile with 0.1% formic acid as mobile phase employing a linear gradient of 20% to 45% acetonitrile over 12.5 min followed by 5 mins isocratic run with 45% acetonitrile then a wash of 100% acetonitrile for 10 min. Ginsebactin is eluted at 17.6 min. HPLC fractions contain ginsebactin were pooled and dried under nitrogen to yield 1.4 mg of pure compound. Ginsecidin containing fraction (160 mg) was subjected to reverse phase HPLC using water and acetonitrile with 0.1% formic acid as mobile phase employing a linear gradient of 40% to 60% acetonitrile over 17.5 min followed a wash of 100% acetonitrile for 10 min. Ginsecidin was eluted at 19.9 min with a total of 2.1 mg. Both ginsebactin and ginsecidin were then dissolved in DMSO-d6 and subjected to NMR analysis. Detailed NMR spectra and assignment of chemical shifts are detailed in supplementary Figures A1.10-A1.18.

Screening of antimicrobial activity

To test if streptopeptin, ginsebactin and ginsecidin possess antimicrobial activity, a broth microdilution assay was conducted on three indicator strains, *P. aeruginosa* PAO1, *S. aureus* Newman, and *C. albicans* ATCC 90028 grown in CAMH, TSB, PDB, media respectively. The indicator strains were grown in at 37 °C and shacked at 200 rpm for 24 h except for *C. albicans* was grown without shaking. The actively grown cultures were diluted to 1:10,000 using the same media used for the growth of each strain. To 96-well microtiter plate, 196 µL inoculated medium was added and mixed with 4 µl of either streptopeptin, ginsebactin or ginsecidin (20-5 µg/mL final concentration were used for a preliminary screen). Polymyxin, erythromycin, and amphotericin were used as positive control for *P. aeruginosa* PAO1, *S. aureus* Newman, and *C. albicans* ATCC 90028, respectively, at final concentration of 2 µg/mL. Negative control consisted of inoculated media and DMSO as solvent used to dissolve streptopeptin. Blank consisted of non-inoculated and DMSO. The entire experiment was repeated in triplicate with 3 independent replicates each time.

Bibliography

[1] Arnison, P.G. et al. Ribosomally synthesized and post-translationally modified peptide natural products: Overview and recommendations for a universal nomenclature. *Natural product reports*, 30(1):108–160, January 2013. ISSN 0265-0568. doi: 10.1039/c2np20085f.

[2] Blin, K. et al. antiSMASH 4.0—improvements in chemistry prediction and gene cluster boundary identification. *Nucleic Acids Research*, 45(Web Server issue):W36–W41, July 2017. ISSN 0305-1048. doi: 10.1093/nar/gkx319.

[3] Dejong, C.A. et al. Polyketide and nonribosomal peptide retro-biosynthesis and global gene cluster matching. *Nature Chemical Biology*, 12(12):1007–1014, December 2016. ISSN 1552-4469. doi: 10.1038/nchembio.2188.

[4] Flinspach, K. et al. Heterologous Expression of the Thiopeptide Antibiotic GE2270 from Planobispora rosea ATCC 53733 in Streptomyces coelicolor Requires Deletion of Ribosomal Genes from the Expression Construct. *PLoS ONE*, 9(3), March 2014. ISSN 1932-6203. doi: 10.1371/journal.pone.0090499.

[5] Frattaruolo, L. et al. A Genomics-Based Approach Identifies a Thioviridamide-Like Compound with Selective Anticancer Activity. *ACS Chemical Biology*, 12(11):2815–2822, November 2017. ISSN 1554-8929. doi: 10.1021/acschembio.7b00677.

[6] Gustafsson, C., Govindarajan, S. and Minshull, J. Codon bias and heterologous protein expression. *Trends in Biotechnology*, 22(7):346–353, July 2004. ISSN 0167-7799. doi: 10.1016/j.tibtech.2004.04.006.

[7] Ibrahim, A. et al. Dereplicating nonribosomal peptides using an informatic search algorithm for natural products (iSNAP) discovery. *Proceedings of the National Academy of Sciences of*

the United States of America, 109(47):19196–19201, November 2012. ISSN 0027-8424. doi: 10.1073/pnas.1206376109.

[8] Ikeda, H., Kazuo, S.y. and Omura, S. Genome mining of the Streptomyces avermitilis genome and development of genome-minimized hosts for heterologous expression of biosynthetic gene clusters. *Journal of Industrial Microbiology & Biotechnology*, 41(2):233–250, February 2014. ISSN 1476-5535. doi: 10.1007/s10295-013-1327-x.

[9] Jasial, S. et al. Activity-relevant similarity values for fingerprints and implications for similarity searching. *F1000Research*, 5, April 2016. ISSN 2046-1402. doi: 10.12688/f1000research.8357.2.

[10] Johnston, C.W. et al. Assembly and clustering of natural antibiotics guides target identification. *Nature Chemical Biology*, 12(4):233–239, April 2016. ISSN 1552-4469. doi: 10.1038/nchembio.2018.

[11] Kimura, K. et al. Propeptin, a new inhibitor of prolyl endopeptidase produced by Microbispora. I. Fermentation, isolation and biological properties. *The Journal of Antibiotics*, 50 (5):373–378, May 1997. ISSN 0021-8820.

[12] Komatsu, M. et al. Genome-minimized Streptomyces host for the heterologous expression of secondary metabolism. *Proceedings of the National Academy of Sciences of the United States of America*, 107(6):2646–2651, February 2010. ISSN 1091-6490. doi: 10.1073/pnas.0914833107.

[13] Li, H. et al. Accurate identification of mass peaks for tandem mass spectra using MCMC model. *Tsinghua Science and Technology*, 20(5):453–459, October 2015. doi: 10.1109/TST.2015.7297744.

[14] Lian, J. et al. Design and construction of acetyl-CoA overproducing Saccharomyces cerevisiae strains. *Metabolic Engineering*, 24:139–149, July 2014. ISSN 1096-7184. doi: 10.1016/j.ymben.2014.05.010.

[15] Luo, Y., Enghiad, B. and Zhao, H. New tools for reconstruction and heterologous expression of natural product biosynthetic gene clusters. *Natural Product Reports*, 33(2):174–182, February 2016. ISSN 1460-4752. doi: 10.1039/c5np00085h.

[16] Malone, C.L., Boles, B.R. and Horswill, A.R. Biosynthesis of Staphylococcus aureus Autoinducing Peptides by Using the Synechocystis DnaB Mini-Intein. *Applied and Environmental Microbiology*, 73(19):6036–6044, January 2007. ISSN 0099-2240, 1098-5336. doi: 10.1128/AEM.00912-07.

[17] Medema, M.H. et al. Pep2Path: Automated Mass Spectrometry-Guided Genome Mining of Peptidic Natural Products. *PLOS Computational Biology*, 10(9):e1003822, 04-Sep-2014. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1003822.

[18] Mohimani, H. et al. Automated genome mining of ribosomal peptide natural products. *ACS chemical biology*, 9(7):1545–1551, July 2014. ISSN 1554-8937. doi: 10.1021/cb500199h.

[19] Nguyen, D.D. et al. Indexing the Pseudomonas specialized metabolome enabled the discovery of poaeamide B and the bananamides. *Nature Microbiology*, 2:16197, 10 31, 2016. ISSN 2058-5276. doi: 10.1038/nmicrobiol.2016.197.

[20] Okada, M. et al. Structure of the Bacillus subtilis quorum-sensing peptide pheromone ComX. *Nature Chemical Biology*, 1(1):23–24, June 2005. ISSN 1552-4469. doi: 10.1038/nchembio709.

[21] Potterat, O. et al. BI-32169, a Bicyclic 19-Peptide with Strong Glucagon Receptor Antagonist Activity from Streptomyces sp. *Journal of Natural Products*, 67(9):1528–1531, September 2004. ISSN 0163-3864. doi: 10.1021/np040093o.

[22] Renard, B.Y. et al. NITPICK: Peak identification for mass spectrometry data. *BMC Bioinformatics*, 9:355, August 2008. ISSN 1471-2105. doi: 10.1186/1471-2105-9-355.

[23] Rogers, D. and Hahn, M. Extended-Connectivity Fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742–754, May 2010. ISSN 1549-9596. doi: 10.1021/ci100050t.

[24] Shen, B. A New Golden Age of Natural Products Drug Discovery. *Cell*, 163(6):1297–1300, December 2015. ISSN 0092-8674. doi: 10.1016/j.cell.2015.11.031.

[25] Skinnider, M.A. et al. Genomic charting of ribosomally synthesized natural product chemical space facilitates targeted mining. *Proceedings of the National Academy of Sciences*, 113(42): E6343–E6351, October 2016. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1609014113.

[26] Skinnider, M.A. et al. Comparative analysis of chemical similarity methods for modular natural products with a hypothetical structure enumeration algorithm. *Journal of Cheminformatics*, 9(1):46, December 2017. ISSN 1758-2946. doi: 10.1186/s13321-017-0234-y.

[27] Smith, C.A. et al. XCMS: Processing Mass Spectrometry Data for Metabolite Profiling Using Nonlinear Peak Alignment, Matching, and Identification. *Analytical Chemistry*, 78 (3):779–787, February 2006. ISSN 0003-2700. doi: 10.1021/ac051437y.

[28] Tietz, J.I. et al. A new genome-mining tool redefines the lasso peptide biosynthetic landscape. *Nature Chemical Biology*, 13(5):470–478, May 2017. ISSN 1552-4469. doi: 10.1038/nchembio.2319.

[29] Weber, W. et al. Anantin–a peptide antagonist of the atrial natriuretic factor (ANF). I. Producing organism, fermentation, isolation and biological activity. *The Journal of Antibiotics*, 44(2):164–171, February 1991. ISSN 0021-8820.

[30] Wenzel, S.C. and Müller, R. Recent developments towards the heterologous expression of complex bacterial natural product biosynthetic pathways. *Current Opinion in Biotechnology*, 16(6):594–606, December 2005. ISSN 0958-1669. doi: 10.1016/j.copbio.2005.10.001.

[31] Yang, Y. et al. Regulating malonyl-CoA metabolism via synthetic antisense RNAs for enhanced biosynthesis of natural products. *Metabolic Engineering*, 29:217–226, May 2015. ISSN 1096-7184. doi: 10.1016/j.ymben.2015.03.018.

[32] Zhou, M. et al. Sequential deletion of all the polyketide synthase and nonribosomal peptide synthetase biosynthetic gene clusters and a 900-kb subtelomeric sequence of the linear chromosome of Streptomyces coelicolor. *FEMS microbiology letters*, 333(2):169–179, August 2012. ISSN 1574-6968. doi: 10.1111/j.1574-6968.2012.02609.x.

Chapter 3

In situ systemic algorithm reveals functional peptides exclusively encoded within the human microbiome

3.1 Preface

The human body is a host to a wide assortment of bacterial residents. While we have extensively studied those that are deleterious to human health, relatively little is understood about those bacteria which may play a protective role. Due to advances in metagenomic sequencing, several studies have correlated the presence and abundance of various bacteria to numerous ailments. Yet still, much is still unknown regarding the various mechanistic interactions in which these bacteria engage with the human body and among themselves. Previous works had demonstrated that antimicrobial peptides were abundantly encoded within the human microbiome. However, apart from those studied extensively *in vitro*, not much is understood about the role these metabolites play within the polymicrobial human environment. To answer this question, we developed an informatics platform, AMPLIFY, which firstly identifies those peptide families uniquely present within the human microbiome. Further, AMPLIFY aims to characterize the potential antimicrobial activity of these products through correlating their expression to the constant flux in microbial population dynamics seen within the GIT. As an application of this pipeline, we sought to identify peptides relevant against the infectious pathogen *Clostridium difficile*. In this endeavour, AMPLIFY identified several candidates, one of which was produced by *Scardovia wiggisae*, a bacterium previously implicated in early childhood dental caries. This peptide, termed scardovicin, was synthesized to evaluate its biological activity *in vitro*. We show here its potent and multimodal activity in protecting against *C. difficile*. Informatic strategies such as these are essential in understanding the complexities and multitudes of chemical mediators underlying human biology. The software that I have built to tackle this problem specifically uses large scale surveillance data of the varying microbial populations and peptide expressions in the GIT. While this data is publicly available for research, no such applications have been developed before to specifically link bacterial metabolites to putative functions.

The following chapter is formatted as a manuscript which is in preparation for journal submission. I developed AMPLIFY to analyse genomic, metagenomic and metatranscriptomic data. I curated data, built HMMs, contributed to study design and wrote the manuscript. Walaa Mousa developed assays to test the immunomodulatory effects of scardovicin and the antibacterial activity of scardovicin against *C. difficile* in the context of cell growth, spore germination and spore formation. Walaa Mousa also curated data, built HMMs, contributed to study design and wrote the manuscript. Bilal Athar, Keshav Dial and Mathusan Gunabalasingam curated data and built HMMs. Waliul Khan helped in designing the cell line assay and Huaqing Wang helped in performing the assay. Professor Nathan Magarvey contributed to study design and wrote the manuscript.

3.2 Abstract

The human microbiome is a complex ecosystem with a diversity of species, genes and metabolism. Systematic methods to define the uniquely encoded chemistry of the microbiome has been lacking. Defining such metabolites is timely as links between the human microbiota and human biology

are increasingly made and elaborating the mechanistic underpinnings would further both the basic and applied sciences. To date, we still lack systematic platforms to integrate genomics, metagenomics and transcriptomics data to expose unique microbiome chemistry as it relates to human health and disease. Herein, we present AMPLIFY, A tool to enrich for Microbial Peptides Linked to In-situ FunctionaliTy. AMPLIFY defines novel microbiome exclusive antimicrobial peptides *in silico* based on the fluctuation in their expression profile in accordance with microbial dynamic shifts. AMPLIFY is an integration of multi-omics data to compare *in situ* expression of genetically encoded peptides and subsequent definition of novel entities correlated with biological outcomes. As a proof of concept, we use AMPLIFY to find a human microbiota-unique molecule which, according to analytics, is inferred as an agent antagonistic with *Clostridium difficile* growth. We report the discovery of scardovicin, a peptide encoded within *Scardovia wiggisae* with nano-molar potency toward *C. difficile* both as an antibiotic, anti-sporulation agent, and as a host immune modulator dampening the host response to *C. difficile* toxins. We see scardovocin as a system construct where such modes of action have not been observed previously from a synthetic or natural product. This example highlights the use of this technology and its capabilities to define important evolved human microbiome exclusive mediators with beneficial effects.

3.3 Introduction

The human microbiota forms a dynamic consortium with a poorly understood interactive network thought to be mediated by evolved secreted molecules. Our knowledge of the chemistry that mediates the interaction between microbes and with the host is sparse yet at the core gaining knowledge through advancement of new tools may facilitate bridging of such knowledge gaps [9, 33]. At current, strong correlative links are noted between microbiota shifts with the onset or progression of human diseases such as diabetes [45], ulcerative colitis [31], arthritis [39], Alzheimer's disease [50], depression [20], and autism [27].

Several examples of microbiome products with immunomodulatory activity are noted such as polysaccharide A [30], pyro-dipeptides [59], and indole-3-aldehyde [56]. Recently identified, N-acyl amides of the human microbiome interact with human cellular receptors and are suggested to mimic native ligands [4]. The microbiome product, 4-ethyl phenyl sulfate mediates autism spectrum disorder in mice, an effect that is neutralized by *Bacteriodes fragiles*, an effect claimed to be related to elevated level of N-acetylserine. Other chemical mediators are thoughts to cause cancer such as putative colibactin of *E. coli* Nissle 1917 [49, 40]. Specialized enzymes encoded in the microbiome are recently blamed to mediate inactivation of chemotherapeutics. Example is a long form of cytidine deaminase produced by *Mycoplasma* bacteria which has been proven to degrade anticancer drug, gemcitabine resulting in drug resistance [14]. An interesting class of functional molecules produced by the microbiota are antibiotics. In all microbial ecosystems production of antibiotics is seen as a weapon to enhance ecological fitness of the producer [8]. To date, dozens of antimicrobial molecules are discovered from the human microbiota and their bioactivity spectrum against body site specific pathogens emphasizes the co-evolution hypothesis. Examples of these molecules include Lugdunin [58], lactocillin [10], humimycins [3], and others as reviewed [33]. Among these antimicrobials, unmodified peptides, known as bacteriocins, are believed to exert lethal and selective killing by variety of mechanisms [26, 38]. Known bacteriocins from the human microbiota include gassericin A, pediocins, leucocins with activity against some human pathogen such as *Listeria monocytogenesis* [33].

Discovery of microbiome-exclusive and previously unknown chemical mediators in among the human microbiota, either through traditional bioactivity screening or modern genome mining is extremely challenging, shaded by uncertainty, and mostly fails to establish a link between *in situ* microbial fluctuations and the corresponding shift in metabolites expression. Here, we attempt to develop a platform to combine multi-omics tools to annotate functional unmodified peptides *in silico*. We present, AMPLIFY, A tool to enrich for Microbial Peptides Linked to *In-situ* FunctionaliTy. AMPLIFY integrates genomics, metagenomics and metatranscriptomic data into a sole platform designed to enrich for accurate functional annotations of these peptides using the relative gene expressions of these peptides in correlation with the varying population abundances of diverse microbial species *in situ*.

AMPLIFY identified 189 peptides unique to the microbiome that exhibit over two thousand antagonistic relationships towards 275 microbes including numerous human pathogens. Given the challenges associated with targeting toxigenic spore-forming pathogens, we selected one of the leading cause of nosocomial infections worldwide, *Clostridium difficile*, as a model to validate our platform [25, 6]. We report for the first time the discovery of novel peptide, we named it scardovicin, encoded in the genome of *Scardovia wiggisae* with unprecedented actions against *C. difficile* ameliorating the growth and health consequences corresponding with its infection. Scardovicin combines potent direct inhibitory activity on the active growing cells of *C. difficile* with a powerful inhibition of sporulation, spore germination, and strong anti-inflammatory effect.

We see AMPLIFY as an enabling tool to discover microbiome exclusive functional peptides and leverage our knowledge of microbiome chemical mediators as they relate to human diseases.

The AMPLIFY workflow

In the hunt for new chemical mediators encoded in the human microbiome, we developed AMPLIFY, which integrates genomic, metagenomic and transcriptomic data to predict a functional annotation for microbial peptides (Figure 3.1).

The input to AMPLIFY is DNA-predicted sequences of peptides, here we used 202 peptides that we identified as unknown and unique to the human microbiome, as detailed below. AMPLIFY measures the expression of each peptide in transcriptomic data compiled from more than 700 fecal samples available from the integrated human microbiome project, iHMP [18, 17]. AMPLIFY then generates a matrix to correlate the fluctuation in peptide expression to shifts in microbial populations detected in each metagenomic sample. Through this process, AMPLIFY annotates a putative *in situ* antimicrobial function to each peptide with significant negative correlation to the abundance of a target pathogen. The output from AMPLIFY is candidate antimicrobial peptides with a validated accuracy of 71%.

Identification of encoded peptides in the human microbiome

The first step in AMPLIFY pipeline starts with identification of sequences of DNA-encoded peptides in a given set of bacterial genomes sharing some degree of homology to characterized antimicrobial peptides. In this analysis we used a sample consisting of 9,953 bacterial genomes

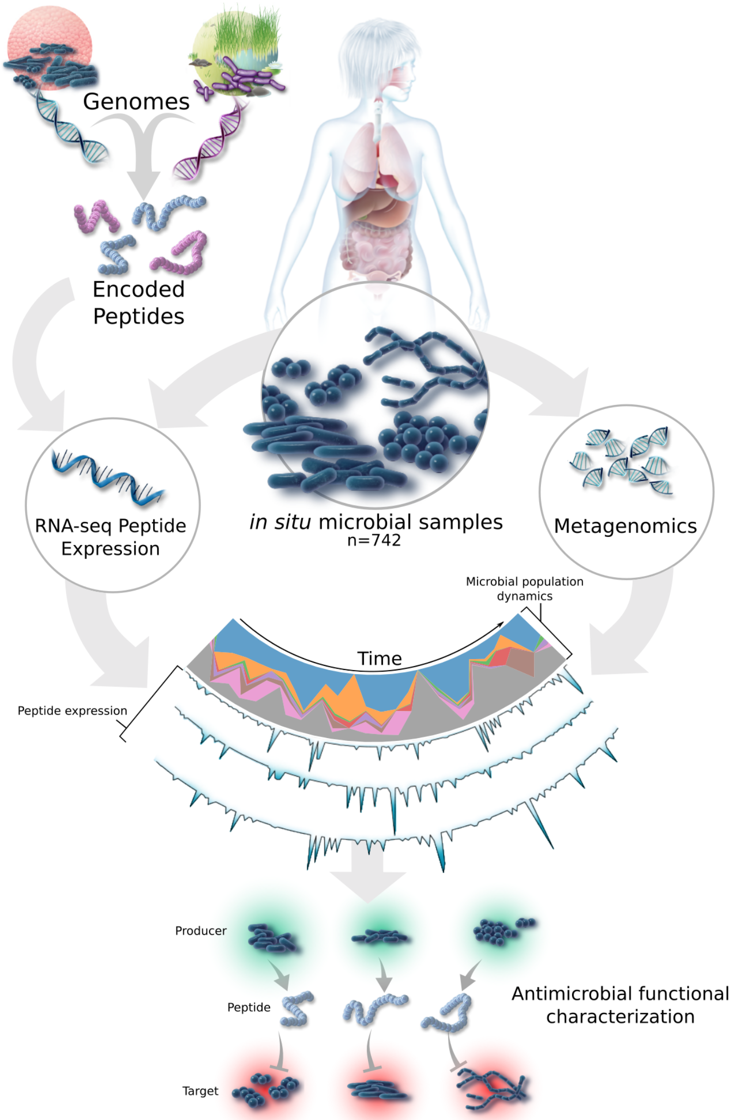
Genomes
Encoded Peptides
RNA-seq Peptide Expression
in situ microbial samples
n=742
Metagenomics
Microbial population dynamics
Time
Peptide expression
Producer
Peptide
Target
Antimicrobial functional characterization

FIGURE 3.1: **Protocol for *in situ* functional annotation of microbial peptides using AMPLIFY.** The pipeline starts with identification of microbial peptides encoded in the analyzed genomes using an updated version of PRISM. A new developed algorithm, BARLEY is then employed to assign sequence similarities between identified peptide and dereplicate known entities. The committed step in the pipeline is to measure the expression of each peptide as it correlates to the shift in microbial population in each metagenomics sample. Using these antagonisms correlation, we can annotate targeted antimicrobial function to analyzed peptides with 71% confidence.

isolated from both human and environmental sources. These genomes were processed through PRISM v4.2.0, which has been implemented with 123 new pHMMs for the identification of putative antimicrobial peptides. We constructed these pHMMs through the curation, grouping and subsequent alignment of the 156 known antimicrobial peptides. We identified a total of 2815 peptides from the entire genome set used in this study (Supplementary table 3). From the 2274 microbiome genomes analyzed, a total of 955 peptides were identified encoded within 514 genomes, isolated from microbes distributed over 42 genera (Figure 2). This represents 22.6% of the analyzed microbiome genomes encoding for at least a single of these peptides (Figure 3.2). *Enterococcus* and *Streptococcus* represent the highest peptide containing genomes with 306, and 255 peptides detected over 72.1%, and 52.11% of the tested genomes for these genera, respectively. While we did not detect any of these peptides in 162 genera including *Helicobacter* and *Fusobacterium*. Some individual genomes were especially rich in putative antimicrobial peptides such as *Streptococcus oralis* SK255 isolated from a blood sample with 8 encoded peptides, *Enterococcus* sp. HMSC063H10, *E. coli* subsp. MS 16-3, *Lactobacillus plantarum*, with 8, 5 and 6 encoded peptides in each respectively.

Apparently most enriched genomes were sequenced from GIT microbes, with 499 identified peptides. However, 65% of the analyzed genomes are sequenced from microbes isolated from the GIT, after normalizing this effect, the skin shows the highest enrichment of peptides (186 in 112 genomes). To dereplicate known peptides from the identified sequences, we used our recently developed alignment tool, BARLEY, Basic Alignment of Ribosomally Encoded products locally. BARLEY aligns predicated sequences to a curated database of known antimicrobial peptides with a sequence similarity threshold of 0.85 (Supplementary table 1). We matched 21% of identified peptides to previously characterized antimicrobial peptides. Interestingly all identified peptides within the genera of *Cutibacteria* and *Staphylococcus*, 71 and 30 peptides respectively, are unknown.

Revealing peptide families unique to the human microbiome

To identify peptide families exclusive to the human microbiome, a set of 7,679 bacterial genomes, represents 30 phyla and 960 genera, were curated from NCBI with isolation sources broadly

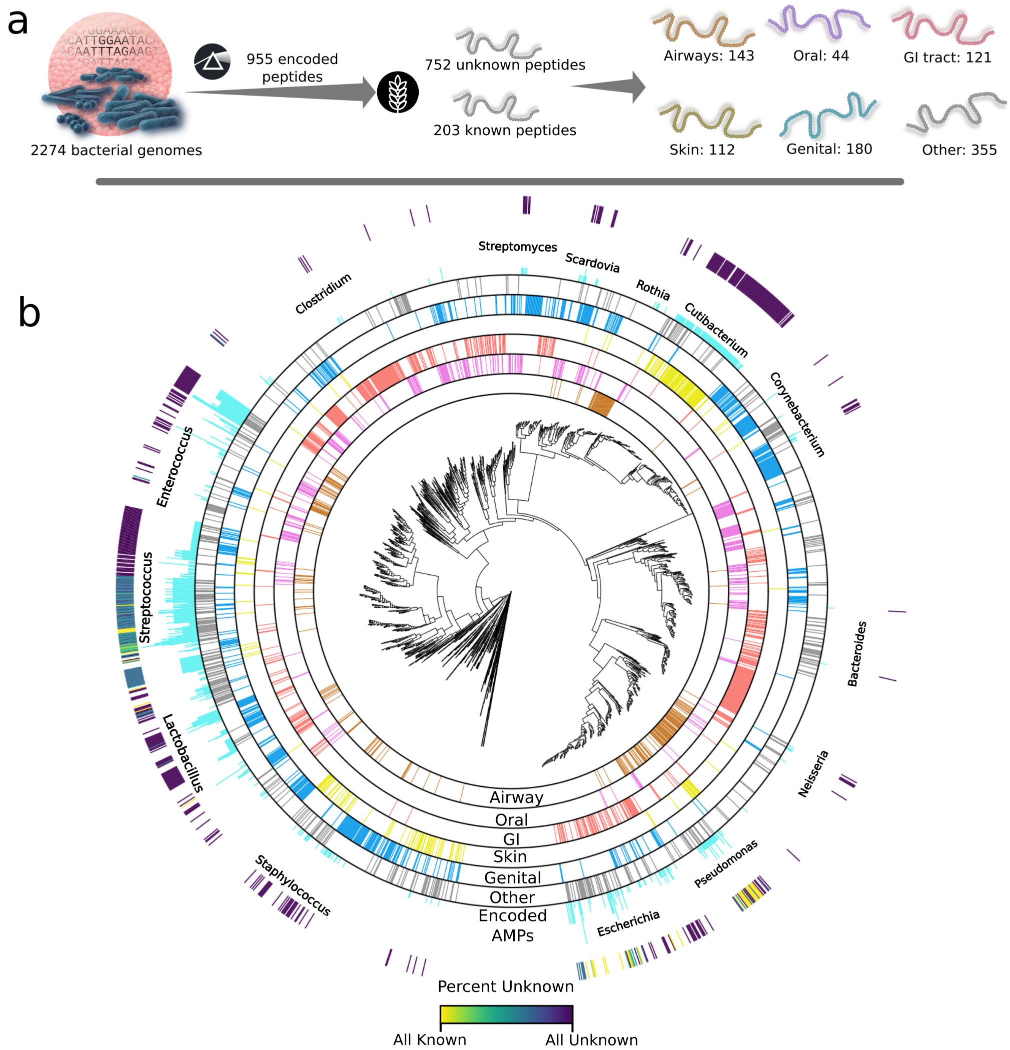

FIGURE 3.2: **Distribution of encoded candidate antimicrobial peptides among the representative 2274 human associated microbes.** (a) Schematic overview of antimicrobial peptide detection engine using pHMMs within PRISM and dereplication using a database of characterized antimicrobial peptides using BARLEY. (b) Phylogenetic tree of all microbial genomes generated via PhyloPhlAn [41]. Outer rings in brown, magenta red, yellow, blue and grey represent the isolation site of each individual microbe analysed where "Other" is comprised of bacterial isolates from wounds, blood, nose, bone, eye, spinal cord, brain, ear, heard abdomen, limb and liver samples. In cyan, the total number of peptides encoded by each strain is shown where the max value represents six from *Streptococcus oralis* SK255. The furthest ring, represents the percent of encoded bacteriocins per strain that were identified as previously characterized. Selected genera abundant in encoded peptides are labelled.

annotated as environmental (e.g. soil, plant marine, etc.). Having analyzed this taxonomically diverse dataset, we identified 1,860 putative antimicrobial peptides. Using a similarity threshold of 0.55, we generated a pairwise similarity score between all peptides which resulted in a total of 380 families (Figure 3.3-a). Of these, 33 families were shared between the human microbiome and the environment, while a total of 124 were exclusive to the human microbiome (Figure 3.3b-c). Interestingly, one of the largest families observed is an overlap between environmental and the human microbiota (Figure 3.3b). This family consists of 274 peptides from 67 genera is distantly related to lincocin M18, originally isolated from *Brevibacterium linens* [46]. Among the 124 microbiome-exclusive families (Figure 3.3-c), only 17 shared sequence similarity to at least one known peptide with experimentally validated antimicrobial activity. Surprisingly, 71% of the families exclusive to the human microbiome consists of only one member with distinct sequence. (Figure 3.3-c). Of the remaining larger peptide families, 89% percent were noted to be shared between members of the same genera with only few exceptions. Interestingly, 80% of peptides exclusive to the human microbiome are exclusive to one body site. While exclusive microbiome peptides that are shared between different body sites are mostly encoded within microbes of the same genera. Of the peptides relevant in multiple body sites, a notable example is a thermophilin A derivative encoded within the genomes of several *Streptococci* isolated from blood, airway, oral and GIT samples (Figure 3.3-c).

Functional annotation of new unmodified peptides unique to the human microbiome

Having identified 202 peptides exclusive to the human microbiome, the next step in the pipeline was to algorithmically assign them a putative antimicrobial function. AMPLIFY employs a novel protocol to infer the functional spectrum of peptides based on metagenomics and transcriptomic data. Using the inflammatory bowel disease multi'omics databases (IBDMDB) which consists of 742 fecal samples collected at varying time points from 109 patients denoted as either healthy or diseased with ulcerative colitis or Chron's disease [17]. The dataset collected and released by the iHMP project is currently the only resource where this analysis is amenable, both due to the

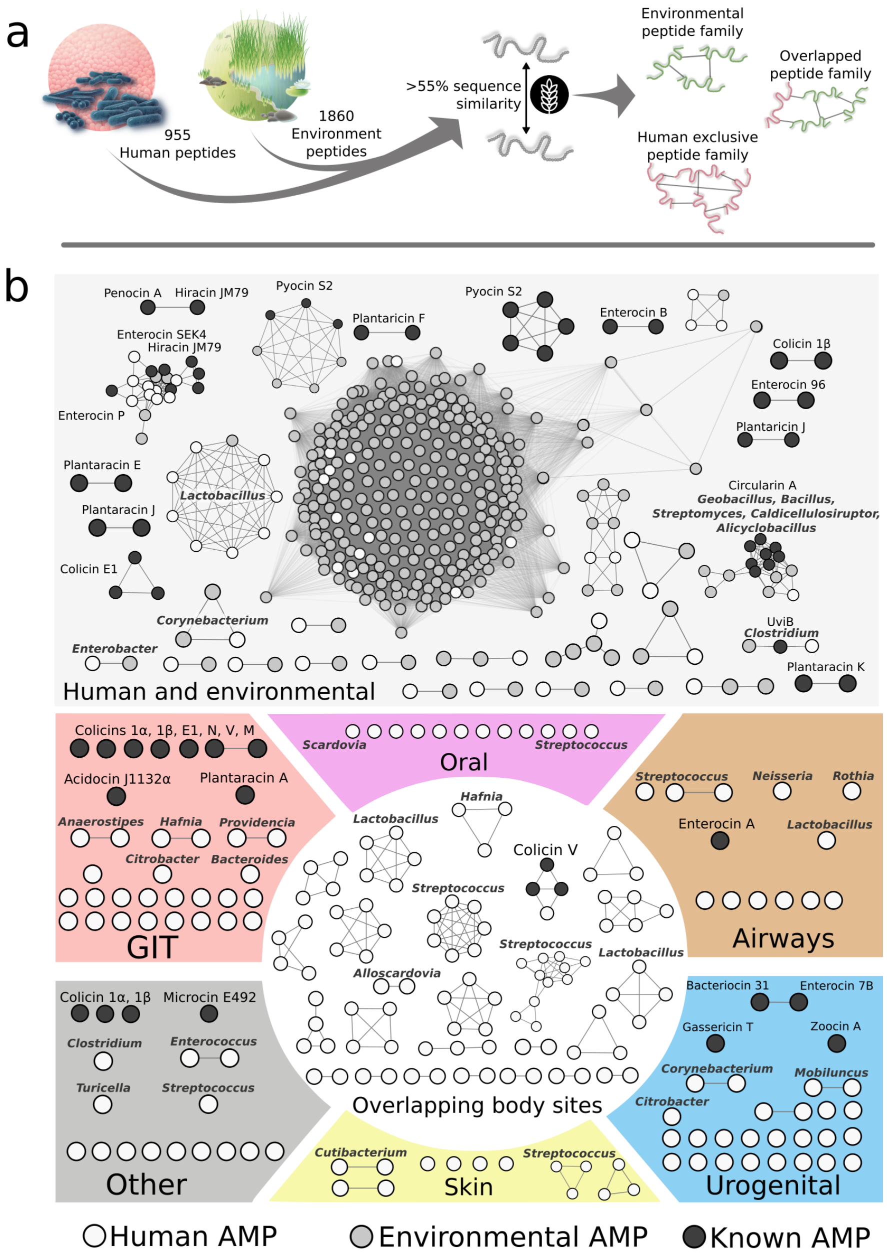

FIGURE 3.3: **Families of peptides observed in human microbiota, their relative distribution among environmental and human body sites.** (a) All peptides observed in environmental and human microbiomes were assigned into families as depicted. (b) Peptide families found exclusive to human associated microbiota are shown within specific body sites, or shown in the middle as overlapping between multiple body sites. Peptide families found in both human and environmental genomes are shown in top grey section. Circles represent unique peptide sequences, while edges between circles represent a shared sequence similarity. Peptides sharing a high sequence similarity to characterized antimicrobial peptides are shaded black and labelled, while select uncharacterized peptide groups are labelled with the producer genera in italics.

large number of samples collected and the varieties of multi-omic technologies used to experimentally investigate these samples. In particular, the availability of metagenomic sequence data from each of these samples allowed for accurate species level characterization of relative abundance of microbes using MetaPhlAn2 [44], and bypassed the need to use 16s rRNA sequencing platform with relatively low resolution at species and strain levels [23]. Further, the microbial metatranscriptomic sequencing data available for each of these samples allowed us to query the relative expression of any gene of interest.

Across all samples, metagenomic analysis revealed a total of 499 microbial species. The most dominant species in this analysis was *Faecalibacterium prausnitzii* with a mean abundance of 12%, but this was highly dynamic as seen by a standard deviation of 12.4%. Metatranscriptome sequence data corresponding to the totality of bacterial RNA expression in these samples were used to identify the relative abundance of microbiome exclusive peptides. For the 202, there were 209 unique RNA segments due to codon variance. From these, a total of 197 candidate peptides were observed in at least 10% of samples and were chosen for further functional annotation.

Giving the challenges associated with this approach mainly due to the strong amounts of noise and variation in population dynamics, we performed two sets of validation. First, we applied the pipeline to four constitutively expressed housekeeping genes. The genes were curated as PFAM models (Figure 3.4-a). These genes were then identified in all microbiome genomes used in this study, mapped to metatranscriptomic data, and correlated to the microbial abundance values. In total, 628 genes from 226 species were identified in at least 10% of samples. As a control, these same genes were compared to randomly assigned species in metagenomic data. Data supports that the relative expression of these genes is strongly associated with the abundance of the bacteria which not only validates AMPLIFY pipeline, but also bolsters the quality of this dataset (Figure 3.4-a).

As a next step of validation, we conducted this analysis on 16 peptides, identified in some microbiome genomes in this study, with exact sequence similarity to known antimicrobial peptides. For each of the validated targets of these peptides, 10 cases were found to be amenable to testing according to the selection criteria described earlier. A total of 52 antagonistic relationships were identified for this training set. We randomly selected some of these relationships as a comparison representing the inherent noisiness of this dataset (Figure 3.4-b). Antagonistic relationships between known antimicrobial peptides and their target species do seem to present a slightly, yet significant shifted spearman correlation when compared to a random selection of relationships, with an increasing likelihood of true antagonism present with more negative correlations (Figure 3.4-b). The student's *t*-test reveals this shift as significant (p=0.007). Assuming a 50% prior likelihood of antagonism and using a cutoff of -0.05, a positive predictive value of 71% can be obtained while predicting on 28% of cases. This cutoff represents the bottom 20.7% of this distribution and can be used as a guide for subsequent analysis. Given these layers of validation, we confidently demonstrate that metatranscriptomic associations to microbial population dynamics is a valid strategy to computationally annotate antagonism relationships between a given peptide and specific microbial target.

We identified 275 target bacterial species present in the metagenomics data samples, present at least in 1% of the samples. This represents a total of 54175 potential tests. Considering the top 5% of negative correlations (a maximum spearman correlation of -0.086), we identified 2708 antagonistic relationships between microbiome exclusive peptides and various target species (Supplementary table 8). Using the results of the validation experiment discussed above, alongside a more stringent cutoff, we can be at least 71% confident that each of these relationships are in fact true. AMPLIFY reveals at least one antagonistic relationship for 95% of the peptides tested with a total of 266 target species across 90 genera (Supplementary table 8).

As a proof of concept, we thought to pursue a targeted search for new unique microbiome exclusive peptide with antagonism potential against a defined target. We choose *C. difficile* as a target pathogen, given the challenges associated with discovery of molecules with effective anti-*C. difficile* activity. *C. difficile* is a spore-forming and toxigenic pathogen with complex disease etiology leads to sever complications that might include pseudomembranous colitis, colon rupture, sepsis, and even death [24, 42].

A total of 12 candidate peptides were identified with confident antagonizing relationships to *C. difficile* (Figure 4c). The candidate peptides are encoded in genomes from the genera *Enterococcus*, *Streptococcus*, *Bacteroides*, *Lactobacillus*, *Hafnia*, and *Scardovia*. Of these, we selected a candidate peptide, we named it scardovicin, encoded within two strains of *Scardovia wiggisae*, an inhabitant of the oral cavity, for further synthesis and i activity profiling. Scardovicin, was primarily chosen due its relatively small size that allows for total synthesis. The primary sequence of scardovicin consists of 53 amino acids, MGAFFRLLSILARYGARAVQWAWAHRGTVLRWIGAGQAIDWVIKQIKRLLGIR

Bioactivity profile of scardovicin

To test the antimicrobial activity of scardovicin, we conducted a broth microdilution assay. We selected a virulent *C. difficile* strain from ribotype 027 group [54]. Scardovicin was proven

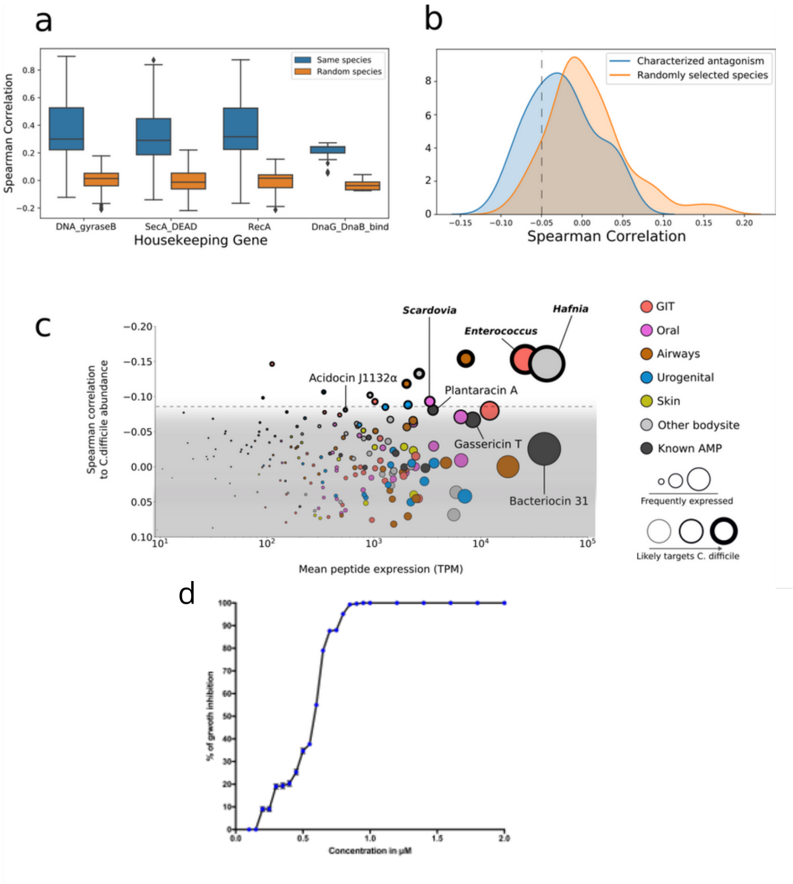

FIGURE 3.4: **Validating the correlations of metatranscriptomic data to metagenomic data within this dataset and the identification of putative anti-clostridial peptides.** Full caption on following page.

Figure 3.4: (Previous page.) (a) Validation of correlations between metatranscriptomic data and metagenomic data using housekeeping genes. Boxes are drawn between the first and third interquartile range (IQR) with whiskers extending 1.5 IQR and outliers beyond this are plotted as points. (b) Distribution of spearman correlations between characterized antimicrobial peptides and their targets alongside these same peptides and randomly selected targets. A total of 16 peptides were identified with exact sequence similarity to characterized peptides. Of these, a total of 10 were identified with unique RNA sequences that could be used as individual test cases. Through literature search, a total of 18 target species were identified for these cases, representing 52 pairwise comparisons with experimentally verified antimicrobial activity. For each peptide, a randomly selected list of bacterial target species were obtained of equal length to its known targets. A significant negative shift is seen in the case of true antagonism (p=0.007, two sample *t*-test). Using a cutoff of -0.05, a positive predictive value of 71% can be obtained by predicting on 28% of true antagonistic cases. (c) Identification of 12 peptides with confident antagonizing relationships with *C. difficile*. (d) Graph shows percent of growth inhibition of *C. difficile* DSM 27147 upon treatment with serial dilutions of scardovicin (2 µM to 0.1 µM). Scardovicin shows antibacterial activity against *C. difficile* with MIC_{100}=0.85 µM, MIC_{50}=0.35 µM. Data points in d represent averages of 6 independent biological replicates, and error bars represent the standard deviation from the mean.

effective in inhibition of C. difficile with MIC_{50}= 0.35 µM and MIC_{100} = 0.85 µM. (Figure 3.4-e). Figure 3.4-e-inset shows preliminary agar well diffusion assay against *C. difficile* using observed MIC_{100}.

However, the main challenge associated with *C. difficle* infection is its ability to form resistant spores that are easily spread, in particular within hospital settings. To test the abilities of scardovicin in this context, transmission electron microscopy (TEM) and scanning electron microscopy (SEM) imaging were conducted to visualize spore formation and spore germination in C. difficile when treated with scardovicin in comparison to control non-treated cells or spores.

SEM results demonstrated that sporulation was aborted completely with 100 nM and continues with 50 nM of scardovicin (Figure 3.5c-f), while sporulation is gradually restored at lower scardovicin concentrations (25 nM and 10 nM) albeit with smaller immature spores (Figure 3.5g-j). Total reversion to normal sporulation was noticed with 5 nM concentration of the peptide (Figure 3.5k-i). Concerning inhibition of spore germination, SEM imaging revealed that scardovicin inhibits spore outgrowth and subsequent germination upon treatment of *C. difficile* spores with a concentration as low as 1X of MIC_{100} compared to control untreated cells (Figure 3.5m-t). However, there was no significant effect observed on early germination steps albeit a slightly delayed germination with high concentration of the peptide at 80 and 100x MIC (Supplementary Figure A2.5).

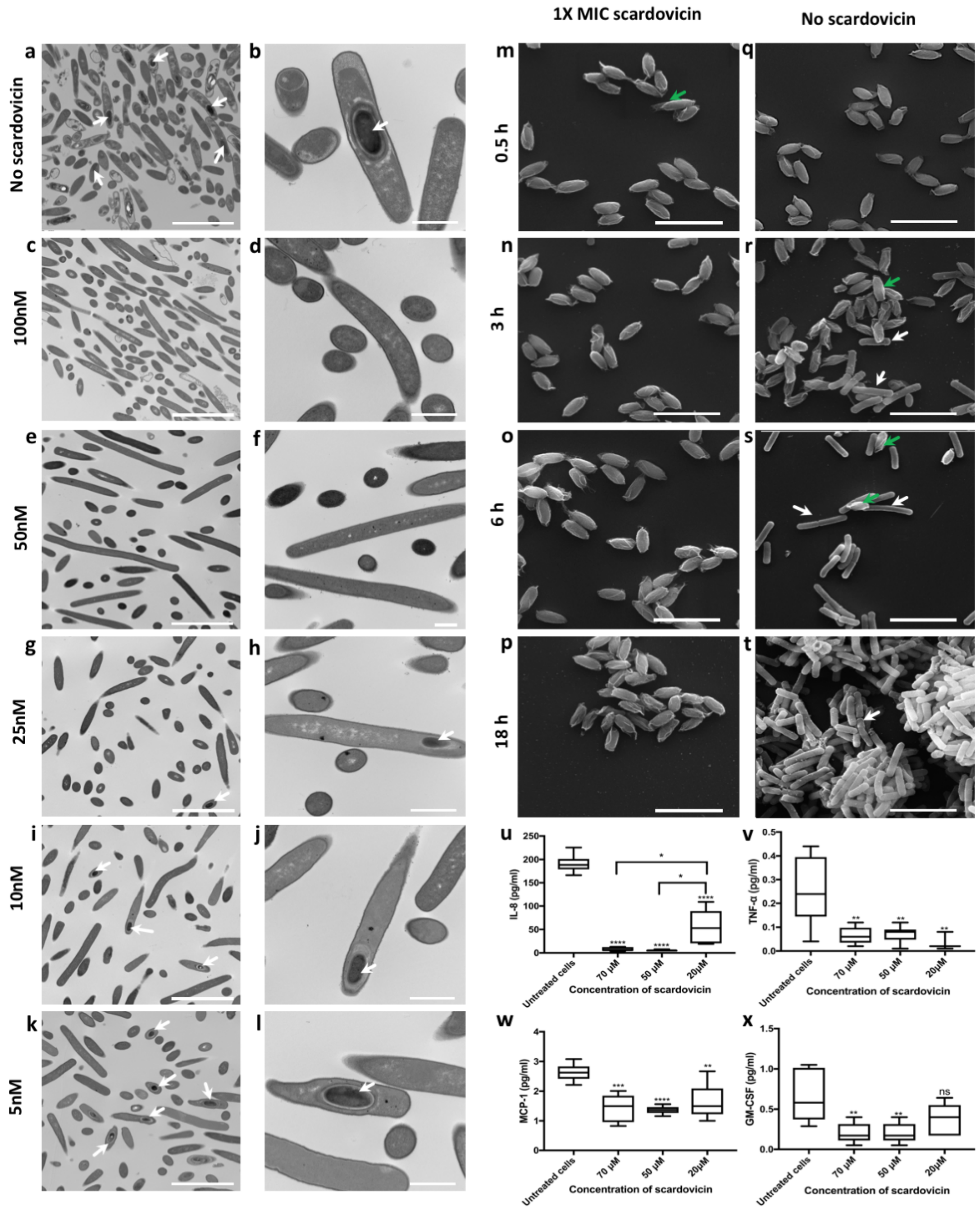

FIGURE 3.5: **Bioactivity profile of scardovicin including inhibition of sporulation, spore outgrowth and antiinflammatory activity.** Full caption on following page.

Figure 3.5: (Previous page.) a-l, Transmission electron microscopy imaging show inhibition of sporulation in *C. difficile* upon cultured on sub-MIC concentrations of scardovicin. a, b show *C. difficile* vegetative cells grown on 70:30 medium without scardovicin. c-d, e-f, g-h, i-j, k-l show *C. difficile* grown on decreasing concentration of scardovicin, 100 nM, 50 nM, 25 nM, 10 nM, and 5 nM, respectively. White arrows point to *C. difficile* spore which is formed inside the mother cell. Number of arrows in images a, g, I, and k is proportional to the average frequency of observed spores in each samples. Images shown are representative of six independent biological replicates for each treatment. Scale bar in a, c, e, g, I, and k is equal to 5 µm while in b, d, f, h, j, and l is equal to 1 µm. m-t, Scanning electron microscopy imaging show inhibition of spore outgrowth in *C. difficile* upon treatment with scardovicin. m-p, show *C. difficile* spores treated with 1X MIC of scardovicin in the presence of germination inducer (10% taurocholate) then incubated anaerobically for 0.5, 3, 6, and 18 h, respectively. q-t, show *C. difficile* spore cultured in the presence of germination inducer only and incubated anaerobically for 0.5, 3, 6, and 18 h, respectively. White arrow points to emerging vegetative cells while green arrow point to un-germinating spore. u-x, Immunomodulatory activity of scardovicin on HT-29 cell line. u-x, graphs show the inhibitory activity of three concentrations of scardovicin (70, 50 and 20 µM) on IL-8, TNF, MCP-1, and GM-CSF, respectively. Whiskers represent the range of data points of six independent biological replicates while error bars indicate the standard error of the mean. Data were analyzed using one-way ANOVA test and t-test. ****P¡0.0001, ***P¡0.0002, **P¡0.001, *P¡0.01

One of the main complications of *C. difficile* infection is severe inflammation of the colon due to secreted toxins. An anti-inflammatory assay using a the human colorectal cell line (HT-29) was used to determine the immunomodulatory properties of scardovicin. The concentration of 13 inflammation biomarkers including GM-CSF, IFN-γ, IL-1β, IL-2, IL-4, IL-5, IL-6, IL-8, IL-10, IL-12(p70), IL-13, MCP-1, TNF-α was measured in supernatant of cells incubated for 24 h with different concentration of scardovicin and compared to control untreated cells. Results reveals that scardovicin has inhibitory activity on four of the major inflammation biomarkers IL-8, TNF-α, MCP-1, GM-CSF in a concentration dependent manner (Figure 3.5u-x). The inhibitory activity of scardovicin on IL-8 was most significant (Figure 3.5u) compared to the control (at p¡0.0001, one-way-ANNOVA) with the highest inhibition observed with 70 and 50 µM. Scardovicin at 50 µM causes the highest inhibitory effect on MCP-1 followed by 70 µM then 20 µM (Figure 3.5w) (at p¡0.0001, and p¡0.0002, respectively, one-way-ANOVA). Concerning the inhibitory effect of scardovicin on TNF-α, 20 µM all used concentrations resulted in the same statistically significant effect (Figure 3.5v) (at p¡0.001, one-way-ANOVA). While the inhibitory effect of scardovicin on GM-CSF was only significant at 70 and 50 µM (Figure 3.5x) (at p¡0.005, one-way-ANOVA). The effect of scardovicin on IL-4 and IL-10 was insignificant (Supplementary Figure A2.6. Using spontaneous non-stimulated cells, other biomarkers showed no expression in all treatment including the control. Scardovicin did not exhibit any cellular cytotoxicity on the HT-29 cell line used in this study.

Collectively, these results revealed that scardovicin is a unique inhibitor of *C. difficile* with multifactorial effects starting with direct inhibitory activity, followed by inhibiting spore formation and germination, and suppression of inflammation. The bioactivity of scardovicin is not accompanied by cellular cytotoxicity.

3.4 Discussion

The dynamics of microbial populations and the drivers of these dynamics are increasingly a focal point for basic and translational science in understanding the microbiome [19, 15]. Certain drivers can be advanced through an enhanced understanding of the products derived from individual human microbiome strains and likewise appreciating the expression of these products *in situ*. Here we have detailed a pipeline that leverages the computational capabilities to translate microbiome genomic information into small molecule data, connect with this expression of genomically inferred products and populations of microbes from metagenomic sequence information.

A guiding principle of the work is to define products created by human microbiota and how they correlate to the abundance of other microbiota. We developed AMPLIFY, A tool to enrich for Microbial Peptides Linked to In-situ FunctionalitY (Figure 3.1). We used a comprehensive repository of environmentally sourced genomes to filter for peptide families exclusive to the human microbiome (Figure 3.3). Thereafter, we correlated the expression of each peptide unique to the microbiome to the corresponding shift in microbial population, using transcriptomic and metagenomic datasets available through the iHMP. Our analysis reveals more than two thousand antagonistic relationships with a prediction accuracy of 71%, based on a validation experiment of known antibacterial agents and their corresponding targets. Applying our pipeline to find a unique microbiome antimicrobial peptide against *C. difficile*, we discovered scardovicin, a multifactorial bioactive peptide with strong anti-inflammatory activity.

The AMPLIFY workflow identifies peptides exclusive to the microbiome and maps their distribution over all body sites

Several efforts have been made to curate databases of antimicrobial peptides among all forms of life [52, 43, 47, 16], with some focus to characterize those encoded within the human microbiome using sequence alignment of putative primary sequences obtained from either a collection of reference genomes or metagenomics data to reference peptides from BAGEL3 [57, 51]. These studies only evidenced the presence of encoded peptides in genomes of the host bacteria, without assessing the actual expression and production of these peptides in-situ. Previously, Donia et. al. used both metagenomic and metatranscriptomic data to verify the presence and expression of modified peptide gene clusters in the microbiome [10]. While most previous analyses made a relevant use of these datasets, our analysis moved beyond identification of metabolites to provide a real functional annotation of these encoded peptides *in situ*, and reveal those that might have co-evolved to combat human pathogens relevant to distinct body site. We mapped the distribution of 202 microbiome exclusive peptides over distinct body sites and showed that they follow a non-random distribution pattern. This pattern might contribute a specific function at each body site, mediated by shaping the population structure. In accordance with our hypothesis, a previous study revealed that the human microbiota follows a specific ecological distribution, either a co-occurrence or co-exclusion, when colonizing each body site [12]. Some of these patterns are pre-requisites to perform a given function [12].

AMPLIFY leads to the discovery of a chemical mediator with unique bioactivity profile against *C. difficile*

C. difficile is one of the most frequently occurring infections in hospital setting and blamed for claiming the lives of 30,000 persons in the USA annually, according to estimates from Centers for Disease Control and Prevention [32], with a increasing rates of epidemic outbreaks [22, 37]. Treatment of *C. difficile* is technically challenging, in part due to the ability of *C. difficile* to produce dormant resistant spores which can spread easily between patients, in addition to further complications associated with secreted toxins. *C. difficile* toxins cause severe inflammation which can lead to pseudomembranous colitis, weakened colon membrane, abdomen distension (toxic megacolon), which have in many cases lead to colonic rupture, sepsis and death [24, 42, 34]. Currently, our best small molecule for treatment of C. difficile infection is the macrolide fidoxamicin (lipiarmycin, tiacumicin B) which exhibits 50% lower rates of failure rates and relapse compared to vancomycin and metronidazole [5]. However, giving the complex disease etiology of *C. difficile*, there is no antibiotic discovered to this date that causes full treatment and completely prevent recurrence [7, 48].

C. difficile infections most occurs after prolonged antibiotic treatment, which drastically alters microbial composition of the GIT. These opportunities allow *C. difficile* to colonize in the absence of protective microbial agents, a hypothesis that has been further confirmed by the success of fecal transplants in preventing recurrence of C. difficile infection [2]. Although a mechanistic link is missing, there are thought to be chemical mediators produced by microbes that may play a protective role against *C. difficile*. Extending this, it is interesting to speculate that the disappearance of low abundance protective microbiota through the administration of broad-spectrum antibiotics, is accompanied by the disappearance of these chemical mediators. Thus, we can propose that the expression of these anti-*C. difficile* mediators is likely correlated with a low abundance of *C. difficile*. Among the peptides uniquely found in the human microbiome, we developed AMPLIFY to identify peptides which follow this predicted expression profile against *C. difficile*, leading the the functional annotation of 12 candidates with putative anti-*C. difficile* activity.

We pursued synthesis and *in vitro* activity profiling of the top candidate to identify scardovicin with unique activity profile superior to all known antibiotics. While fidaximicin exhibits a strong cytotoxicity on human cell lines [11], an effect that is neutralized by its minimal absorption from the GIT, scardovicin shows a protective anti-inflammatory effect on GIT cells through potent inhibition of the chemokines IL-8 and TNF-α (Figure 3.5) implicated in *C. difficile* infection [55]. Discovery of scardovicin validates AMPLIFY as a new robust pipeline that surpasses routine natural products discovery protocols, allows for *in situ* functional annotation of unmodified peptides, and exposes a wealth of unique bioactivities of the human microbiome that relates to human disease.

3.5 Methods

Peptide genome detection model

A curated set of 156 antimicrobial peptides together with their microbial producers and functional annotations were gathered from literature including sequences from BAGEL, UniProt and NCBI (Supplementary table 1). The sequences of these peptides were used to build a genome detection tool employing profile hidden markov models (pHMMs). pHMMs were created from sequence alignment of closely related peptides using, first, a hierarchical clustering algorithm, cd-hit, to identify groups of sequences from the reference set with greater than 70% sequence identity [28]. To ensure that each group contains the maximum possible numbers of available related sequences, the representative sequence of each of these cluster families were then compared to the totality of the proteins within the NCBI non-redundant database [36] of sequences using BLASTP with an e-value threshold of 10^{-6}. Matches with greater than 70% sequence identity in accordance to the representative sequence were added to each cluster family. Each of these cluster families were then aligned using PRANK (v. 140603) [29] to generate a multiple sequence alignment (MSA), which were then compiled to pHMMs using HMMER (v. 2.3.2). HMMMER is a suite of tools used to build profile hidden markov models. These models are built from MSAs, and are used to search for protein families based on the information in the MSAs [13].

To assess the validity of these pHMMs, each pHMM was processed through JackHMMER to iteratively identify sequences in the UniProt database. JackHMMER uses an iterative strategy to rapidly use an HMM against a large database of sequences. It has a higher error rate than typical HMMer, but is essential for large scale searches [21]. A bit-score cutoff for each model was manually chosen at the score in which all matched sequences above this cutoff were either a characterized or a putative antimicrobial peptide as annotated by UniProt. In total, this strategy generated 123 pHMMs (supplementary Table 2) which have been integrated into PRISM v4.2.0 for analysis of genomic data and made available to public: `https://magarveylab.ca/prism`.

Genomic analysis of DNA-encoded unmodified peptides

We then analyzed a total of 9,957 bacterial genomes through PRSIM to identify all putative antimicrobial peptides within these genomes. The analyzed genomes included a total of 2,274 human microbiome genomes together with the associated body site annotations downloaded from NCBI, using all whole genome sequences linked to the Human Microbiome Project (NCBI BioProject ID 28331) [18]. A phylogenetic tree of these bacteria was generated using PhyloPhlAn (v0.99) [41], which examines the homology between the top 400 genes shared between all bacteria.

In addition to the microbiome genomes, a total of 7,683 environmental genomes were obtained from the NCBI genome database. We confirmed the non-human origin of these environmental genomes by gathering the corresponding metadata on isolation sites of these bacterial strains as listed on JGI, NCBI trace, BacDive and PATRIC databases. The purpose of including environmental genomes in the analysis is to use them as filter later on in the analysis to determine peptides families exclusive to the human microbiome. All of these detected peptides are combined and shown in Supplementary table 3.

Determination of novelty of PRISM-identified peptides

All peptides detected by PRISM from human and environmental microbiomes were compared to the previously curated reference dataset of characterized antimicrobial peptides to determine their novelty. A similarity score was calculated by BARLEY using a Smith-Waterman local alignment employing an identity matrix with a match score of 1 and a gap opening and extension penalty of -2. The score of the alignment is normalized between 0 and 1 by dividing this score by the alignment score of the detected peptide to itself. Examining the distribution of the scores among all detected peptides and their highest matched antimicrobial peptide in the reference dataset, the data was determined to be bimodal, where a separation line was drawn in the minima separating known and unknown peptides (supplementary Figure A2.2). BARLEY scores for all identified clusters together with the closest match to known entities are shown in Supplementary table 4.

Genomic clustering to determine exclusivity of predicted peptides

For determining peptides exclusive to the human microbiome, all clusters identified with both environmental and human source annotations were grouped into families, and only families observed entirely from a corresponding human origin were denoted as human exclusive. We obtained a sequence similarity score (described above) for all pairwise comparisons between all peptides and then generated a graph considering all peptides as nodes. Edges between nodes appear only when sequence similarity score fell above a threshold. This graph was then segmented according to connecting groups of nodes to isolate individual unmodified peptides into families. Pairwise similarity scores were calculated by BARLEY using a Smith-Waterman local alignment using the same parameters discussed above. Since this score is directionally dependent, an edge was drawn between peptides when the minimum of either pairwise directional score passed a threshold. This threshold was tested iteratively between 0.01 and 0.99, where each cutoff was scored according to the number of valid families outputted. A valid families is defined as having greater than 1 entity and detected by the same pHMM or having a minimum pairwise similarity greater than 0.7 between all members of the families. This threshold was tested iteratively between 0 and 1 at 0.01 increments. It was found that the percent of valid families according to this criterion was stable between 0.45 and 0.7, so the midpoint at 0.55 was chosen as the optimal cutoff (supplementary Figure A2.3).

Population dynamics informatics

To assess the expression of PRISM-detected peptides, which we denote as human exclusive, a total of 742 fecal samples of single-end raw reads from 109 individuals were collected from the Inflammatory Bowel Disease Multi'omics Database (IBDMDB) (accessed Dec. 2017). We collected both metatranscriptomic and metagenomic data to assess gene expression and microbial abundance in each sample respectively. The single-end raw reads obtained from each sample were trimmed using Trimmomatic with default settings (LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36 ILLUMINACLIP:TruSeq3-SE:2:30:10) to eliminate adapter sequences introduced during RNA sample preparation and low quality reads. To estimate the relative abundance of query peptides in each of these samples, we generated a reference transcriptome that represents the DNA sequences of the query peptides' open reading frames. The reads from each sample were then aligned to this reference transcriptome using Salmon quasi-alignment with a k-cutoff of 11 to generate a transcript per million (TPM) value that reflects the overall relative expression of each gene to the total amount of transcript information in the sample [35]. For the 742 samples, the relative abundance of any given bacterial species was determined via MetaPhlAn2 using metagenomic sequencing data. MetaPhlAn2 uses metagenomic sequences and accurately predicts the relative abundance at the species and strain levels [44]. This data was preprocessed by the The Inflammatory Bowel Disease Multi'omics Database, IBDMDB. In order to determine the correlative relationship between peptides and the corresponding microbial abundance, we calculated a spearman correlation between these two streams of data, considering each sample as a unique observation. To limit the number of comparisons to candidates with sufficient information, only query genes detected in at least 10% of samples, and subject bacterial species present in at least 1% of samples were considered.

Validation of population dynamics informatics

To validate this analysis, the expression of four housekeeping (HK) genes in each transcriptomic sample was measured. HK genes are constitutively expressed so their relative expression from a query organism should be positively associated with the same subject organism in all of these samples. The four HK genes used are key for DNA replication, protein membrane translocation, and DNA repair. We gathered 4 PFAM families representing multiple sequence alignments and HMMs for the selected 4 HK genes (PFAM ids: PF08278, PF00204, PF07517, PF00154). We then identified the ORFs aligned to these PFAMS in the 2274 human microbiome genomes using HMMER, above the trusted cutoff identified on PFAM for each gene family.

The expression of the HK genes was then compared to the relative abundances of individual species in each metagenomic sample. As a control, the same number of pairwise comparisons were generated between each of these genes and a randomly selected species from different genera than the producer organism. Using the strategy outlined above, these pairwise comparisons were generated and the distribution of these correlations were used to ensure data integrity and provide support for the methods and techniques used in this analysis (Figure 3.4-a and Supplementary table 6).

Another validation was performed to examine the difference in spearman correlations between experimentally validated antagonistic relationships and random noise in this dataset. Each peptide from the human microbiome that shared 100% sequence identity to characterized peptide with proven antimicrobial activity were collected. A total of 10 known peptides were queried against 18 different species represent known microbial targets resulting in 52 pairwise comparisons (Supplementary table 6). Multiple peptides with the same protein sequence were allowed, but only genes with unique ORF were used for this validation, thus each candidate peptide represents a unique and independent experiment. Using the same strategy discussed above, all pairwise comparisons were made and the resulting distribution in correlations was used to infer the validity of this pipeline, and guide the accuracy of downstream analyses. Expression values of all analyzed sequences measured in TPM together with their antagonistic correlations are shown (Supplementary table 7).

Antimicrobial activity and determination of minimum inhibitory concentration of scardovicin

C. difficile DSM 27147 [Ribotype 027, producer of toxins A and B (TcdA and TcdB) and the binary toxin (CtdA and CtdB)] was maintained on carbohydrate chopped meat (CCM) agar medium supplemented with 5% defibrinated horse blood (SR0050, ThermoScientific). The medium composition is 30 g/L peptone, 5 g/L yeast extract, 5 g/L K2HPO4, 4 g/L glucose, 1 g/L cellobiose, 1 g/L maltose, 1 g/L starch, 4 ml/L resazurin solution (0.025%), 15 g/L agar. The volume is made up to 1 liter by chopped meat broth composed of 500 g/L fat-free ground beef boiled with 25 ml/L NaOH (1N) and deionized water of up to 1 liter. Scardovicin was synthesized by GenScript (Piscataway, NJ, USA) and primary structure was validated by LC/MS/MS (Supplementary Figure A2.4).

As a preliminary screening to assess if scardovicin processes antimicrobial activity, we conducted agar well diffusion assay. Briefly, 10 µL of overnight actively grown culture of *C. difficile*

DSM 27147 were plated on the top of CCMA plates then holes were punctured in the agar using sterile glass pipette and 20 µL of 1-5 µM scardovicin were applied into the holes. The plates were incubated anaerobically at 37 °C for 24 h. Thereafter, plates were screened for any developed zone of inhibition.

To determine the minimum inhibitory concentration (MIC) of scardovicin, we conducted broth microdilution antimicrobial assay in 96-well microtiter plate. Briefly, a single colony of each of *C. difficile* DSM 27147 grown for 48 h in CCM agar supplemented with 5% defibrinated horse blood was inoculated into CCM broth for 24 h then diluted with the same medium to 1: 10,000. Thereafter, 196 µL of this inoculated medium were added to each well, 4 µL different serial dilutions of scardovicin were added to the well resulting in final concentration range starting from 100 µM to 100 nM. Blank control was wells contains 196 µL non-inoculated CCM broth and 4 µL DMSO (solvent used to solubilize scardovicin). Positive control was wells containing 196 µL inoculated CCM broth and 4 µL DMSO. The FDA-approved antibiotic, fidaxomicin (1µM) was used as a positive control.

The plates were incubated anaerobically at 37 °C. After 24 h, the OD_{600} of each well was measured with a microplate reader. Thereafter, MIC_{100} and MIC_{50}, defined as the lowest concentration of the peptide that results in 100% and 50% growth inhibition, respectively, were measured. Each concentration was tested in triplicates and the entire assay was repeated independently in duplicates. Percent of growth inhibition was determined according to the following equation:

$$\text{Inhibition}_{\%} = 1 - \left(\frac{(OD_{600}\text{ test}) - (OD_{600}\text{ blank medium})}{(OD_{600}\text{ pathogen only}) - (OD_{600}\text{ blank medium})} \right) * 100$$

Purification of *C. difficile* DSM 27147 spores

To assess if scardovicin has effect on spore germination or spore production, *C. difficile* DSM 27147 spores were purified. A single 48 h *C. difficile* colony grown on CCM medium supplemented with 5% defibribnated horse blood was inoculated into brain heart infusion supplement (BHIS) medium (composed of 37 g brain heart infusion extract and 5 g yeast extract per liter) supplemented with 10% taurocholate (86339, Sigma-Aldrich) and incubated anaerobically at 37 °C for 24 h. The cells were then diluted with BHIS (1:100) and 200 µL of the diluted actively growing culture were plated on 70:30 sporulation agar plates (composed of 63 g bacto peptone, 3.5 g proteose peptone, 0.7 g ammonium sulphate, 1 g tris base, 11 g brain heart infusion extract, 1.5 g yeast extract, 15 g agar, 3 ml cysteine (10%W/V), 10 mM taurocholate. Five plates were then incubated anaerobically at 37 °C for 5 days. Thereafter, colonies were suspended in ice cold sterile deionized water and the suspension was then removed from the anaerobic chamber. The cells were centrifuged at 2000 rpm for 10 min at room temperature, the supernatant contains vegetative cells and debris was decanted and pellets were washed up to 10 times in ice cold water at the same speed. The washed pellets were then suspended in 10 mL sterile deionized water and incubated at -20 °C for 48 h to help lysis of the mother cells and release of mature spores. Thereafter, the spore suspension was centrifuged at 13,000 rpm for 2 min at 4 °C followed by at least 10 washes in sterile water. The pellets were suspended in 3 ml water and applied slowly on top of 10 ml 50% nonionic density gradient medium (Histodenz™, D2158, Sigma-Aldrich) in 15 ml polypropylene conical tube and centrifuged in swinging bucket rotor at 6000 rpm for 40 min at 4 °C. The vegetative cells and debris were collected through the gradient and on the interface while spores formed a pellet at the bottom. The spore pellet was then washed at least 10 times in 1X filter-sterilized phosphate buffer Saline (PBS, composed of 2.5 g/L Na2HPO4.7H20, 8 g/L NaCL, 0.2 g/L KCL, 0.2 g/L KH2PO4) at 13 rpm for 2 min at 4 °C. The final spore preparation was kept in 1X PBS+1% bovine serum albumin. Throughout the entire purification steps, spores were checked for purity by imaging under light microscopy using malachite green/safranin counterstain, where spores stained green and vegetative cells stained in magenta color. To count the number of the viable spores, spores were heat activated at 80 °C for 15 min in sterile deionized a water immediately before germination then a series of 10 fold dilutions were prepared. To allow spore germination, 100 µL of each dilution was plated on pre-reduced BHIS agar plates supplemented with 10 mM taurocholate and incubated anaerobically at 37 °C for 24 h then the number of spores per ml was calculated following the equation:

$$\frac{\text{CFU}}{mL} = \frac{\text{Number of colonies}}{\text{Dilution factor} * \text{Volume plated (mL)}}$$

Spore count was performed using light microscopy and staining malachite green and safranin counterstain. The working spore count differs according to the downstream experiment as detailed later.

Testing the activity of scardovicin on initiation of germination of *C. difficile* spores

The early steps in spore germination involves release of small molecules such as Ca2+-dipicolinic acid upon sensing some germination triggers such bile acid, followed by hydration of the cortex and core. These reactions could be measured as a drop in OD_{600} to 60-70% [53]. To determine if scardovicin has effect on this step, we measured the change in OD_{600} of a purified spore suspension upon treatment with the peptide.

Spores were suspended in BHIS medium supplemented with 50mM lactate, 100 mM alanine and 10 mM taurocholate and adjusted to OD_{600} of 1 (equivalent to 5000 spores per mL) and different dilutions of scardovicin were added to a final concentration ranging from 100X MIC to 1X MIC in 96 well-plate format. The negative control use was the spore suspension in BHIS only without inducer and the positive control was spore suspension treated with 10 mM taurocholate. There were three replicates from each concentration and the entire experiment was performed in three independent replicates. GraphPad-PRISM software (GraphPad 7.0d, USA) was used to plot and analyze data.

Spore outgrowth inhibition assay

To test if scardovicin can inhibit spore outgrowth and emergence of vegetative cells, anaerobic germination followed by scanning electron microscopy (SEM) imaging was conducted. Heat activated spore (1000 spore per mL) suspended in germination solution consisted of CCM broth supplemented with 1% taurocholate, 100 mM L-alanine and 50 mM lactate and different concentration of scardovicin (from 100 nM to 2 µM) and allowed to incubate anaerobically at 37 °C. Spores germinating on medium without scardovicin were used as negative control while 2 µM final concentration of fidaxomicin antibiotic was employed as positive control known to inhibit spore outgrowth [1]. At 30 min time interval, 50 µL of the spore suspension were withdrawn, fixed with equal volume of 2% glutradhyde in 0.1M phosphate buffer (pH 7.4) and left for 2 h at 4 °C. To further prepare samples for SEM, samples were centrifuged at 4000 rpm for 20 min at 4 °C, supernatant was decanted and pellets were suspended in 1X PBS and washed up to three times. 10 µL were then applied to poly-lysine coated cover slips, air dried for 1 h, dehydrated through a graded ethanol series (70%, 95%, and 100% (2 x 2min) and then transferred to the critical point dryer. The samples were kept immersed in 100% ethanol, placed into wire baskets and transferred to the chamber of a Leica EM CPD300 critical point dryer (Leica Mikrosysteme GmbH, Wien, Austria). The chamber was sealed and then flushed 12 times with liquid CO2. The CO2 filled chamber was heated to 35 °C and pressure increased in chamber to above 1100 psi so that CO2 was changed from liquid phase to gaseous phase. The gas was vented slowly from the chamber until atmospheric pressure was reached and the samples were dehydrated without surface tension damage. The dried samples were mounted onto SEM stubs with double-sided carbon tape. The samples on stubs were then placed in the chamber of a Polaron Model E5100 sputter coater (Polaron Equipment Ltd., Watford, Hertfordshire) and approximately 20 nm of gold was deposited onto the stubs. The samples were viewed in a Tescan Vega II LSU scanning electron microscope (Tescan USA, PA) operating at 20kV.

Inhibition of sporulation experiment

To test if scardovicin can inhibit sporulation at sub-MIC concentration, *C. difficile* DSM 27147 was grown on sporulation medium supplemented with different concentration of the peptide then followed by transmission electron microscopy (TEM) imaging to visualize the produced spores inside the mother cell. Briefly, a single colony of *C. difficile* DSM 27147 grown on CCM medium for 48 h was used to inoculate CCM broth which is then incubated anaerobically for 24 h at 37 °C. Thereafter, 100 µL of the actively grown culture were plated on 70:30 agar medium supplemented with different sub-MIC concentrations of the peptide starting from 100 nM to 5 nM and incubated anaerobically for 48 h at 37 °C. Produced colonies were then washed out in 2% glutradhyde in PBS and kept at 4 °C for 2 h, centrifuged at 4000 rpm for 10 min. Pellets were re-suspended in 1X PBS and washed three times. To prepare cells for TEM, the samples were then post-fixed in 1% osmium tetroxide in 0.1M phosphate buffer for 1 hour. The samples were dehydrated through a graded ethanol series (50%, 70%, 70%, 95%, 95%, 100%, 100%). The final dehydration for the TEM samples was done in 100% propylene oxide (PO). Infiltration with Spurr's resin was through a graded series (2:1 PO:Spurr's, 1:1 PO:Spurr's, 1:2 PO:Spurr's, 100% Spurr's, 100% Spurr's, 100% Spurr's) with rotation of the samples in between solution changes. The samples were transferred to embedding moulds which were then filled with fresh 100% Spurr's resin and polymerized overnight in a 60°C oven. Thin sections were cut on a Leica UCT Ultramicrotome and picked up onto Cu grids. The sections were post-stained with uranyl acetate and lead citrate and then viewed in a JEOL JEM 1200 EX TEMSCAN transmission electron microscope (JEOL, Peabody, MA, USA) operating at an accelerating voltage of 80kV.

Immunomodulation and cytotoxicity assay

To test if scardovicin has immunomodulatory activity, we developed a cell-based assay using HT-29 cell line. HT-29 cells were obtained from Dr. Bruce Vallance (BC Children's Hospital, The University of British Columbia, British Columbia, Canada). HT-29 cells were cultured in T-75 tissue culture flasks (Costar, Cambridge, MA, USA) in Dulbecco's modified Eagle medium–nutrient mixture F-12 (DMEM–F-12, Gibco BRL Life Technologies, Burlington, Canada). DMEM was supplemented with 10% fetal bovine serum, 100 U/mL of penicillin, 100 µg/mL of streptomycin, 1% MEM and 20 mM HEPES (Invitrogen Life Technologies). Cells were maintained at 37°C in a humidified incubator at 5% CO_2. Culture medium was replaced with prewarmed medium every 2 days. Confluent cultures (¿80%) were harvested using trypsin-EDTA. Cells from passages 13 to 15 were used in this study.

500 µL of HT-29 cells were added to 24-well tissue culture plate. Each well was treated by either one of three tested concentrations of scardovicin (70 µM, 50µM, 20µM) or DMSO (solvent used to dissolve scardovicin). The plate was then incubated at 37°C in a humidified incubator at 5% CO_2 for 24 h. Thereafter, supernatant was harvested by centrifugation at 4000 rpm for 15 min, filtered, kept frozen on dry ice and shipped to Eve Technologies (Calgary, Canada) for an inflammatory-focus 13-custom plex discovery assay to determine the spontaneous secretion

of granulocyte-macrophage colony-stimulating factor (GM-CSF), IFN-γ, IL-1β, IL-2, IL-4, IL-5, IL-6, IL-8, IL-10, IL-12(p70), IL-13, MCP-1, and TNF-α. For cytotoxicity assay, HT-29 cell lines were used under the same conditions detailed above, except for a longer incubation period of 72 h, were treated with different concentrations of scardovicin (70 µM, 50 µM, 20 µM) and then stained with Resazurin solution (500 µM) for 10% of the final volume (v/v) and incubate at 5% CO2 for 5 h. Dead cells remain purple while live cells turned pink.

Statistical analyses

Statistical analysis was conducted using GraphPad Prism (GraphPad Software, Inc., La Jolla, USA; version 7.0d). Statistically significant differences were calculated by appropriate statistical methods as indicated in each experiment.

Bibliography

[1] Allen, C.A. et al. Both fidaxomicin and vancomycin inhibit outgrowth of Clostridium difficile spores. *Antimicrobial Agents and Chemotherapy*, 57(1):664–667, January 2013. ISSN 1098-6596. doi: 10.1128/AAC.01611-12.

[2] Carlucci, C., Petrof, E.O. and Allen-Vercoe, E. Fecal Microbiota-based Therapeutics for Recurrent Clostridium difficile Infection, Ulcerative Colitis and Obesity. *EBioMedicine*, 13: 37–45, November 2016. ISSN 2352-3964. doi: 10.1016/j.ebiom.2016.09.029.

[3] Chu, J. et al. Discovery of MRSA active antibiotics using primary sequence from the human microbiome. *Nature Chemical Biology*, 12(12):1004–1006, December 2016. ISSN 1552-4469. doi: 10.1038/nchembio.2207.

[4] Cohen, L.J. et al. Commensal bacteria make GPCR ligands that mimic human signalling molecules. *Nature*, 549(7670):48–53, September 2017. ISSN 1476-4687. doi: 10.1038/nature23874.

[5] Cornely, O.A. et al. Treatment of first recurrence of Clostridium difficile infection: Fidaxomicin versus vancomycin. *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America*, 55 Suppl 2:S154–161, August 2012. ISSN 1537-6591. doi: 10.1093/cid/cis462.

[6] Crobach, M.J.T. et al. Understanding Clostridium difficile Colonization. *Clinical Microbiology Reviews*, 31(2):e00021–17, January 2018. ISSN 0893-8512, 1098-6618. doi: 10.1128/CMR.00021-17.

[7] Crook, D.W. et al. Fidaxomicin versus vancomycin for Clostridium difficile infection: Meta-analysis of pivotal randomized controlled trials. *Clinical Infectious Diseases: An Official*

Publication of the Infectious Diseases Society of America, 55 Suppl 2:S93–103, August 2012. ISSN 1537-6591. doi: 10.1093/cid/cis499.

[8] de Vos, M.G.J. et al. Interaction networks, ecological stability, and collective antibiotic tolerance in polymicrobial infections. *Proceedings of the National Academy of Sciences of the United States of America*, 114(40):10666–10671, October 2017. ISSN 1091-6490. doi: 10.1073/pnas.1713372114.

[9] Donia, M.S. and Fischbach, M.A. HUMAN MICROBIOTA. Small molecules from the human microbiota. *Science (New York, N.Y.)*, 349(6246):1254766, July 2015. ISSN 1095-9203. doi: 10.1126/science.1254766.

[10] Donia, M.S. et al. A systematic analysis of biosynthetic gene clusters in the human microbiome reveals a common family of antibiotics. *Cell*, 158(6):1402–1414, September 2014. ISSN 1097-4172. doi: 10.1016/j.cell.2014.08.032.

[11] Erb, W. and Zhu, J. From natural product to marketed drug: The tiacumicin odyssey. *Natural Product Reports*, 30(1):161–174, January 2013. ISSN 1460-4752. doi: 10.1039/c2np20080e.

[12] Faust, K. et al. Microbial Co-occurrence Relationships in the Human Microbiome. *PLOS Computational Biology*, 8(7):e1002606, 12-Jul-2012. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1002606.

[13] Finn, R.D. et al. HMMER web server: 2015 update. *Nucleic Acids Research*, 43(Web Server issue):W30–W38, July 2015. ISSN 0305-1048. doi: 10.1093/nar/gkv397.

[14] Geller, L.T. et al. Potential role of intratumor bacteria in mediating tumor resistance to the chemotherapeutic drug gemcitabine. *Science (New York, N.Y.)*, 357(6356):1156–1160, 09 15, 2017. ISSN 1095-9203. doi: 10.1126/science.aah5043.

[15] Gilbert, J.A. et al. Current understanding of the human microbiome. *Nature Medicine*, 24 (4):392–400, April 2018. ISSN 1546-170X. doi: 10.1038/nm.4517.

[16] Hammami, R. et al. BACTIBASE second release: A database and tool platform for bacteriocin characterization. *BMC Microbiology*, 10:22, January 2010. ISSN 1471-2180. doi: 10.1186/1471-2180-10-22.

[17] Integrative HMP (iHMP) Research Network Consortium. The Integrative Human Microbiome Project: Dynamic Analysis of Microbiome-Host Omics Profiles during Periods of Human Health and Disease. *Cell host & microbe*, 16(3):276–289, September 2014. ISSN 1931-3128. doi: 10.1016/j.chom.2014.08.014.

[18] Integrative HMP (iHMP) Research Network Consortium. The Integrative Human Microbiome Project: Dynamic analysis of microbiome-host omics profiles during periods of human health and disease. *Cell Host & Microbe*, 16(3):276–289, September 2014. ISSN 1934-6069. doi: 10.1016/j.chom.2014.08.014.

[19] Jeffery, I.B. et al. Categorization of the gut microbiota: Enterotypes or gradients? *Nature Reviews. Microbiology*, 10(9):591–592, September 2012. ISSN 1740-1534.

[20] Jiang, H. et al. Altered fecal microbiota composition in patients with major depressive disorder. *Brain, Behavior, and Immunity*, 48:186–194, August 2015. ISSN 1090-2139. doi: 10.1016/j.bbi.2015.03.016.

[21] Johnson, L.S., Eddy, S.R. and Portugaly, E. Hidden Markov model speed heuristic and iterative HMM search procedure. *BMC Bioinformatics*, 11:431, August 2010. ISSN 1471-2105. doi: 10.1186/1471-2105-11-431.

[22] Johnson, S. Evidence-Based Approach to Clostridium difficile Infection. *Gastroenterology & Hepatology*, 13(4):238–241, April 2017. ISSN 1554-7914.

[23] Jovel, J. et al. Characterization of the Gut Microbiome Using 16S or Shotgun Metagenomics. *Frontiers in Microbiology*, 7, April 2016. ISSN 1664-302X. doi: 10.3389/fmicb.2016.00459.

[24] Kelly, C.P. and Kyne, L. The host immune response to Clostridium difficile. *Journal of Medical Microbiology*, 60(8):1070–1079, 2011. doi: 10.1099/jmm.0.030015-0.

[25] Kwon, J.H., Olsen, M.A. and Dubberke, E.R. The morbidity, mortality, and costs associated with Clostridium difficile infection. *Infectious Disease Clinics of North America*, 29(1):123–134, March 2015. ISSN 1557-9824. doi: 10.1016/j.idc.2014.11.003.

[26] Lazdunski, C.J. Pore-forming colicins: Synthesis, extracellular release, mode of action, immunity. *Biochimie*, 70(9):1291–1296, September 1988. ISSN 0300-9084.

[27] Li, Q. et al. The Gut Microbiota and Autism Spectrum Disorders. *Frontiers in Cellular Neuroscience*, 11, April 2017. ISSN 1662-5102. doi: 10.3389/fncel.2017.00120.

[28] Li, W. and Godzik, A. Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics (Oxford, England)*, 22(13):1658–1659, July 2006. ISSN 1367-4803. doi: 10.1093/bioinformatics/btl158.

[29] Löytynoja, A. Phylogeny-aware alignment with PRANK. *Methods in Molecular Biology (Clifton, N.J.)*, 1079:155–170, 2014. ISSN 1940-6029. doi: 10.1007/978-1-62703-646-7_10.

[30] Mancuso, G. et al. Bacteroides fragilis-Derived Lipopolysaccharide Produces Cell Activation and Lethal Toxicity via Toll-Like Receptor 4. *Infection and Immunity*, 73(9):5620–5627, January 2005. ISSN 0019-9567, 1098-5522. doi: 10.1128/IAI.73.9.5620-5627.2005.

[31] Manichanh, C. et al. The gut microbiota in IBD. *Nature Reviews Gastroenterology & Hepatology*, 9(10):599–608, October 2012. ISSN 1759-5053. doi: 10.1038/nrgastro.2012.152.

[32] McAlpine, J.B. The ups and downs of drug discovery: The early history of Fidaxomicin. *The Journal of Antibiotics*, 70(5):492–494, May 2017. ISSN 1881-1469. doi: 10.1038/ja.2016.157.

[33] Mousa, W.K. et al. Antibiotics and specialized metabolites from the human microbiota. *Natural Product Reports*, 34(11):1302–1331, 2017. doi: 10.1039/C7NP00021A.

[34] Napolitano, L.M. and Edmiston, C.E. Clostridium difficile disease: Diagnosis, pathogenesis, and treatment update. *Surgery*, 162(2):325–348, August 2017. ISSN 1532-7361. doi: 10.1016/j.surg.2017.01.018.

[35] Patro, R. et al. Salmon: Fast and bias-aware quantification of transcript expression using dual-phase inference. *Nature methods*, 14(4):417–419, April 2017. ISSN 1548-7091. doi: 10.1038/nmeth.4197.

[36] Pruitt, K.D., Tatusova, T. and Maglott, D.R. NCBI Reference Sequence (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research*, 33(suppl_1):D501–D504, January 2005. ISSN 0305-1048. doi: 10.1093/nar/gki025.

[37] Riedel, T. et al. High metabolic versatility of different toxigenic and non-toxigenic Clostridioides difficile isolates. *International journal of medical microbiology: IJMM*, 307(6):311–320, September 2017. ISSN 1618-0607. doi: 10.1016/j.ijmm.2017.05.007.

[38] Sand, S.L. et al. Plantaricin A, a cationic peptide produced by Lactobacillus plantarum, permeabilizes eukaryotic cell membranes by a mechanism dependent on negative surface charge linked to glycosylated membrane proteins. *Biochimica Et Biophysica Acta*, 1828(2): 249–259, February 2013. ISSN 0006-3002. doi: 10.1016/j.bbamem.2012.11.001.

[39] Scher, J.U. and Abramson, S.B. The microbiome and rheumatoid arthritis. *Nature Reviews Rheumatology*, 7(10):569–578, October 2011. ISSN 1759-4804. doi: 10.1038/nrrheum.2011.121.

[40] Schwabe, R.F. and Wang, T.C. Bacteria Deliver a Genotoxic Hit. *Science*, 338(6103):52–53, October 2012. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1229905.

[41] Segata, N. et al. PhyloPhlAn is a new method for improved phylogenetic and taxonomic placement of microbes. *Nature Communications*, 4:2304, August 2013. ISSN 2041-1723. doi: 10.1038/ncomms3304.

[42] Shen, A. Clostridium difficile Toxins: Mediators of Inflammation. *Journal of Innate Immunity*, 4(2):149–158, February 2012. ISSN 1662-811X. doi: 10.1159/000332946.

[43] Thomas, S. et al. CAMP: A useful resource for research on antimicrobial peptides. *Nucleic Acids Research*, 38(suppl_1):D774–D780, January 2010. ISSN 0305-1048. doi: 10.1093/nar/gkp1021.

[44] Truong, D.T. et al. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nature Methods*, 12(10):902–903, October 2015. ISSN 1548-7105. doi: 10.1038/nmeth.3589.

[45] Upadhyaya, S. and Banerjee, G. Type 2 diabetes and gut microbiome: At the intersection of known and unknown. *Gut Microbes*, 6(2):85–92, 2015. ISSN 1949-0984. doi: 10.1080/19490976.2015.1024918.

[46] Valdés-Stauber, N. and Scherer, S. Isolation and characterization of Linocin M18, a bacteriocin produced by Brevibacterium linens. *Applied and Environmental Microbiology*, 60(10): 3809–3814, October 1994. ISSN 0099-2240.

[47] van Heel, A.J. et al. BAGEL3: Automated identification of genes encoding bacteriocins and (non-)bactericidal posttranslationally modified peptides. *Nucleic Acids Research*, 41(Web Server issue):W448–W453, July 2013. ISSN 0305-1048. doi: 10.1093/nar/gkt391.

[48] Vardakas, K.Z. et al. Treatment failure and recurrence of Clostridium difficile infection following treatment with vancomycin or metronidazole: A systematic review of the evidence. *International Journal of Antimicrobial Agents*, 40(1):1–8, July 2012. ISSN 1872-7913. doi: 10.1016/j.ijantimicag.2012.01.004.

[49] Vizcaino, M.I. and Crawford, J.M. The colibactin warhead crosslinks DNA. *Nature chemistry*, 7(5):411–417, May 2015. ISSN 1755-4330. doi: 10.1038/nchem.2221.

[50] Vogt, N.M. et al. Gut microbiome alterations in Alzheimer's disease. *Scientific Reports*, 7 (1):13537, October 2017. ISSN 2045-2322. doi: 10.1038/s41598-017-13601-y.

[51] Walsh, C.J. et al. In silico identification of bacteriocin gene clusters in the gastrointestinal tract, based on the Human Microbiome Project's reference genome database. *BMC Microbiology*, 15(1):183, December 2015. ISSN 1471-2180. doi: 10.1186/s12866-015-0515-4.

[52] Wang, G., Li, X. and Wang, Z. APD3: The antimicrobial peptide database as a tool for research and education. *Nucleic Acids Research*, 44(D1):D1087–D1093, January 2016. ISSN 0305-1048. doi: 10.1093/nar/gkv1278.

[53] Wang, S. et al. Characterization of the Dynamic Germination of Individual Clostridium difficile Spores Using Raman Spectroscopy and Differential Interference Contrast Microscopy. *Journal of Bacteriology*, 197(14):2361–2373, July 2015. ISSN 0021-9193. doi: 10.1128/JB.00200-15.

[54] Yakob, L. et al. Mechanisms of hypervirulent Clostridium difficile ribotype 027 displacement of endemic strains: An epidemiological model. *Scientific Reports*, 5:12666, July 2015. ISSN 2045-2322. doi: 10.1038/srep12666.

[55] Yu, H. et al. Cytokines Are Markers of the Clostridium difficile-Induced Inflammatory Response and Predict Disease Severity. *Clinical and vaccine immunology: CVI*, 24(8), August 2017. ISSN 1556-679X. doi: 10.1128/CVI.00037-17.

[56] Zelante, T. et al. Tryptophan catabolites from microbiota engage aryl hydrocarbon receptor and balance mucosal reactivity via interleukin-22. *Immunity*, 39(2):372–385, August 2013. ISSN 1097-4180. doi: 10.1016/j.immuni.2013.08.003.

[57] Zheng, J. et al. Diversity and dynamics of bacteriocins from human microbiome. *Environmental Microbiology*, 17(6):2133–2143, June 2015. ISSN 1462-2920. doi: 10.1111/1462-2920.12662.

[58] Zipperer, A. et al. Human commensals producing a novel antibiotic impair pathogen colonization. *Nature*, 535(7613):511–516, 07 28, 2016. ISSN 1476-4687. doi: 10.1038/nature18634.

[59] Zvanych, R. et al. Small molecule immunomodulins from cultures of the human microbiome member Lactobacillus plantarum. *The Journal of Antibiotics*, 67(1):85–88, January 2014. ISSN 0021-8820. doi: 10.1038/ja.2013.126.

Chapter 4

Significance and future prospective

Microbial natural products (NPs), as secondary metabolites, are facets of chemistry that have been evolutionarily honed towards highly specific functions. It is this property that has made NPs immensely valuable small molecules with direct relevance as antibiotics and other therapeutics [6]. While these products were the basis of many drug discovery programs, the emphasis on classical methodologies for their discovery has lead to ever diminishing returns [4]. Historically, the field of natural products has been rapidly advanced through the advent of new technologies. Examples include the development of NMR in the 1950s accelerating structure elucidation [2], and the advancement of molecular biology techniques in the 1980s allowing us to better understand NP biosynthesis [5]. Currently, biology is facing a strong paradigm shift due to the emergence of new large scale methods of collecting data [7]. From the explosion in genome sequencing, to the advancements in metabolomics, these platforms facilitate a new wave of untargeted experimentation where generalizable data under standard conditions can be collected and published. Through the development of software platforms, cognizant and specialized for accelerating NP discovery, we can leverage these rich datasets to guide future endeavours. The work demonstrated here shows some examples in which I have successfully crafted informatic platforms targeted towards the discovery of diverse, novel and bioactive peptidic chemical scaffolds.

Taxonomically, the golden era of NP discovery was largely fueled by the *en masse* collection and culturing of *Streptomyces*. Through routine screening of these bacteria, it was revealed that common secondary metabolites such as streptomycin may be present in 1% of soil isolates [4] leading to false positives and successive wasted efforts in isolation and structural characterization. A large risk in the strategies of this era was the relatively blind nature in which active fractions were further pursued. In contrast, the platforms presented here aims to target leverage genomic data to infer key structural elements of resulting natural products. Specifically, when looking into ribosomally and post-translationally modified peptides (RiPPs), I have developed a platform, BARLEY, that can infer differential chemistry through genomic data (**Chapter 2**). While efforts such as these have been attempted in other classes of modularly encoded natural products (PKS and NRPS) [1], this represents the first and currently only effort in quantifying the diversity and novelty among genomically encoded RiPPs. As a further step, I present a targeted isolation strategy that leverages high resolution mass spectral datasets. As a demonstration of this workflow, we were able to isolate a novel and highly dissimilar lasso peptide. More importantly, this workflow not only revealed the wealth of genomically encoded RiPP diversity, but is built to provide researchers with a clear avenue for targeted discovery. Tools such as BARLEY are essential in the modern data-rich world. With recent evaluations demonstrating we may have only recognized 1% of the total bacterial population [8], we are heading towards a future with an even greater abundance of information.

The function of natural products has largely been guided through reductionist experimental strategies. In the early days of this work, it was often noticed that a particular biological extract had an interesting property, which was subsequently purified to result in a single compound. However, the evolved function of these NPs was honed in a likely polymicrobial environment. In this context, there has been relatively little study, likely due to the difficulties in monitoring this process [9]. However, due to the recently developed platforms for metagenomic and metatranscriptomic sequencing, we can now peek into the role these metabolites may play *in situ*. With this in mind, I have developed AMPLIFY, A tool to enrich for Microbial Peptides Linked to In-situ FunctionaliTy (**Chapter 3**). Here, we look to find chemical mediators evolved specifically to tackle the niche environment of the human gastrointestinal tract. Further, through examining the interplay between peptide expression and overall microbial composition, we were able to extract an interesting relationship between antimicrobial peptides and their targets. Specifically, we found that negative correlations between peptide expression and target microbial abundance can be used to infer an antimicrobial relationship. In this context, we mined this dataset to find peptides likely to antagonize the invasion of foreign pathogens. Against the infectious microbe, *Clostridium difficile*, this analysis revealed several peptides with putative antimicrobial, not only indicating that our GI residents may provide an innate protective role, but also that nature, by default, implements redundancy. To further validate this protocol, we synthesized a candidate antimicrobial peptide from *Scardovia wiggisae*, and evaluated its effects *in vitro*. This peptide, scardovicin, demonstrated anti-clostridial effects in a multi-faceted manner, inhibiting growth, preventing spore formation and spore germination. Further, this peptide also demonstrated significant anti-inflammatory activity on GI cell lines, an effect that is ideal in tackling the pro-inflammatory conditions induced by *C. difficile* infection [3]. Discoveries such as these would not be possible through reductionist approaches, and are solely enabled through large scale data analytics as demonstrated here.

These projects in conjunction demonstrate a small portion of what is capable given the technologies available. The goal of these works is to enable scientists to better leverage the technologies and data available in a manner amenable to NP discovery and characterization. As our technological and data collection capabilities continue to increase, as we have seen within the last decade, software development and evaluation will inevitably play a much larger role in the scientific process. The works demonstrated here effectively have shown how these computational platforms, built with the goal of accelerating NP discovery, can successfully guide research and lead to a better understanding of the chemical mediators underlying our biology.

Bibliography

[1] Dejong, C.A. et al. Polyketide and nonribosomal peptide retro-biosynthesis and global gene cluster matching. *Nature Chemical Biology*, 12(12):1007–1014, December 2016. ISSN 1552-4469. doi: 10.1038/nchembio.2188.

[2] Freeman, R. A short history of NMR. *Chemistry of Heterocyclic Compounds*, 31(9):1004–1005, September 1995. ISSN 0009-3122, 1573-8353. doi: 10.1007/BF01165047.

[3] Kelly, C.P. and Kyne, L. The host immune response to Clostridium difficile. *Journal of Medical Microbiology*, 60(8):1070–1079, 2011. doi: 10.1099/jmm.0.030015-0.

[4] Li, J.W.H. and Vederas, J.C. Drug discovery and natural products: End of an era or an endless frontier? *Science (New York, N.Y.)*, 325(5937):161–165, July 2009. ISSN 1095-9203. doi: 10.1126/science.1168243.

[5] Malpartida, F. and Hopwood, D.A. Molecular cloning of the whole biosynthetic pathway of a Streptomyces antibiotic and its expression in a heterologous host. *Nature*, 309(5967):462–464, May 1984. ISSN 1476-4687. doi: 10.1038/309462a0.

[6] Newman, D.J. and Cragg, G.M. Natural Products as Sources of New Drugs from 1981 to 2014. *Journal of Natural Products*, 79(3):629–661, March 2016. ISSN 0163-3864. doi: 10.1021/acs.jnatprod.5b01055.

[7] Palsson, B. In silico biology through ”omics”. *Nature Biotechnology*, 20(7):649–650, July 2002. ISSN 1087-0156. doi: 10.1038/nbt0702-649.

[8] Rappé, M.S. and Giovannoni, S.J. The uncultured microbial majority. *Annual Review of Microbiology*, 57:369–394, 2003. ISSN 0066-4227. doi: 10.1146/annurev.micro.57.030502.090759.

[9] Wiener, P. Experimental studies on the ecological role of antibiotic production in bacteria. *Evolutionary Ecology*, 10(4):405–421, July 1996. ISSN 0269-7653, 1573-8477. doi: 10.1007/BF01237726.

Appendix A

Chapter 2 Supplement

A1 Supplementary Files

Due to size, these tables are not presented here, but are submitted during peer review. Please correspond with Nishanth Merwin for access.

- **Supplementary table 1: All characterized RiPP structures associated with names and families curated and used in this analysis.**
- **Supplementary table 2: All RiPP post-translational modifications recognized by GRAPE/BARLEY.**
- **Supplementary table 3: Characterized RiPP BGCs with associated families and structures.**
- **Supplementary table 4: List of reactions encoded within BARLEY using PRISM detected genes.**
- **Supplementary table 5: All genera analysed and their associated diversity indices across multiple RiPP families.**
- **Supplementary table 6: List of RiPP post-translational modifications recognized by BARLEY as similar.**

A2 Supplementary Figures

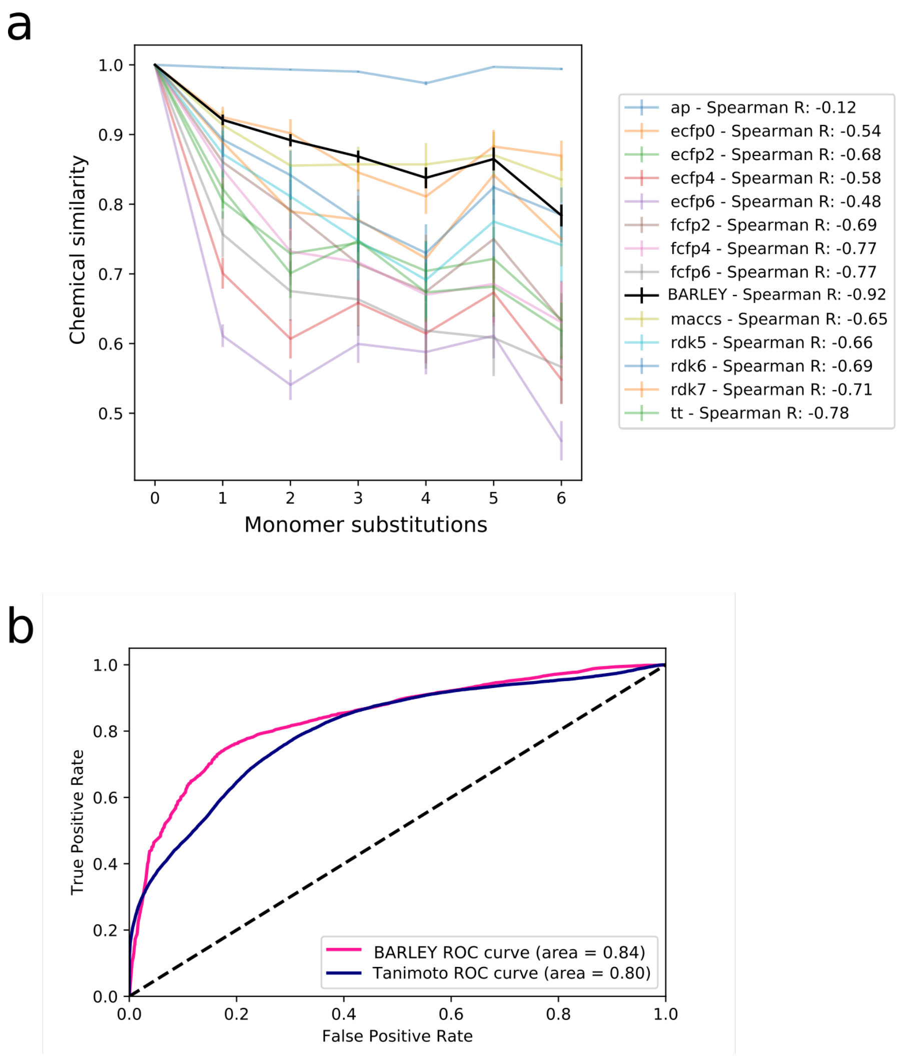

FIGURE A1.1: **Evaluating chemical similarity comparison tools in the context of RiPPs.** (A) BARLEY demonstrates a more consistent ranked relationship across peptide libraries with increasing monomer substitutions. Using LEMONS, a library of theoretical class I lantipeptides were created with randomly substituted monomers. A similarity index was calculated between the derived

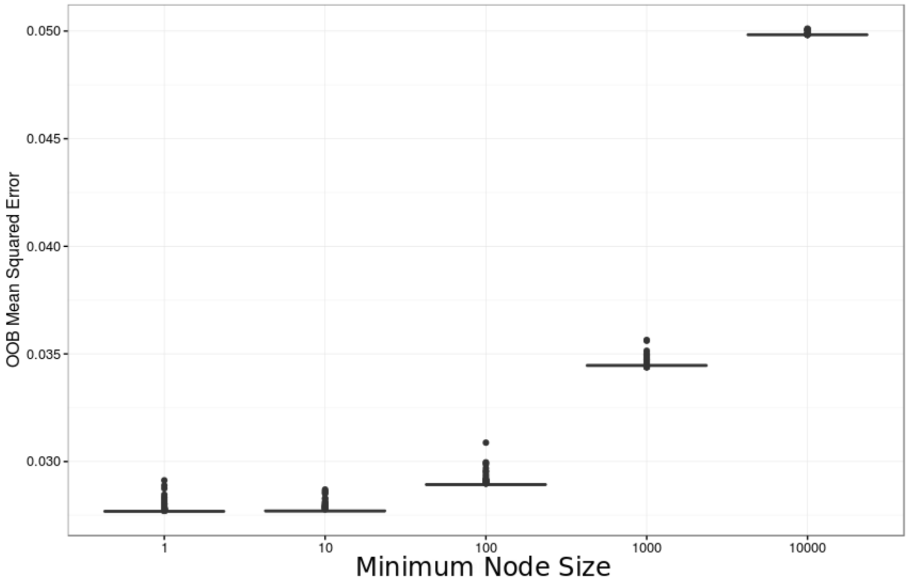

Figure A1.2: **Tuning BARLEY minimum node size.** Minimum node size represents the size of the terminal node in a regression tree where smaller sizes develop deeper and more complex trees. This parameter was tuned using a ten fold cross validation within the traning set.

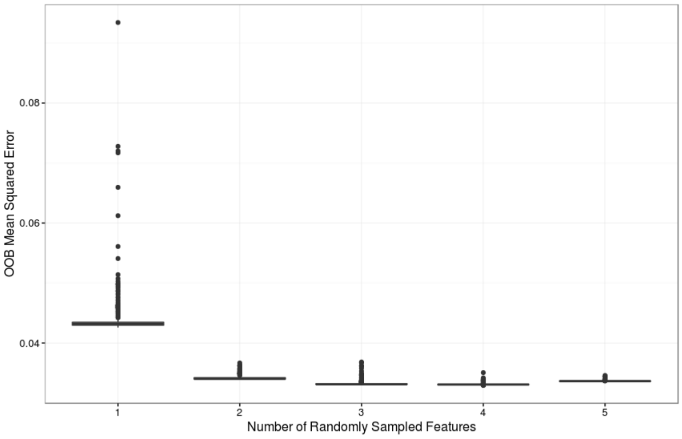

Figure A1.3: **Tuning BARLEY number of randomly sampled features.** In a random forest, each tree can be randomly assigned a set of features to predict upon. Of a total of five features, this model performed optimally when each decision tree was able to access all variables.

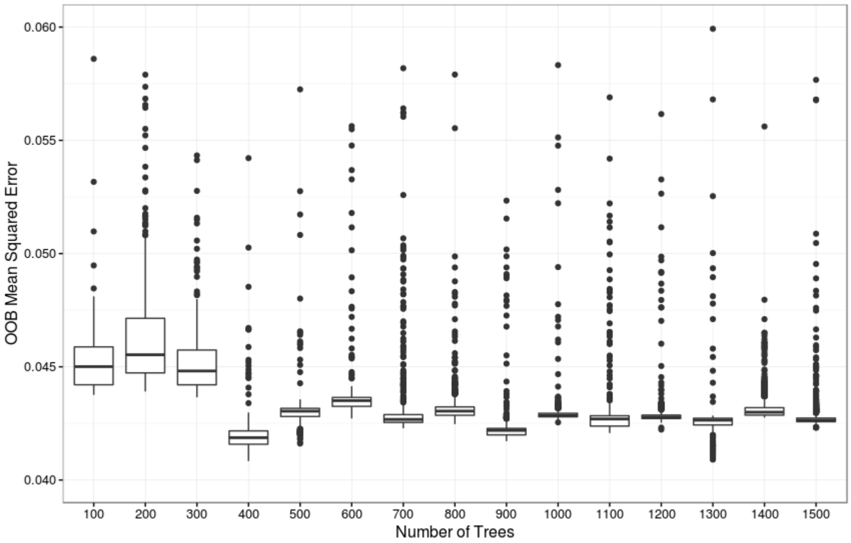

FIGURE A1.4: **Tuning BARLEY number of base estimators.** Random forests are ensemble methods which use many less accurate estimators to generate a consensus. Through iterating through the number of base estimators, it was found that accuracy mostly reaches a minimum around 400 trees.

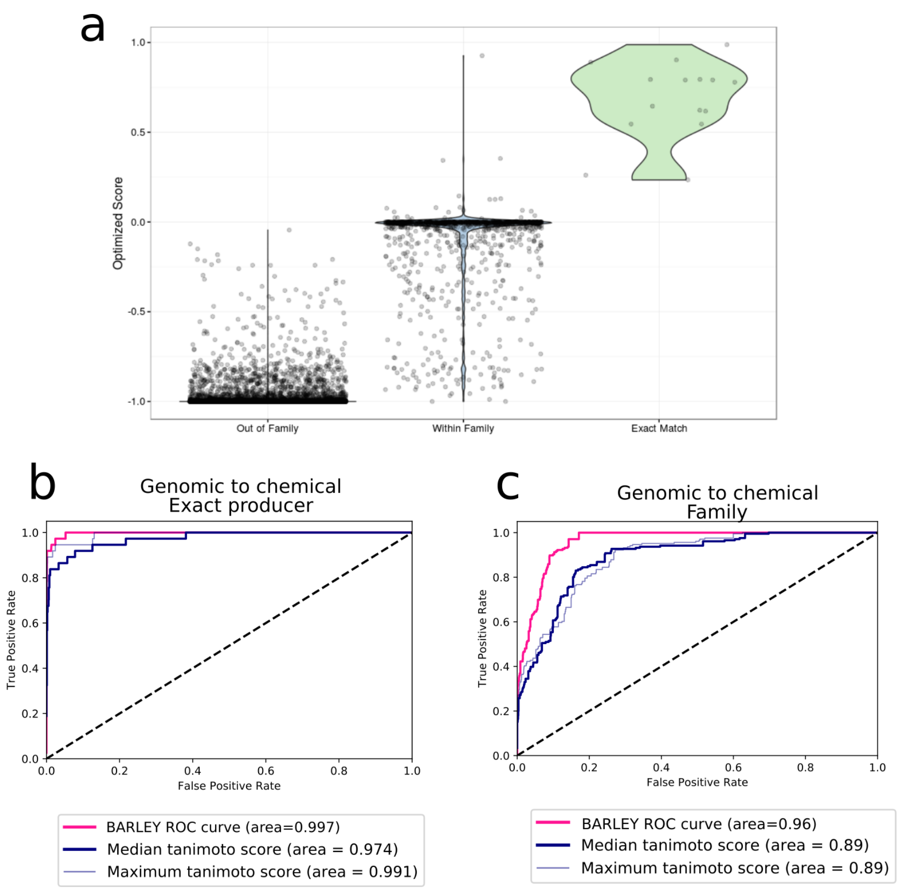

Figure A1.5: **Accuracy of BARLEY novelty index.** (a) A model was trained on 75% of the data and tested on the remaining to classify according to whether a particular comparison between an encoded RiPP and a chemical scaffold was an exact match, within the same family, or from a different family. Using a cutoff of 0.2, there is a 99.7% accuracy in classifying exact matches from other comparison types. In (b) and (c), this same model on the test data was compared to PRISM's structural library using median and maximum ECFP6-Tc. (b) Accuracy in classifying comparisons as exact matches or other. (c) Accuracy in classifying ccomparisons as same-family or other.

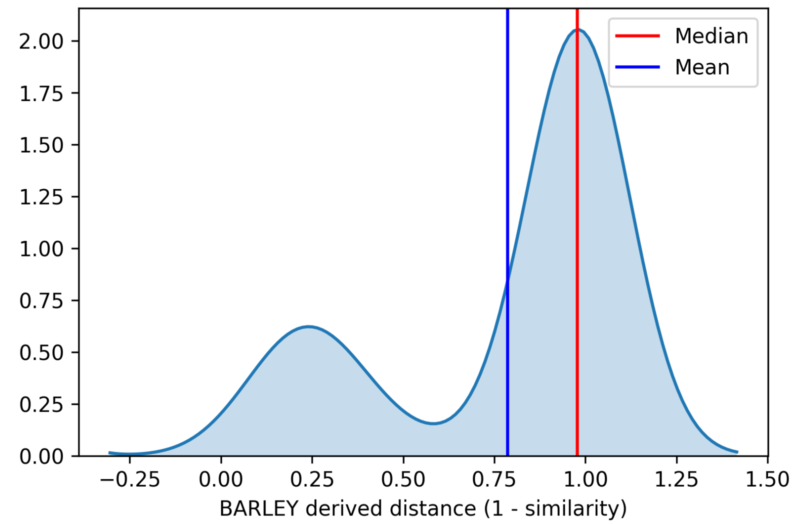

FIGURE A1.6: **Evaluating summary statistics for measuring average distance.** Shown are the distribution of BARLEY derived distances across all genomically encoded RiPPs. Median and mean values are shown in red and blue respectively. In an unbalanced multimodal distribution such as this, the median is heavily weighted towards the dominant mode, while the mean is able to capture the shift associated with minor modes.

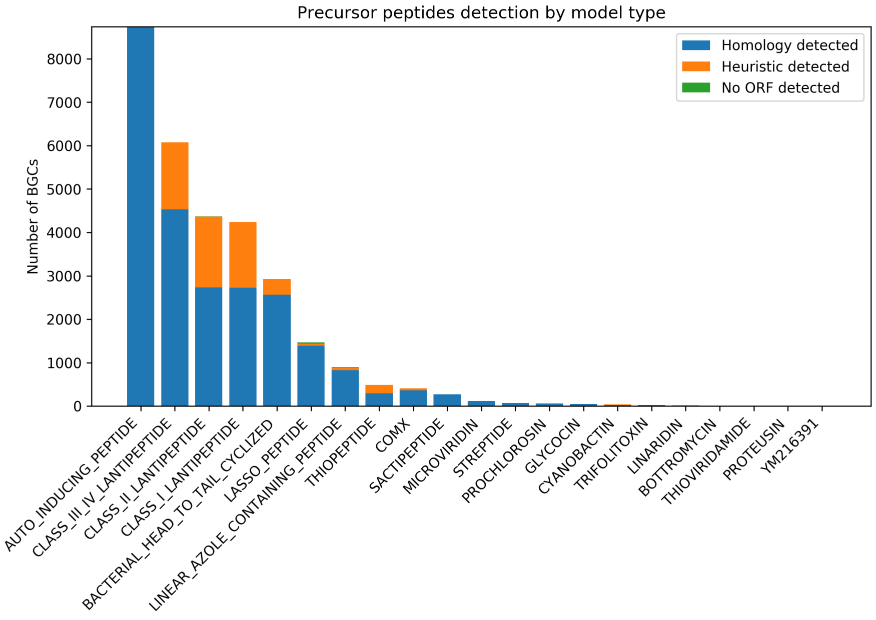

FIGURE A1.7: **Precursor peptide detection type across RiPP families.** Distribution of RiPP BGCs across 65 thousand prokaryotic genomes sorted according to the presence of a precursor peptide, and the model used for detection.

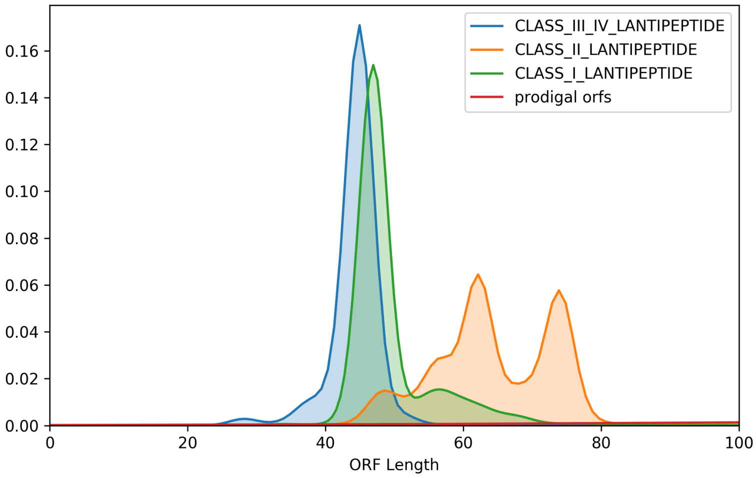

Figure A1.8: **Distribution of lantipeptide precursor peptide lengths.** Precursor peptides ORF sizes are shown here across all detected lantipeptide BGCs with successful motif identification for subsequent cleavage.

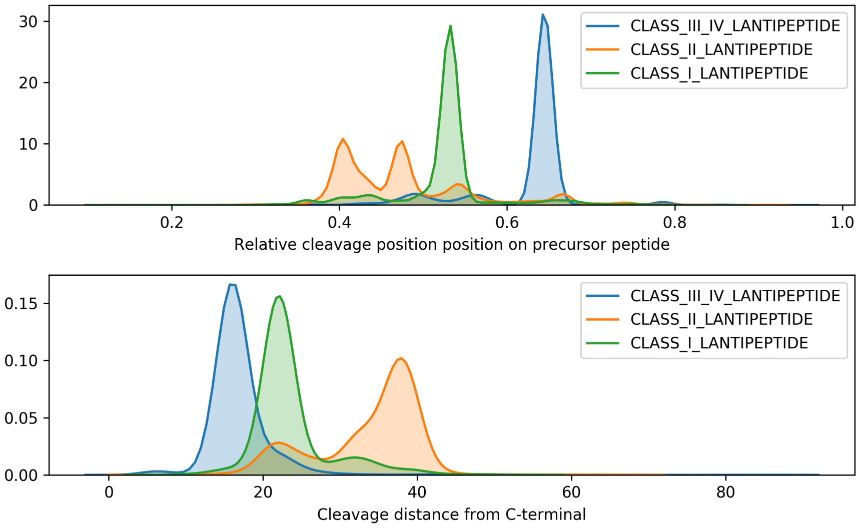

Figure A1.9: **Lantipeptide precursor cleavage sites.** All detected lantipeptide precursors through PRISM with predicted cleavage sites via homology based motifs. (a) Relative position of cleavage site, where x-axis represents the indexed cleaveage position divided by ORF length. (b) Cleavage site as a measure of distance from C-terminal.

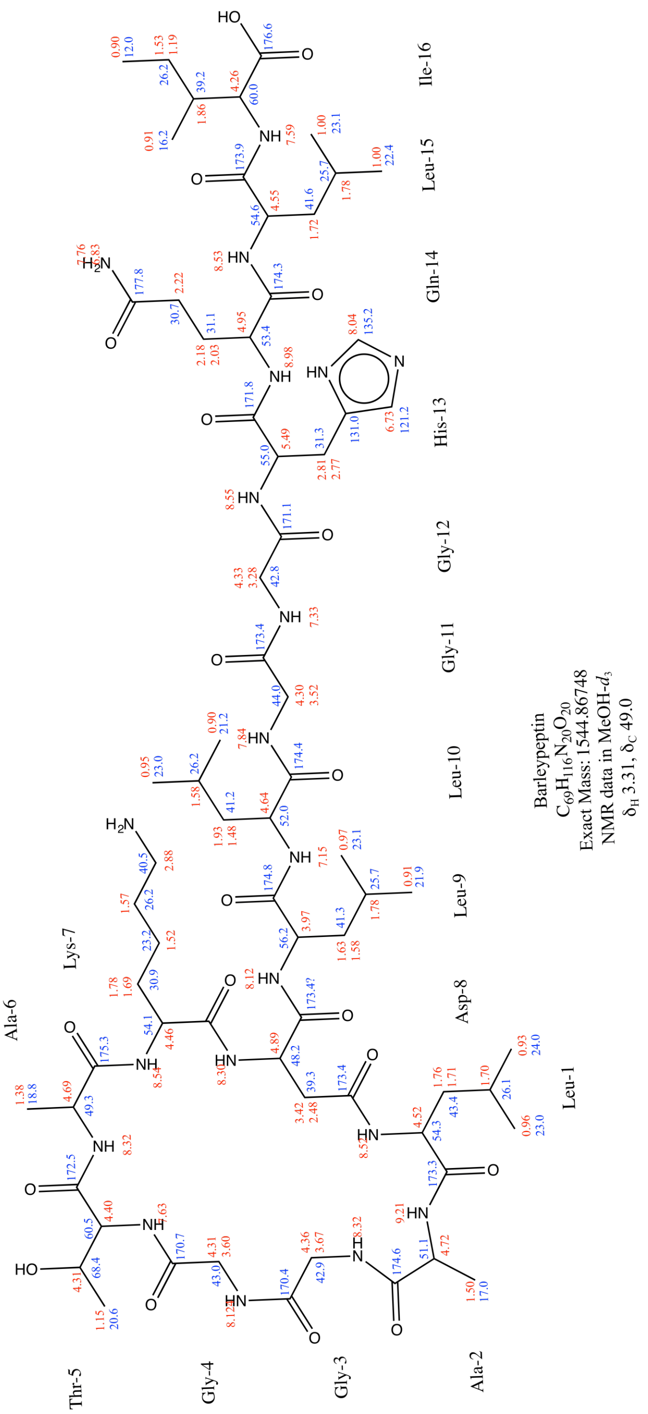

FIGURE A1.10: **Structure of streptopeptin with assigned chemical shifts.** Proton chemical shifts are shown in red while carbon chemical shifts are shown in blue.

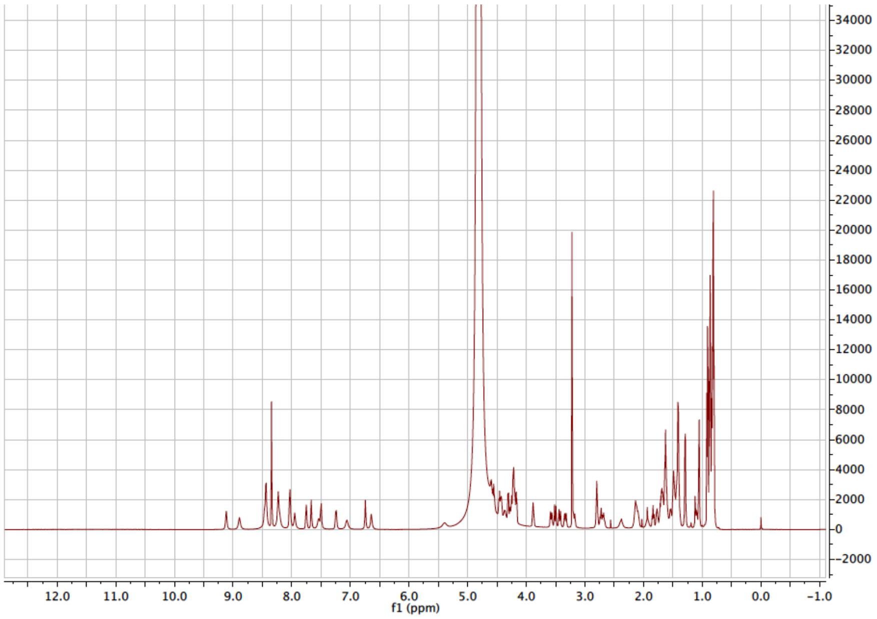

FIGURE A1.11: **^{1}H-NMR spectrum of streptopeptin in methanol-d3.**

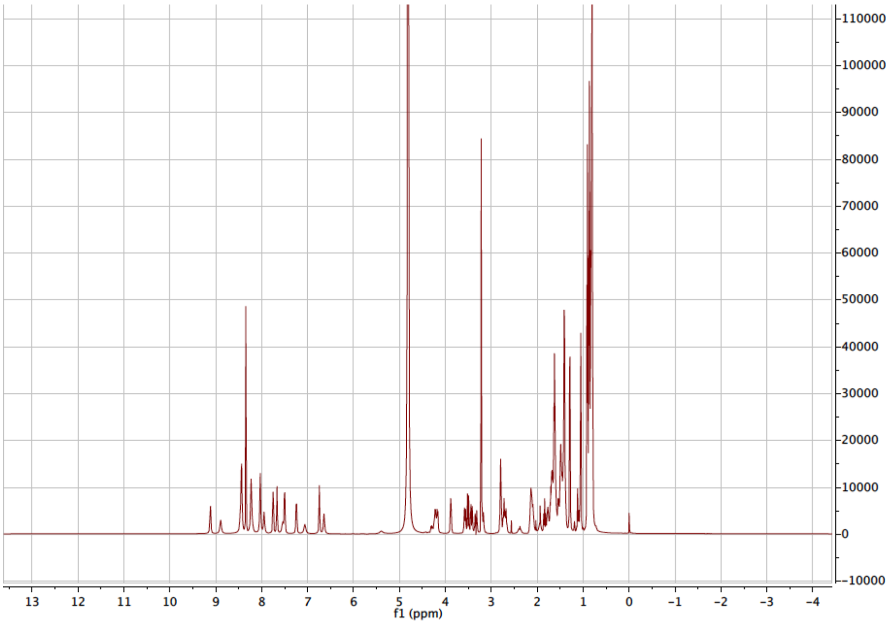

FIGURE A1.12: **^{1}H-NMR (water suppression) spectrum of streptopeptin in methanol-d3.**

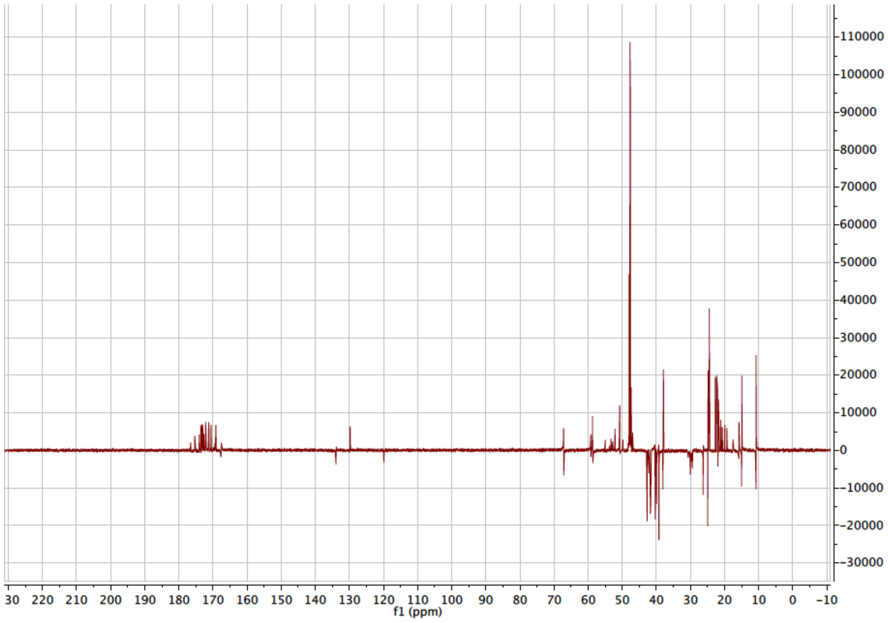

Figure A1.13: **DEPTq spectrum of streptopeptin in methanol-d3.**

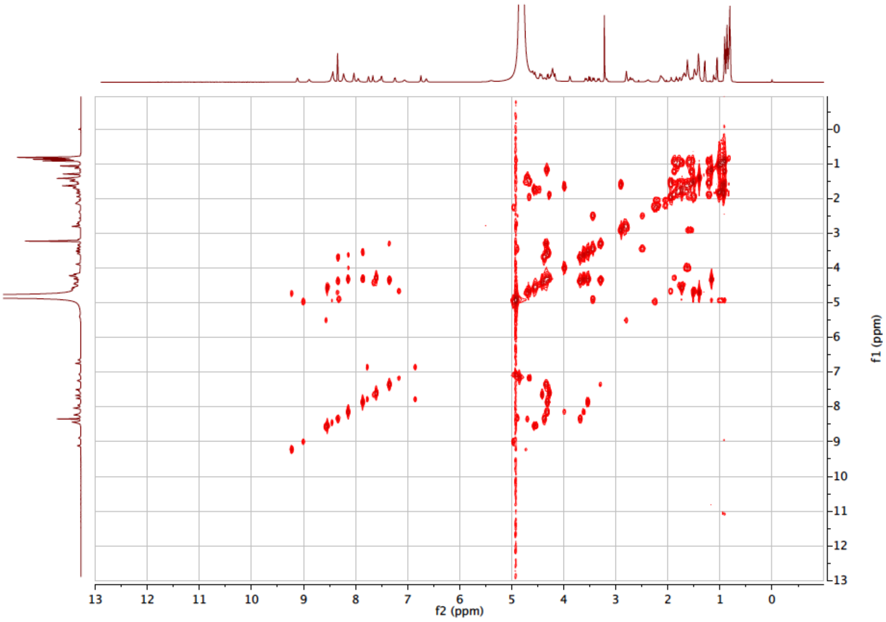

Figure A1.14: **^{1}H-^{1}H COSY spectrum of streptopeptin in methanol-d3.**

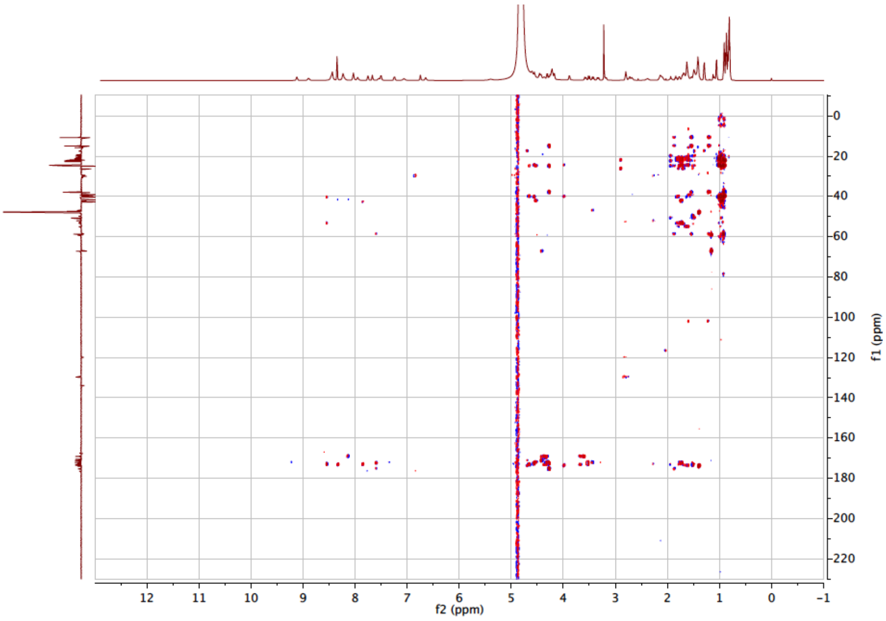

FIGURE A1.15: ^{1}H-^{1}C HMBC spectrum of streptopeptin in methanol-d3.

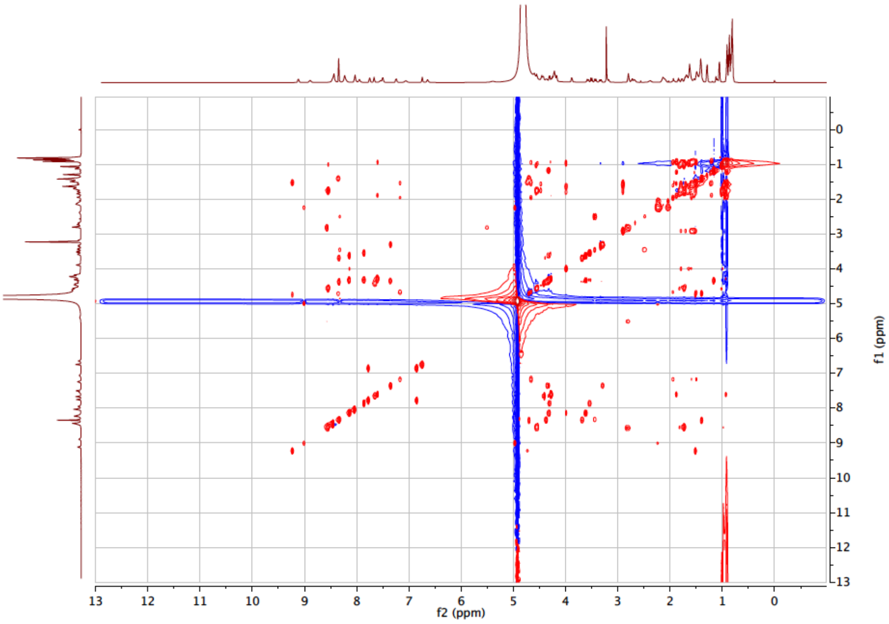

FIGURE A1.16: ^{1}H-^{1}H TOCSY spectrum of streptopeptin in methanol-d3.

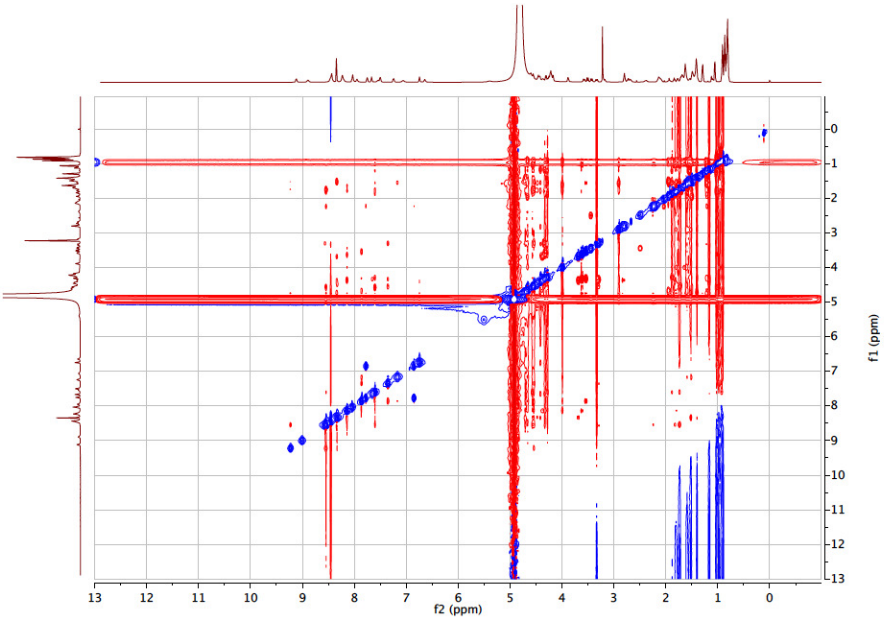

FIGURE A1.17: **^{1}H-^{1}H ROESY spectrum of streptopeptin in methanol-d3.**

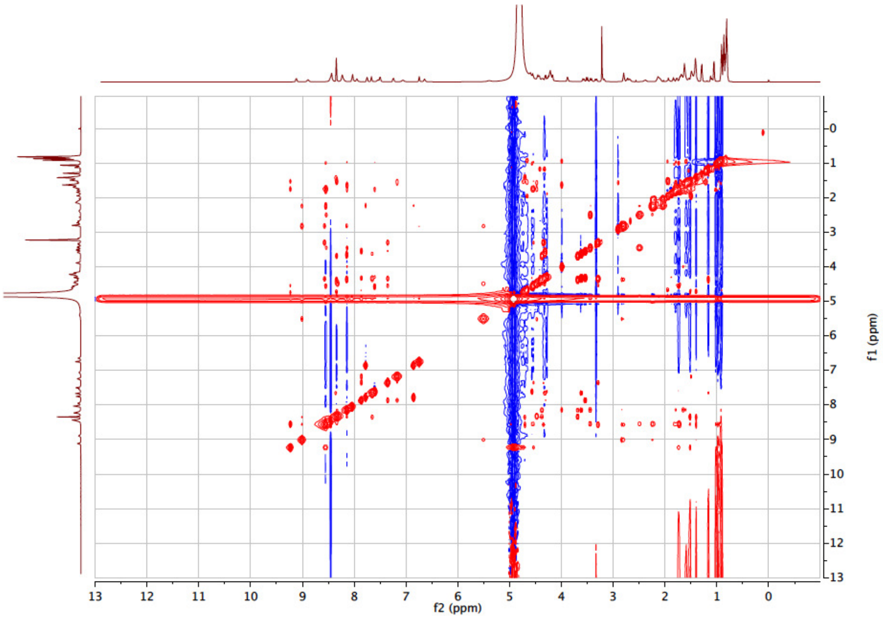

FIGURE A1.18: **^{1}H-^{1}H NOESY spectrum of streptopeptin in methanol-d3.**

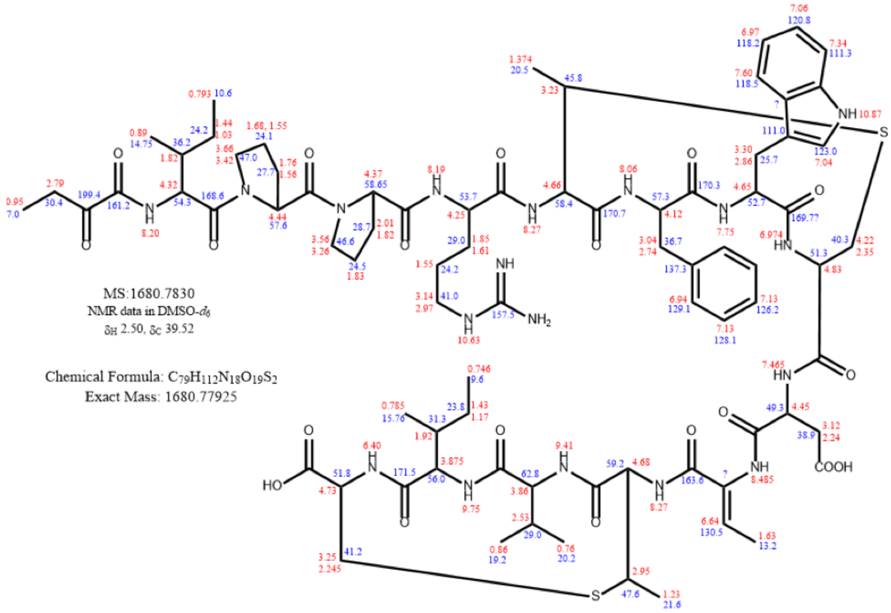

FIGURE A1.19: **Structure of ginsebactin with assigned chemical shifts.** Proton chemical shifts are shown in red while carbon chemical shifts are shown in blue.

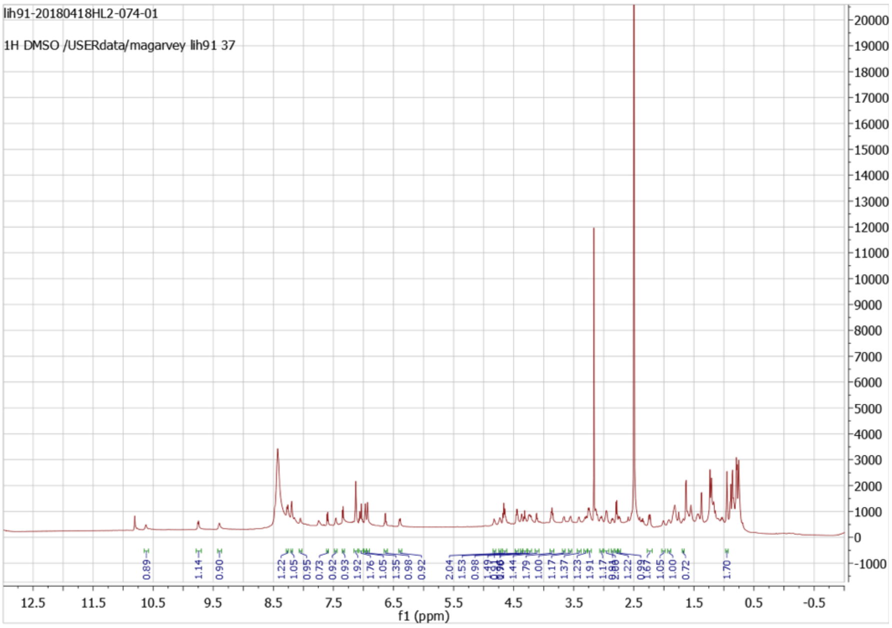

FIGURE A1.20: **^{1}H-NMR spectrum of ginsebactin in DMSO-d6.**

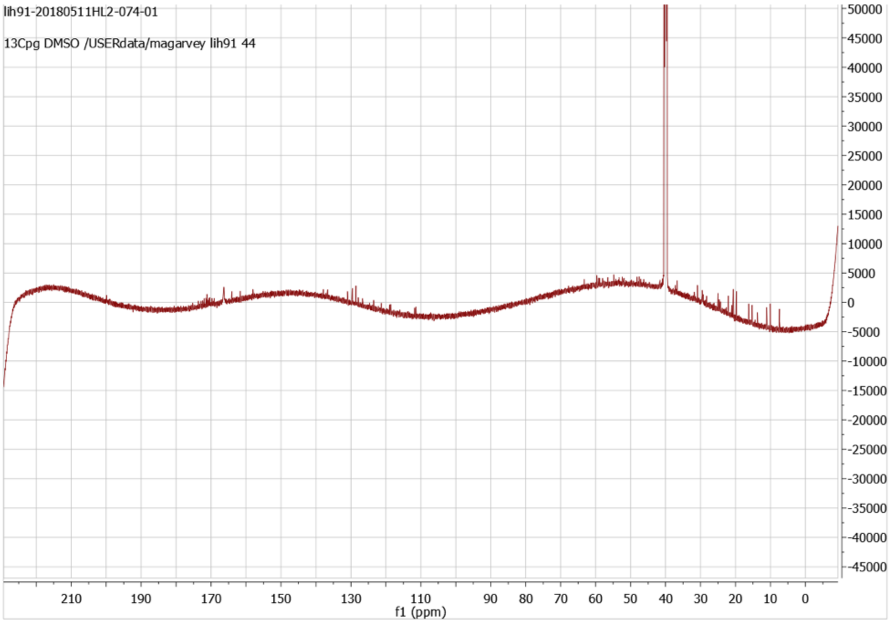

FIGURE A1.21: **^{13}C-NMR spectrum of ginsebactin in DMSO-d6**

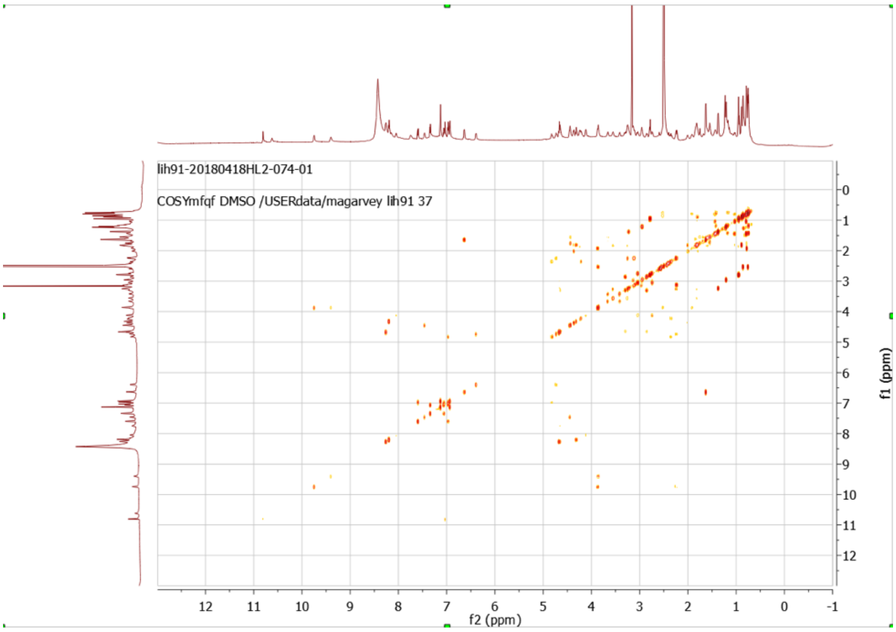

FIGURE A1.22: **^{1}H-^{1}H COSY spectrum of ginsebactin in DMSO-d6.**

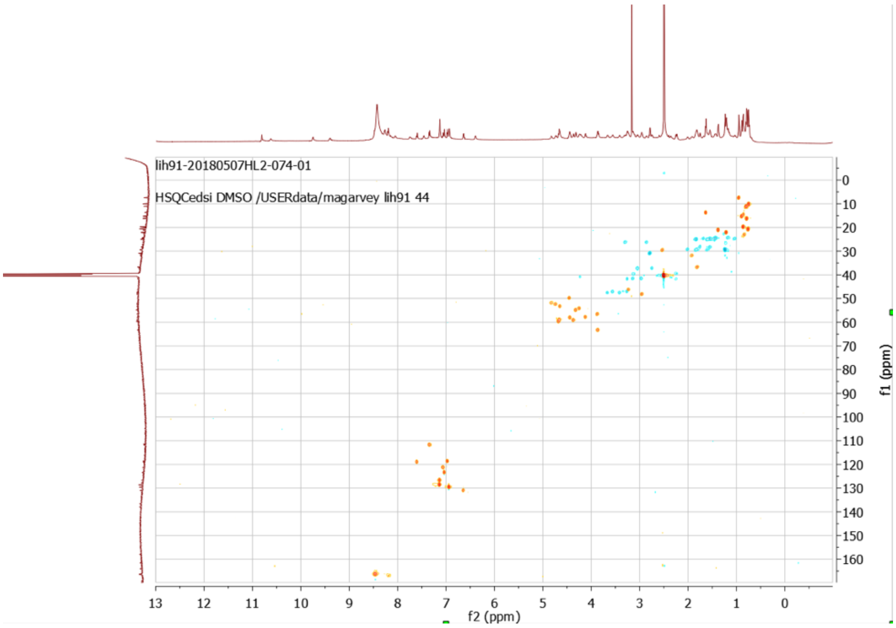

FIGURE A1.23: ^{1}H-^{1}C HSQC spectrum of ginsebactin in DMSO-d6.

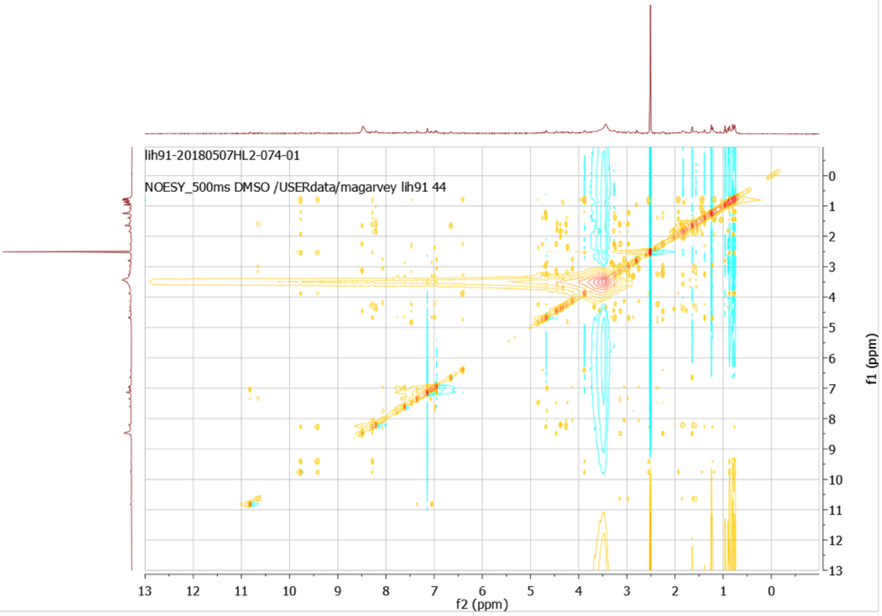

FIGURE A1.24: ^{1}H-^{1}H NOESY spectrum of ginsebactin in DMSO-d6.

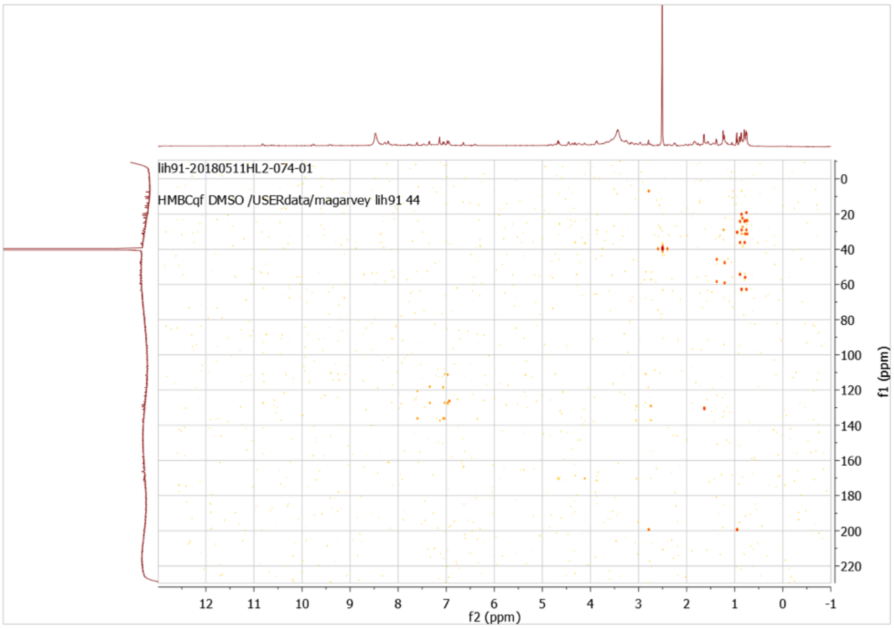

Figure A1.25: **^{1}H-^{13}C HMBC spectrum of ginsebactin in DMSO-d6.**

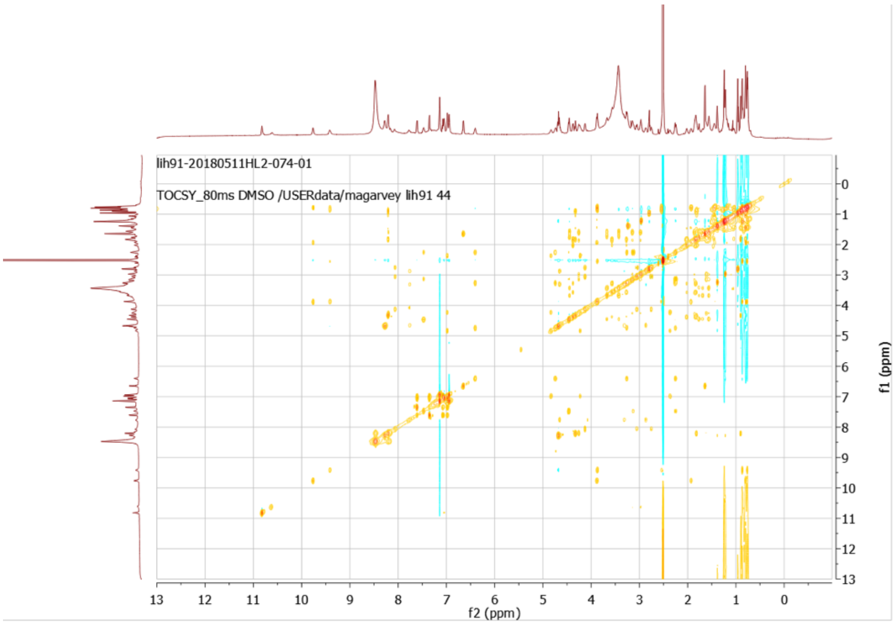

Figure A1.26: **^{1}H-^{1}H TOCSY spectrum of ginsebactin in DMSO-d6.**

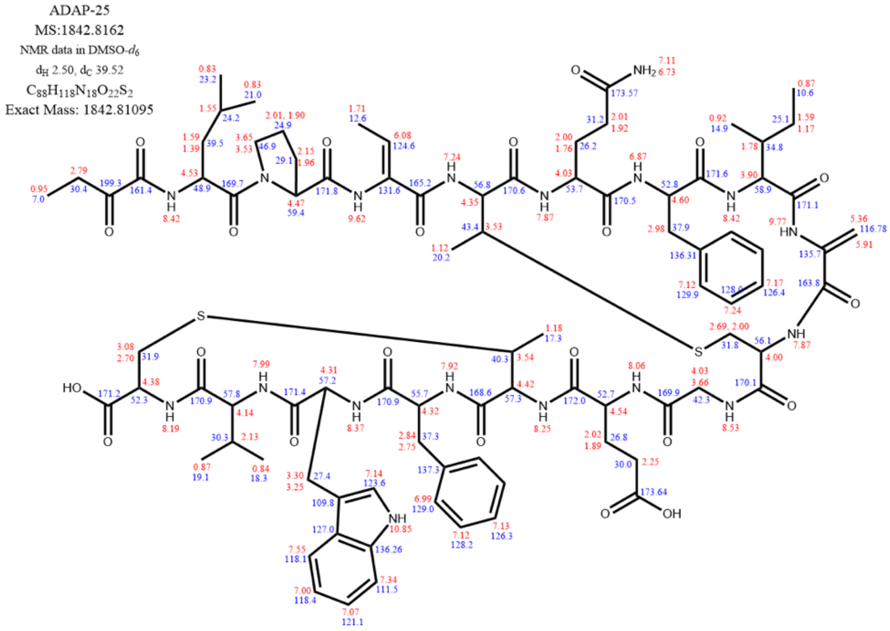

FIGURE A1.27: **Structure of ginsecidin with assigned chemical shifts.** Proton chemical shifts are shown in red while carbon chemical shifts are shown in blue.

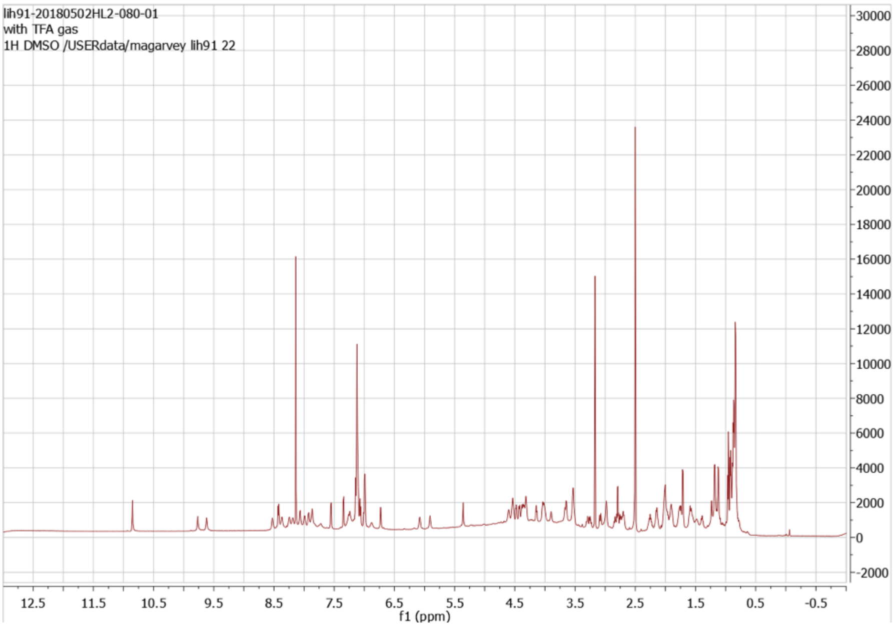

FIGURE A1.28: **^{1}H-NMR spectrum of ginsecidin in DMSO-d6.**

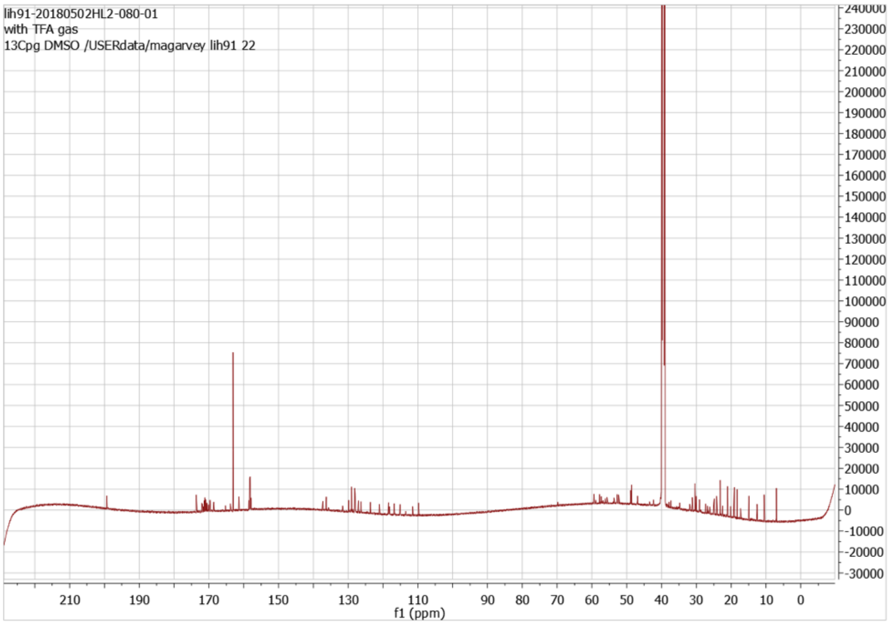

FIGURE A1.29: ^{13}**C-NMR spectrum of ginsecidin in DMSO-d6**

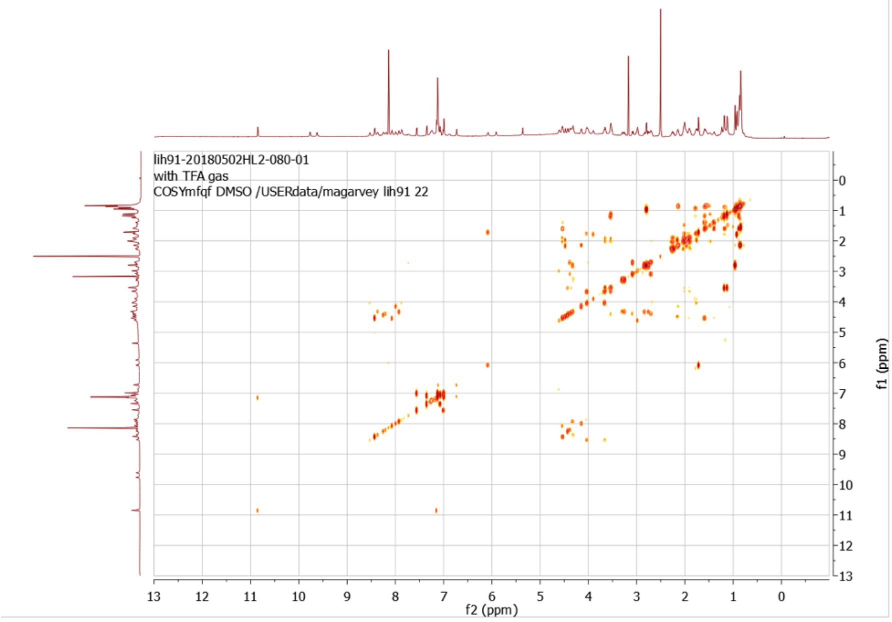

FIGURE A1.30: ^{1}**H-**^{1}**H COSY spectrum of ginsecidin in DMSO-d6.**

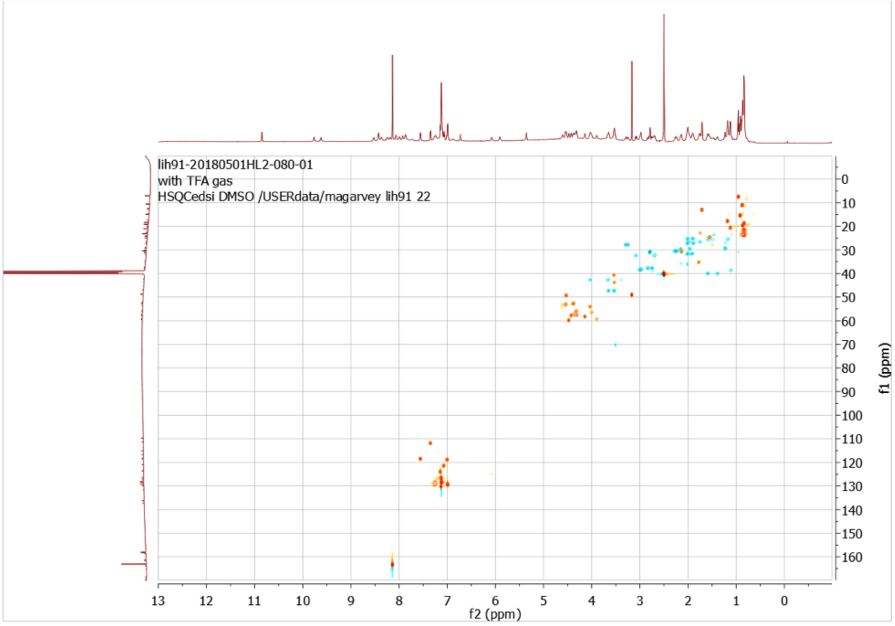

FIGURE A1.31: **^{1}H-^{1}C HSQC spectrum of ginsecidin in DMSO-d6.**

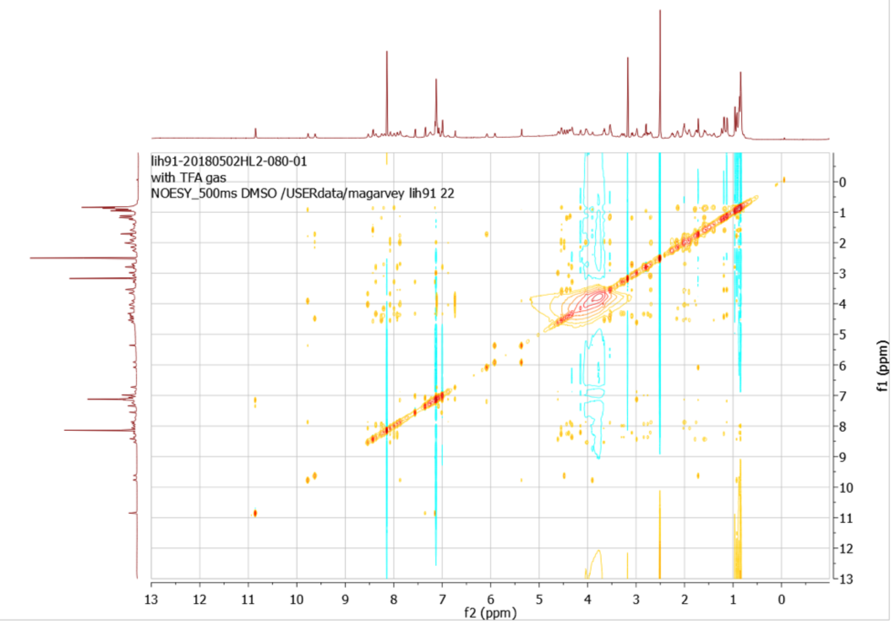

FIGURE A1.32: **^{1}H-^{1}H NOESY spectrum of ginsecidin in DMSO-d6.**

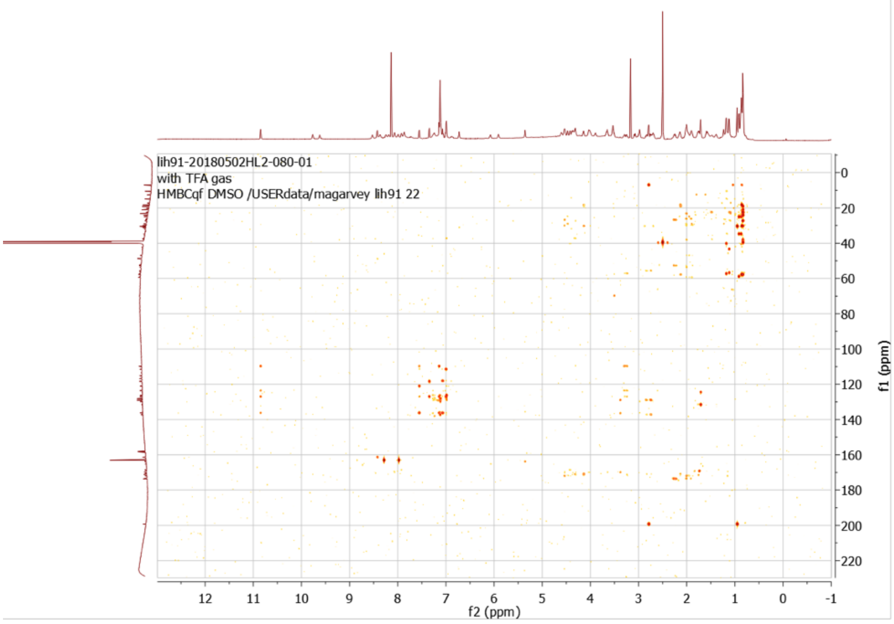

FIGURE A1.33: **^{1}H-^{13}C HMBC spectrum of ginsecidin in DMSO-d6.**

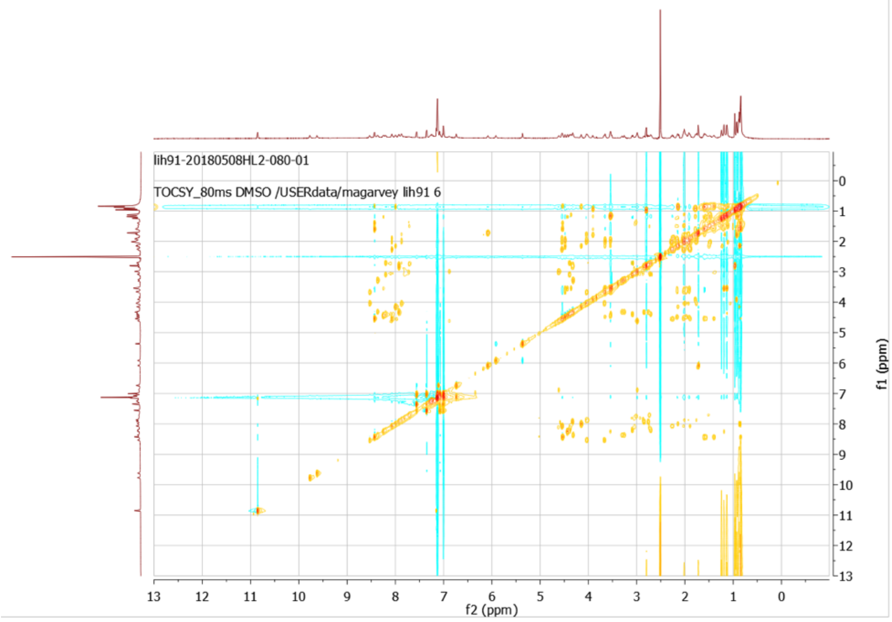

FIGURE A1.34: **^{1}H-^{1}H TOCSY spectrum of ginsecidin in DMSO-d6.**

Appendix B

Chapter 3 Supplement

A1 Supplementary Files

Due to size, these tables are not presented here, but are submitted during peer review. Please correspond with Nishanth Merwin for access.

- **Supplementary table 1: Reference dataset of curated bacteriocins.**
- **Supplementary table 2: pHMMs generated from reference dataset.**
- **Supplementary table 3: All predicted peptides, associated organism, and isolation site.**
- **Supplementary table 4: Microbiome predicted peptides, associated organism, and isolation site and BARLEY score to reference dataset.**
- **Supplementary table 5: Peptide families exclusive to human microbiome.**
- **Supplementary table 6: Pairwise correlations of housekeeping genes.**
- **Supplementary table 7: Pairwise correlations of known antimicrobial peptides.**
- **Supplementary table 8: All confident antagonistic relationships among peptides unique to the human microbiome.**

A2 Supplementary Figures

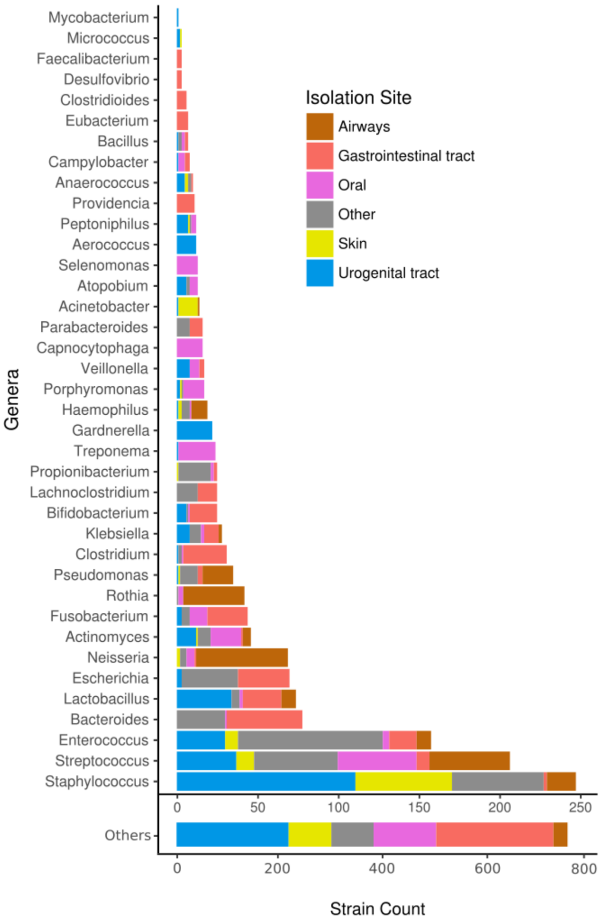

FIGURE A2.1: **Distribution of collected human microbiome genomes across genera and body sites.** (A) A total of 2274 genomes were collected and annotated belonging to 17 body sites, represented here according to colour where other is comprised of bacteria isolated from wounds, blood, nose, bone, eye, spinal cord, brain, ear, head, abdomen, limb and liver samples. The top 40 most sequenced genera are represented here, with the remaining 161 genera represented as “Others”.

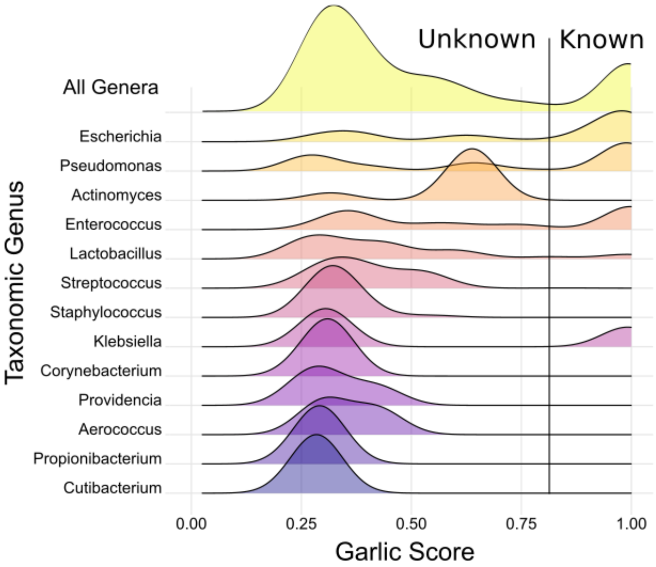

FIGURE A2.2: **Novelty distribution of peptides among enriched genera.** All genera shown at top demonstrate a bimodal distribution, where characterized peptides typically fall above a sequence similarity threshold of 0.85. These genera were sorted according to the median sequence similarity score.

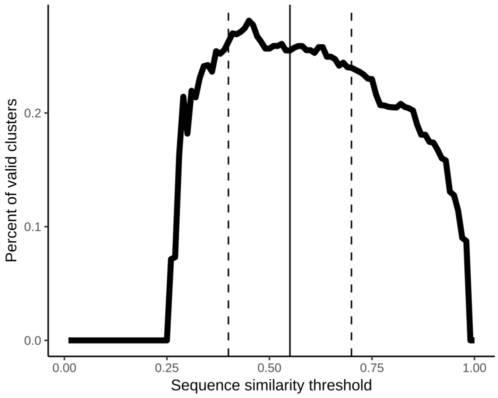

FIGURE A2.3: **Determining the optimal sequence similarity threshold for peptide clustering.** All threshold values were tested between 0 and 1 at 0.01 intervals, and the percent of valid clusters observed were recorded. A stable flat region in this distribution was observed between 0.45 and 0.7 (dashed lines). As such, the midpoint, representing a threshold of 0.55 was chosen as the sequence similarity threshold when identifying peptide families.

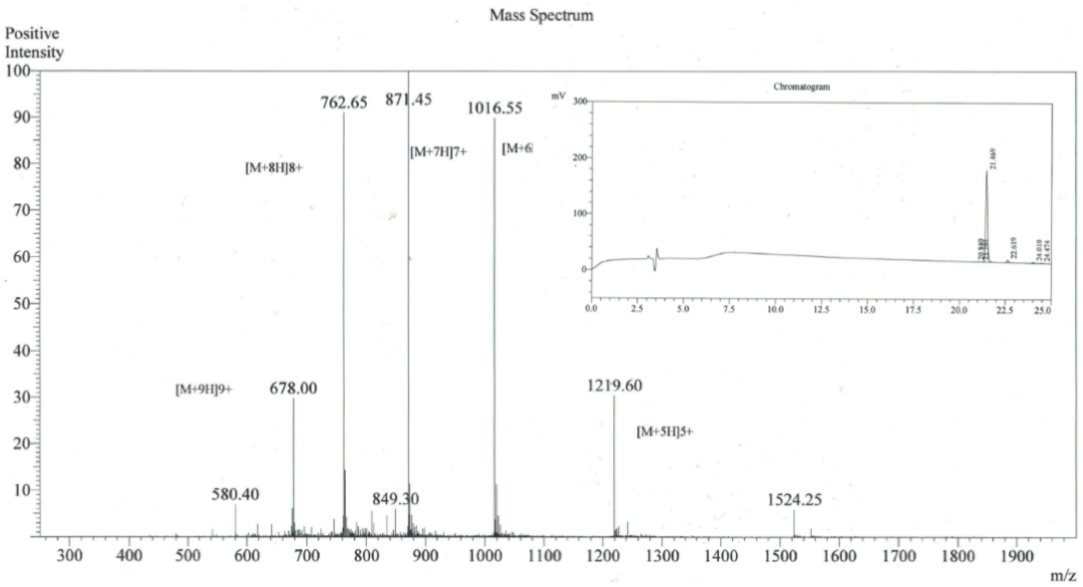

FIGURE A2.4: **LC/MS/MS chromatogram of scardovicin.** Scardovicin is a 53mer peptide with an exact mass of: 6088.437 Da, and sequence: MGAFFRLLSILARYGARAVQWAWAHRGTVLR-WIGAGQAIDWVIKQIKRLLGIR.

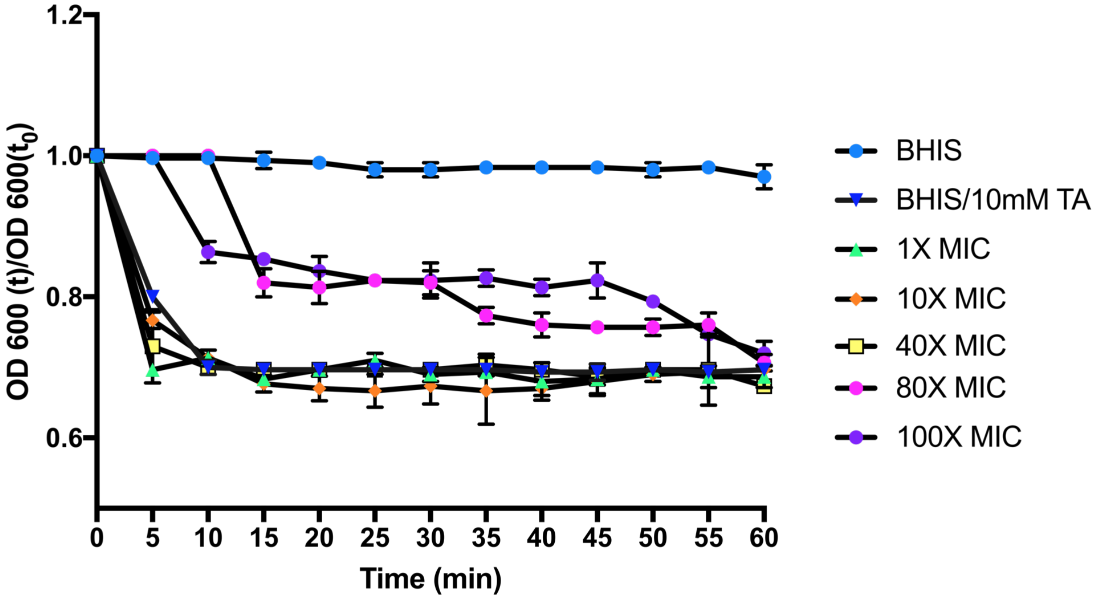

FIGURE A2.5: **Effect of scardovicin on initiation of spore germination in *C. diffcile.*** Shown is graphical representation of OD_{600} (t)/OD_{600} (t_0) of purified spore suspension treated with different folds of scardovicin MIC concentration and measured over time points for a total period of 60 min. Data points represent mean of 3 independent biological replicates while error bars represent standard error of the mean.

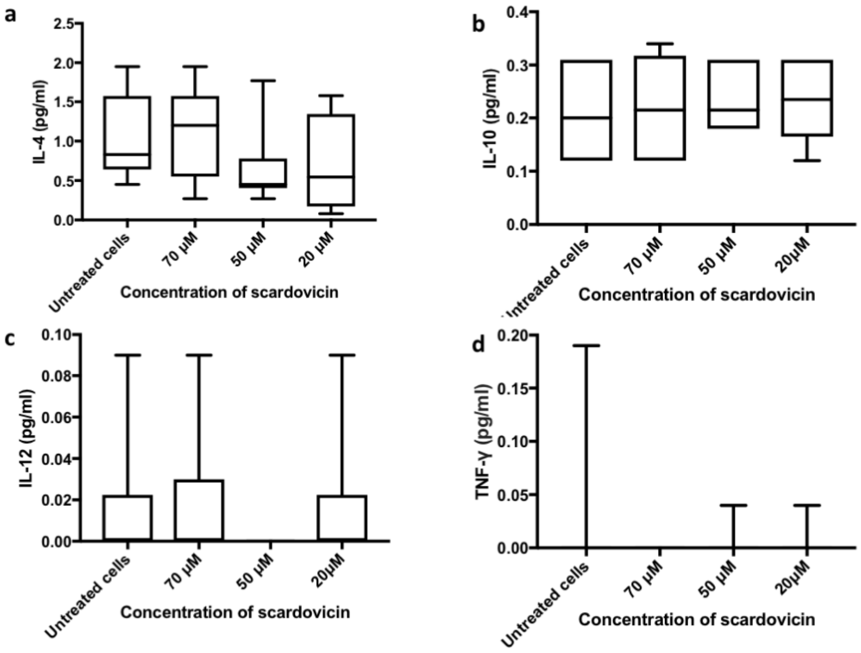

Figure A2.6: **Immunomodulatory activity of scardovicin on HT-29 cell line.** a-d, graphs show activity of three concentration of scardovicin (70, 50 and 20 µM) on IL-4, IL-10, IL-12, and TNF-γ, respectively. Whiskers represent the range of data points of six independent biological replicates while error bars indicate the standard error of the mean. Data were analyzed using one-way ANOVA test and t-test. No significant difference was found between all treatments.