Metagenomic exploration of *Mycale hentscheli* microbiome secondary metabolite biosynthesis.

By

Mathew Ambrose Storey

A thesis submitted to the Victoria University of Wellington in fulfilment of the requirements for the degree of Doctor of Philosophy In Biotechnology

Victoria University of Wellington

(2022)

Abstract

The natural environment is replete with potent bioactive organic molecules that are produced by the microorganisms that inhabit it. These so-called secondary metabolites provide niche advantages to the producers and have been exploited by mankind for their medicinal and industrial value, especially as antibiotic and antiproliferative agents. One particular environment that has proven to be a rich source of secondary metabolites are certain species of marine sponge, or more specifically, their microbial symbionts. The New Zealand marine sponge *Mycale hentscheli* is the source of at least three secondary metabolite like molecules, pateamine, peloruside and mycalamide which have spurred much research interest due to their potent bioactivity. However, the exact origin and genetic factors responsible for the production of these molecules remains unknown.

In this study undertake a metagenomic drug discovery campaign employing multiple sequencing strategies to reveal that these compounds are produced by multiple members of the sponge's symbiont microbiome. This is in contrast to previously studied sponges where the secondary metabolites are produced by a single species of microorganism. By analysing the biosynthetic gene clusters in their wider genomic context we found the production of the secondary metabolites were achieved by canonical and non-canonical biosynthetic mechanisms. This approach also shed light on an additional repertoire of biosynthetic potential within this already secondary metabolite rich sponge and provided the complete genomes of novel taxa.

In addition to furthering our understanding of secondary metabolite biosynthesis, we anticipate that the elucidation of the genetic factors responsible for producing the secondary metabolites of *M. hentscheli* will foster research into their sustainable production by contributing to advances in synthetic biology, heterologous expression and culturing technologies.

Dedication

I would to dedicate this work to two people that will always be beloved to me. In memory of my late father Ross Alastair Storey and my late grandmother Noelene Mary Barnett, who both passed this year while I was writing this thesis. Thank you for your unceasing care, encouragement and sacrifices that will continue to benefit me forever. I'm sorry I couldn't finish this before you could see it, but I know you would be proud.

Acknowledgements

I would also like to sincerely thank and acknowledge the large team of people who have made this research project not only possible but a rewarding and valuable experience.

Thank you so much to my primary supervisors, Dr Jeremy Owen and Prof David Ackerley, being a recipient of years of your expertise, wisdom and patience has been a privilege for which I am truly grateful. Thank you to Prof John Miller, who I have always admired, your passion and encouragement has always been immensely influential.

I would like to heartfeltly acknowledge the Faculty of Science and School of Biological Sciences staff members who went above and beyond to help me navigate the many processes of academia and keep everyone pointing in the right direction. I would like to specifically mention a few individuals that I will always remember and be indebted to. The always amazing Mary Murry, your smile brightened my day so many times. The incredible Patricia Stein for you unfaltering guidance and kindness. The brilliant Charlotte Ansel who I would not have survived without. None of this would be possible without you and your teams.

I had the pleasure of always being surrounded by the highest calibre of lab mates, fellow students and post docs alike. There are far too many to mention, so thankyou to all the members of the Owen and Ackerley labs groups and the Metamax team. You have gifted me with so many fond memories and you all really taught me everything I know. A special thanks to Assoc. Prof Robert Keyzers and the members of his lab group who graciously assisted with chemical analysis and samples.

Finally, a huge thank you to the Cancer Society of New Zealand for the PhD training scholarship that partially funded this work.

Table of Contents

1 Introduction 1

1.1 Underpinning principles of metagenomic drug discovery 1

1.1.1. 'Microbial Dark Matter' 2

1.1.2. Bacterial bioactive natural products as pharmaceuticals 3

1.1.3. Biosynthetic gene clusters 5

1.1.4. A platform for metagenomic based drug discovery. 6

1.2 Secondary metabolite biosynthesis 9

1.2.1. Polyketides 9

1.2.2. NRPS 16

1.2.3. Tailoring modifications 19

1.2.4. Expanded biosynthetic diversity 21

1.3 Microbial genome mining 21

1.3.1. AntiSMASH 23

1.3.2. Genome mining summary 27

1.4 Sponge derived secondary metabolites 28

1.4.1. Drugs from sponges 28

1.4.2. Mycale hentscheli secondary metabolites 30

1.5 Aims and objectives 34

1.6 Components of the research and contributions to knowledge 35

1.6.1. Components of research 35

1.6.2. Components of research 36

1.6.3. Contributions to knowledge 36

2 Methods and Materials 38

2.1 Sponge sample collection 38

2.1.1. Sponge subsampling 38

2.2 Microbiological methods 39

2.2.1. Laboratory *Escherichia coli* strains used in this study are detailed in Table 2-1. 40

2.2.2. Plasmid vectors used in this study are detailed in Table 2-2. 40

2.2.3. Growth media 40

2.2.4. Media supplements 40

2.2.5. Microbial growth and maintenance 41

2.2.6. Electroporation 41

2.3 Molecular Biology methods 41
2.3.1. Oligonucleotide primers 41
Primers used in this study are detailed below (Table 2-3) 41
2.3.2. Plasmid and cosmid DNA extraction and isolation 41
2.3.3. High molecular weight (HMW) DNA extraction and isolation 42
2.3.4. Commercial enzymes 44
2.3.5. Commercial phage packaging extracts 44
2.3.6. Magnetic bead DNA purifications 44
2.4 Metagenomic cosmid library construction 45
2.4.1. Lab-made phage packaging extract preparation 45
2.4.2. pWEB::TNC vector preparation 46
2.4.3. Cosmid library packaging 47
2.5 DNA Sequencing 48
2.5.1. Illumina 48
2.5.2. Pacbio 48
2.5.3. ONT 48
2.5.4. Sanger sequencing 48
2.6 Bioinformatic analysis 48
2.6.1. Read trimming, decontamination, error correction and read-merging. 49
2.6.2. Short read assembly 49
2.6.3. Hybrid assembly 50
2.6.4. Read mapping 50
2.6.5. Taxonomy 50
2.7 Chemotyping 51
3 Marine sponge metagenomic clone library construction and validation. 52
3.1 Introduction 52
3.1.1. Metagenome library functional screening 53
3.1.2. Metagenome library sequence homology screening 54
***3.2* Chemotypes of *M. hentscheli* 56**
3.2.1. Sponge chemotypes 57
3.3 Extraction of HMW-DNA from *M. hentscheli* 58
3.3.1. HMW-DNA extraction results 60
3.4 Building the *M. hentscheli* metagenomic large-insert library 61
3.4.1. Trial library building results. 62

3.4.2. Commercial and 'lab-made' phage packaging extracts 63
3.4.3. pWEB vector preparation 64
3.4.4. HMW-DNA cloning and packaging 64
3.4.5. Insert validation 65
3.4.6. Direct sequencing of the 48 test cosmid pool 69
3.4.7. Assembly and analysis of the SpA_1-48 cosmid pool sequence data 71
3.4.8. Metagenomic library construction summary. 74

4 Direct metagenomic sequencing of M. hentscheli 75

4.1. Direct metagenomic sequencing introduction 75

4.2. *M. hentscheli* direct metagenomic sequencing pilot 77

4.2.1. *M. hentscheli* Illumina paired-end sequencing. 77
4.2.2. Illumina short-read *M. hentscheli* metagenomic assembly. 78
4.2.3. Metagenomic assembly sequence mining for mycalamide BGC. 79
4.2.4. BGC mining of the short-read *M. hentscheli* metagenome assembly 81
4.2.5. Metagenome binning of the short-read *M. hentscheli* assembly 83
Binning approaches 83
4.2.6. Assembly plotting 84

4.3. Long-read sequencing of *M. hentscheli* 86

4.3.1. *M. hentscheli* PacBio long-read sequencing 87
4.3.2. Hybrid *M. hentscheli* metagenomic assembly 89

4.4. Annotation of Pat and Myc BGCs contigs 92

4.4.1. Annotation of a complete Pateamine BGC 92
4.4.2. Annotation of a complete Mycalamide BGC. 95

4.5. Sequence mining the hybrid *M. hentscheli* metagenome assembly. 96

4.5.1. Additional BGCs 97

4.6. *M. hentscheli* direct metagenomic sequencing summary 98

5 Peloruside biosynthetic gene cluster exploration 100

5.1 Introduction 100

5.2 Peloruside positive sponge samples 101

5.3 DNA extraction and sequencing 102

5.3.1. Additional sponge sample sequencing results. 102

5.4 MH_PEL metagenomic cosmid library attempt. 103

5.5 Metagenome assemblies 104
5.5.1. Data pre-processing 104
5.5.2. Initial MH_PEL assemblies. 104
5.5.3. Co-assembly 105
5.5.4. Pel BGC origin 108
5.6 Summary 109
6 M. hentscheli symbiont secondary metabolite biosynthesis 110
6.1 Introduction 110
6.2 Mycalamide biosynthetic model 111
6.2.1. Myc gene cluster description 111
6.2.2. Myc biosynthetic mechanism 112
6.3 Pateamine biosynthetic model 113
6.3.1. Pat gene cluster description 113
5.1.1. Pat biosynthetic mechanism 113
6.4 Peloruside biosynthetic model 115
6.4.1. Pel gene cluster description 116
6.4.2. Pel biosynthetic mechanism 116
6.5 A new polytheonamide-like gene cluster 118
6.6 Summary 120
7 M. hentscheli microbiome binning and taxonomy 122
7.1 Introduction 122
7.1.1. Binning of sponge metagenomes 122
7.1.2. Ensemble binning approach 123
7.2 Binning Results 124
7.2.1. Raw binning 124
7.2.2. MAG quality filtering 124
7.2.3. MAG dereplication 125
7.3 Recovered MAGs 125
7.3.1. MAG assembly statistics 125
7.3.2. MAG Taxonomic classifications 128
7.4 MAG abundance profiles and comparison of samples. 128
7.4.1. Core microbiome 130

7.4.2. Biosynthetic MAGs proposed nomenclature 131
7.4.3. Phylum-level microbiome composition 132
7.5 Summary 133
8 Summary, conclusions and future directions 135
8.1 Research motivation 135
8.2 Advancing field 136
8.3 Key findings 138
8.3.1. Metagenomic clone library survey. 138
8.3.2. Direct metagenomic sequencing. 139
8.3.3. Multiple bacterial producers. 139
8.3.4. Peloruside BGC recovery. 140
8.3.5. Biosynthetic mechanisms 141
8.3.6. Core microbiome 141
8.4 Future directions 142
8.5 Data availability 143
8.6 Concluding remarks 143
9 References 145

1 Introduction

1.1 Underpinning principles of metagenomic drug discovery

Bacteria have been our best source of antibiotics, yet the vast majority cannot be grown in a laboratory setting, metagenomics allows access to the uncultivated majority for drug discovery.[1–6] Metagenomic studies aim to explore the entire collection of microbial genomes present in an environmental sample by directly sequencing and/or cloning the total genetic material extracted from the sample (environmental DNA - eDNA).[7,8] Metagenomic based drug discovery methods are proving to be a potential solution to the gradual decline of output from traditional methods of microbial-natural-product discovery.[9–12] Microbial natural products have provided almost all of our antimicrobial therapies and many effective anticancer chemotherapies in clinical use.[13–15] Traditional discovery approaches involved screening lab cultivated strains of bacteria for the production of bioactive molecules. Importantly, the employment of this approach led to discoveries that proved the exceptional value of microbially derived natural products as beneficial medicinal compounds. However, over time the returns from this method have diminished, due to some inherent limitations of the method, namely the requirement for cultivation, low throughput capacity and increasing rates of compound rediscovery.[16–18] In light of the diminishing returns caused by these limitations, the research and investment into microbial natural product discovery using such methods stagnated for some time.[9,15,19] More recently, renewed interest in discovering new microbial natural products has been garnered by a frightening increase in occurrences of antimicrobial resistance to even our “last resort” therapies, threatening modern health care as we know it. Tackling this issue requires new strategies that leverage the advances in our understanding of the microbial world, enabled by modern sequencing technologies, to overcome the limitations of traditional discovery methods. Metagenomics based drug discovery is one such strategy.[11,20–23]

The motivations and practices of metagenomic drug discovery campaigns are based on a foundational set of core principles and assumptions within modern microbiology and microbial genetics, which are discussed below. A growing understanding of the implications

and interplay of these principles has spurred a renaissance of natural product drug discovery from a vast and previously inaccessible resource, the so-called 'microbial dark matter'. The following sections will introduce the overarching concepts that underpin metagenomic drug discovery in order to frame the aims and design of this study and give the reader a "lay of the land" for metagenomic driven drug discovery in general.

1.1.1. 'Microbial Dark Matter'

The overwhelming majority of the tremendous bacterial and archaeal diversity on the Earth is yet to be cultivated by humans, despite decades of effort to characterise and catalogue as many species as possible. Typical estimates in the literature cite that at present, only 0.1 -1% of observable microbial diversity can been cultivated in a laboratory setting. The first notion of this principle was described as 'the great plate count anomaly', when, in 1985, early microbiologist noticed that the direct enumeration of the microbial diversity seen through a microscope far outweighed what they could successfully grow on nutrient agar plates for a given sample.[24,25] Subsequent metagenomic studies of environmental 16S rRNA gene sequences found that a multitude of environments were teeming with uncultivated novel phyla.[26] Although microbial culturing technology has advanced enormously since those pioneering agar plate experiments, with the advent of many specialised devices that precisely reproduce specific environmental conditions and cues[21], the divide between the cultivated and uncultivated continues to widen. Our ability to detect microbes by direct genetic analysis of environmental samples, independent from culturing, is becoming ever more sophisticated and sensitive, and is overstocking our taxonomy and sequence databases with records of uncultivated members. Although the true expanse of the diversity of uncultivated microbes is certainly much higher than what we have managed to characterise thus far in the ever-growing tree of life.[27,28]

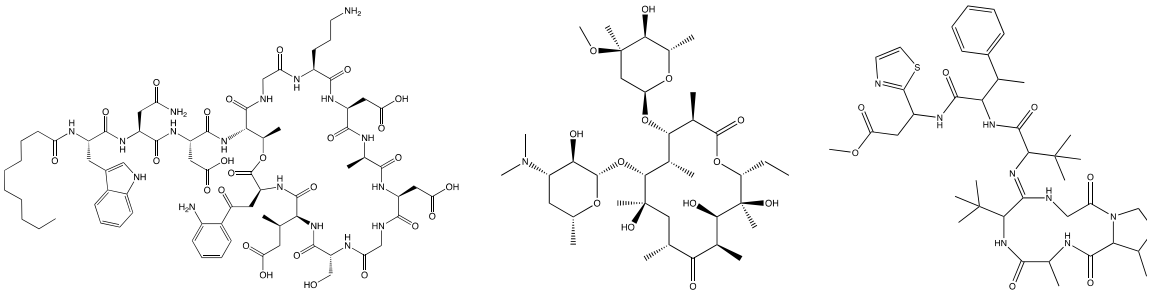
It is this yet to be uncultivated majority that comprise what has become known as the 'microbial dark matter'. Advances in metagenomic techniques are starting to shed light on 'microbial dark matter' as culturing is now no longer a requirement to gain full access to the complete genomic sequences of environmental microbes. This is progressing our understanding of environmental microbes from just simply detecting them and estimating diversity to allowing the comprehensive analysis of ecological functions and lifestyle.

Such analysis has revealed that perhaps the most intriguing members of the 'microbial dark matter' are the obligate symbionts of higher organisms. These symbionts are very rarely cultivated owing to the extremely specific growth requirements and are often dependant on

the host for life support in return for a mutual exchange of specialised functions or metabolites, such as macro-nutrient (e.g. carbon or nitrogen) fixation or defensive compounds.[29,30] Genomes of obligate symbionts, usually much smaller than their known free-living counterparts, can be dominated by genes responsible for the production of specialised metabolites or proteins with very distant matches to any known sequence, making them of particular interest to researchers.[31–34] Further understanding of the 'microbial dark matter' may be key to understanding important ecological functions and informing novel biotechnology inventions. Successful efforts in this area of research are producing findings which challenge our understanding of canonical microbiology and expose new bacterial traits.[35]

1.1.2. Bacterial bioactive natural products as pharmaceuticals

Bacteria are one of the principle sources of clinical medicines and organic compounds with potential pharmaceutical applications known,[36] due to their production of potently bioactive natural products. These bioactive products are specialised secondary metabolites with structural diversity (see Figure 1.1) that have evolved to be highly efficient at interacting with specific biological targets. The range of biological activities and medical applications of these natural product molecules is broad, but in their native context, it is believed that they provide a competitive ecological advantage to the producing organism. This may be by increasing nutrient acquisition (siderophores), acting as surfactants[37] or as signalling molecules.[38–42] Another proposed role for these molecules is defending against, or attacking, competing organisms.[31,34,43–45] This role is elicited by targeting and interrupting essential primary metabolic functions of the competing organism leading to growth inhibition or cell death. The inherent physiochemical and biological properties evolved to confer such competitive advantages, such as target selectivity and specificity, avoidance of resistance, access across cellular membranes and potent cytotoxicity make natural products promising drug candidates.[46,47] As such, natural products offer an abundant source of chemical compounds for drug discovery.[10]



Daptomycin Erythromycin Bottromycin

Figure 1.1 – Structural diversity of microbial metabolites. This figure showcases a fraction of the structural diversity of bacterial natural products. These three well characterised bacterial natural products have useful or promising pharmacological application, and all three are derived from the single genus *Streptomyces*. Daptomycin is a clinically used non-ribosomal peptide antibiotic. Erythromycin is a clinically used macrolide class of antibiotic with a glycosylated polyketide structure.

Medicines inspired by or directly derived from natural products constitute ~65% of approved small molecule drugs,[48] including numerous classes of antibiotics, anticancer chemotherapies, diabetes and hypocholesterolemic drugs, as well as immunomodulatory agents.[14,49] Microbes are the most prolific producers of natural products used to treat infectious diseases and cancers, or as immunosuppressants.[48,50] There are at least 33,000 – 35,000 known bioactive microbial natural products with over half displaying antibiotic activity. This collection of compounds has yielded over 400 approved drugs.[51,52] Thus far, the majority of bacteria identified producing these known natural products are members a single phylum Actinomycetes, and especially species of the readily cultivable genus *Streptomyces.*[53]

Showcasing the utility of bacterial natural products as pharmaceuticals is their prominence as clinical antibiotics. The majority of antibiotics in use for medicine (~90%) are derived from bacterial natural products, either used directly or as semi-synthetic derivatives. [51,54] These include (but are not limited to) compounds form the classes; penicillins, carbapenems, cephalosporins, aminoglycosides, ansamycins, lincosamides, lipopeptides, glycopeptides, polypeptides, tetracyclines, chloramphenicol, and macrolides (e.g. erythromycin).

Bacterial derived antitumor antibiotics also make significant contributions to clinical chemotherapeutic agents, with members of the following natural product classes in use; actinomycins (Dactinomycin), ansamycins (e.g. the rifamycins), anthracyclines (e.g. doxorubicin), bleomycin, epothilones, epoxomicins (e.g. Carfilzomib), rapamycins and staurosporins. The parent compound of the rapamycins, rapamycin, is also implicated for use

as an immunosuppressive agent for preventing tissue rejection in patients receiving organ transplantation. Bacteria have also supplied natural products with antimycotic (e.g. the polyene Nystatin) and anti-viral activity. [52] The discovery of these lifesaving compounds is largely thanks to efforts commencing in the 1940s that involved bioactivity-based screening of organic extracts from thousands of cultures of microorganisms collected from around the World.[3,18]

1.1.3. Biosynthetic gene clusters

The genetic basis for the biosynthetic pathways of the aforementioned bacterial secondary metabolite natural products are organised as biosynthetic gene clusters (BGCs). That is to say, all of the genes required for a bacterium to orchestrate the biosynthesis of a given natural product are primarily co-localised in a single continuous region of the chromosome called a BGC.[55] Genes required for regulation, resistance, cellular export and horizontal transfer of the BGC are also commonly located within the boundaries of the BGC.[56]

There are many different classes of microbial natural products, the best studied of which are the nonribosomal peptides (NRPS) and type I polyketides (PKS), each of which is described in detail in section 1.2. The natural product molecules produced by these different classes of BGCs differ primarily in structure due to the subunits from which they are derived and the resulting chemical bonds catalysed to join incorporated subunits; the non-ribosomal peptides being derived from a diverse pool of amino-acids joined via peptide bond formation, and the polyketides are derived from acetyl-coenzyme-A and malonyl-coenzyme-A precursors joined via C-C bonds formed by catalysed decarboxylative Claisen condensation. However, the general biosynthetic strategy of each of these classes is very similar, both have molecules tethered to, and assembled by, a set of multimodular enzymes encoded in BGCs.

In each of these classes, the core genes that make up the BGCs of NRPSs and PKSs encode for large multimodular enzymes[55,57] which are comprised of discrete catalytic units called modules that operate in succession in an assembly-line like mechanism.[58,59] These core enzymes construct the secondary metabolite's core structure by appending specific building block subunits to an enzyme-tethered molecular chain that is extended as the maturing molecule is passed sequentially from module to module along the enzyme(s). Each module in a core enzyme is further subdivided in to several discrete catalytic domains that function individually to select, condense and modify the incoming subunit. The core genes in modular biosynthesis frequently catalyse chemical modification of the growing molecule, and additional genes located *cis* to the core enzymes genes may produce tailoring enzymes that

can further modify the intermediate secondary metabolite by introducing additional functionally or decorating the molecule to give the final bioactive secondary metabolite.

Another important class of bacterial natural products that has increasingly been the subject of DNA sequence based discovery efforts are the 'ribosomally synthesised and post-translationally modified peptides' (RiPPs), these are some of the largest and most complex secondary metabolites discovered to date (e.g. polytheonamide B at 5.0 kDa with 48 post-translational modifications). These secondary metabolites are not produced in the multimodular assembly-line like fashion, but conversely, as the name suggests, are highly modified ribosomal peptides. However, analogous to the non-ribosomal peptides and the polyketides, a RiPP's biosynthetic pathway genes are typically clustered together on the chromosome as distinct BGCs. [60]

An important feature of BGCs are signature amino-acid sequences that are consistently present in particular biosynthetic classes and can be used as markers to detect and classify gene clusters. A variety of molecular and bioinformatic techniques can be used to exploit these markers to aid the discovery of new BGCs.[61]

1.1.4. A platform for metagenomic based drug discovery.

Taken together, the concepts outlined in sections 1.1.1, 1.1.2 and 1.1.3 build a foundation on which the discovery of new bioactive secondary metabolites can be achieved *via* identification of their corresponding BGCs in DNA sequence data, followed by heterologous expression.[62,63] This is a 'reverse genetic' approach, progressing from sequence to phenotype,[64] and can be applied to both cultivated and uncultivated bacterial strains, workflows that are referred to as genome-mining and metagenome-mining respecitively.[10]

Culture independent direct metagenome sequencing allows researchers to investigate the genomic sequences of the vast abundance of uncultivated bacteria from diverse environments. Novel BGCs can be identified within these sequences (see Section – 1.3 Microbial genome mining) to inform the discovery of new drug like compounds, or to provide as sustainable source of a compound that is produced by an uncultivated bacterial symbiont. This can be executed in a targeted approach, BGCs of a particular class may be sought to find related compounds within the class, or the specific BGC of a known compound may be targeted to identify the producing microbe or guide production via synthetic biology, chemical extraction or specialised culture conditions. Untargeted approaches may seek to identify novel classes of putative BGCs by searching for non-canonical arrangements of

biosynthetic genes or evolutionally distant sequences, to expand the chemical space of known secondary metabolites with desired bioactivity.

To date, the majority of new compounds discovered in metagenome mining studies have used clone libraries to capture the immense genomic diversity in soil microbiomes.[63,65–68] In these studies, rather than sequencing the eDNA directly, it is cloned into a cosmid vector, packaged into phage heads, and delivered to *E. coli*. This results in a library of high-molecular-weight fragments (35-45 Kb) that redundantly covers the microbial diversity present in the original sample. The cloned fragments can then be interrogated *via* sequence similarity or functional analysis to identify and rapidly isolate sequences of interest. This large fragment cloning technique is particularly useful when heterologous expression of BGCs is desired and/or when access to the DNA sample is limited, as sequences of interest can be readily propagated and manipulated.

Continuing technical advances in DNA sequencing, such as the reduction of cost and improvements to data output, read length and quality, coupled with the ongoing progress of bioinformatic analysis tools and computing recourses, will further strengthen metagenomics as a platform for BGC discovery by reducing the costs and time to results.

1.1.4.1. Heterologous expression

To access the pharmaceutical potential from the systematic discovery and characterisation of novel BGCs an essential step is expression of the enzymes in a BGCs such that they are able to collectively biosynthesise their product molecule. As the target bacterial strains in metagenomic studies are typically recalcitrant to cultivation, the production of secondary metabolites is usually achieved by the heterologous expression of BGC in a cultivation amenable heterologous expression host strain. Even in cases where target bacterial strains are cultivable, about 90% of BGC are "silent", these BGCs can be identified bioinformatically but are not expressed under standard conditions, [69,70] thus heterologous expression is also proven to be a fruitful method for discovery of new compounds from cultivable isolates.[71] Heterologous expression systems aim to activate an introduced BGC to produce the corresponding secondary metabolite in useful quantities. In some cases this can be achieved by simply introducing the heterologous DNA, in other cases modifying either the biosynthetic pathway or the expression host to permit higher levels of production of the desired compound is required.[72]

By far the most common expression hosts to date have been *Streptomyces sp.*, due to the availability of tools to cultivate and genetically manipulate them, their known biosynthetic capability and compatibility, and historical precedent.[65,73] One metabolic engineering approach has been to produce host strains that are devoid of native secondary metabolite background to be used as a "chassis" for BGC expression. For example, Myronovskyi et al., constructed *Streptomyces albus* chassis strains (*S. albus* Del14 and derivatives) with 15 native BGCs removed from the chromosome. These strains showed higher yields of heterologously expressed secondary metabolites compared to other commonly used expression hosts and led to the discovery of a new compound fralnimycin.[74] Advances in genome editing tools, including CRISPR/Cas, have allowed previously undomesticated strains to be engineered to optimise the production of natural products. This permits the use of host strains that are more closely related to a BGC donor. Wang et al., developed a chassis-independent recombinase-assisted genome engineering (CRAGE) system to allow high efficiently single step introduction of BGCs into bacteria across diverse phyla and provide evidence of improved efficiency of BGC activation in those strains that were phylogenetically similar to the donor. Using the CRAGE system, the production of secondary metabolites (ririwpeptides A-C) from a previously uncharacterised BGC in selected heterologous hosts was achieved by the researchers.[75]

The activation and increased expression of BGCs in heterologous, as well as naive hosts, can also be achieved by the refactoring of BGCs.[76,77] Systematic approaches to introduce inducible or constitutive promotors into regulatory regions of BGCs has proven useful in accessing novel molecules.[78–80] In this manner, Montiel et al., transcriptionally activated the *Lzr* gene cluster, a previously uncharacterised silent indolotryptoline BGC derived from an eDNA library, by replacing each bidirectional *Lzr* promoter region with a synthetic promoter cassette and successfully expressing Lazarimides in *S. albus*.[80] This molecule was so called because the gene cluster was considered by the authors to be "naturally dead". Another fruitful method has been targeted overexpression or deletion of regulatory proteins to induce expression from native promoters. For example, Kallifidas *et al.*, activated an environmentally derived type-II polyketide BGC by introducing a path-way specific regulatory protein under constitutive promotion to produce the MRSA active antibiotic Tetarimycin A in a *Streptomyces albus* host strain.[81]

Yamanaka *et al.,* activated a transformation-associated recombination cloned BGC form *Saccharomonospora* sp. CNQ-490. By removing a LuxR-type negative regulator gene from

the cloned BGC, the lipopeptide Taromycin A was produced heterologously in *S. coelicolor* M1146.[82]

Advances in culture conditions have also expanded the scope of NPs accessible from newly discovered BGCs, in native and heterologous hosts. Methods to increase expression include introducing a physical scaffold into the growth media, the addition of small molecule elicitors, and co-culture with other microbes to awaken quiescent secondary metabolism.[83]

1.2 Secondary metabolite biosynthesis

BGCs covering a wide range of classes were discovered and investigated in this study. However, the primary focus was to elucidate the BGCs of three known potent bioactive compounds with promising clinical applications; pateamine, peloruside, and mycalamide. These compounds were originally isolated from the marine sponge, *Mycale hentscheli,*[84] and were suspected by us to be secondary metabolites of microbial symbiont origins.

Based on the known structures and established biosynthetic logic, the BGCs of these three secondary metabolites were expected to be members of the bacterial *trans*-AT type I PKS class, specifically hybrid NRPS/*trans*-AT PKS-I BGCs. This specific class of BGC is a sub-class of type-I PKS which fall under the broader PKS class of BGCs and are particularly highly represented in sponge and other marine invertebrate symbionts.[85,86] The fundamental understanding of these classes of biosynthesis, as they relate to the molecules of interest in this study, will be outlined below for the purpose of proving context for later chapters.

1.2.1. Polyketides

Polyketides are a diverse class of often complex natural products constructed from simple building blocks. Their biosynthesis is controlled by the PKS class of BGCs. In polyketide biosynthesis, simple α-carboxyacyl coenzyme-A linked (acyl-CoA) monomers are assembled into functionalised polyketide chains *via* a Claisen type condensation reaction. Three general classes of PKS pathways are known (types I, II and III) with sub-classes recognised for each type.[87,88] Each class is mechanistically related, but vary due to the differing arrangement and utilisation of catalytic domains, and are phylogenetically distinct at the sequence level. This is paralleled by the distinctive molecular scaffolds produced by each class. The most important general class in regards to this study were the modular type I PKSs.

1.2.1.1. Type I PKS

Modular type-I PKS pathways synthesise natural product molecules that often have large macrocyclic lactone scaffolds, the antibiotic erythromycin (see Figure 1.1) being the model example in this class.[56,87,89,90] These pathways contain very large type-I PKS enzymes that have a repeating multimodular arrangement. Each module in the enzyme consists of several covalently linked functional catalytic domains that act in concert to extend a polyketide chain by a single two-carbon monomer.[91–93] The resulting chain can then be further modified by tailoring enzymes to generate the final molecule.[94,95] Typically, each module in the enzyme acts in order of its position in the enzyme, in what is termed 'co-linearity', thus the enzyme acts as template and biosynthetic machinery at the same time.[96,97]

At the beginning of the pathway, a single 'loading module' initiates the chain by tethering an acyl-CoA PKS starter unit (usually acetyl-CoA or propionyl-CoA, others are known[98]) to the enzyme to produce a β-ketothioester, then presenting it to the enzymes following extension module for subsequent elongation.[99–101] The enzymes multiple extension modules each possess a core of biosynthetic domains[58] that are required to extend the polyketide chain by two carbons derived from a single acyl-CoA extension unit (usually malonyl-CoA or methylmalonyl-CoA)[102] with concomitant loss of CO_2.[90] The extension modules function to elongate the chain and then present it to the following PKS module in the BGC. Each extension module can be adorned with additional catalytic domains that modify the growing chain *in situ* via iteratively reducing the carbonyl group of the previously incorporated ketone extender unit into the respective hydroxyl, alkenyl or alkyl moieties.[103,104] Thus, it is the inclusion or exclusion of modifying reductive domains in a module, and the global order of the extension modules in the gene cluster that dictate the general structure of the final polyketide chemical scaffold. The selection of methylated extender units or the possible presence of a methylation domain within an extension module also contribute the structural diversity of the polyketide scaffold by introducing methyl branches into the growing chain.[92]

The final module of a type-I PKS pathway contains a thioesterase (TE) domain that functions to liberate the polyketide chain by hydrolysing the thioester bond that tethers it to the enzyme.[105] The TE can alternatively be coupled with a cyclization event to form a macrocyclic lactone product.[106,107] Rarely other release mechanisms are observed in non-canonical type-I PKS, such as reductase domain-mediated release.[108] To complete the mature secondary metabolite, the polyketide scaffold chain can undergo intramolecular

cyclisation[105,109,110] and a variety of additional pathway dependent modifications.[94,111–113] Usually, the core PKS modules in a BGC are divided over several enzymes, with each enzyme containing a block of one or more of the pathways modules.[93]

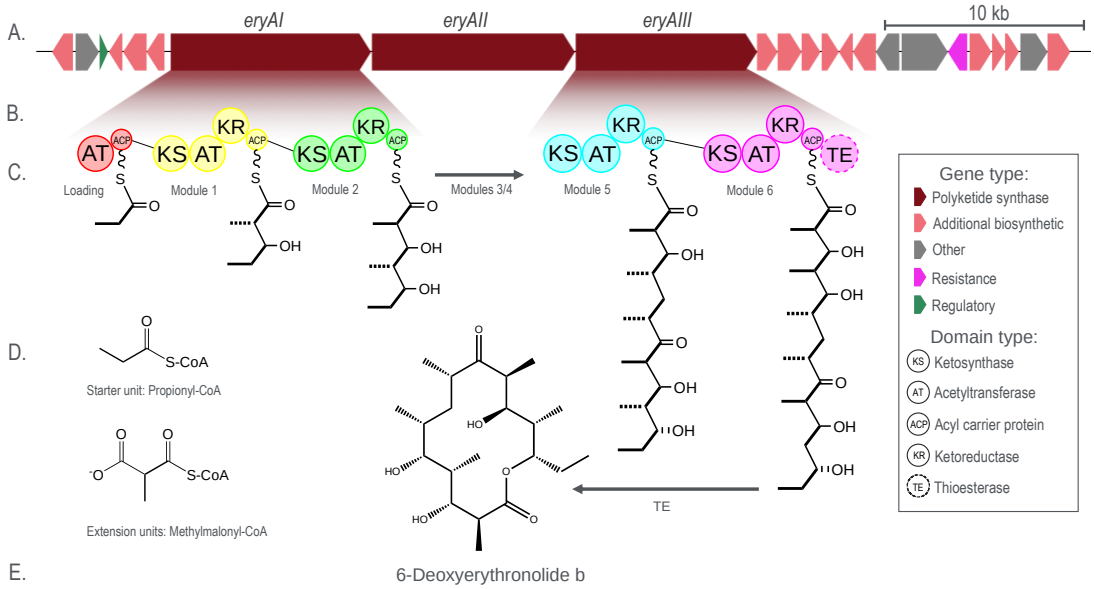


Figure 1.2 – Diagrammatic representation of a type-I PKS gene cluster - the DEB-6 biosynthetic pathway. DEB-6 is the biosynthetic non-glycosylated precursor to the macrocyclic antibiotic Erythromycin.

A) The arrangement of biosynthetic genes as a cluster in the biosynthetic pathway. Central here are the three large PKS genes (brown), surrounded by additional biosynthetic genes (pink). **B)** Two of the three PKS genes (eryAI and eryAIII) have been expanded to show the arrangement of PKS modules within these genes (groups of coloured disks), starting with a loading module and followed by six extension modules. For clarity, the two extension modules of eryII are not shown. **C)** The arrangement of catalytic domains within the modules. The individual catalytic domains in each are represented as a labelled disk. The acetyltransferase (AT) domains load the starter or extension units on the acyl-carrier-proteins (ACP). The ketosynthase domains (KS) are responsible for carrying out the condensation reaction of the extension units to the growing chain. The ketoreductase (KR) domains reduce the ketone to the hydroxy, further reduction is carried out by reductase domains in modules three and four not shown in this diagram. **D)** The starter unit (propionyl-CoA) and six extension units (methylmalonyl-CoA) are used for each DEB-6 molecule produced by the biosynthetic pathway. **E)** The growing linear polyketide chain is extended in length by two carbons at each module. After liberation from EryAIII by the terminal TE domain the complex linear polyketide is cyclised to DEB-6. DEB-6 is further tailored by additional biosynthetic genes, for example glycosyltransferases, to yield Erythromycin (not shown).

The minimal biosynthetic core necessary for polyketide chain extension by a prototypical type-I PKS extension module consist of three domains; an AT-domain, an acyl-carrier protein (ACP) and a KS-domain.[89,91] For each modules extension cycle, the AT-domain is the first to act, this domain initially selects the appropriate electrophilic acyl-CoA extension unit[102] as a specific substrate and then loads it onto to a nucleophilic pantetheine arm attached to the ACP, to give a an ACP bound extension unit.[91] The ACP (also denoted as T-domain) is a dedicated domain in each module that is modified by the attachment of a 4`-phosphopantetheine (PPT) prosthetic flexible arm derived from Co-A, the attachment of which is catalysed by a phosphopantetheine transferase (PPTase) enzyme in *trans*.[58] The task

of this prosthetic arm is to flexibly tether the growing chain intermediates at each module of the PKS enzyme through a thioester bond while they are being progressed to the next elongation module for extension. To complete an extension, the ACP bound intermediate from the previous module is transthiolated to a cystine in the active site of the KS-domain[114] in the current extension module, the ACP bound extension unit of the current module then reacts with the incoming intermediate in a CO_2 evolving Claisen condensation catalysed at the KS-domain[103] (Figure 1.3). This liberates the chain from the KS-domain and covalently attaches it to the ACP bound extension unit, extending the chain by a two-carbon ketide unit at the previously KS-domain bound terminal. Thus, in a model modular type-I PKS, the number of extension modules determines the length of the chain. [96,103] The loading module is responsible for the addition of the first ketide unit and need only consist of an AT-domain and an acyl-carrier protein domain.

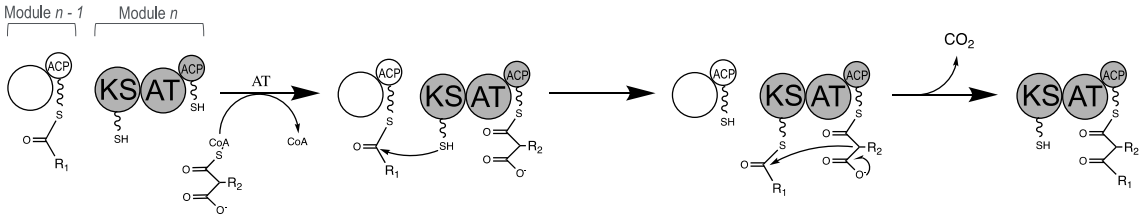


Figure 1.3 – Polyketide elongation by minimal PKS. A minimal PKS module with the catalytic domains required to process a single PKS extension reaction.

In this figure the active module (Module *n*) is shaded grey. The domains required in the minimal PKS are; a ketosynthase (KS) domain, an acetyltransferase (AT) domain and an acyl-carrier-protein. The first step of the reaction is the activation of the module where the acyl-CoA extension unit is loaded on to the ACP by the AT domain. Secondly, the pre-extended chain intermediate is transthiolated from the previous module's (Module n - 1) ACP to the KS domain of the active module. Finally, the KS domain catalyses the claisen condensation of the chain intermediate to the extension unit, with the extension unit acting as the nucleophile, liberating CO2 in the process. Only the domains taking part in these reaction steps are labelled.

To introduce structural variation in the nascent poly-β-keto chain, the extension of the chain at each module can be followed by some degree of β-ketoreduction at the terminal β-ketone of the incoming intermediate in an enzymatically controlled stepwise manner.[115] Each extension module may be equipped with a partial or complete set of defined reducing domains that carry out these ordered reactions. The presence or omission of domains in the set determines the level of reduction carried out at each module. A fully reducing extension module consists of three additional domains, a β-ketoreductase (KR-) domain, a dehydratase (DH-) domain and an enoyl reductase (ER-) domain.[116] The KR-domain reduces the β-ketone to a hydroxyl group. The hydroxyl group can be reduced further to an unsaturated enoyl

group by the DH-domain. Finally, the presence of an ER-domain reduces the enoyl group to a saturated alkyl group.[58]

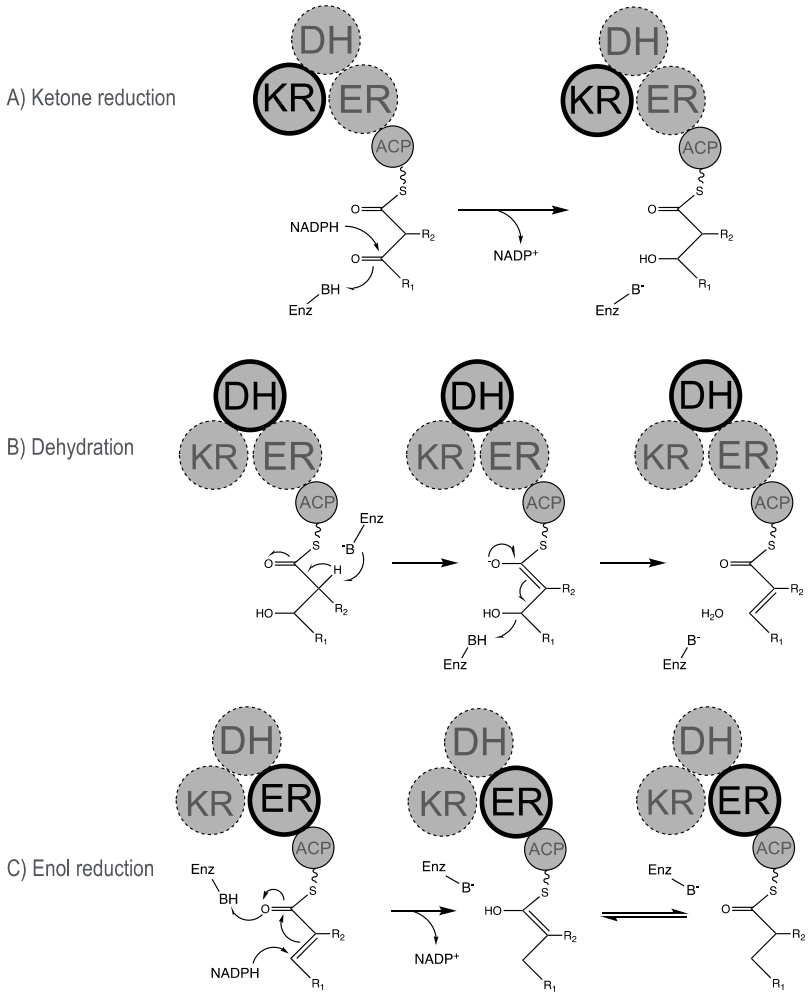


Figure 1.4 – PKS reduction domains and reaction mechanism.

A) A reaction scheme for the reduction of a ketone in a growing polyketide chain to a hydroxy by the ketoreductase (KR) domain of a keto-reducing extension module. Following this, B) the dehydration of the hydroxy to an enol can carried out by a dehydratase (DH) domain. C) The enol can further reduced to saturation by an enol reductase (ER) domain in a fully reducing extension module. The active domain in each reaction scheme is shown in bold.

To introduce further functional group complexity into polyketides, the scaffold chain can be modified by the addition of methyl branching groups at either the α- or β-positions relative to the carbonyl group. Several distinct mechanisms are known to make such modifications. Methylation at the α-position can be mediated by the inclusion of a *S*-adenosyl methionine (SAM)-dependent carbon methyltransferase (cMT-) domain in PKS extension modules.[117–119] More common in the *cis*-AT PKS sub-class of type I PKSs, substitution at the α-position can occur during chain extension, where by a 2-branched malonyl-CoA derived extender unit (e.g. Methylmalonyl-CoA) can be incorporated to introduce a methyl group or other

functionality to the molecule mediated by phylogenetically distinct KS- and AT-domains.[98,102] While virtually absent in *cis*-AT PKSs, methyl branching at a β-position is common in the *trans*-AT PKS sub-class of PKSs, and involves a more complex mechanism, including the synthesis of the specialised precursors carried out by distinct biosynthetic cassettes and auxiliary enzyes.[120–122] In either clade of PKSs, the hydroxyl group of a partially reduced β-ketone can also be methylated by an oxygen methyltransferase (oMT-) domain within a module to give a methoxy group.[123]

1.2.1.2. *trans*-AT type I PKS

This phylogenetically distinct clade of type I PKS BGCs is characterised by a unique acyl-transfer mechanism that is provided by one, or a small number of, discrete free-standing acyltransferase enzymes that serves all extension modules in the pathway, as opposed to a *cis*-AT PKS where each module has a local integrated AT-domain.[124,125] The *trans*-AT PKSs also exhibit other specialised features and unusual cluster architecture compared to wider known "conventional" *cis*-AT PKS architecture commonly found in the thoroughly researched cultivated free living actinomyces.[125] The idiosyncratic architecture of *trans*-AT PKSs can include cluster fragmentation and repeated regions spread over large spans of the genome, non-canonical module order or type and unique, repeated or non-functional domains within modules, and other *trans*-acting enzymes functioning during polyketide assembly.[85,120] Seemingly, this flexibility of cluster configuration allows for a much-expanded diversity of *trans*-AT PKS BGCs. Naturally, the diversity of possible cluster arrangement and composition is reflected by the structural diversity of the chemical products produced by these biosynthetic gene clusters.[85,103,126]

A common distinguishing structural characteristic of trans-AT produced polyketides, afforded by the flexible architecture,[125] and despite rarity in *cis*-AT PKS products, is β-branching, the substitution of a carbon branch at a β-carbonyl position relative to the thioester.[120,122,127] *Trans*-acting enzymes from a conserved five-enzyme hydroxymethylglutaryl-CoA synthase (HMGS) gene cassette is in charge of mediating the β-branching reactions.[120,122,128–130] The enzymes from this cassette provision and transfer a carbon nucleophile required for addition to the electrophilic β-carbon, and then processes the substitution to the final β-alkyl group or functionalised sidechain. This five-member cassette nominally consists of: (i) a free standing 'donor' ACP (ACP_D) usually associated with the main BGC; (ii) a non-catalytic, free standing KS lacking the conserved active cysteine

residue (KS_0); (iii) a freestanding 3-hydroxy-3-methylglutaryl (HMG)-CoA synthase homolog (HMGS) which catalyses the carbon-carbon bond formation; (iv) a dehydrating enoyl-CoA hydratase analogue (EHC_1); (v) a second, decarboxylating, EHC (ECH_2).[120,122,130] The position of a β-branching reaction is encoded within the context of the core *trans*-AT PKS BGC and can be determined by the presence of consecutive ACP/T-domains as doublets or triplets.[122] The model mechanism of β-branching (see Figure 1.5), elucidated for *pksX* from *Bacillus subtilis*,[131–133] starts with the loading of the ACP_D with a CoA derived malonyl group by the *trans*-AT. The freestanding KS_0 then decarboxylates the malonyl-ACP_D to acetyl-ACP_D. This acetyl group is then transferred to a cysteine in the HMGS and will serve as nucleophilic acetyl-enolate donor after deprotonation by a histidine residue of the HMGS. The HMGS then catalyses the β-C-C bond through aldol addition of the reactive acetyl-enolate to the electrophilic ketone of the ACP-bound β-ketothioester of the growing chain forming a crosslinked intermediate. Hydrolysis liberates the HMGS resulting in 3-hydroxy-3-methylglutaryl-S-ACP (*S*-HMG-ACP). This is dehydrated by the EHC_1 to give methylglutaconyl-ACP (MG-ACP), which is decarboxylated by EHC_2 to the final ACP tethered β-branched intermediate.[120,122,130]

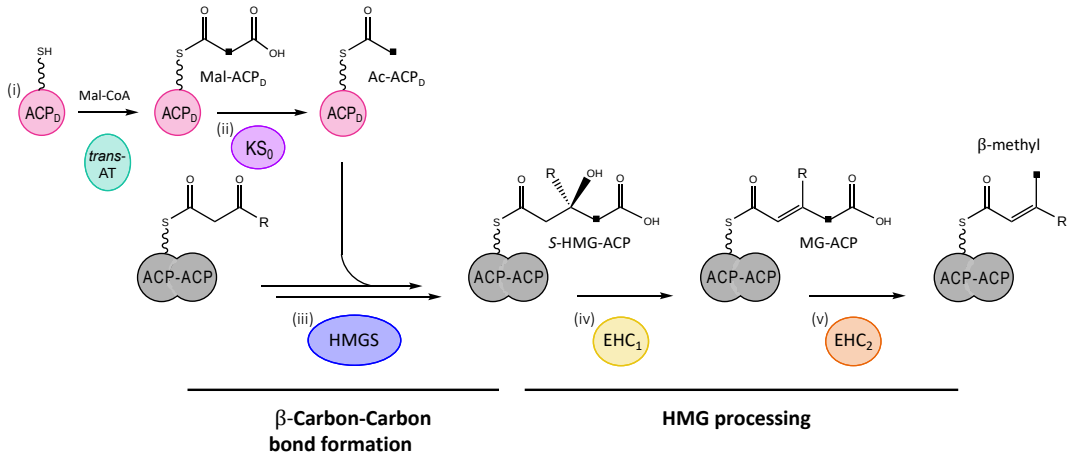


Figure 1.5 – β-branching incorporation in *trans*-AT PKSs.

The donor ACP (i; ACP_D) is loaded with malonyl-CoA (Mal-CoA) by the *trans*-AT to give Mal-ACP_D. A KS homologue (ii; KS_0) decarboxylates the Mal-ACP_D to acetyl-ACP_D (Ac-ACP_D). A nucleophile is generated from Ac-ACP_D by the 3-hydroxy-3-methylglutaryl (HMG)-CoA synthase homologue (iii; HMGS) which attacks the ketone of the ACP-bound β-ketothioester of the growing polyketide to form *S*-HMG-ACP. The tandem ACP domains are depicted in this diagram as ACP-ACP. Two enoyl-CoA hydratases (iv; EHC_1, v; EHC_2) process the *S*-HMG-ACP to the final β-methyl. The added carbon is shown as a black square.

The process of methylation at the α-position in *trans*-AT PKSs is generally driven by the distinctive SAM-dependent cMT mechanism as the incorporation of α-substituted acyl-CoA extension units is limited by the lack of distinct AT-domains at each module. This

mechanism can be extended to allow for dimethylation resulting in *gem*-dimethyl moieties.[134,135]

Another mechanism that can contribute to the structural complexity of type I PKSs is the possible incorporation of nonribosomal peptide synthetase (NRPS) modules in a hybrid NRPS/PKS BGC,[136] resulting in the product containing one or more of an exceptional variety of amino-acid residues in the molecule.[104,137,138]

1.2.2. NRPS

The NRPSs are another class of bacterial biosynthetic pathway that assemble polymeric non-ribosomal peptide (NRP) molecules by enzymatically linking a variety of amino-acid monomers,[139] *via* peptide bonds, independent of the cellular ribosome complex. The amino-acids are selected, and the peptide bonds catalysed by large multimodular NRPS enzymes. Analogous to the PKS biosynthesis described above, the enzymes in NRPS pathways are organised as discrete modules. Each module in an NRPS enzyme is programmed to select the appropriate amino-acid, append it to a growing peptide chain tethered to the module's PPT arm *via* a thioester bond, and pass it along to the next module for further extension in a co-linear process. Likewise, these processes are catalysed by individual domains in each module.[140]

1.2.2.1. NRPS biosynthetic module architecture and mechanism

The typical conventional NRPS extension module includes three catalytic domains that work in concert to extend the peptide chain by one amino-acid[141]: an adenylation (A-) domain[142], a thiolation (PCP-) domain (also denoted as T-domain) and a condensation (C-) domain.[143] NPR extension starts with the A-domain which is responsible for the selection and ATP-dependent activation, by adenylation (AMP addition), of an amino-acid monomer extension unit that will be utilised in the extension cycle.[142,144] Once activated, the amino-acid-AMP monomer is transferred to the PCP-domain. The PCP-domain features the same PPT modification as the ACP of type-I PKSs[145], a flexible tether used to secure and shuttle the developing peptide chain down the biosynthetic assembly line.[146] The C-domain then acts to catalyse the peptide bond[147] between the T-domain tethered activated monomer of the current module and the T-domain tethered nascent peptide chain incoming from the previous up-stream module.[148] Once a T-domain has donated its peptide to the down-stream C-domain, it can again accept a new amino-acid from the A-domain and participate in the next cycle of

biosynthesis.[58] In a free standing NRPS, the entry and exit points of the pathway are defined by specialised modules[149]. The 'initiation' module at the start of the pathway usually lacks a C-domain as its singular role is to activate the initial monomer of the chain[150]. The final 'termination' module performs extension and is also responsible for release of the peptide chain from the biosynthetic machinery.[151] The most common mechanism of peptide release is a macrocyclisation reaction, in which an internal nucleophile, typically a hydroxyl or primary amine, performs nucleophilic attack on the terminal thioester, releasing a cyclic product.[58] This reaction is catalysed by a thioesterase (TE) domain (see Figure 1.6).[59,152] Less commonly, TE-domains will catalyse hydrolysis of an enzyme linked peptide, resulting in a linear product.[141] However, a number of alternative termination strategies are also known to exist.[153] As with type-1 PKS pathways, the number of modules in an NRPS is indicative of the scaffold chain length. One key difference between these two pathways is the ability of the NRPS modules to select form a much wider range of extension units, the selection of which is determined by the phylogeny of the A-domain.[154]

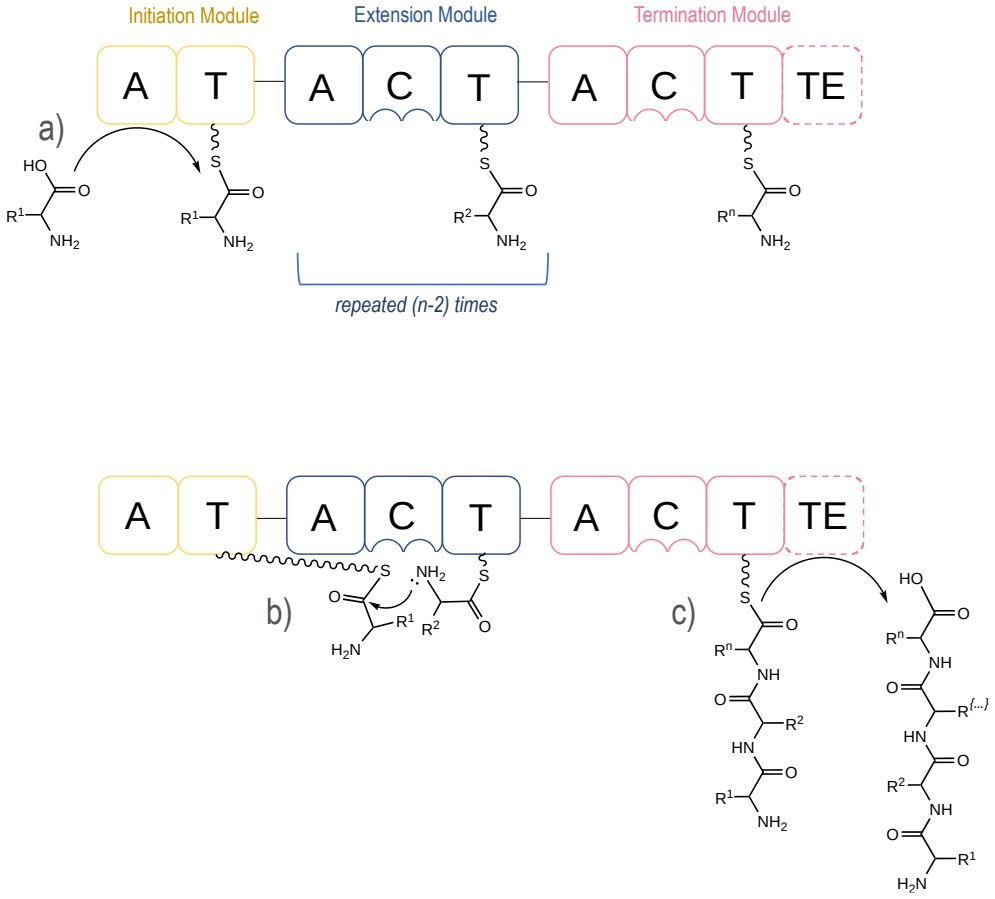


Figure 1.6 - NRPS module architecture and activity.

A hypothetical NRPS pathway showing the domains of simplified 'Initiation' (yellow), 'Extension' (blue) and 'Termination' (pink/green) modules. A = adenylation domain (confers amino-acid substrate recognition and activation), C = condensation domain (catalyses peptide bond formation), T = thiolation domain (peptide carrier), TE = thioesterase domain (final release of peptide from PPT arm). The peptide synthesis pathway begins at a), the A-domain mediated selection and activation of the amino-acid monomers followed by loading onto the PPT arm of the T-domains via the formation of a thioester bond. Each module is loaded with a specific activated amino-acid. The length of the final chain is determined by the number of 'Extension' modules in the pathway, which will be equivalent to *n-2* the number of monomers in the final chain. Following loading, b) the flexible arms allows the C-domain access the tethered substrates from two adjacent modules and to perform the condensation reaction, resulting in the addition of the substrate tethered to the upstream module to the N-terminal of the monomer tethered the downstream module. A scheme is shown for the condensation reaction between the 'Initiation' module and the first 'Extension' module. The final step of this pathway, c) is the release of the peptide chain from the 'Terminal' module. This is effected by the hydrolytic activity of the TE domain in this module.

1.2.2.2. NRP chemical diversity

The ribosomal independence of NRPS peptide biosynthesis allows for greatly increased structural complexity over standard ribosomal peptides due a vastly increased pool of specialised amino-acids substrates available to NRPSs.[155,156] In addition to the 20 standard proteinogenic amino-acids, over 500 NRPS substrate monomers are known to be

incorporated into NRPS produced secondary metabolites.[139] Further complexity is endowed to the secondary metabolites produced by NRPSs by tailoring modifications and molecular cyclisation carried out during synthesis by additional biosynthetic enzymes encoded in the gene cluster.[113,153] Additional domains have been identified in NRPS pathways including but not limited to, methyltransferase (MT-), oxidation (Ox-), heterocyclisation (Cy-) and epimerisation[157] (E-) domains.[140,153,158] Post-translational modifications of NRPs by stand-alone enzymes encoded in the pathway are also possible,[153] e.g., the addition of sugars by glycosyltransferases to produce glycopeptides, a useful class of antibiotics.[159,160]

1.2.3. Tailoring modifications

Both NRPS and PKS core scaffolds can be transformed by the action of tailoring modifications.[113] This imparts further structural diversity to the mature molecule and is often important for biological activity.[153] Such modifications can occur *in cis* during the assembly of the molecular core and are imparted by accessory or specialised catalytic domains in the core biosynthetic machinery. Tailoring modifications can also occur *in trans* post-release of the molecular core from the biosynthetic machinery and are catalysed by enzymes associated with the biosynthetic cluster but independent of the core biosynthetic NRPS or PKS enzymes. Both PK and NRP chains can be tailored by alkylation, acylation, glycosylation, halogenation, hydroxylation and cross linking by oxidoreduction. These modifications can occur at various positions and by various conserved mechanisms;[113,153] the scope of all of these modifications is too broad to cover in detail here. However, some important tailoring modification domains present in BGCs investigated in this thesis were predicted to cause β-methylations of *trans*-AT-PKSs (discussed above - section 1.2.1.2) and the generation of a thiazol from a cysteine residue by an NRPS module which will be discussed below (section 1.2.3.2).

1.2.3.1. Subunit biosynthesis

A distinguishing feature of NRPs is the inclusion of the non-proteogenic amino-acids (AAs), including D-AAs.[161] The genes encoding the enzymes that synthesise these distinctive monomers are usually clustered within the BGC.[162] For example, in the biosynthesis of vancomycin which includes several nonproteinogenic amino-acids,[111] a group of five enzymes (DpgA, B, C, D and HpgT) make up a precursor biosynthetic sub-pathway to

produce one of the amino-acid precursors, 3,5-dihydroxyphenylglycine (DPG), required for vancomycin biosynthesis, from four molecules of the primary metabolite malonyl-CoA.[163] Non-proteogenic amino-acids can also be synthesised from proteogenic amino-acids precursors. Modified versions of the proteogenic amino-acids include the methylated, hydroxylated, and D-forms.[139] D-amino-acids can be specifically incorporated into NRPs and are generated by epimerization (E-) domains from readily available L-amino-acids during elongation.[164] The incorporation of β-amino-acid and β-hydroxy-amino-acid[104] is also common in NRPs. β-amino-acids result from the action of aminomutase enzymes converting proteogenic amino-acids to the β-amino from.[165] This is seen, for example in, Taxol biosynthesis in which a Phenylalanine is converted to a β- Phenylalanine prior to activation by the A-domain.[166] Heterocycles can also be embedded into NRPs by the conversion of the proteogenic amino-acids cysteine, serine, and threonine during chain elongation by heterocyclisation (Cy) -domains. These principles can be used to identify the putative BGCs of molecules containing a particular known structural motif arising from the incorporation of rare amino-acids,[167] or reciprocally, can be used to guide the isolation and structural elucidation of a molecule from a newly discovered BGC.

1.2.3.2. Thiazole biosynthesis

Thiazole functionality is seen in the NRP scaffolds of several secondary metabolites, such as the antitumor agents bleomycin[104] and epothilone A.[164,168] The antibiotic bacitracin contains a thiazoline moiety[169,170]. Characterisation of the BGCs that produce these molecules[115] show the thiazole is formed during chain elongation when a cysteine side chain is intramolecularly cyclized and dehydrated to a thiazoline by a NRPS Cy-domain, which is a specialised variant of the C-domain,[168] the thiazoline is then oxidised to a stable thiazole structure. In this process, the cysteine incorporation is specified by a conventional cysteine specifying A-domain in the participating thiazole extension module, the Cy-domain of the same module performs the condensation reaction, cyclization and dehydration to form the five-membered ring structure of thiazoline (See Figure 1.7). An Ox-domain then finishes the transformation of thiazoline to thiazol.[164] The presence and arrangement of these specialised domain in an NRPS module is indicative of a thiazole in the molecular scaffold of the molecule.

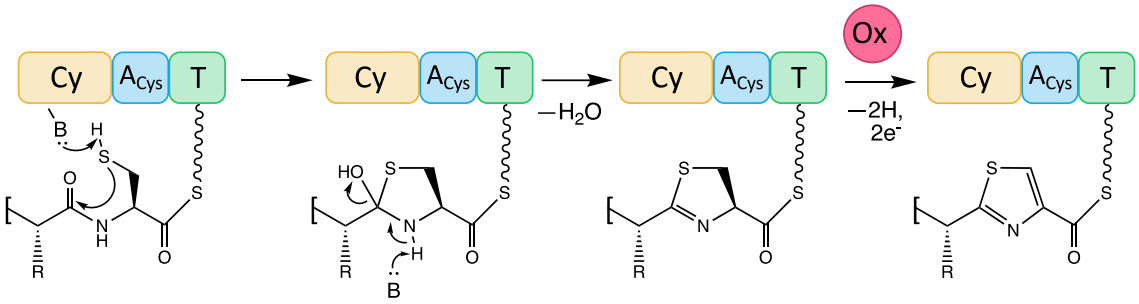


Figure 1.7 – Proposed mechanism for thiazole biosynthesis by NRPSs.
An A-domain (A_{Cys}) and a Cy-domain incorporate a cysteine in to the growing molecule that is tethered to the thiolation domain (T-). The Cy-domain then base catalyses the cyclisation of the cysteine's thiol with the carbonyl group of the downstream peptide bond with the loss of water to produce a thiazoline. An oxidation domain (Ox-) acting in *trans* from another module oxidises the thiazoline to the final thiazole.

1.2.4. Expanded biosynthetic diversity

With increasing frequency, BGCs and new molecules are now being discovered from further afield using the technologies of metagenomics and heterologous expression systems. Reports of secondary metabolite biosynthesis that break the rules of the elementary biosynthetic models outlined above are being published.[137] This is hardly surprising, as the classic models were derived by exploring only a phylogenetically and environmentally restricted set of bacteria, namely the free living, soil dwelling actinomycetes. For example, the characterisation of BGCs responsible for producing the thalassospiramides[171] in marine α-proteobacteria demonstrate that these biosynthetic assembly lines do not strictly abide by the rules of co-linearity. The intriguing biosynthetic machinery that produce these immunosuppressive cyclic lipodepsipeptides employs a pass-back chain extension mechanism where the growing intermediate chain is passed back up the assembly line for repeated extension by an upstream module.[137]

1.3 Microbial genome mining

Genome mining is a computational approach used to discover, classify and dereplicate biosynthetic pathways and their natural products based on microbial genome sequence data without requiring knowledge of chemical structures.[172] This is achieved by identifying BGC sequences in microbial genomes and using this data to make structural predictions and potentially produce novel compounds via native or heterologous expression.[173] Retrospective genome mining can also be useful in answering longstanding questions about

the biosynthesis of previously identified natural products of interest where the biosynthetic pathway is unknown. [174,175]

Generally, genome mining detects BGCs in sequence data by virtue of the conservation of amino-acid sequences exclusively shared by class specific core biosynthetic enzymes or other conserved BGC signature sequences motifs, which can be efficiently identified *in silico* by comparison to libraries of verified sequences.[176] Other genome mining methods are emerging that do not rely on specific signature sequences motifs but leverage global sequence patterns, often arising due to distinctive evolutionary mechanisms acting on BGC regions,[177] and have proven useful in detecting novel classes of BGCs.[172]

The genome mining discovery technique is not restricted to discovery of compounds that are expressed at high levels during laboratory fermentation, and thus addresses a fundamental limitation of the traditional *ad hoc* approach of detecting only the secondary metabolites that are present in high abundance in a sample[2,76]. In traditional discovery, colloquially referred to as "the grind and find method",[178] compounds produced under lab cultivation conditions are directly detected, by either chemical analysis or bioactive screening of fermentation extracts of cultured microbes.[179] Although historically this methodology produced many useful lifesaving medicines, in recent times it has fallen out of favour, as fermentation-based screening efforts suffer from diminishing returns due to throughput limitations and high rates of compound rediscovery, ultimately deleting the natural product drug discovery pipeline.[54,180] This pipeline has been instrumental in proving the world with antibiotics and its eventual failings have contributed to the antibiotic-crisis currently threating society. [181] To overcome this crisis, new classes of antibiotics are desperately needed, requiring a reinvigoration of the discovery pipeline with innovative methods to find lead compounds. [182]

In recent years, genome mining for BGCs has become a key methodology to identify new molecules and has led to the discovery of dozens of novel compounds.[76,172] The inception of the genome mining discovery paradigm occurred when two well studied and, at the time, recently sequenced species of actinomycetes were analysed in the early 2000s. At this time these model actinomycetes, *Streptomyces coelicolor*,[183] and *Streptomyces avermitilis*,[184] were already known to be the producers of the bioactive secondary metabolites actinorhodin and avermectin respectively. However, when the full genome sequences of these bacteria were analysed,[185,186] it became apparent they were potentially genetically capable of producing ~10 times the number of secondary metabolites than was suggested by the

observed chemotype. As microbial sequence data from more natural product producers became available, similar patterns of "silent" BGCs were observed in many more strains.[73,83] This pattern was especially typical for the genomes of *Streptomyces*, a commonly cultivated genus of soil dwelling actinobacteria which commonly harbour more than 30 BGCs, and readily extended to members of many other phyla as subsequent genome data became available. This led researchers to postulate that the bioactive compounds detected in fermentation extracts of lab cultivated bacteria did not always accurately reflect the true biosynthetic potential latent in the genomes of these strains. [187]

Thus, the obvious advantage of a genome mining approach is the ability to realise a much broader range of biosynthetic capability of a given strain by capturing BGCs that would be missed by the traditional compound discovery methods, without the laborious cultivation and screening requirements.[188] Genome mining can also be conducted in a completely culture independent nature where BGCs are retrieved from metagenomic sequence data produced directly from environmental samples giving researches unprecedented access to the natural products of the microbial dark matter.[189] Additionally, the computational methods underpinning genome mining can be scaled and automated[190] to match pace with the ever-increasing trends in the production and improvement of microbial genome data in an ongoing genomic revolution.[172] Computational genome mining is also easily extensible; algorithms to predicted new or even unknow classes of BGCs can and have been developed[191] (e.g., DeepBGC[64]), and systems to analyse metagenomic data sets are commonplace and openly available for researches to integrate into existing computational pipelines.[192] The discovery and validation of new compounds feeds back into curated databased to further improve genome mining discovery methods in an iterative way.[193] This is exemplified by one of the most common, broadly adopted and generalised genome mining tools antiSMASH (see section 1.3.1). The initial release of this genome mining software suite in 2011 was capable of detecting a modest 17 classes of BGC,[194] while the most resent version (antiSMASH 5), as of 2019 (antiSMASH 6 is in beta-release), boasts the ability to detect 52 classes of BGC[195] and integrates advanced biosynthetic prediction modules.

1.3.1. AntiSMASH

AntiSMASH[195] is a pipeline of *in silico* tools that has been designed to identify and annotate the broadest possible range of secondary metabolite BGCs classes in genomic data. This open source software was released in 2011,[196] and this year (2019) the fifth updated iteration

was released,[195] demonstrating the ongoing development and implementation this tool. The pipeline detects and annotates the core genes of biosynthetic loci as well as many accessory genes known to be associated with BGCs. It then analyses the configuration of biosynthetic domains in the newly identified BGC to make predictions regarding the chemical structure of the biosynthetic end product. The newly identified BGC is also compared to a database of experimentally validated BGCs to evaluate the evolutionary similarity to known clusters and infer possible gene functions by homology.

At the core of the antiSMASH pipeline is a rule-based algorithm, in which the presence of particular combinations of protein types within a sequence window is used to infer the presence of a particular class of BGC. For example, the presence of an A-domain and a PCP-domain within a 20 kb window is a condition that will result in identification of an NRPS cluster. AntiSMASH draws extensively on well-established algorithmic procedures for gene finding and functional annotation. Briefly: To detect a BGC in the submitted query DNA sequences, the antiSMASH pipeline first identifies the open reading frames (ORFs) in the input sequence data using established open source software (Glimmer3[197]). ORFs are then given a functional annotation using HMMER3; the translated ORFs are searched with a manually curated database of profile hidden Markov models (pHMMs; see section 1.3.1.1) that can detect distant homology to signature amino-acid sequences exclusive to genes found in BGCs. The type, class and validity of potential BGCs are then determined by applying a rule-based logic to the pHMM hits detected. Hits known to cause false positive are discarded. The logic for detecting positive BGCs is based on locating proximal clusters of known minimal core components typical for each gene cluster type within a predefined sequence length window. Each biosynthetic class is classified by a unique set of pHMM hit requirements and bit score thresholds, for example, a *trans*-AT type I PKS (section 1.2.1.2), would be characterised when a *trans*-AT docking domain is detected with a HMM bit score over 65, a KS domain scoring over 50, and no matches to rules for a 'normal' type I PKSs are all met, all within a10 kb stretch of sequence.

The compiled pHMM database and cluster type detection rules central to BGC detection by antiSMASH are sourced from multitude of established methods and models or were produced specifically by the developers of antiSMASH. This integrative approach currently supports the rule-based detection of 52 different classes of BGCs making antiSMASH a powerful tool for detecting many types of BGCs and streamlining discovery workflows.

AntiSMASH performs a second round of pHMM analysis on detected BGCs based on the resulting classification to determine the arrangement and substrate specificity of the

biosynthetic domains within the genes where applicable. The outcome of this analysis informs the structural predictions of the natural product produced by the respective biosynthetic gene cluster. A consortium of third-party analysis modules are also employed to aid in domain architecture and chemical structural predictions for appropriate BGCs, e.g. SANDPUMA[154] is used in some versions of antiSMASH for predicting the substrate specificities of NRPS adenylation domains. Accessory genes, such as resistance elements, transport machinery and tailoring enzymes are then functionally annotated from another library of pHMMs based on secondary metabolite clusters of orthologous groups (smCOGs). The smCOGs for this process were derived from all biosynthetic gene cluster proteins extracted from the NCBI nr database and clustered using OrthoMCL,[198] consensus gene function annotations are then assigned.

1.3.1.1. pHMMs

Efficient and sensitive sequence homology detection using pHMMs is the central principle behind antiSMASH and many other genome mining tools. As such, a brief overview of this technology and the origin of the specific pHMM databases used for BGC identification is warranted.

A pHMM is a probabilistic sequence profile model (Figure 1.8) obtained from a multiple sequence alignment of known members of a family of proteins that can then be used to search a database for other related members of the family. The structure of these models encodes subtle similarities in positional sequence variation and conservation between the sequences in the family, allowing for the sensitive detection of distantly related members. pHMM perform better than pair wise alignments in this task mainly due to the systematic way that position dependent alignment gaps are modelled, as it is not rare that two homologous sequences will have less than 30% overall identical amino-acids per position.[199] Additional layers of information, beyond just the frequency profile of amino-acid residues at a given position, are extracted from the multiple sequence alignment of family members and are used to determine the likelihood of encountering insertions and deletions (indels) in a sequence at given positions. For example, at highly variable positions in the alignment, indels in the query sequence may not be heavily penalised by a pHMM, whereas in a pairwise alignment any gaps encountered are scored with a fixed penalty independent of position.

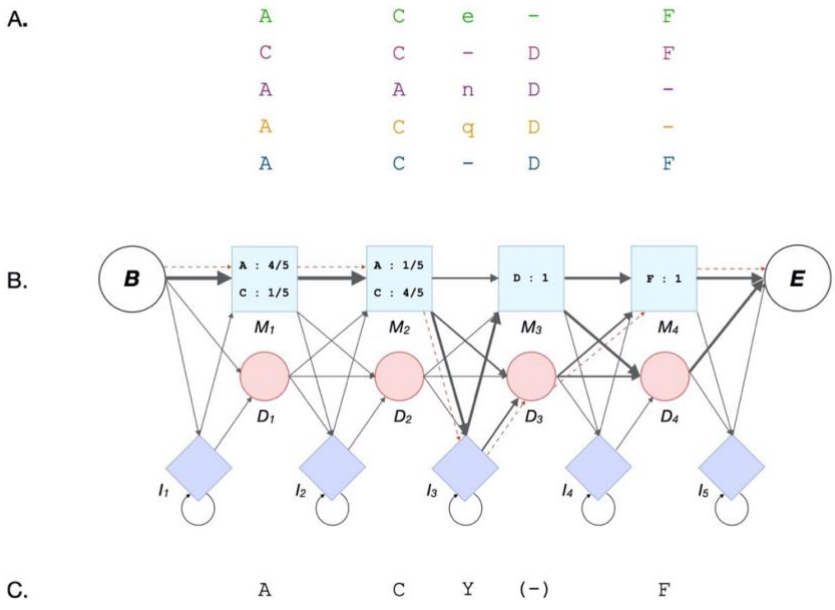


Figure 1.8 – pHMM. The structure of a general pHMM built from a mock alignment of five example sequences with a possible path taken through the model by aligning a query sequence.

A.) A multiple alignment of five short amino-acid sequences. Each sequence in the alignment is shown in a different color. Columns of conserved positions are shown as uppercase letters and are used to build the model. Dashes represent gap positions in the alignment.

B.) This diagrammatic representation of the pHMM built from the multiple alignment shows the layers that encode possible states for matches (M_x - blue), deletions (D_x - red) and insertions (I_x - lilac) at each position in a query sequence. Arrows show the possible transitions between the states of the model. The weights of the arrows are proportional to the 'transition probability' (actual values not shown), that is the probability of transitioning from state to state. States '**B**' and '**E**' represent the beginning and end states of the model. Fraction values in the 'match' states correspond to the 'emission probabilities' for a match and are calculated from the amino-acid profile for that position. In this model, the first transition is from state '**B**', with the highest probability of a transition (heavy arrow) to match state M_1 which would emit an 'emission probability' of 0.8 for an '`A`' amino-acid match and 0.2 for a '`C`' amino-acid match. Other possible, but less-likely, transitions from '**B**' are to the deletion D_1 or insertion I_1 states. The red-dashed arrows show a possible path through the model taken by a query sequence to produce a probability score.

C.) The query sequence accompanying the red-dashed path in B). On this path, the first transition is from '**B**' to M_1, a high probability event as all the first positions in the alignment are a match (i.e., no deletions or insertions in this alignment column in A.), this match state then emits a probability of 0.8 as the query residue is '`A`'. The next transition in the path is to M_2, also a high probability event and emitting p(0.8) as position two in the query is '`C`'. Following this, the path transitions to insertion state I_3, to resolve the insertion of '`Y`' at this position in the query. Transitioning to an insertion state at this point holds a strong likelihood as insertions are common in the alignment at this position. Next is a transition with modest probability, to D_3, to cover the conserved '`D`' of the alignment which is missing in the query. Then a transition to M_4 emits p(1) to account for the match of '`F`' at this position. Finally, the path terminates at the '**E**' state. The product of the probabilities encountered along the path at each transition and emission is used to determine the probability that the query sequence belongs to the family of aligned sequences used to build the model. Many valid paths may transverse the model for a given query, the most likely of which, called the Viterbi path, can be efficiently calculated by the Viterbi algorithm[200,201] and is used to score the query.

Comparison of a query sequence to a pHMM conveys the probability of its fit to the multiple sequence alignment of the protein family members that the pHMM was built from, which can then be converted into a score for determining the relatedness of the sequence to the family. Thus, the choice of protein family members to be included in the multiple sequence

alignment when building a pHMM is the major determinant of what the model will confidently detect. The pHMMs used in antiSMASH's primary BGC's detection phase are acquired from a variety of sources, where the models have been built from multiple sequence alignments of validated protein family members. Where available, the models used are from publicly accessible, standardised, protein family databases that are actively curated and validated, such as the Pfam,[202] TIGRFAMS[203] and SMART[204,205] databases. In other cases, the models have come from specialised projects aimed at mining or analysing specific classes of BGC (e.g. BAGEL2/3:[206,207] bacteriocins). Where no appropriate pHMMs were already available for a protein family or signature sequence of interest, the developers of antiSMASH use a method outlined in the antiSMASH publication[194] to select sequences to align and create a suitable model. Invariably for all sources, the models are generated from the multiple sequence alignments using the HMMer3[208,209] software. As such, the sources are dependent on the class of BGC being detected, for example, of the four the signature pHMMs used for detection of the NRPS class of BGCs three are taken from the Pfam database and one was built specifically by the AntiSMASH developers, whereas the pHMMs used for the PKS classes are taken from a combination of the SMART6[210] database, the Pfam database, work published by Yadav et al.,[211,212] and the antiSMASH developers. Several other pHMM sources are drawn upon for the detection of specific BGC classes to allow antiSMASH to detect its full complement of BGCs. The complete list of pHMMs and their sources can be found in the supplementary materials associated with the relevant antiSMASH release publications. A key strength of the antiSMASH pipeline is the ability to incorporate additional pHMM models, and cluster detection rules, thus allowing advanced users to design custom searches for specific sub-classes of BGC, or even expand the search capabilities to capture newly discovered BGC types.

1.3.2. Genome mining summary

Genome mining is a powerful tool that complements metagenome sequencing to facilitate the discovery of novel BGCs.[188,213] The true potential for bacteria to produce secondary metabolites was initially underestimated until the genomes of antibiotic producing actinomyces were first fully sequenced and analysed, revealing many silent or 'cryptic' BGCs. With the current state of high-throughput sequencing and the development of specialised bioinformatic tools novel BGCs are now being found at an unprecedented rate. These technological improvements have extend the search space of BGCs to uncultivated bacteria by allowing the routine analysis of eDNA. This is useful as BGCs from

environmental bacteria that are genetically distant from anything in the present collection of cultivable strains and can provide novel chemical scaffolds with unseen mechanisms of action. Novel chemical classes with new mechanisms of action will likely be required to overcome the growing resistance to current therapeutics. [181]

There are continuing efforts in the development of genome mining software and databases, as well as metagenomic assembly software to leverage the advancements seen in the availability and quality of sequencing data.[214] The successes of genome mining provides opportunities for the development of molecular techniques and synthetic biology[215] to convert novel BGCs into novel bioactive molecules. Efforts in this area have been met in kind to support genome mining as a productive way to yield new drug-like compounds to supply the drug discovery pipeline.[62,216]

1.4 Sponge derived secondary metabolites

1.4.1. Drugs from sponges

Chemical exploration has shown the marine environment to be a rich source of bioactive natural products displaying a range of activities of pharmaceutical importance.[217,218] A report published in February of 2020 discusses 1554 new marine natural product compounds from 469 papers,[219] a 4% increase from the 1490 new compounds in 477 papers reported for 2017.[220] Sponges are particularly rich in bioactive natural products, during this two-year period, sponges were the source material for 453 new or new to the marine environment compounds and validated bacterial sources accounted for 482 new or new to the marine environment compounds. One of the most prolific sponge sources of known marine natural products is the well-studied *Theonella swinhoei.*[221] Sub-species of this sponge can be found with at least three different phenotypes each displaying variations in the colour of the exterior (ectosome) and interior of the sponge body. These different coloured sub-species display unique largely non-overlapping bioactive chemical profiles.[222,223] Like many sponges, the interior harbours hundreds of uncultivated bacterial phylotypes.[224] Of these phenotypes, the sub-species termed *T. swinhoei* 'Y'[225] (referring to a yellow interior) collected from the ocean around the Hachijō-jima island of Japan has been extensively studied for its bioactive metabolites.[226] This phenotype is associated with an onnamide and polytheonamide chemotype, and more than 40 other bioactive compounds including peptides and polyketides have been isolated from this single phenotype.[227] The polytheonamide cytotoxins

(polytheonamides A-C) are structurally complex 49-reside linear peptides composed of many non-proteogenic amino-acids. Initially thought to be produced by an NRPS, it was subsequently discovered by metagenomic cloning experiments that these compounds are in fact highly modified ribosomally synthesised and post-translationally modified peptides (RiPPs).[228] The polytheonamide BGC, isolated in metagenomic cloning experiments, comprises 12 open reading frames, seven of which encode the genes required to carry out the 49 modifications to the precursor peptide which is also encoded within the BGC.[229,230] The architecture of the BGC is polycistronic and lacks any introns which was suggestive of a bacterial origin for this BGC. This observation led researchers to explore the bacterial symbionts hosted by *T. swinhoei*.[222]

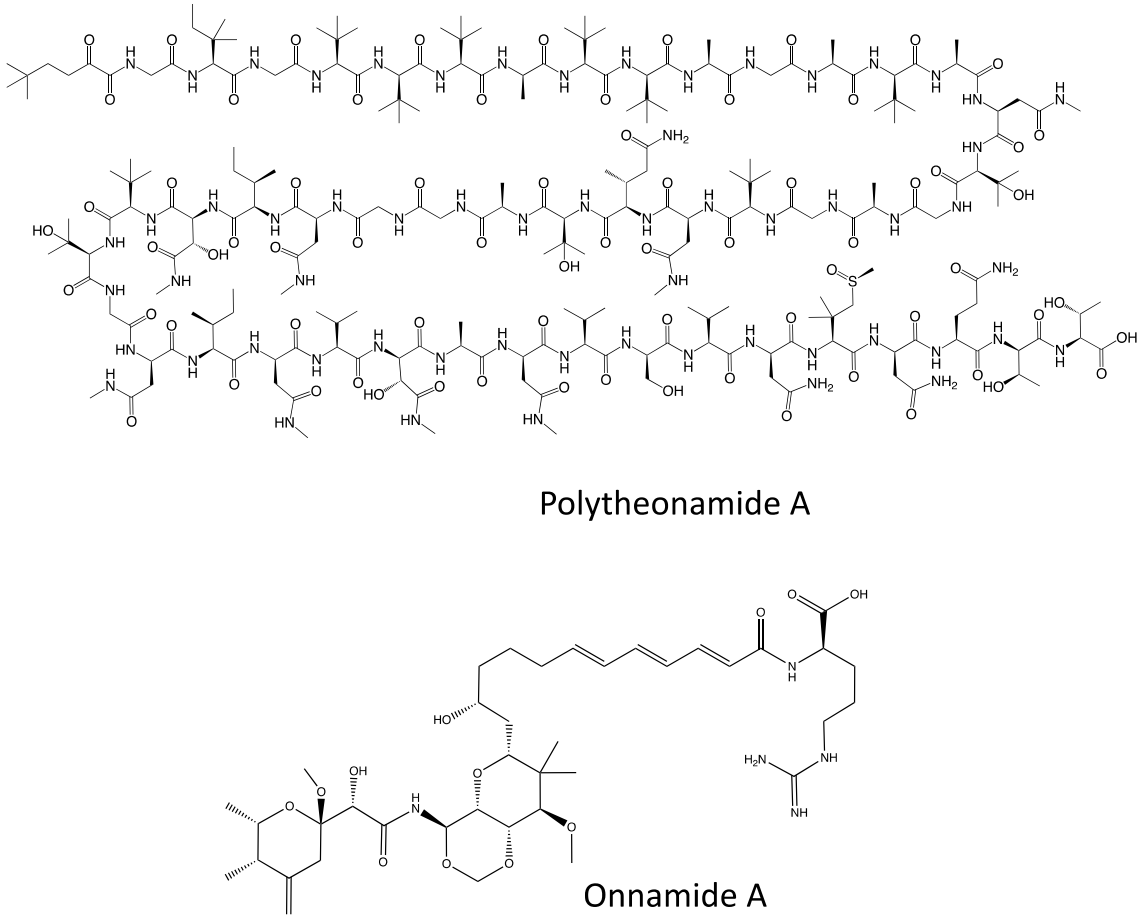


Figure 1.9 – Chemical structures of *T. swinhoei* secondary metabolites.
A) The RiPP, polytheonamide A, produced in a sub-species of the sponge *T. swinhoei* with a yellow interior. B) The *trans*-AT type I polyketide, onnamide A, also produced in this sub-species.

Metagenomic and cell isolation studies revealed that the majority of the compounds isolated from this sponge are indeed of bacterial symbiont origin, moreover, all isolates were found to be produced by only two species of the bacterial genus *Entotheonella* with the majority

produced by "*Candidatus* Entotheonella factor".[30] The comparatively large 9 Mb genomes of these bacteria contain a plethora of BGCs, many of yet unknow function, but may synthesise compounds that are thought to play a role in chemical defence of the sponge host.[231] *Theonella swinhoei* sponge samples from diverse geographic locations all accommodate species of symbiotic *Entotheonella* and show overlapping as well as unique chemical profiles. Species of *Entotheonella* have also been found in other genus of sponge hosts that display bioactive chemical profiles.[232,233] The lithistid sponge *Discodermia dissoluta* contains the powerful anticancer polyketide discodermolide.[234] This sponge also has *Entotheonella sp.* as an abundant symbiont found predominantly in the mesohly.[232] This symbiont was discovered to be the producer of Calyculin A, another bioactive secondary metabolite isolated from this sponge.[235,236] This suggests a link between natural product rich sponges species and the widely distributed endosymbiont genus *Entotheonella*.[30,86,225,227] Microbial symbionts can make more than 40% of an animal's wet weight in some "bacteriosponges".[237] In many cases symbiotic microbes are suspected or proven to be the true produces of the rich source of natural products derived from marine sponges.[238,239] An impressive example of this is the production of crystalised polybrominated secondary metabolites that make up to 12% dry weight of the sponge *Dysidea herbacea* is actually by a symbiont cyanobacteria *Oscillatoria spongeliae.*[240] Although certain sponge species may be rich in the diversity of their natural product repertoire and endosymbiont community, accessing target compounds can be problematic. For example, one metric ton sponge was harvested to produce only 300mg of compound in the case of the sponge-derived actin inhibitor halichondrin B. [241]

The marine sponge *Mycale hentscheli* is endemic to New Zealand and crude extracts of this sponge are usually exceedingly potently cytotoxic to mammalian cell lines.[242] This has attracted researches to investigate the natural products produced by this sponge.[84]

1.4.2. Mycale hentscheli secondary metabolites

M. hentscheli produces at least three classes of biologically active compounds with pharmaceutical potential, namely the mycalamides (section 1.4.2.1), the pateamines (section 1.4.2.3) and the pelorusides (section 1.4.2.2). Considerable efforts, including multistep-synthesis and aquaculture,[243] have been invested in accessing these compounds in significant quantities to allow further investigations into the bioactivity and to allow for clinical trials.[244] The definitive source (i.e., sponge, symbiont or other) and biosynthetic pathways of these compounds was unknown.[245] However, the mycalamides have marked structural similarity

to the secondary metabolite pederin produced by terrestrial symbiont bacteria of the South American *Paederus* beetle.[246] This suggests the biosynthetic pathway to mycalamides is encoded by a functional BGC that resembles bacterial BGCs and potentially harboured by a bacterial symbiont of this sponge. The pateamines and pelorusides also display structural scaffolds and elements that are reminiscent of bacterial NRPs and polyketides, respectively.

1.4.2.1. Mycalamide A

Mycalamide A (MycA) is a potent inhibitor of protein synthesis and a sub-nanomolar inducer of apoptosis in some tumorigenic cell lines.[35] On the basis of chemical structure alone, mycalamide belongs categorically to the Pederin-like family of cytotoxins (See Figure 1.10). Pederin was one of the first *trans*-AT polyketides to be described and is produced by a bacterial symbiont of the rove beetle (genera *Paederus*) which uses the compound as chemical defence [36]. This group of *trans*-AT polyketides also includes many other close structural congeners that are produced by bacterial symbionts from various species of marine sponges, for example onnamide A and theopederin A from *T. swinhoei*[43]. The metagenomic analysis confirming the association between cytotoxin production and the bacterial symbionts has in turn provided the sequences of the BGCs which show a corresponding homology to one another as do the chemical structures of the family members. This foundation of experimental characterisation trivialises the identification of new BGCs belonging to this family when analysing even complex metagenomic data sets. Sequence homology between a KS-domain from the pederin BGC and a KS-domain amplified from a mycalamide chemotype *M. hentscheli* sample was reported by Fisch, K. in 2009.[247] However, the rest of the mycalamide BGC was not recovered in that study.

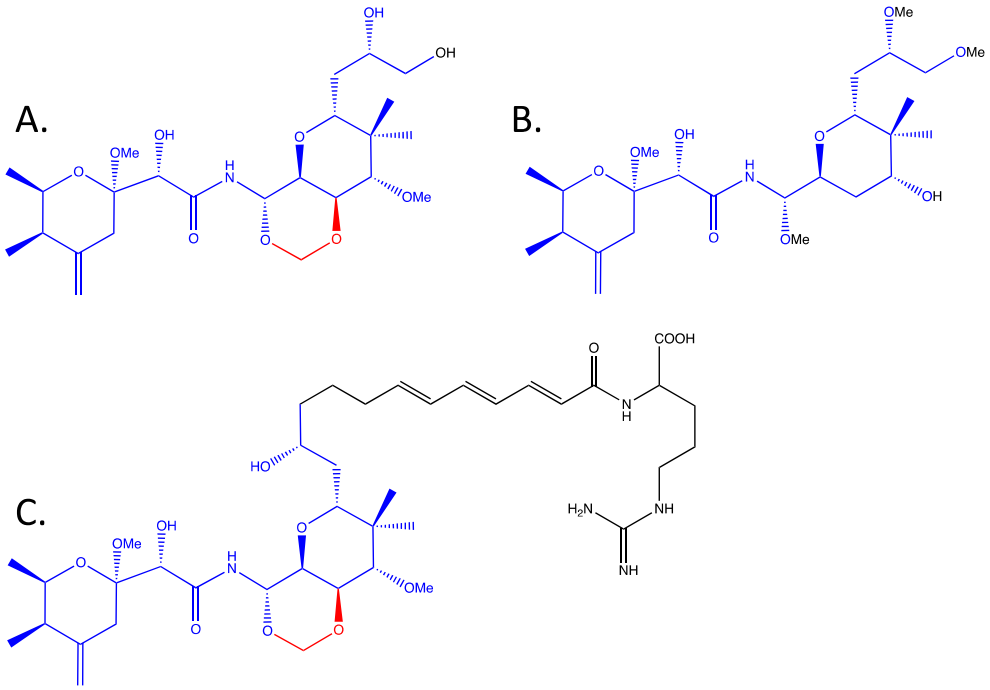


Figure 1.10 – Mycalamide A and related structures in the Pederin-like family.
Structure A. is that of mycalamide A, structure B. is pederin and C. is onnamide A. The blue skeleton highlights the structural elements of mycalamide A common between all three molecules. The red skeleton highlights the structural elements shared between mycalamide A and onnamide A. The black skeleton is unique to each molecule.

1.4.2.2. Peloruside A

Peloruside A (PelA) is a macrocyclic polyketide and a powerful microtubule stabilising agent (MTA) discovered in 1989 from analysis of *M. hentscheli* extracts.[31,38] A member of an important and effective class of cancer chemotherapeutics, MTAs, PelA has a clinically unique non-taxoid binding site on the β-tubulin subunit.[39] This unique binding site facilitates the use of peloruside synergistically with taxoid site targeting MTAs and for it to remain effective in cancer cell lines that have developed taxoid site mutation mediated resistance.[40] Reports of therapeutic efficacy with low toxicity in preclinical xenograft animal models and other promising anti-neurodegenerative and immunomodulatory effects have garnered continuing research interests in PelA many years after its discovery.[41] This ongoing research has focused in part on providing meaningful access to the molecule as current total chemical synthesis and natural sources have proven impractical for both economic and sustainability considerations. These issues plague the clinical advancement of many otherwise promising NPs. Like many sponge derived bioactive NPs comprised of complex polyketides, PelA shows close structural relationships, especially in the appearance of β-alkylation, to other *trans*-AT NPs from symbiont bacterial origins. This

highlights the need for the elucidation and understanding of the NP BGCs and the genomic context of the producers as a means to develop a useful supply via native or heterologous expression systems.[42]

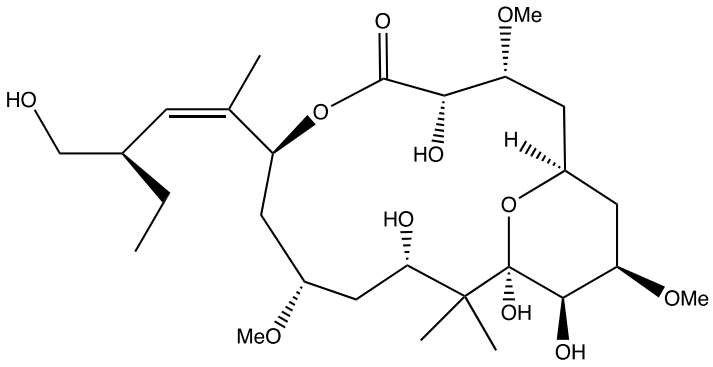


Figure 1.11 - Peloruside A. Structure of cytotoxic ribosome inhibiting compound peloruside A, isolated from the marine sponge *M. hentscheli*

1.4.2.3. Pateamine A

Pateamine A (PatA), isolated form *M. hentscheli* in 1991, blocks translation initiation in eukaryotic protein synthesis by interfering with the activity of the helicase eIF4A.[43] As well as making PatA a highly potent antiproliferative cytotoxin the therapeutic repertoire of this NP also includes immunosuppressive and anti-cachectic activity.[44,45] PatA is uniquely and highly selective for eIF4A activity modulation thereby rendering it a valuable molecular tool for probing the regulation and complexities of translation initiation, a critical cellular function and therapeutic target.[46] Total chemical synthesis of PatA, a thiazole-containing dilactone macrolide, was published as early as 1998[45] but this multi-step process was complex and prohibitively low yielding. Spurred on by the promising therapeutic properties of PatA, work has been on going to simplify and optimise the synthesis with a new synthetic route reported as resent as 2018.[47] While significant improvements were reported, this NP remains problematic and expensive to synthesis due to the molecule's structural complexity, as such a solution to the commercial scale supply problem is yet to be realised.[48,49] A retro-biosynthetic analysis of the structure of PatA suggests this molecule is producible by a hybrid *trans*-AT PKS/NRPS BGC. Several structural features of note, known to be synthesised by this biosynthetic class, informed the predictions of cognate components that should be represented in the BGC architecture namely, polyene β-branching, thiazol incorporation,

amino incorporation and methylation, and the unusual dilactone structure. Predicting biosynthetic components based on the patterns of specific structural handles in target secondary metabolites aids in the analysis of genomic data to discover BGC of interest.[248]

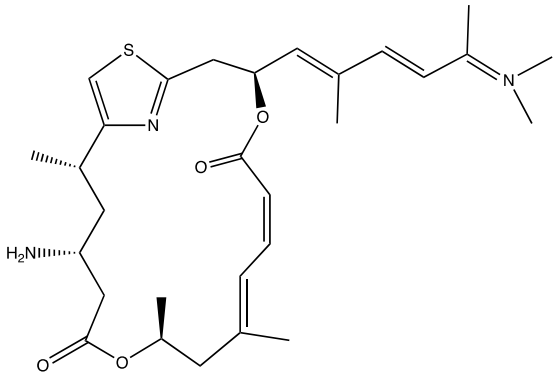


Figure 1.12 - Pateamine A. Structure of cytotoxic translation inhibiting compound pateamine A, isolated from the marine sponge *M. hentscheli*

1.5 Aims and objectives

The work coved in this thesis primary sought to use metagenomic techniques to discover the biosynthetic gene clusters responsible for the production of the three main bioactive secondary metabolites, mycalamide, peloruside and pateamine, that have been isolated from extracts of *M. hentscheli.* Second to this, would be the use of the metagenomic data produced for an investigation of the microbiome of *M. hentscheli* to understand the secondary metabolism in the context of the symbiont community. A stretch goal was the recovery of biosynthetic clusters to allow enzymatic and heterologous expression studies to be conducted. Specific project goals:

1) Extraction and isolation of high quality, high molecular weight (HMW) DNA from *M. hentscheli.*
2) Build a metagenomic cosmid library from the HMW DNA.
3) Generate high quality short read and long read sequence data from the HMW DNA.
4) Metagenome assemblies of sequencing data.
5) Interrogation of metagenomic assemblies to identify target and off target BGCs.
6) Analyse microbial communities from assembly data.
7) Targeted analyses of cosmid libraries to recover BGCs.

1.6 Components of the research and contributions to knowledge

1.6.1. Components of research

This research project covered two main components. The first component was concerning the nature of the bio-chemical and bio-synthetic genetics of the secondary-metabolites produced by the target sponge samples. This was motivated by the potential of eventually accessing the secondary-metabolites produced by the sponge in a sustainable and scalable way and in furthering out understating of the production of complex secondary-metabolites in general. Research was undertaken to extract and analyse both secondary-metabolite constituents and the genetic material of the host sponge and the residing microbial population. The chemical extracts containing the secondary-metabolites were used to confirm the presences of the target metabolites in the samples at hand and were also fractionated and stored to allow for continued analysis as new information was gleaned from the genetic analysis. The extracted genetic material was initially cloned into metagenomic cosmid libraries to allow for the isolation and expansion of specific fragment for analysis, and also for archival purposes. The genetic material was also DNA sequenced deeply using multiple complementary short and long read technologies. The analysis conducted in this primary research component reviled partial answers towards understanding the bio-synthetic nature of the target secondary-metabolites and highlighted several new potential targets of investigation.

The second major component of the research, spurred on by the findings in the former, was centred in the computational and bioinformatic analysis of the DNA sequence data generated from the metagenomes of the sponge samples. While this research component also shared the objective of further understanding the production of the sponge secondary-metabolites, a comprehensive metagenomic analysis of this sponge species had not previously been published. This presented the opportunity to also investigate new hypotheses about the make-up of the microbial population and bio-synthetic potential of sponge metagenome. The analysis here ultimately provided a wider context about the structure of secondary-metabolite gene clusters and their relationship to the microbial population.

1.6.2. Components of research

This research project covered two main components. The first component was concerning the nature of the bio-chemical profile and bio-synthetic genetics of the secondary-metabolites produced by the target sponge samples. This was motivated by the potential of eventually accessing the secondary-metabolites produced by the sponge in a sustainable and scalable way, and in furthering our understating of the production of complex secondary-metabolites in general. Research was undertaken to extract and analyse both the secondary-metabolite constituents and the genetic material of the host sponge and the residing microbial population. The chemical extracts containing the secondary-metabolites were used to confirm the presences of the target metabolites in the samples at hand and were also fractionated and stored to allow for continued or follow up analysis as new information was gleaned from the genetic analysis. The extracted genetic material was initially cloned into metagenomic cosmid libraries to allow for the isolation and expansion of specific fragment(s) for analysis, and also for archival purposes. The genetic material was also DNA sequenced deeply using multiple complementary short and long read sequencing technologies. The analysis conducted in this primary research component revealed partial answers towards understanding the bio-synthetic nature of the target secondary-metabolites and highlighted several new potential targets of investigation.

The second major component of the research, spurred on by the findings in the former component, was centred in the computational and bioinformatic analysis of the DNA sequence data generated from the metagenomes of the sponge samples. While this research component also shared the objective of further understanding the production of the sponge secondary-metabolites, a comprehensive metagenomic analysis of this sponge species had not previously been published. This presented the opportunity to also investigate new hypotheses about the make-up of the microbial population and the bio-synthetic potential of the sponge metagenome(s). The analysis here ultimately provided a wider context about the structure of secondary-metabolite gene clusters and their relationship to the microbial population.

1.6.3. Contributions to knowledge

This research has furthered the understanding of the production of several classes of complex bacterial secondary-metabolite compounds by fully or partially resolving the genetic sequence of large biosynthetic gene clusters that produce three known bioactive compounds. The raw source material used in this analysis is limited and transient. This research examined

and catalogued a snapshot of this niche environment to a breadth and depth that had not previously been possible. Entire chromosomal sequences of uncultivated putative microbial species were also resolved in this research, broadening the known diversity of the tree-of-life and focusing a light on the complex sponge-host microbiome interaction. The raw DNA sequence datasets from several sponge samples, the assembled contigs and putative microbial genomes, and the annotated biosynthetic gene cluster sequences have all been made publicly available. The appropriate accession numbers for access these resources can be found accompanying the relative results sections of this thesis.

1.6.3.1. Publications

The main findings surfaced by this research have been published under the title "*Metagenomic Exploration of the Marine Sponge Mycale hentscheli Uncovers Multiple Polyketide-Producing Bacterial Symbionts*" in the journal *mBio* in March 2020 (doi: 10.1128/mBio.02997-19).[249] This was accompanied by the simultaneous publication of a related work[250] in alliance with a collaborating lab group also studying the production of bacterial secondar metabolites including those from the same sponge sample. Both works showed highly concordant results where subject matter was overlapping. Minor differences in interpretation of the details of some results are addressed in the relevant sections.

2 Methods and Materials

2.1 Sponge sample collection

The marine sponges of the species *M. hentscheli* (*n* = 7) used in this study were collected by scuba at two locations in the Marlborough Sounds, South Island, New Zealand. Sample `MH_PAT` was collected from a depth range of 5 to 15 m at Capsize Point in November 2014. The six additional specimens of *M. hentscheli* included in this study (designated `MH_PEL`, `S1`, `S2`, `S3`, `S4` and `S5`) were collected during a separate expedition from a depth range of 5 to 15 m at Pelorus Sound in May 2003. All sponge samples were stored below at least -20 °C until subsampled for organic and DNA extractions. The provenance of the samples since collection is unknown. Collection details were provided from archived diving logbooks and personal communication.

2.1.1. Sponge subsampling

Where possible, sponge samples were processed while still frozen by taking longitudinal sections through an entire leuconoid branch to include the mesohyl, atrium (spongocoel), choanocytes and epidermal tissues to capture the widest variety of symbiont habitats. Solid frozen sponges were cut with a 50mm wide bevel chisel (Irwin Maples) and mallet. The sectioned tissue was then ground under liquid nitrogen to a fine powder with caution to avoid skin contact and inhalation with the use of appropriate personal protective equipment.

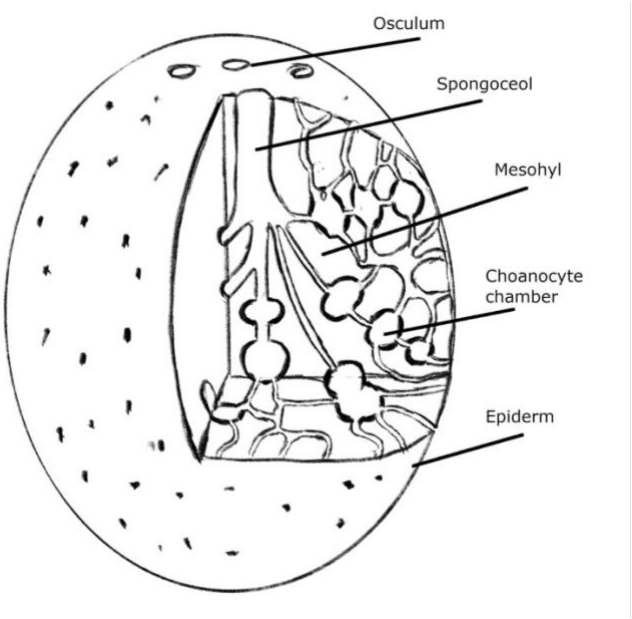


Figure 2.1 – Leuconoid branch sponge anatomy. Diagram of a leuconoid type sponge showing the main structures and tissue types targeted in the sponge sampling process.

2.2 Microbiological methods

All purchased reagents were of analytical or molecular biology grade, obtained from Sigma Aldrich unless otherwise noted. Dilutions were performed using distilled de-ionised water (ddH_2O) unless otherwise stated. Solutions pH were adjusted using 10 M HCl or 10 M NaOH. Sterilisation was performed by autoclave, or by filtration through 0.22 μm syringe filter. Media supplements and antibiotics were added to media following autoclaving.

2.2.1. Laboratory *Escherichia coli* strains used in this study are detailed in Table 2-1.

Strain	Description	Source
E. coli DH10β (EC100)	F-*mcrA*, Δ(*mrr-hsd*RMS-*mcr*BC) φ80d*lac*ZΔM15 Δ*lac*X74 *rec*A1 *end*A1 *ara*D139 Δ(*ara*, *leu*)7697 *gal*U *gal*K λ-*rps*L *nup*G	Epicentre
E. coli NM 759	*rec*A56 Δ(*mcr*A) e14° Δ(*mrr-hsd-mcr*) (λ*imm*434 *clts b2 red*3 *Dam*15 *Sam*7)/λ (D^- extract)	[251]
E. coli BHB 2688	N205 *recA-(λimm434 cIts b2 red3 Eam4 Sam7)*/λ (E^- extract)	[252]

Table 2-1 – Bacterial strain used in this study

2.2.2. Plasmid vectors used in this study are detailed in Table 2-2.

Vector	Description	Source
pWEB::tnc	Cosmid vector. ColE1ori, Chl^R, Amp^R, *cos*	Epicentre
pWEB	Cosmid vector. ColE1ori, Kan^R, Amp^R, *cos*	Epicentre

Table 2-2 - Plasmid vectors used in this study

2.2.3. Growth media

Media was sterilised by autoclaving for 20 minutes at 121 °C. Supplements and antibiotics were added to cool media post sterilisation. Solid media followed the below liquid media recipes with agar added prior to sterilisation to either 1.5 % (w/v) or 2 % (w/v). Standard cultivation in liquid and solid media was in Low salt lysogeny broth (LB) medium. Solid media was made with the addition of 1% w/v agar prior to autoclave sterilisation.

Low salt lysogeny broth (LB) medium : Tryptone 10 g/L, NaCl 5 g/L, yeast extract 5 g/L

2.2.4. Media supplements

Supplements were dissolved in ddH_2O unless otherwise stated. Antibiotic stock concentrations were made to 1000x the working concentration in media. Supplements were sterilised by filtration through 0.22 μm syringe filter.

Working stock dilutions:

1000x ampicillin: 200 mg/mL

1000x kanamycin: 50 mg/mL

1000x chloramphenicol: 34 mg/mL in 95 % ethanol

2.2.5. Microbial growth and maintenance

LB liquid media was used for routine growth and maintenance of *E. coli* strains. Standard cultures were grown to stationary phase by 16 h incubation in 5 ml of LB in sterile 15 ml centrifuge tubes capped loosely with aluminium foil to allow for gas exchange. Culture incubation was at 37 °C with 200 rpm orbital shaking. For long-term maintenance of cultures, an aliquot of stationary phase culture was mixed 1:1 with 80 % glycerol and stored at -80 °C.

2.2.6. Electroporation

Electroporation was carried out at 2.5kV, 2.5mm for 5.0 ms after which 1ml of pre-warmed SOC was immediately added by gentle resuspension. Cells were then incubated (37°C, 200rpm, 30 min) for recovery. After recovery, cells were pelleted by centrifugation (5,000g, room temperature, 5 min), resuspended in 100µl of supernatant and plated out on LB agar plates with appropriate selective agents using sterile plating beads.

SOC: 0.5% (w/v) yeast extract, 2% (w/v) tryptone, 10mM NaCl, 2.5mM KCl, 20mM MgSO4, 20mM glucose

2.3 Molecular Biology methods

2.3.1. Oligonucleotide primers

All primers used were synthesised by Macrogen Inc. (Korea).

Primers used in this study are detailed below (Table 2-3)

Primer	Sequence
T7_promoter	5'-TAATACGACTCACTATAGGGAGA-3'
M13F	5'-CGCCAGGGTTTTCCCAGTCACGAC-3'

Table 2-3 - Primer sequences

2.3.2. Plasmid and cosmid DNA extraction and isolation

Plasmid and cosmid DNA was isolated from overnight *E. coli* cultures in LB, using the standard Qiagen protocol with EconoSpin® silica spin columns obtained from Epoch Life Science and the following buffers prepared in house.

2.3.2.1. Miniprep Buffers:

Buffer P1: 50 mM Tris-HCl pH 8.0,10 mM EDTA, 100 μg/mL RNaseA

Buffer P2: 200 mM NaOH, 1% SDS

Buffer N3: 4.2 M Gu-HCl, 0.9 M potassium acetate, pH 4.8

Buffer PB: 5 M Gu-HCl, 30% isopropanol

Buffer PE: 10 mM Tris-HCl pH 7.5, 80% ethanol

Buffer EB: 10 mM Tris-HCl pH 8.5

Minipreps were eluted with 60 μL Elution buffer (EB), preheated to 50 °C.

Elution buffer: 10mM Tris-Cl, pH 8.0

2.3.3. High molecular weight (HMW) DNA extraction and isolation

Paramount to the success of cosmid library building and long-read sequencing is the isolation of HMW DNA from the environmental sample. The extraction method used in this study was based on that described in Gurgui & Piel (2010).[253] The entire process can be divided into several major steps which are (1) sample lysis, (2) raw DNA extraction, (3) gel purification and size selection, and (4) final concentration.

2.3.3.1. Sample lysis

Powdered sponge tissue (~1 g wet sponge), prepared as described in 2.1.1, was treated with 10x volume freshly prepared pre-warmed sponge lysis buffer and incubated at 60 °C for at least 20 minutes with gentle agitation every five minutes. Prior to lysis, to help concentrate the microorganisms, large sponge tissue particles and inorganic debris can be separated and removed by differential centrifugation. The sample was suspended in ~10mls spin buffer followed by brief slow speed centrifugation (i.e., 250 g, 1 min). The supernatant was transferred to a fresh tube and then centrifuged at high speed (i.e., >4,000 g, 5 mins), the resulting pellet was then resuspended up to 1ml in spin buffer followed by lysis. Application of the technique was optional and was used where the amount of sample permitted, otherwise bulk tissue was lysed.

Sponge lysis buffer:

8 M urea, 2% sarkosyl, 1 M NaCl, 50 mM EDTA, 50 mM Tris–HCl. Adjust to pH 7.5 with HCl.

Spin buffer:

1 M NaCl, 50 mM EDTA. Adjust to pH 7.5.

2.3.3.2. Raw DNA extraction

The raw lysate from previous the step was extracted three times with freshly distilled tris-buffered phenol/$ChCl_3$. This was done by adding an equal volume of the phenol solution to the lysate and inverting until a suspension is formed. The phases are then split by centrifugation. The top aqueous phase was then washed twice again in the same manor. This was followed by a $ChCl_3$ to remove any remaining phenol. The interface between the two surfaces was additional spin buffer and re-extracted to increase yield. The clear aqueous fractions were combined, and the bulk nucleotides were precipitated with one volume of isopropanol (100%, chilled) and the addition one-tenth volumes of 3 M sodium acetate (pH 5.2). The tube was very gently inverted and placed in ice for >30 minutes prior to high-speed (>400 g) centrifugation for 15 minutes. The pallet was redissolved in 0.5 mL of TE (pH 8.0). To this, 25ul of DNase free RNase A solution (10mg/mL) was added.

The freshly distilled phenol was found to be beneficial for maintaining the integrity of the HMW DNA. This was produced using a vacuum distillation apparatus fitted with a Vigreux condenser and heated with an electric mantle and heat gun. All distillation procedures were preformed using personal protective equipment and in a ventilated fume cabinet.

2.3.3.3. Gel purification and size selection.

The raw DNA was then purified from co-precipitating contaminants and low molecular weight DNA by low density gel electrophoresis and recovered by electroelution.

An agarose gel of 0.8% made with TAE buffer (Tris-Acetate EDTA) was set with a single large loading lane. The entire sample was loaded with the aid of bromophenol blue and glycerol loading buffer (6x) prior to filling the electrophoresis chamber with the TAE buffer level with the top of the gel to avoid loss of the buoyant sample. Electrophoresis was carried out in two stages. The first stage lasted for ~1h with the voltage clamped at 100V and was monitored to avoid melting or deformation of the gel due to overheating. Once the sample had visibly migrated into the gel matrix the voltage was clamped to 20V, and electrophoresis was continued for ~16h. Several complete changes of the TAE buffer in the electrophoresis chamber were completed over the duration of the run to dilute any liberated contaminants and refresh the buffer.

TAE buffer (50x): 2M Tris, 1M acetic acid, 50mM EDTA.

To make: Dissolve 242 g Tris-base in ~700 mL. Add 100 mL 0.5M EDTA and 57.1 mL glacial acetic acid. Bring up to 1 L total volume.

Final working concentration: 40 mM Tris, 20 mM acetic acid, 1 mM EDTA

2.3.3.4. Final DNA concentration

On completion of electrophoresis the HMW DNA band was excised from the gel as described in figure 3.3. The DNA was then electroeluted from the gel slice using a CBS scientific electroelution tank with the electroelution blocks fitted with Spectra/Por 3 dialysis membranes (MWCO: 3.5 kDa). Electroelution buffer was recirculated with a peristaltic pump and the voltage was clamped to 100V. The entire buffer volume of the electroelution block was collected and the DNA was concentrated using an Amicon Ultra centrifugal filter unit Ultra-15 (MWCO: 30 kDa).

Electroelution buffer: 5mM Tris-Ac pH 7.5, 1mM EDTA.

2.3.4. Commercial enzymes

Enzymes used for the production of cosmid libraries. End-It DNA End-Repair Kit and Fast-link DNA ligase were obtained from Epicentre. Restriction endonucleases, recombinant shrimp alkaline phosphatase (rSAP) and Cas9 were obtained from New England Biolabs.

2.3.5. Commercial phage packaging extracts

The commercially sourced MaxPlax™ phage packaging extracts used for the production of cosmid libraires were purchased from Lucigen. Extracts were stored at -80 °C until thawed on ice for use.

2.3.6. Magnetic bead DNA purifications

Carboxyl-coated superparamagnetic iron oxide nanoparticles (SeraMag SpeedBeads (FisherScientific #09-981-123)) were prepared for solid-phase reversible immobilization (SPRI) of DNA for routine purification and concentration tasks. One volume of DNA SPRI mix with 1:50 commercial bead suspension was added to the sample to be purified. The beads were then coalesced on a magnet and washed with excess 75% EtOH. The bound DNA was then eluted in an appropriate amount of EB or water.

DNA SPRI mix: 10 mM Tris, 1 mM EDTA, 2.5 M NaCl, 20% PEG 8000, 0.05% Tween 20, pH 8.0

2.4 Metagenomic cosmid library construction

The following methods were employed for the construction of metagenomic cosmid libraries from extracted HMW sponge eDNA using either lab-made or commercially sourced phage packaging extracts.

2.4.1. Lab-made phage packaging extract preparation

The use of particular high-grade reagents appeared to be critical for success in this protocol, the following suppliers were used for each reagent are listed below.

- Agar (Becton, Dickinson and Company (BD) Bacto™ Agar (Ref 214010))
- Casein digest (Becton, Dickinson and Company (BD) Difco™ Casein Digest (Ref 211610))
- Yeast extract (Becton, Dickinson and Company (BD) Bacto™ Yeast Extract (Ref 212750))ATP, pH 7.0
- β-mercaptoethanol (Sigma-Aldrich 2-Mercaptoethanol (M3148))
- EDTA (Sigma-Aldrich EDTA (EDS))
- Lysozyme (Sigma-Aldrich Lysozyme from chicken egg white (62970))
- $MgCl_2$ .6 H_2O (Sigma-Aldrich Magnesium chloride hexahydrate (M2670))
- Putrescine dihydrochloride (Sigma-Aldrich (P5780))
- Spermidine tryhydrochloride (Sigma-Aldrich (85578))
- Sucrose (Sigma-Aldrich (S7903))
- Tris-base (Sigma-Aldrich Trizma® Base BioXtra (T6791))
- ATP (Sigma-Aldrich Adenosine 5′-triphosphate disodium salt hydrate (A7699))
- HCl (Sigma-Aldrich HCL 36.5-38% (H1758)
- NaOH (Scharlau, Reagent Grade (SO 0425))
- Chloroform (does not need to be from a specific supplier)

Using the above reagents, the protocol described in Techniques in Aquatic Toxicology, Volume 2, chapter 38, appendix 1: Preparation of λ packaging extracts (Winn and Norris 2005)[254] was followed. Minor changes to the protocol were made to suit laboratory specific conditions.

2.4.2. pWEB::TNC vector preparation

Liquid LB (Amp_{200} Chl_{25}) was inoculated with the pWEB::tnc transformed EC100 cell stock and grown to stationary phase. The pWEB::tnc cosmid was then miniprepped (section 2.3.2), quantified and checked for purity (A260/A230 > 1.6 and A260/280 > 1.8) using a NanoPhotometer NP80. Aliquots of 40 µg magnetic bead purified pWEB::tnc were then digested overnight using 100U SmaI in a final volume of 500µl at 25°C. A further 60U of SmaI was added the next morning and incubated (25°C, 2 h). Next, the digested pWEB::tnc was dephosphorylated by adding 8U of rSAP and incubating (37°C, 2 h). Enzymes were then heat-inactivated at 65°C for 5 minutes. The prepared pWEB::tnc was then precipitated using 0.7x vol. isopropanol and 0.1x vol 3M NaOAc (pH 5.2) and pelleted by centrifugation (16,000g, 4°C, 30 min). The supernatant was discarded and the DNA pellet washed twice with 70% ethanol. After letting the pellet air dry for approximately 3 minutes, pWEB::tnc was dissolved in 80 µl of EB by incubating at room temperature overnight. The next morning pWEB::tnc was again quantified and checked for purity (A260/A230 > 1.8 and A260/280 > 2.0) using the NanoPhotometer NP80. Where applicable, pWEB::tnc was further diluted using EB to a concentration less than 400 ng/µl. Successful digest of pWEB::tnc was confirmed by running a 1% agarose gel in TAE. Ligation efficiency of digested and dephosphorylated pWEB::tnc was quantified by electroporating equal molar amounts of digested pWEB::tnc, digested pWEB::tnc after a ligation reaction using the Quick Ligation™ Kit, undigested pWEB::tnc as a positive control and 1x Quick Ligation Reaction Buffer as a negative control. Each electroporation consisted of 50 ng of DNA (no more than 5 µl volume) and 50µl EC100 electrocompetent cells in cold electroporation cuvettes. Plates were incubated (37°C, overnight) and colonies counted the next day. Only digested and dephosphorylated vector stocks, which resulted in 3 or more orders of magnitude less colonies from the "digested vector" and "digested vector after ligation reaction using the Quick Ligation™ Kit (NEB)" treatments compared to the "undigested vector" treatment, were used for further experiments.

2.4.3. Cosmid library packaging

The eDNA substrate for metagenomic cosmid library construction was first extracted from sponge samples and size selected as detailed in 2.3.3. The DNA purity was quantified by nanodrop spectrophotometer and the concentration was confirmed by Qubit™ fluorometer using the dsDNA HS Assay Kit. The sample was then end-repaired for blunt-end ligation using End-It DNA End-Repair Kit (Epicentre) according to the manufacturer's instructions. The repaired DNA was isopropanol precipitated to remove the end-repair enzymes, and resuspended in EB.

The cosmid vector pWEB::tnc was prepared for blunt-end ligation as detailed in section 2.4.2. Alternatively, pWEB was supplied blunt-end ligation ready from Epicentre as part of the MaxPlax kit. Prepared cosmid vector (250 ng) and eDNA substrate (125 ng) were then ligated in a final reaction volume of 5 μL using Fast-link DNA ligase (Epicentre) according to the manufacturer's instructions.

λ-phage packaging of ligated eDNA for cosmid library preparation was carried out using a modified version of the method described by Brady (2007).[7]

One tube of each complementary packaging extract strain were thawed, and combined (45 μL strain E^-, 60 μL strain D^-) to create a complete functional packaging extract mixture. Ligated DNA was packaged by adding 33 μL of packing extract per ligation reaction with incubation at 30 °C for 90 minutes, followed by the addition of a further 33 μL packaging extract and repeated incubation step. Then, 250 μL of phage dilution buffer was then added, followed by 7 μL chloroform. The sample was gently mixed and the phases separated by brief centrifugation. The diluted packaged phage particles were adhered to an ice chilled day culture of *E. coli* EC100 grown to OD_{600} of 1.0 in LB 10 mM $MgSO_4$ at a 1:10 ratio. This mixture was incubated at room temperature for 20 minutes, aliquoted across 96 wells and then incubated at 37 °C 200 rpm for 75 minutes. The recovered, phage adhered *E. coli* cells were then diluted to 5 mL in LB with appropriate antibiotics for cosmid selection. At this stage, a dilution series was made and plated on to selective media from at least three of cultures to measure the efficiency of the packing reaction and estimate total library size. The entire 96 cultures of the library was expanded by culturing to stationary phase at which time a samples (500 μL) of each well was taken for cryogenic storage (2.2.5) and the remaining culture was miniprepped (2.3.2) and stored for later analysis.

Phage dilution buffer: 10 mM Tris-HCl pH 8.3, 100 mM NaCl, 10 mM $MgCl_2$.

2.5 DNA Sequencing

Illumina, PacBio and Sanger sequencing were provided by third party off-shore service providers. Oxford Nanopore (ONT) sequencing data was generated inhouse.

2.5.1. Illumina

Macrogen Inc. (Korea) performed the 250bp PE sequencing on a four-color chemistry HiSeq 2500 instrument and provided the sequencing library building service using and the TruSeq PCR free kit.

GeneWiz Inc. (China) provide the Illumina HiSeq 150 bp PE sequencing service and TruSeq library preparation.

2.5.2. Pacbio

The PacBio sequencing was provided by Macrogen Inc. on the PacBio Sequel System and sequenced on a single M1 SMRTcell.

2.5.3. ONT

ONT sequencing was performed using the ONT Rapid sequencing kit (SQK-RAD002) and a single SpotON flow cell Mk-I R9 Version (FLO-MIN106). Base calling was processed with Albacore (v 1.0.1).

2.5.4. Sanger sequencing

Sanger sequencing was provided by Macrogen Inc. Premixed sequencing reactions were prepared inhouse. A 10 µL reaction volume containing 5 µL of template cosmid DNA (20-200 ng/ µL) and 2.5 pmol/µL primer in 5 µM betaine was submitted per reaction.

2.6 Bioinformatic analysis

Example scripts for the following standard bioinformatic analysis used in this study are available at the GitHub repository: Mattstorey/thesis_scripts

Bioinformatic processing was conducted on a variety of local and remote/cloud computing platforms including single node workstations, multi-node managed high performance computing (HPC) clusters and Amazon web service (AWS) EC2 instances. Each resource presented differing computation environments. The following will show generalised implementations of the commands used to perform the various tasks were appropriate. Further details can be found in the code repository files. All other file manipulations and calculations were done in the UNIX shell or with Python (=> v3.6).

2.6.1. Read trimming, decontamination, error correction and read-merging.

Prior to assembly and mapping, short read data were pre-processed to remove low quality and contaminating sequences or improve read quality for assembly purposes by error correction and read-merging overlapping reads into a single longer read.

2.6.1.1. Skewer

Skewer (v0.2.2)[255] was used for in paired-end mode trimming adaptors and quality filtering 2x250 PE reads prior to assembly. Quality threshold was set at Q30. TruSeq adaptor sequences for removal were provided by the software package.

```
$ skewer -m pe -q 30\
-o trimmed_reads.fq.gz -z -t 16 /path/to/reads1.fq.gz\
/path/to/reads2.fq.gz
```

2.6.1.2. Trimmomatic

Trimmomatic (v0.36)[256] was used in paired-end mode for trimming operation on the 2x150 PE reads used in this study.

```
$ trimmomatic PE -threads 18 /path/to/reads1.fq.gz\
 path/to/reads1.fq.gz -baseout trimmed_reads.fq.gz\
 ILLUMINACLIP:adapters/TruSeq3-PE.fa:2:30:10 LEADING:3\
 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36
```

Decontamination of human and lab strain sequences was performed with BBDuk of the BBTools[257] suite. Error correction and read-merging was also carried out with BBTools as detailed in the code repository.

2.6.2. Short read assembly

Two iterative de bruijn graph assemblers were used for the assembly of the pre-processed short-read data sets based on the reported ability to handle wide variations in sequence coverage.

2.6.2.1. IDBA_UD

The IDBA-UD[258] assembler was used for to calculate the initial short-read assembly of `MH_PAT`. The source code was slightly modified and recompiled to allow for the assembly

of longer kmers (k=124 to k=240). The fastq sequencing reads were converted to merged fasta format using fq2fa tool supplied by the package.

```
$ idba_ud -r reads_1_2_merged.fa –pre_correction --num_threads\
 16 --mink 20 --maxk 240 -o IDBA_assembly.fa
```

2.6.2.2. SPAdes

The SPAdes[259] assembler was also used for short-read assembly in this project. The latest version of SPAdes released at the time the data was available for assembly was used for, with the latest version employed in the study being v3.13. This running of this version was invoked with the merged paired-reads function.

```
$ spades.py -k 21,41,71,101,127 -o SPAdes_assembly.fa\
 --12 unmerged.fq –merged merged.fq -t 16
```

2.6.3. Hybrid assembly

Hybrid assembly in this study was performed with the MaSuRCA[260] genome assembler (v3.2.8). The hybrid assembly of `MH_PAT` was calculated from a combination of 2x150 PE, 2x250PE Illumina TruSeq and PacBio SMART-Bell sequencing reads.

2.6.4. Read mapping

Mapping short reads to contigs to estimate coverage and sequencing depth was carried out using BBMap in fast mode. SAM/BAM alignment map files were manipulated with SAMtools (v1.6).[261]

2.6.5. Taxonomy

High-level taxonomic classification (phylum rank) of metagenomic contigs was carried out following the method described in Albertsen (2103)[262]. Briefly, ORFs were detected in contigs using Prodigal[263] and a set of single copy marker proteins HMMs were used to detect genes for taxonomic analysis. The identified marker genes were blastp searched against the refseq protein database. The top five blast hits and MEGANs LCA algorithm was used to get the taxonomic assignment of each protein sequence. The final taxonomic assignment was derived by majority vote of the total markers where multiple marker genes were present in a contig.

To ascertain the taxonomy of completed metagenomic bins, 16S rRNA genes were extracted and aligned to the SILAV ARB database (SSU Ref NR 99 release 132) for classification with SINA (v1.3.1). Full MAGs were also classified using the GTDB-Tk pipeline (v 0.3.2).

2.7 Chemotyping

Frozen powdered sponge samples (~1.5 g dry weight) were then extracted with 80% MeOH-H_2O (20 ml) and then MeOH (20 ml) for 10 min each. The first extract, followed by the second extract, was passed through a polystyrene divinylbenzene (PSDVB) column (2 ml). The combined eluents were diluted and reapplied to the column a total of three times, using H_2O for dilution (2 ml twice, then 80 ml). The column was eluted with H_2O (10 ml) and then 55% Me_2CO-NH_4 acetate (0.2 M, adjusted to pH 4 with acetic acid). The latter fraction was neutralized with NH_4OAc (0.2 M, 20 ml) and then loaded onto another PSDVB column (0.5 ml) to desalt and remove water. The fraction was passed through the column twice, followed by elution with H_2O (10ml) and then acetone (6ml). The resulting acetone fractions were dried and analysed by ^1H NMR and LC-MS. ^1H NMR spectra were acquired using a 600-MHz Varian Direct Drive spectrometer. Spectra were recorded in $CDCl_3$ and referenced to the residual solvent peak (δ_H 7.26). LC-MS data were acquired with an Agilent 6530 accurate-mass quadrupole time of flight (Q-TOF) LC-MS mass spectrometer equipped with a 1260 Infinity high-pressure liquid chromatography (HPLC) system using positive-mode electrospray ionization. The instrument parameters were set as follows: gas temperature, 275 °C; drying gas, 9 litres/min; nebulizer, 30 lb/in2; sheath gas temperature, 300 °C; sheath gas flow, 10 litres/min; capillary voltage, 4,000 V; nozzle voltage, 500 V. Masses were recorded between 100 and 2,000 m/z at a rate of 3 spectra per second. Chromatographic separation was achieved with a reversed-phase C18 column (Kinetex; 50 mm by 2.1 mm by 2.6 μm), set to 35 °C. Sample elution was achieved using eluent A (H_2O 0.1% formic acid) and eluent B (acetonitrile 0.1% formic acid) with a gradient from 5% B to 100% B over 11 min at a flow rate of 0.4 ml/min. Samples were adjusted to a concentration of 0.1 mg/ml in methanol and an injection volume of 10 μL was used.

3 Marine sponge metagenomic clone library construction and validation.

3.1 Introduction

The metagenomes derived from environmental samples are inherently complex due to the numerous community members they are comprised of, spanning the domains Bacteria, Archaea and Eukaryota.[264] Adding to this complexity is the variability in abundance of any given species in a sample and strain heterogeneity within species.[265] Furthermore, the majority of the population cannot be cultivated, making the analysis of a target species or functional genetic unit in isolation difficult. This complexity of environmental samples usually makes direct shotgun-sequencing of environmental metagenomes intractable as the required sequencing depth is cost prohibitive and the accurate assembly of useful genomes is computationally challenging due to the homologous regions sheared between community members. A proven metagenomic natural product discovery technique that is used to harness this complexity is the construction of large-fragment (>35 kbp) cosmid libraries from extracted environmental DNA (eDNA) (See Figure 3.1).[253] This process can redundantly capture the entire genetic diversity of the microbial population as discrete overlapping molecules that can then be analysed in isolation. With this approach, large portions of BGCs from their native produces may be captured in a single high molecular weight (HMW) DNA clone which can then be cultivated in a lab amenable strain, usually *Escherichia coli*.[253] This is especially advantageous for rare sample types or for accessing BGCs from low abundance organisms as the library can be infinitely expanded, easily stored and analysed. The sequences contained in these isolated clones can be reviewed individually without interference form the background of the entire metagenome sequence diversity. It also facilitates the isolation of the actual DNA molecules which can be employed in biotechnological applications, such as the reconstitution of entire BGCs for homologous expression. For example, the isolation and reconstruction of the entire BGCs for pateamine and peloruside would be the first crucial step in engineering large scale production of these difficult to synthesize molecules.[266,267]

This chapter describes the generation of a large metagenomic clone library from eDNA extracted from a *M. hentscheli* specimen as well as the initial findings from sequencing and analysis of the library that led to employing a supplementary strategy to facilitate the search for the target BGCs.

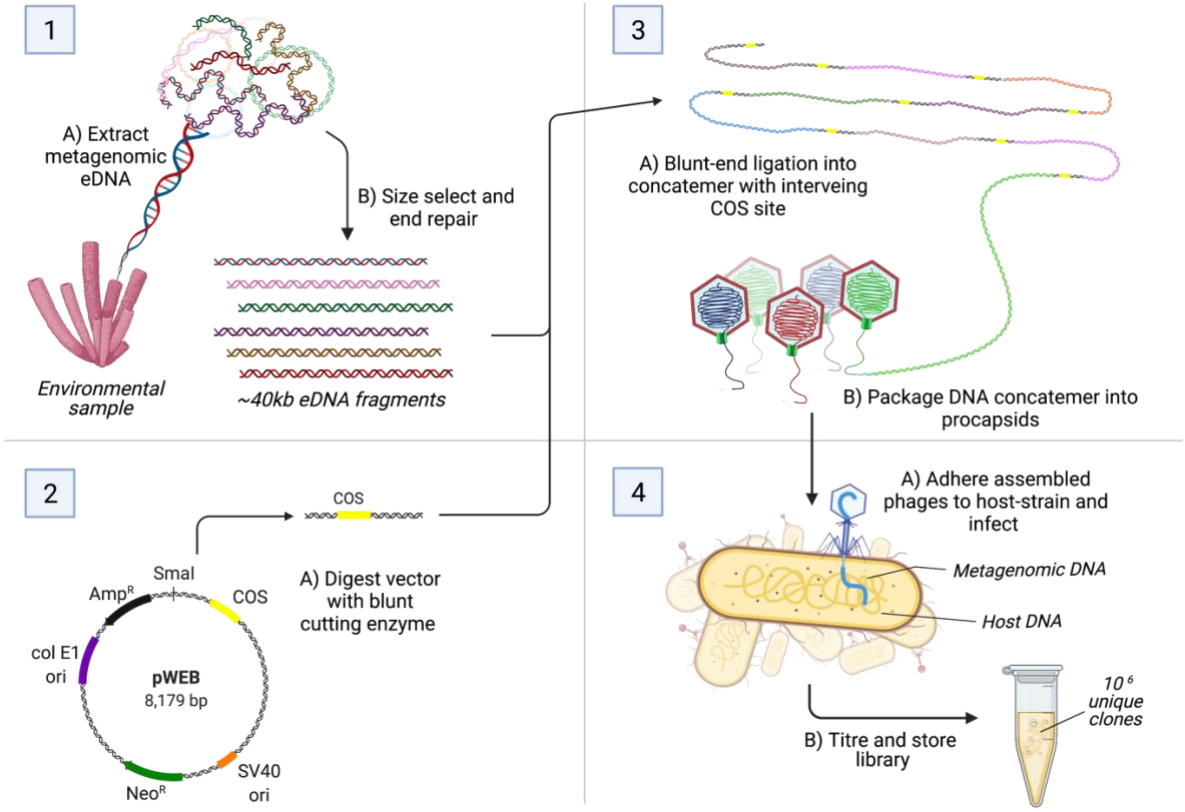


Figure 3.1 – Construction of metagenomic cosmid library from eDNA.
This figure shows the main steps and reactions used to construct a metagenomic cosmid library. The steps are divided into four panels and reactions within each step are labelled. **Panel – 1**; **A**) Metagenomic DNA is comprehensively extracted from the environmental sample, in this case a marine sponge sample. **B**) The crude extract is then carefully purified, size selected and enzymatically end-repaired to make it suitable for blunt- end ligation. **Panel – 2**; **A**) An appropriate cosmid vector encoding a COS site (yellow bar), a replication origin (purple bar) and antibiotic resistant markers (black/green bars) is digested with a blunt-cutting restriction enzyme to prepare it for blunt-end ligation. For clarity, only the COS site is shown in the digested vector icon. **Panel – 3**; **A**) The prepared vector and end-repaired eDNA fragments are blunt-end ligated to form concatemers of eDNA and vector backbone. **B**) The concatemers are packaged into purified procapsids and self-assemble into mature phage particles. **Panel – 4**; **A**) The mature phage particles are then used to infect a susceptible host strain which will maintain the metagenomic DNA fragments. **B**) The number of unique clones are enumerated before the library is expanded and stored cryogenically for future screening.

3.1.1. Metagenome library functional screening

Functional screening relies on clone captured BGC fragments to retain expression of some phenotypic trait in the library carrier strain that can be used to screen for clones of interest.[268] For example, screening for clones that have phosphopantetheinyl transferase (PPTase) activity can be indicative of the presence of an NRPS or PKS BGC in the clone.[269] A PPTase enzyme is required by these BGCs to perform an essential post-translational modification of

the thiolation domains of the encoded megasynth(et)ase.[270] The required PPTase gene is usually found within the bounds of the BGC.[271] By coupling PPTase activity to host cell viability or the production of a coloured product[272] large metagenomic libraries can be effectively enriched for clones with PPTase function very efficiently. The recovered 'functional' clones can then be analysed in isolation for biosynthetic activity or the captured DNA fragment can be sequenced and queried for the expected associated biosynthetic genes. Indeed, this approach has been used successfully to isolate BGCs from very complex soil metagenome libraries, as demonstrated in 2013 by Charlop-Powers *et. al*.[270] In addition to the high-throughput that functional screening enables, another important advantage is this method does not require any prior knowledge of target sequences, which allows the discovery of novel biosynthetic gene families.[268] However, this method does not have the resolution to accurately differentiate between the PPTases associated with primary or secondary metabolism, or between different classes of BGC, which could result in a high rate of undesirable or redundant hits.

Other functional screens identify pigmentation or antibiosis from individual clones expressing small coloured or cytotoxic molecules. These approaches are highly host dependant and require active molecules to be expressed from a single cosmid in the screening host which is limited by condon bias, substrate availability, gene regulation and clone size.[273,274]

These untargeted functional screening methods preclude the isolation of BGCs where the screening trait is non-functional in the library host strain resulting in the possible absence of desired hits. For this reason, a functional screening approach was not considered a suitable means for the targeted recovery of the BGCs expected to be found in the metagenome of *M. hentscheli.*

3.1.2. Metagenome library sequence homology screening

By virtue of the sequence homology between core components of BGC families a semi-targeted approach can be employed to identify and isolate BGC types of interest from metagenomic clone libraries. The phylogenetic relationship between evolutionarily distinct sequences of the core components can be used to further target this isolation strategy.[275] As outlined in section 1.2, the biosynthetic modules of each BGC family are built around a minimal and conserved core of genes that encode for the enzymes that join and modify precursors to yield bioactive products. By selecting for clones that harbour sequence motifs known to be found in genes from the minimal biosynthetic core, clones capturing BGCs of

the target family can be recovered. In order to select for such clones, sets of degenerate PCR primers (primer set that can amplify a board range of related sequence targets)[276] are used to identify clone pools of ever decreasing complexity until a single PCR positive clone is recovered. This technique does require the prior knowledge of sequence information pertaining to the BGC family of interest and assumes some degree of relatedness of the target BGCs to this family, as such the discovery of divergent novel BGCs is precluded under this method.

As the sequences of many secondary metabolite biosynthetic pathways have been characterised, target-sequences to base the design of degenerate primers targeting a BGC family of interest can deduced and validated readily.[277] Although each round of this selection technique may not capture as broad of a range of novel BGC as a functional screening approach might achieve because of its targeted nature, it does generalise well to capture novel BGCs within the target family due to the capacity of degenerate primer sets to cover a wide range of homology sequence space.[278,279] The successful isolation of novel PKS BGCs from complex terrestrial and marine metagenome clone libraries by focusing on the PKS KS-domain as a target has been demonstrated by Owen (2015)[61] and Fisch (2009)[247] respectively. The sequences homology screening approach has also been applied to adenylation (A) domains from NRPS BGCs.[280]

3.1.2.1. Targeted homology screening

A structural-evolutionary relationship has been demonstrated by phylogenetic analysis of *trans*-AT KS-domains of know biosynthetic output which can provide information about the chemical structure of the incoming intermediate substrate that a KS-domain condensates by examination of the KS-domain sequence.[125,275,281] In turn, this allows the targeting of specific PKS genes based on a known polyketide structure. In the case of Fisch (2009)[247], KS domain that were specific to *trans*-AT PKS BGCs were identified in public sequence databases, this information was then used to design a degenerate primer set targeting a specific motif present in 84% these evolutionarily distinct KS-domains but absent from 'background' KS like sequences found in primary fatty acid metabolism and *cis*-AT PKS. A nested PCR strategy was then used whereby KS-domains of diverse PKS types are first amplified in a degenerate PCR step to reduce the overall complexity of the template. This was followed by a second, nested, PCR using the identified *trans*-AT targeting primer in combination with a clade-specific reverse primer targeting PKS KSs with substrates of interest present in target

molecules. In the study Fisch (2009)[247], this strategy was in fact applied to *M. hentscheli* to target a clade of KS-domains that elongate αß-saturated intermediates, with the aim to identify a KS-domain predicted to be present in the mycalamide BGCs. Interestingly, such a αß-saturated chemical moiety is not present in the mycalamide structure, but inclusion of a KS-domain from this clade in the mycalamide BGC was predicted due to the structural relationship between mycalamides, onnamides and pederin. The onnamides do have the αß-saturated moiety and the pederin BGC has a PKS module (PedH) with a KS-domain in this clade. However, the saturated moiety is lost from the mature pederin molecule when the onnamide-like terminus is oxidatively cleaved off. This terminal cleavage event is also expected to occur during the biosynthetic maturation of the mycalamides, based on this expectation, a KS domain of the respective clade was also predicted to be present in the mycalamide BGC. A KS-domain of the expected clade was indeed discovered in a mycalamide positive chemotype of *M. hentscheli* in the study, however the full gene cluster was not recovered in this case.

Based on the successful indication of target BGCs in these studies and others[279], screening of a metagenomic clone library derived from a biosynthetically rich chemotype of *M. hentscheli* by screening for *trans*-AT PKS gene sequence homology was considered to be a viable approach to recover BGCs for at least mycalamide, and potentially pateamine and peloruside.

3.2 Chemotypes of *M. hentscheli*

In total, this study had access to six archived sponge samples that had previous been collected for chemical analysis studies over various expeditions and kindly donated by Assoc. Prof. Peter Northcote. These samples were identified as *M. hentscheli* at the time of collection. This species of sponge is known to present with differing chemotype profiles that can be positive or negative for any composition of the bioactive metabolites that have been isolated from the species. This variation can be seasonal and site specific.[84,244] One of the donated samples was a large (>1 kg) cluster of intact leuconoid branches that had tested strongly positive for presence of all three target metabolites prior to donation and had been constantly stored at -80º C from time of donation. Due to the sample's abundant biomass, highly preserved appearance and reported potency, it was chosen to be the main focus of this investigation. As this was the only sample with detectible levels of pateamine at the time of the study, it was denoted as `MH_PAT`. The remaining samples were much smaller (<25 g).

These additional samples were included for comparative analysis and completeness of metabolite coverage. These samples were denoted as S1, S2, S3 and S5. The sample pertaining to S4 failed to yield suitable DNA for this study so was dropped from the study. An additional sample, denoted as MH_PEL, introduced later in the study was not available for chemotyping in this study due to limited sample size.

3.2.1. Sponge chemotypes

A common observation in studies of marine sponge secondary metabolites is the variation in chemotypes between sponges of the same species. For example, specimens of the sponge *Theonella swinhoei* growing side-by-side can display distinct chemotypes and also vary in the colour of their interior.[43] *M. hentscheli* also displays distinct chemotypes profiles for the three main metabolites, with variation across geographical regions and seasons, even for genetically identical clones of the same individual.[243] Although the cause of this intra-species variation in *M. hentscheli* remains cryptic, studies of microbial diversity and composition of six distinct sponge chemotypes, using a coarse-grained 16S rRNA gene fragment analysis, showed a spatial and temporal clustering of bacterial composition and by extension grouped some of the chemotypes that shared a similar spatial and temporal variation. However, where different chemotypes co-occurred at the same location at the same time the same grouping patterns did not resolve.[245] This may suggest that factors beyond the mere presence of a bacterial species, or group of species, contribute to the observed secondary metabolite profile in *M. hentscheli.*

The *M. hentscheli* samples used in this current study covered three distinct chemotypes which were determined during the study by NMR analysis of ethanoic extracts and comparing these results to previously published data[242,282,283] (See Figure 3.2). The results of this analysis indicated that the sample MH_PAT was chemotype negative for peloruside A, this was an unanticipated finding as it was contrary to the expected chemotype suggested to us by the samples previous custodians who had kindly donated the sample for our research. However, the sample was strongly positive for pateamine A, and the additional samples (S1-S3) were all positive for peloruside A. All samples included in this study were positive for mycalamide A. Analysis of the 'S' samples and MH_PEL are discussed in later chapters, all available chemotypes shown here for comparison.

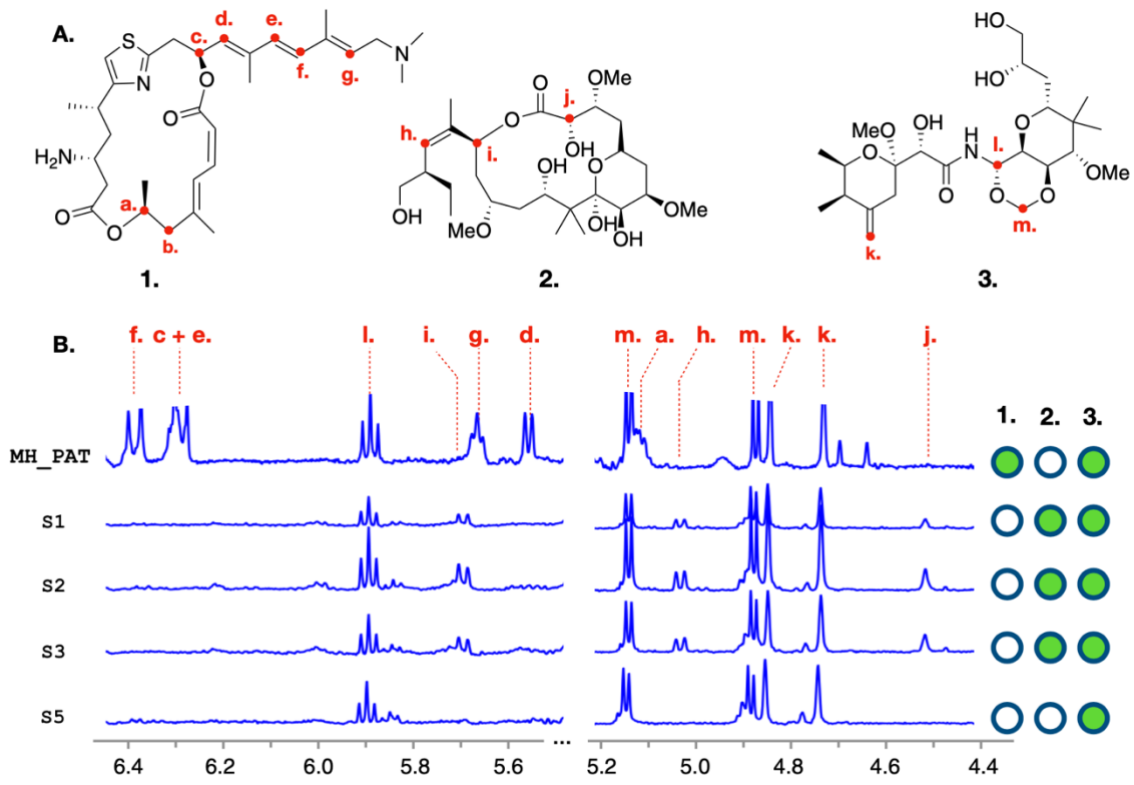


Figure 3.2 – Structures of major metabolites and chemotypes for *M. hentscheli* samples in this study.

(A) Structures for each of the cytotoxic polyketides previously isolated from *M. hentscheli* are shown. These are pateamine A (compound **1**), peloruside A (compound **2**), and mycalamide A (compound **3**). Red labels indicate positions of protons whose shifts were diagnostic of compound presence during chemotyping experiments. (B) Selected regions of ^{1}H NMR spectra for each of the five specimens examined in this study. Diagnostic peaks for the presence of each compound are labelled with dashed red lines, and the letters above these match the positions of protons in panel A. The right-hand panel indicates the compounds that were determined to be present in each specimen.

3.3 Extraction of HMW-DNA from *M. hentscheli*

Critical to the success of building a metagenomic clone library is the extraction and purification of HMW-DNA from the sample under investigation. The DNA must be extracted carefully to avoid mechanical shearing or chemical degradation as DNA fragments below a molecular size threshold of ~35kb will not result in the generation of viable phage particles. The extraction method should also aim to be as comprehensive as possible to ensure all species of microbes are adequately represented in the extract. Routine extraction methods based on silica binding or ionic exchange perform poorly in extracting eDNA suitable for building clone libraries, resulting in either DNA that is overly fragmented or contaminated with impurities that inhibit the library building reactions. To obtain suitable *M. hentscheli* eDNA for clone library construction, an extraction and purification method based on that previously published by Gurgui and Piel (2010)[253] was used with some

modification (see section 2.3.2 for details). Briefly, a longitudinal section including both surface and inner parts of the frozen *M. hentscheli* specimen `MH_Pat` was ground under liquid nitrogen into a fine powder. This was conducted in a chemical fume hood with appropriate PPE, as exposure to *M. hentscheli* tissue can cause severe skin reactions. All exposed surfaces were cleaned with a dilute bleach solution to decontaminate. Dissociation of microbial cells by homogenisation followed by differential centrifugation was used to separate the sponge tissue debris from the microbial cells. Lysis buffer based on high concentrations of urea and sarkosyl was used to liberate the DNA of the 'microbial fraction' into solution. This crude lysate was then extracted several times with freshly distilled tris-buffered phenol until the aqueous phase was clear. The DNA was then precipitated from the separated aqueous phase by the addition of 2-propanol. This crude DNA was gently reconstituted in tris-buffer (~1 ml) and loaded into a single well spanning the entire width of an 0.8% agarose gel large enough to accommodate the entire sample (See - Figure 3.3). Current was applied and electrophoresis was carried out for 12 hours with the voltage clamped at 20V. As expected the DNA migrated through the gel as a function of molecular weight, some contaminants were also separated during electrophoresis, with some visibly migrating opposite to the DNA, so the electrophoresis buffer was replaced several times over the 12-hour course to remove any contaminants leached from gel and main buffer solution integrity. At the conclusion of the electrophoresis period, two narrow sections from either side of the gel were removed for staining and visualisation of the DNA. These sections were marked and used to guide the excision of the remaining main HMW-DNA band from the gel without exposing it to UV light or staining reagents. The above outlined electrophoresis and gel-excision procedure acts as a size-selection and purification process sufficient for obtaining library quality DNA from sponge samples. The DNA contained in the excised gel slice was electroeluted from the agarose into a TE solution inside an acrylic "H-chamber" fitted with semipermeable membranes to trap the DNA while allowing current to pass through. This electroeluted HMW-DNA in solution was then concentrated using a molecular weight cut-off membrane centrifuge filter where the HMW-DNA was trapped by the filter in the column and the bulk buffer solution passed through. The filter was flushed through with several rounds of fresh tris-buffer to remove any remaining contaminants and to replace the TE buffer to something appropriate for the downstream molecular cloning processes. Further concentration was achieved by precipitation with 2-propanol. The size of the final HMW-DNA was compared against a HindIII digested λ-DNA standard, a band above the 23.1 kb marker was considered for library construction (See - Figure 3.4).

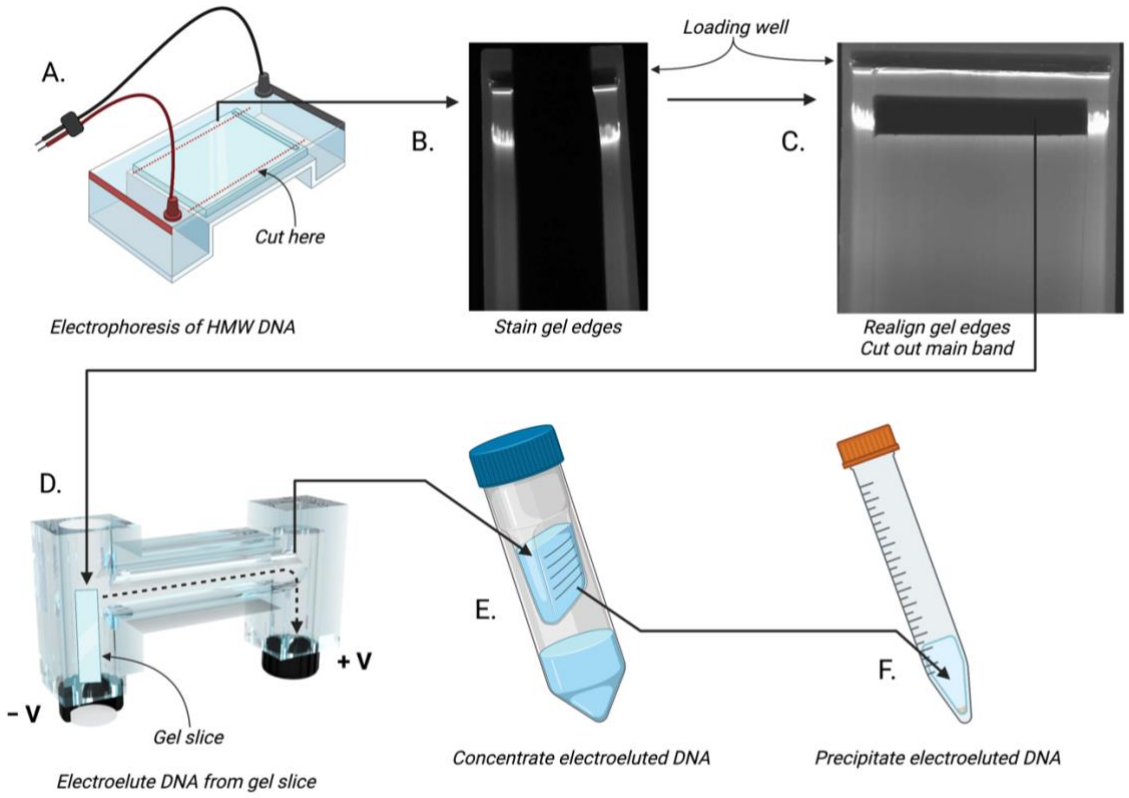


Figure 3.3 – HMW DNA size selection process. Vital to successful cosmid library construction is size selection and purification of HWM DNA. This diagram show the process of size selection used in this study. A) The crude DNA extract is loaded into an agarose gel prepared with a large single lane. B) After electrophoresis, the sides of the gel are cut off and stained to visualise the location of the DNA. The position of the main HMW-DNA band is marked on the slices (marks not visible). C) The stained gel-sides are realigned to the main body of the gel and used to guide the excision of a gel-slice containing the main HMW-DNA band without exposure to short wavelength UV. D) The gel slice is loaded into electroelution "H-chamber" to draw the DNA out of the gel and into solution. The DNA follows the dashed path of the electric potential through the channel and collects in the positive side of the chamber. E) The DNA in solution is aspirated from the "H-chamber" and loaded into a molecular weight cut-off filter to concentrate. F) The enriched HMW-DNA solution can be further concentrated by 2-propanol precipitation if required.

3.3.1. HMW-DNA extraction results

Several other less labour and time intensive HMW-DNA preparation methods were trialled during the optimisation of the one outlined above, including the omission of gel-size selection or the substitution of phenol extraction with other purification methods. None performed well for the purpose of obtaining suitable DNA for cosmid library construction from `MH_Pat` tissue. The final extraction and size selection procedures were performed twice with `MH_Pat` tissue samples prior to building a clone library. In the first instance 1.5 g of tissue was used as input and resulted in the recovery of 5ug of HMW-DNA at a concentration of 25 ng/ul. In the second round, 2 g of tissue was used as input and resulted in 10 ug of HMW-DNA at a concentration of 50 ng/ul. The second sample also registered higher on a size comparison gel run with both final products (See - Figure 3.4). These yield and size

differences were attributed not only to the higher input amounts that were used, but also to extra diligence during the critical separation of the aqueous phase from the phenol extraction, longer electroelution times and omission of a final 2-propanol precipitation step.

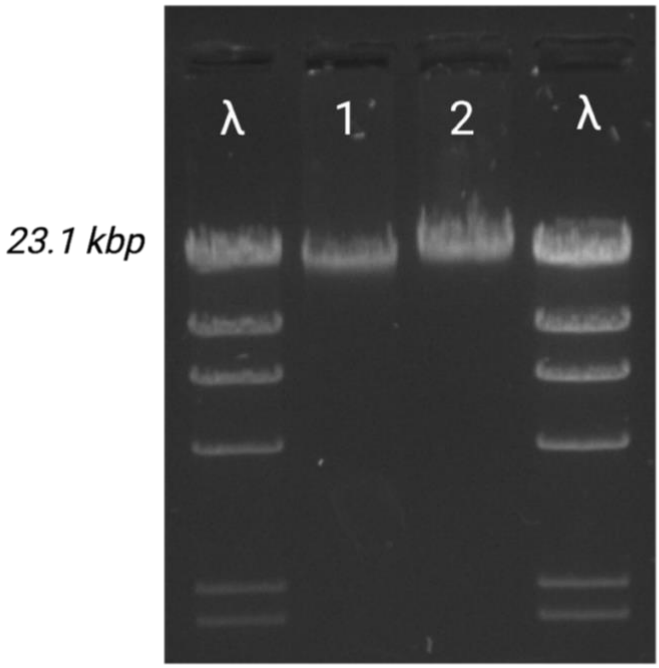


Figure 3.4 – Size selected HMW-DNA purified from *M. hentscheli.* An agarose gel of two (1 and 2) HMW-DNA samples prepared using the optimised size selection and purification process. λ-HindIII DNA markers are used either side to assess the molecular weight of the DNA samples. Sample '2' registers above the 23.1 kbp marker.

3.4 Building the *M. hentscheli* metagenomic large-insert library

The value of a metagenomic cosmid library can be gauged by the breadth of the microbial community's DNA sequence that is captured. Similar to the 'coupon collector problem'[284], the number of clones required for complete coverage grows at a rate approaching $n(log(n))$ in relation to the sequence diversity of the community. This results in an ever-increasing library size as coverage requirements increase. The ideal library would redundantly cover the entire metagenome to capture the entire diversity in overlapping fragments. In practical terms, studies report building library sizes of sponge metagenomes ranging from (2.5×10^4) [285] to (4.0×10^5) [253] individual unique clones. The construction of a comparable size library was desired in this study. The quality of the HMW-DNA to be used strongly affects efficiency of library construction. As such, a final assessment of the HMW-DNA

extracted and purified from *M. hentscheli* (MH_Pat) was done before committing resources to a full-scale metagenome library construction campaign. A test scale library, using all the prepared components and the HMW-DNA sample '2' (See - Figure 3.4), was constructed to ensure building the desired library size was tenable. The two limiting resources for building this library were the amount of HMW-DNA available, and the cost and availability of phage packaging extracts.

3.4.1. Trial library building results.

An initial trial of the 'packaging reaction', the process of packaging the end-repaired HMW-DNA into phage-heads and infecting the host strain, failed to produce any clones with an end-repaired *M. hentscheli* HMW-DNA sample measuring 12.5 ng/µL. This prompted going back to the original HMW-DNA sample to repeat the end-repair and repurification reaction. In this second attempt, ~5 µg of sample '2' was end-repaired and repurified, resulting in ~3 µg of packaging ready HMW-DNA at a measured concentrating of ~153 ng/µL.

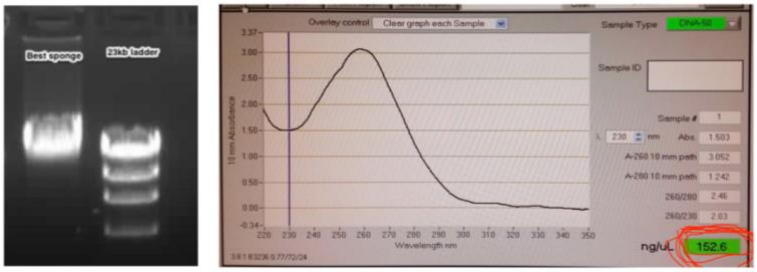


Figure 3.5 – Best sponge end-repaired sample. A successful end-repair and repurification effort of the HMW-DNA sample '2' extracted from MH_Pat was recorded on an agarose gel (left panel) and nanodrop spectrophotometer (right panel). This HMW DNA sample was colloquially referred to as 'best sponge'.

A one-fifth scale 'packaging reaction' was set up consuming 100 ng of the HMW-DNA of the newly end-repaired and purified sample '2'. The number of clones produced by this completed trial reaction was enumerated by counting colony forming units (CFUs) arising from plating a serial dilution of the infected culture on selective media. A one one-hundredth volume of the reaction was plated out neat, followed by serial dilution to 1×10^{-4}.

Volume	1/100	1/1000	1/10000
CFUs	52	6	0

Table 3-1 – Titre results from *M. hentscheli* (MH_Pat) trial library. Top row shows the volume of the total trial reaction that was plated. The bottom row shows the number of CFUs that were counted as a result of plating out the respective volume.

The trial reaction results indicated that a full-scale packaging reaction should yield approximately 25,000 – 30,000 unique clones. This was considered to fall at the low-end of the acceptable range needed to attain a library of target size, but a tractable result given the total available HWM-DNA. It was expected that 10 full-scale packaging reactions would suffice to complete a target library size of >250,000 unique clones. An observation from these tests and previous/subsequent metagenome library building work was, compared to other sources of HMW-DNA (e.g., soil eDNA or positive control phage DNA), libraries built from sponge derived HMW-DNA tend to be an order of magnitude smaller (based on personal communications and data not shown here). The underlying cause of this was never made apparent, but I suspected it may have been caused by the DNA fragments being too large to package efficiently.

3.4.2. Commercial and 'lab-made' phage packaging extracts

The initial *in vitro* phage packaging tests mentioned above employed commercially available certified "lambda packaging extracts" and ligation ready cosmid vector DNA (Epicentre® MaxPlax™ Lambda Packaging Extracts). Due to the high volume of metagenomic library projects that were being anticipated or were underway in the laboratory at this time, we determined it would be beneficial and practical to produce the required reagents for phage packaging in-house. Access to a more economic supply of packaging extracts allowed the construction of libraries from modestly performing HMW-DNA samples. There were also concerns at this time about supply issues from the providing company. Preparing these extracts was a stringent multiday process and required the use of specifically sourced high quality reagents. A robust protocol optimised for our laboratory facilities was painstakingly developed for this and other studies by Dr Mark Calcott. When the "lab-made'" packaging extracts were tested with *M. hentscheli* HWM-DNA an equivalent titre was achieved compared to the commercial reagents. Subject to availability, one-third of the *M. hentscheli* large insert metagenomic library discussed here was built with commercial packaging extract and the remaining with "lab-made" packaging extract. At least two additional sponge libraries, not discussed in this thesis, were able to be successfully built using the "lab-made" reagents due to the reduced cost this approach afforded.

3.4.3. pWEB vector preparation

The COS-site containing vector chosen to be the backbone of the library clones was cosmid vector pWEB (Epicentre Biotechnologies). This high copy vector can be maintained in *E. coli* using the selective markers of kanamycin and ampicillin resistance. It also contains a multiple cloning site (MCS) flanked either side by the M13 forward and T7 priming sites to facilitate Sanger sequencing of the inserts. The vector stock used for this library was prepared by miniprep from *E. coli* (EC100). The extracted vector was then digested with *smaI* to produce a blunt-ended cut in the MCS. The digested vector was dephosphorylated using a thermosensitive shrimp alkaline phosphatase. The enzymatically treated vector is then size selected and repurified in the same process used to prepare the HWM-DNA. These efforts are to reduce the background of insert-less vector contaminating the final library. The prepared vector was quality checked for the expected fragment size and migration pattern and on an agarose gel, and concentration was verified with a Qubit™ DNA analyser.

3.4.4. HMW-DNA cloning and packaging

With the verified reagents and HMW-DNA prepared, the full-scale metagenomic library construction commenced. The entire library was built over three consecutive rounds of full-scale packaging reactions. For each round, multiple reactions were set up. For each reaction, 500 ng of ligation ready *M. hentscheli* (`MH_Pat`) HMW-DNA and 1 µg of ligation ready pWEB vector were ligated, and then packaged into 60 µL of phage packaging extract. The total volume of mature phage particles from each round was then adhered to 8 ml of a prepared *E. coli* host cell (EC100) culture. Prior to any clonal expansion the adhered culture was then evenly arrayed over 96 tubes, charged with 5 ml of selective media, so each tube contained a unique set of clones. A sample of 50 µL from five randomly selected tubes was plated to enumerate the number of unique clones expected from the entire packaging reaction (See 3.4.4.1). The arrayed library was then expanded overnight in a shaking incubator (200 rpm at 37° C), the next day a 500 µL sample of each tube is taken for cryogenic storage and the remaining was miniprepped to extract the clone's cosmid DNA. The first round was build using the commercial packaging extracts and the last two used the 'lab-made' extracts. Surprisingly, the 'lab-made' extracts resulted in significantly higher clone tires than commercial extracts, indicating that preparation of fresh reagents in-house was preferable in terms of both cost and cloning efficiency.

3.4.4.1. Enumeration

Each round of phage-packaging was enumerated to appraise the packaging efficiency and record the total library size. At each round, CFUs arising from quadruplicate samples representing 1/10,000th of the total reaction each were counted by plating serial dilutions on selective media. The sub-libraries built at each round were referred to as "MH-Sponge-Lib-A", "MH-Sponge-Lib-B" and "MH-Sponge-Lib-C". Each sub-library was arrayed over 96 pools.

Packaging round	# Reactions	Clones per pool	Clones per sub-library
"Lib-A"	2	800	78,600
"Lib-B"	2	1300	125,000
"Lib-C"	4	2000	192,000
Total	**8**	**-**	**395,600**

Table 3-2 – Metagenomic Clone Library Phage Packaging Results. Sub-library sizes were enumerated by counting CFUs. The number of clones expected per pool was calculated by averaging the CFU counts across the four samples taken for each round.

These enumeration results show the full scale-packaging reactions were slightly more efficient than what was expected from the packaging trial results (See 3.4.1). After three rounds of packaging the total library was approaching 400,000 unique. This had nearly exhausted the current supply of extracted and prepared HMW-DNA and was considered to be an acceptable library size to proceed with analysing the library quality, and sequence homology screening for target BGCs.

3.4.5. Insert validation

Before undertaking intensive screening efforts, the newly constructed *M. hentscheli* metagenome cosmid library three quality metrics were assessed to ensure the library was of high quality. To facilitate the assessment, 48 random clones were isolated from the sub-library "MH-Sponge-Lib-A" and the harboured cosmid was extracted and purified from a broth culture of each of these clones.

The metrics assessed were: **1.** the "insert-ratio" (3.4.5.1) – the proportion of the 48 isolated cosmids that either contained an expected large-insert or were empty and had failed to capture an insert; **2.** the "sequence diversity" (3.4.5.2) – assessed by a restriction digest fragment analysis of each of the isolated 48 cosmids; and **3.** the "taxonomic profile" (3.4.5.3) – assessed by end-sequencing the metagenomic insert of each of the 48 cosmids. Overall,

these metrics allowed us to determine whether the library was low in empty vector (metric 1), of high diversity (metric 2) and capturing diverse bacterial phyla, with the absence of background sponge DNA (metric 3). Obtaining these metrics was important for guiding screening strategies and the optimisation of future library construction efforts if required.

3.4.5.1. *M. hentscheli* metagenome cosmid library insert proportion

Analysing the insert-ratio of the library was readily archived by measuring the approximate size of the cosmids extracted from the 48 random test-clones. It was expected that insert containing cosmids would be > 30 kb in size and would be readily discernible from empty vector (~ 8 kb) using agarose gel electrophoresis. The proportion of large cosmids (>30 kb) to cosmid empty vector back bone (~ 8 kb) gives the insert ratio. For this group of test cosmids, all 48 had a MW >30 kb, indicating that the library had a high insert ratio approaching complete insert capture. This implied the library had only a low-level background of empty cosmids which is beneficial for screening and target clone isolation.

3.4.5.2. M. hentscheli metagenome cosmid library sequence diversity

The sequence diversity of the same 48 test-clones used above was measured to assess if the library was largely comprised of unique clones as desired. It was a risk that a single clone could dominate during the clonal expansion of the library, or for a DNA fragment from another purified cosmid to contaminate the ligation cloning reaction, and adulterate the library, leading to an imbalance in the expected sequence diversity within the library. To assess this, the 48 purified test cosmids were subject to a restriction digest (EcoRI) and the resulting fragments were again analysed by agarose gel electrophoresis (Figure 3.6). A cursory examination of the fragment analysis showed that although the library was diverse, there were apparent insert-sequence duplications within the 48 test-clones analysed here. Somewhat strikingly, these duplications were found grouped adjacent within the random order that they were analysed, which might indicate an artefact of the clone selection process such as carry-over contamination between adjacent wells. In a manual all-vs-all comparison of the fragment patterns, only ~30 unique patterns were could be visually identified out to the 48 analysed. The plausible explanations considered to explain this duplication were, a sub-set of clones dominated during clonal expansion of the library, or cross contamination during the sampling of the test-clones. The low resolution of this method meant that it was difficult to extrapolate from this result to draw inference about the overall library quality and

possible source of duplication, however it was deemed that the library was of sufficient quality to continue screening.

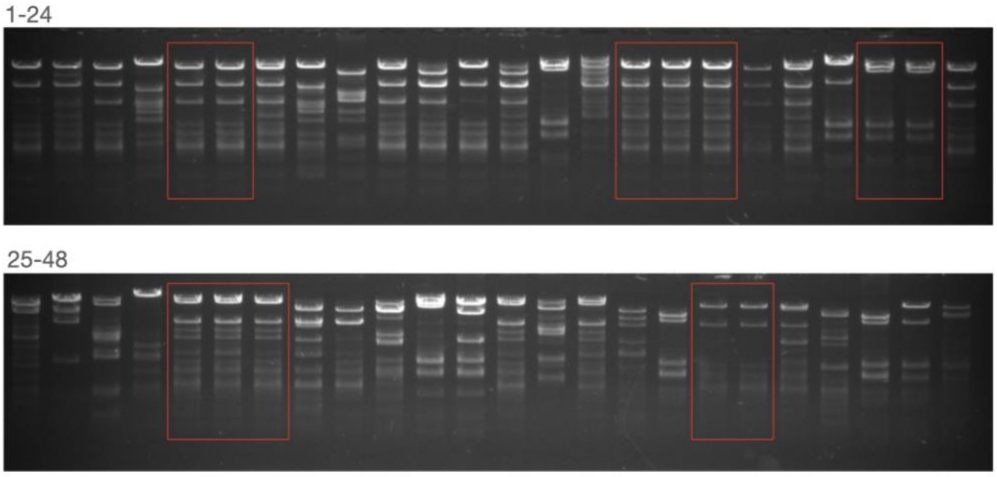


Figure 3.6 – Restriction Digest of 48 test cosmids from the M. hentscheli metagenome library.
Upper panel shows the EcoRI digest fragment analysis results for test cosmids 1-24 and the lower panel shows results for the remaining test cosmids 25-48. The red boxes highlight some apparent groups of repeating identical fragmentation patterns.

3.4.5.3. *M. hentscheli* metagenomic cosmid library taxonomic profile

To gauge the taxonomic makeup of the test clones, as a proxy for the taxonomic distribution of the entire library, the cosmids were prepared for Sanger based "end-sequencing". The successful "end-sequencing" of an isolated cosmid will return a short sequence fragment from both the 5` and 3` extremities of the cosmid's insert-sequence. This is done by taking advantage of the known universal primer sites on the vector back-bone that flank the insert cloning site, in the case of the pWEB vector used to build this library the primers sites were the M13 and T7-promotor priming sequences. The results of this sequencing were analysed to give a rough taxonomic classification of the insert sequence and to determine if the inserts between cosmids are identical or unique. The results that might be expected from the taxonomic analysis of these sequence fragments would be, at least, the distinction between either eukaryotic or prokaryotic origin of the insert and, in the case of prokaryotic sequence, the distance from any known organisms. This allows the proportions of host (eukaryotic)/microbiota (prokaryotic-unknown)/contaminant (prokaryotic-known) to be estimated.

Two sequencing reaction were prepared for each of the 48 test cosmids, one reaction for each of the M13 and T7-promotor priming sequences per cosmid, by premixing the purified

cosmid DNA with the respective primers and sequencing buffer (see Methods) for a total of 96 reactions. After several partially failed sequencing runs were returned by the external provider and manual sequence trimming and quality control (QC) the cumulative data received, a total of 51of the 96 expected sequencing reactions yielded useable sequence data. After manual trimming and quality assessment, the 51 usable reads had a quality score >Q40 over >70% of the read. Of the 48 cosmids submitted for end sequencing only 19 were successfully sequenced at both the M13 and T7 ends. The end-sequences for three of the submitted cosmids mapped directly back to the pWEB vector backbone, an unexpected result as all the cosmids appeared to contain a large insert. A BLASTX search query of the end sequences was used to estimate the taxonomy of each quality passing read. A single pair of end sequence reads originating from the same cosmid aligned to *Amphimedon queenslandica*, the only sponge reference genome in the database at the time. The remaining sequences aligned to various taxonomic ranks of bacteria with pairwise identities below 90%. This assured us that the majority of the cosmid inserts in the library were derived from sponge symbiont or otherwise incidental bacterial species in the original sample.

In analysing for duplication among the cosmid inserts, a sequence clustering analysis, using CD-HIT (cd-hit-est v4.6.6), of the high-quality reads at a conservative 95% identity cut-off returned 23 unique clusters for the T7-end reads, and 22 unique clusters for the M13-end reads, with partial congruency between the clusters. This suggested there were at least 22 unique clones within the end-sequenced subset and there was in fact some duplication. Although much effort was expended, the incompleteness of the Sanger based end-sequencing of the 48 test cosmids precluded any definitive analysis of the apparent repeating pattens seen in the restriction-digest analysis. We moved forward noting that this metagenomic library may not be as large as the previous raw titre results would suggest. We concluded from these results that only a small proportion of the cosmid library was derived from the host genome and large proportion was derived from the targeted microbiome with some redundancy. In an attempt to alleviate end-sequencing data gaps and to assess the potential of directly sequencing pools of entire cosmids with emerging long-read technologies, the DNA of the 48 test cosmids were prepared for long-read Oxford nanopore technology (ONT) sequencing and complementary Illumina short-read paired-end sequencing.

3.4.6. Direct sequencing of the 48 test cosmid pool

A necessary endpoint of the large insert library screening efforts planned for this, and other projects being conducted in the laboratory at the time, was the complete sequencing of isolated cosmids of interest. As such, we were interested in assessing possible routes to efficiently and effectively sequence induvial or small pools of isolated cosmids. As the 48 test cosmids were already catalogued and there was some existing end-sequencing data available which could aid in assessing assembly results, they were considered as reasonable candidates for whole cosmid sequencing experiments. The 48 test cosmids were pooled together in equal amounts and this pool of cosmids was referred to in experiments as SpA_1-48. Both long-read (ONT) and short-read (Illumina) sequencing technologies were employed in this experiment to allow for a comparison between the technologies for whole cosmid sequencing, and to take advantage of the complementary sequencing characteristic of each read type. At the time of sequencing, the reported ONT raw read accuracy was ~85%, while the short Illumina raw read accuracy was >99.9%.

3.4.6.1. Nanopore sequencing of the SpA_1-48 cosmid pool

The pooled 48 test cosmids DNA was diluted appropriately and then prepared for sequencing with the ONT Rapid sequencing kit (SQK-RAD002). The prepared library was sequenced on a Mk-I R9 version (FLO-MIN106 ca. 2017) flow cell with a run time of 48 hours. The raw data was basecalled with Albacore (v 1.0.1), producing a total yield of 488.5 Mb in 18,004 passed reads. The read length distribution of this sequence data was tightly centred around ~ 40kb, with 84.3% (411.6 Mb) of the total yield being between 30kb and 50kb long (Figure 3.7). This suggested that most of the sequencing reads spanned an entire cosmid molecule in a single pass, as the expected size of a cosmid with an insert is 40-45kb.

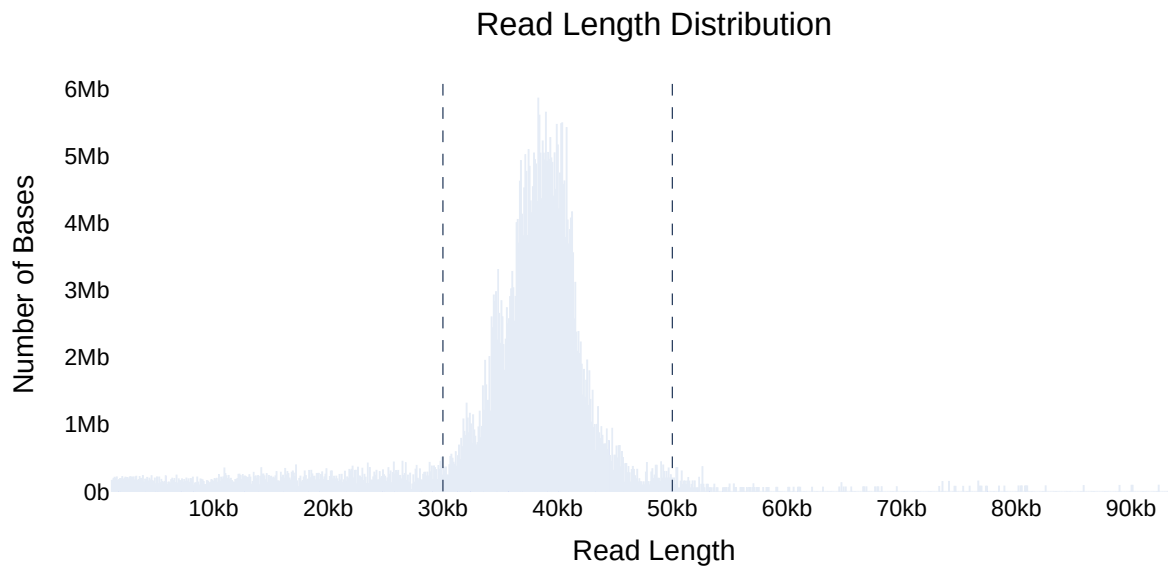


Figure 3.7 – Weighted read length distribution histogram from "SpA_1-48" pooled cosmids ONT rapid sequencing results. Histogram weighted by total bases per bin (1000 bins). The dashed vertical lines are at 30kb and 50kb, 84.3% of the data from this sequencing run falls within the expected range of molecule lengths for intact cosmids.

Assuming the pool was comprised of a maximum of 48 unique cosmids with a length of 40kb, the sequence data between 30-50kb (10,772 reads) gave a theoretical coverage of ~200 fold.

3.4.6.2. Illumina sequencing

The same cosmid pool DNA sample used for the ONT sequencing above was also submitted to an external provider for 101 bp TruSeq paired-end library preparation and sequencing on the Illumina Hi-Seq 2500 platform (Table 3-3). The raw data was quality filtered and sequencing adapters were trimmed to give the "QC passed reads" data set, this data set was used for all subsequent processes.

Data Set	Number of Reads		Total data
	Forward	Reverse	
Raw reads	16,781,608	16,781,608	3.38Gb
QC passed reads	16,254,816	13,681,136	2.94Gb

Table 3-3 – Read metrics for Illumina sequencing of "SpA_1-48" pooled cosmids. The "Raw reads" are the data delivered directly from the sequencing provider. The "QC passed reads" is the resulting data set once filtered for quality and trimmed for sequence adaptors. The "Number of Reads" columns displays the number of forward and reveres reads each data set. The "Total data" is the accumulative read data in giga base-pairs.

The "QC passed reads" had a theoretical coverage of 1.53×10^3 fold over the SpA_1-48 cosmid pool, which exceeded the requirements for this experiment. This excess would adequately make up for any imbalance in abundance in the cosmid pooling, ensuring coverage of unrepresented cosmids.

3.4.7. Assembly and analysis of the SpA_1-48 cosmid pool sequence data

Several bioinformatic approaches were employed to assemble or otherwise generate accurate and, most importantly, unfragmented sequences of the cosmid molecules in the SpA_1-48 cosmid pool. The assembly of pools of cosmids using standard assembly software and algorithms was challenging, as each cosmid contains a large exact disjoined repeat sequence (i.e., the cosmid backbone sequence), as well as overlapping sequence between cosmids and divergent coverage. To my knowledge, no assembly software, for either long or short-reads, was programmed to comprehensively resolve this specific problem. One approach to overcome this challenge was to forgo the assembly of reads sets into contiguous sequences, but rather use the error prone long-reads that spanned the entire cosmid molecules and correct the errors with the accurate short-reads. However, this approach still resulted in large tracts of the target long-read sequences remaining un-corrected and still required extensive further processing of the data to generate an individual sequence for each cosmid molecule. A satisfactory result was not achieved with this method. A similar approach was to use a hybrid assembly method to calculate an assembly using both the short and long-read sets together. To assess this, a hybrid assembler "Unicycler" (v0.3.1), as well as others not discussed here for brevity, was used to generate a hybrid assembly from both short and long read sets. The Unicycler assembly produced 873 contigs in total, only 10 of these were over 25 kb in length, the expected minimal size of a cosmid insert, while the longest of this set was 44.2 kb. These 10 largest contigs all had sequence homology at the extreme 5` and 3` ends to the expected flaking regions of the cosmid vector backbone, confirming they were full insert sequences. Several (5) other shorter contigs had similar homology at only a single end, suggesting a fragmented assembly of the inserts in these cases. The remainder of the data were smaller fragments, which were unusable for our purposes as they could not be reconciled to the original cosmid molecule. Using the Sanger end-sequence data generated from the individual cosmids as query sequences against the Unicycler assembled contigs to recover and assign the full end-to-end insert sequences for each cosmid was too inconsistent to be useful at any scale. The alignments showed cases of infix mapping positions and mismatched mapping of

end-sequence pairs (data not shown). Some contigs had alignment hits with multiple end-sequences as expected considering the duplications seen in the fragments analysis gel. Consequently, it was not possible to show that the direct cosmid sequencing attempts used here could reliably or conveniently reconstitute even a modest pool of cosmids completely, as the expected end-to-end sequences could not be recovered even with extensive manual data analysis. Further efforts to reconstruct these sequences were abandoned. However, many of the intact insert sequences that were recovered exhibited large open reading frames (ORFs) of >10kb in length (Figure 3.8). As large ORFs are prevalent in the megasynthases that this project was targeting it prompted the analysis of these recovered sequences for the presence of BGC like sequences.

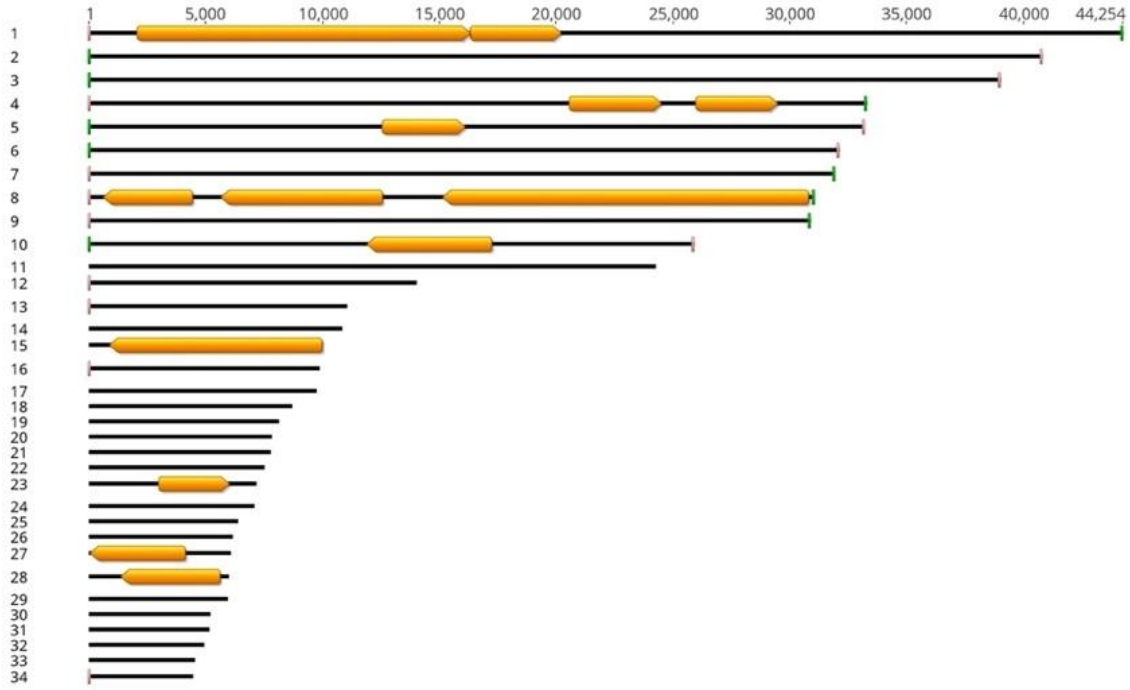


Figure 3.8 – SpA_1-48 cosmid pool assembly. A schematic representation of the top 34 contigs (sorted by length) generated from a hybrid assembly of SpA_1-48 cosmid pool long and short-read data sets. The black horizontal bars represent the span of the contig sequence with number of base pairs shown above. The orange directional bars show ORFs >3 kbp. The vertical-coloured bars show the regions of homology to the T7 (green) and M13 (pink) "ends" of the pWEB vector backbone. The top ten contigs (>25 kbp) show homology to both "ends" of the backbone and are expected to be complete insert sequences. The homology regions are ~30bp in length. The remaining contigs in the assembly are not shown for clarity.

3.4.7.1. antiSMASH analysis of SpA_1-48 cosmid pool assembly

The contigs generated from assemblies of the SpA_1-48 cosmid pool sequencing data (3.4.6) were submitted to the antiSMASH (v4) BGC sequence mining pipeline. For an assembly calculated from the Illumina data set using the IDBA-UD (v 1.1.3) assembler, antiSMASH annotated eight unique contigs (lengths - min. 13.0 kbp, max. 44.2 kbp) with potential

biosynthetic loci, tabulated below (Table 3-4). I have chosen to discuss the results an IDBA-UD assembly of the SpA_1-48 data here, as opposed to the Unicycler assembly discussed above, as it gave the best resolution for the particular BGC highlighted in the following.

BGC type annotation	Number annotated
trans-AT-PKS-like	3
NRPS-like, *trans*-AT-PKS	2
Type-1-PKS	2
Terpene	1

Table 3-4 – antiSMASH annotations for "SpA_1-48" IDBA-UD assembly. antiSMASH results tabulated by annotation type. Annotation types taken directly from antiSMASH results. The "NRPS-like, *trans*-AT-PKS" are hybrid type BGCs.

The annotated BGCs of types "*trans*-AT-PKS-like" and "NRPS-like, *trans*-AT-PKS" aligned with the expected types of the BGCs targeted in this project, i.e., those predicted to encode for the main secondary-metabolites of *M. hentscheli*. On inspection of individual predicted genes or gene fragments in one of the insert fragment sequences (17.0 kbp) that was annotated as a "*trans*-AT-PKS-like" BGCs, BLASTP matches and gene synteny were found with the pederin BGC (Figure 3.9) from symbiont bacterium of *Paederus fuscipes*.[124] These results suggested that the particular cosmid insert that this fragment had originated from had captured part of a pederin-like BGC, presumably the mycalamide A BGC, and that the cosmid library constructed from the metagenome of *M. hentscheli* contained a high proportion of target like secondary-metabolite biosynthetic sequences.

Figure 3.9 – Comparison of a cosmid insert fragments to the pederin BGC. Predicted OFRs of the cosmid insert fragment (top) annotated as a *trans*-AT-PKS-like BGC compared to the annotated pederin BGC (bottom) from the MiBIG database.[193] Directional bars are predicted ORFs or gene annotations. Annotations on top and bottom schematics are not shown at the same scale, the light-pink connected ORFs are the same length (1.4 kpb). Connecting lines indicate BLASTP matches between sequences.

Fortunately, one of the successful end-sequencing reactions mapped accurately and sensibly to the contig, allowing prompt identification of the isolated cosmid. This result implicated that the employed DNA extraction method and library construction were capturing relevant

symbiont organisms. Although only a putative mycalamide BGC fragment could be identified, it was expected that one producer might harbour all of the target BGCs based on previous studies of sponge secondary metabolite production[30].

3.4.8. Metagenomic library construction summary.

A primary aspect of this project was to construct a metagenomic cosmid library from HMW DNA extracted from a biosynthetically rich sample of the sponge *M. hentscheli* and lab-made phage packaging extracts. High quality HMW sponge metagenomic DNA was extracted and its performance in cosmid library building was assessed positively. A full-scale library consisting of ~400,000 cosmid clones was built with the prepared components. A library of this size was deemed expectable for the purposes of this project and was then assessed on several fronts to ensure its suitability for BGC screening. The library was seen to be comprised of the expected large inserts that were predominantly derived from the genomic material of taxonomically board environmental bacterial. An unanticipated degree of apparent insert sequence duplication was observed within a test subset of 48 randomly selected cosmid clones, which may suggest the library was smaller than originally measured. This result was not interpreted as prohibitive for screening. Further investigations with whole cosmid sequencing, using the same 48 test cosmids, failed to demonstrate ideal assembly results, but revealed the biosynthetic richness of the library and a large fragment of a what was likely the BGC for mycalamide, was identified on an isolated cosmid. This metagenomic cosmid library was considered to be fit for its intended purpose. One key insight from these experiments was that *tans*-AT PKS fragments were abundant in our small set of sequence data. Based on previous studies, in which as single talented producer harboured BGCs for the majority of secondary metabolites in a sponge[222], we hypothesised that such a producer might reside within the holobiont of *M. hentscheli*. Furthermore, it seemed likely that such a producer would be present at high abundance, given the large number of BGC fragments in just 48 cosmid clones. These data collectively prompted us to adopt direct metagenome sequencing as an alternative approach for sequencing, assembling and annotating the BGCs for mycalamide, pateamine and peloruside.

4 Direct metagenomic sequencing of *M. hentscheli*

4.1. Direct metagenomic sequencing introduction

The culture independent discovery of novel microbial products from environmental sources, especially soils, has traditionally relied on building large-insert libraries to facilitate the screening and recovery of BGCs.[189,277] This is due, in part, to the difficulty in generating enough sequencing depth with direct shotgun-sequencing approaches from these complex environments to be useful for BGC discovery. In addition, even if you get enough sequence depth, strain level micro-diversity and computational complexity can hinder high quality contiguous assembly.[189,279,286] Compounding this is the difficultly in assembling the highly repetitive BGC from shotgun-sequencing data. A direct sequence-guided approach to these complex environments has been BGC targeted PCR metabarcode sequencing, which can be used to triage environmental samples suitability for metagenomic library building and screening.[279,287] This can allow metagenomic library building efforts to be focused on the most productive samples.[286] The choice of complex soil environments for metagenomic drug discovery has been driven by the abundance of actinomycetes[288] in these environments and their historical precedent for producing secondary metabolites.[15,19] This has seen the development of many systems and tools for building and screening complex metagenomic large-insert libraries which remain useful for BGC discovery[6,7,289,290]. However, working with large-insert cosmid libraries also has some considerable drawbacks. Not only is the construction of these libraries labour intensive and unreliable, but the recovery of target clones can be slow and expensive. As many BGC sequences are larger than the insert sequence held by an individual cosmid, several overlapping cosmids may need to be extracted from the library in multiple rounds of isolation and sequencing to reconcile the final BGC sequence.[291] As such, alternative methods to yield the sequences of BGCs from metagenomic samples would be welcomed.

The use of large-insert libraries has also been the dominant strategy for the discovery characterisation of the BGC responsible for the bioactive compounds isolated from marine sponges.[247] This is presumably carried over from the successes in the application of this method to complex soils environments, as many sponges are also known to have very high phylum-level microbial diversity.[292] During the construction of such a library form *M. hentscheli* metagenomic DNA in the course of this study, it was observed that the sequences we were targeting were in high apparent abundance. We hypothesised that these sequences

were derived from micro-organism(s) that were in high abundance in the sponge microbiome. Based on the "super-producer" hypothesis consistently observed in studies of the sponge *Theonella swinhoei* and the bacterial symbiont *Entotheonella sp*, we expected the target BGCs would be present in the genome of a single metabolically gifted microorganism.[222,225,293] This triggered us to attempt a direct shotgun-sequencing approach of the sponge metagenome, concurrent with the planned library screening campaign. We postulated that it may be tractable to obtain the required sequencing depth of the target organism(s) harbouring the target BGCs to produce a high-quality assembly. At the time of the present study, metagenomic sequencing had seen broad adoption where large continuous sequences (contigs) were not required, especially where the main interest was to resolve the population make up ("metataxonomics") of the environment.[265,294]

Individual microbial genomes were being assembled from lower-complexity mixed populations such and biofilms[295] and activated sludge,[296] as well as some more complex microbiomes from soil[297] and fresh-water.[298] A consistent theme from these studies, especially where sequencing depth was not sufficient, was that the assemblies produced were not suitable for discovery of large BGCs due to very high fragmentation and chimerism.[295,298] The majority of the first genome sequences of uncultured sponge symbionts were obtained from targeted single-cell and whole genome amplification studies.[222,299–302] Although rare, reports of recovering draft genome sequences of uncultured sponge symbionts by direct metagenome sequencing were recently beginning to emerge, however these were typically highly fragmented, and therefore not suitable for contiguous assembly of highly repetitive PKS BGCs.[300,301,303] The use of metagenome sequencing for drug discovery was exceeding rare,[286] although it was shown to be possible in certain ideal circumstances where the producer was found in high abundance.[304,305] These early successes were seemingly driven by the proliferation of new sequencing technologies and the accompanying advances in metagenomic bioinformatic analyses.[301,306] Given the potentially high abundance of the organism(s) we were targeting in our study sample and the maturing field of metagenomic genome assembly,[307] we justified undertaking an explorative short-read sequencing study of total metagenomic DNA extracted from *M. hentscheli*, with the hope of determining whether direct sequencing was likely to be feasible method for discovering the complete BGCs for mycalamide, peloruside and pateamine.

4.2. *M. hentscheli* direct metagenomic sequencing pilot

In the previous chapter, we had detected BGC fragments of interest in a cosmid library constructed from *M. hentscheli* metagenomic eDNA. To assess the feasibility of applying a direct metagenomic sequencing approach to recover further sequence information for these fragments, the same metagenomic eDNA sample (MH_PAT) was submitted for shotgun metagenome sequencing. The aim of this experiment was to generate a pilot scale assembly and investigate this for genome fragments that appeared to cover part, or all our target BGCs. This would then allow us to determine if direct sequencing was likely to be a fruitful strategy and determine the limits of detection, as well as potentially informing the design of subsequent experiments aimed at obtaining higher quality assemblies.

4.2.1. *M. hentscheli* Illumina paired-end sequencing.

A sample of the high quality eDNA used to build the MH_PAT cosmid library was sequenced by Novogene (China) on an Illumina HiSeq2500 platform using the TruSeq DNA PCR-Free sequencing library preparation kit. Over three sequencing runs of this sample, ~10Gbs of total data (250-bp paired-end reads) were returned. The raw data from all runs were combined and processed with Skewer (v0.2.2) to trim or remove low-quality read-pairs and sequencing adaptors to give the QC-passed data set (See Table 4-1). Data set quality was assessed with fastQC (v0.11), which did not indicate any obvious quality concerns with these sequencing runs. Analysis of the read-pair insert length showed the data sets to be highly consistent.

Data Set	Number of read-pairs	Insert length (Ave ±SD)	Total data
Raw data - Run 1	4,581,078	360.9 ±33.0	2.29Gbp
Raw data - Run 2	11,844,229	361.4 ±32.9	5.92Gbp
Raw data - Run 3	3,924,849	361.3 ±33.0	1.96Gbp
Combined	20,350,156	-	10.17Gbp
QC-passed	20,323,215	-	**10.17Gbp**

Table 4-1 Summary of *M. hentscheli* Illumina 250-bp paired-end sequencing data. Total data yields in gigabase pairs for the three Illumina 250-bp paired-end sequencing runs. Combined: the data of all three raw data sets combined to give the total raw data. QC-pass: the data remaining after quality processing. The average insert lengths and standard deviations (SD) were calculated as total fragment lengths minus read length from mapping 1.2M reads to an assembled data set.

The quality control processing scheme discarded both reads of a pair if either read fell below set length and quality thresholds (100bp/Q30). The quality control processing of the total

combined raw-data only removed a tiny fraction of sequencing data (<0.1%). The total yield of ~10Gb of raw data was deemed to be sufficient to achieve the stated aims of this pilot study while balancing resource requirements. The total raw data was submitted to the SRA under accession number: SRX6822952

4.2.2. Illumina short-read *M. hentscheli* metagenomic assembly

The above MH_PAT QC-passed short-read metagenomic data set was assembled with IDBA_UD (v1.1.3/mod) modified to allow for increased kmer length assemblies to accommodate the 250-bp paired-end data. The resultant assembly (MH_PAT_IDBA-UD_10 Gbp_PE-250-bp_Illumina), is summarised in Table 4-1. All contigs shorter than 1000 bp were discarded prior to calculating assembly metrics. Several other available assemblers were also assessed with the same dataset, the results were found to be comparable or inferior based on continuity and total size of the assembly (data not shown for clarity).

Assembly: MH_PAT_IDBA-UD_10 Gbp_PE-250-bp_Illumina			
Assembly metrics		Number of contigs by length	
Total length:	441.2 Mbp	Total :	129,584
N50:	3,502 bp	>= 10 kbp:	4,988
GC (%):	41.14	>= 50 kbp:	348
Largest contig:	3.5 Mbp	>= 100 kbp:	141
Shortest contig:	1000bp	>= 500 kbp:	9

Table 4-1 – Assembly statistics for MH_Pat 10 Gb Illumina assembly. Assembly metrics as assessed by Quast (v5.0.2) for all contigs >1000 bp.

The presence of over one-hundred assembled contigs larger than 100 kbp and several spanning into the mega-base range was a promising result. These larger contigs could represent substantial portions of a symbiont bacterial genome and would potentially encompass full BGCs if the target organism were represented in these contigs. Identifiable fragments of target BGCs may also be located on shorter contigs which could serve as useful anchor points to guide elucidation of the full BGC sequences. The coverage of shorter BGC fragments might also be useful for designing subsequent biosynthetic discovery experiments.

4.2.3. Metagenomic assembly sequence mining for mycalamide BGC.

The putative mycalamide (**Myc**) BGC fragment identified in chapter 3 provided a reference with which to begin searching our metagenome assembly. The sequence fragment recovered from the cosmid clone assembly experiments was recognised as a candidate for **Myc** by close homology to genes in the known pederin BGC, however, only a small portion (~17 kbp) of the **Myc** BGC was recovered from this cosmid clone sequence. A BLASTN search was conducted against the *de novo* metagenomic assembly using the recovered **Myc** fragment (Myc_frag_1) as a query. This returned significant hits (E-value = 0) to two contigs of lengths 33.7 kbp and 37.0 kbp from the assembly. The query sequence spanned the edges of these two contigs such that it joined them together with a 8 bp gap. Manual sequence curation resulted in a 70 kbp contig, extending on the previous **Myc** fragment, Myc_frag_1, by >50 kbp. Joining of the contigs across the assembly breakpoint and inclusion of the 8pb bridge was confirmed by mapping the Illumina sequencing reads across the breakpoint repair using BBMap (figure). This showed a level of heterogeneity at the breakpoint, with a number of reads mapping perfectly in support of the bridge insertion variant and an approximately equal numbers of reads supporting exclusion of the bridge and the following 7bp as a 15bp insertion variant (Figure 4.1). This heterogeneity was likely the cause for the failure of the assemblers to resolve this region and maybe a result of strain micro-diversity, a common issue in metagenomic assemblies.

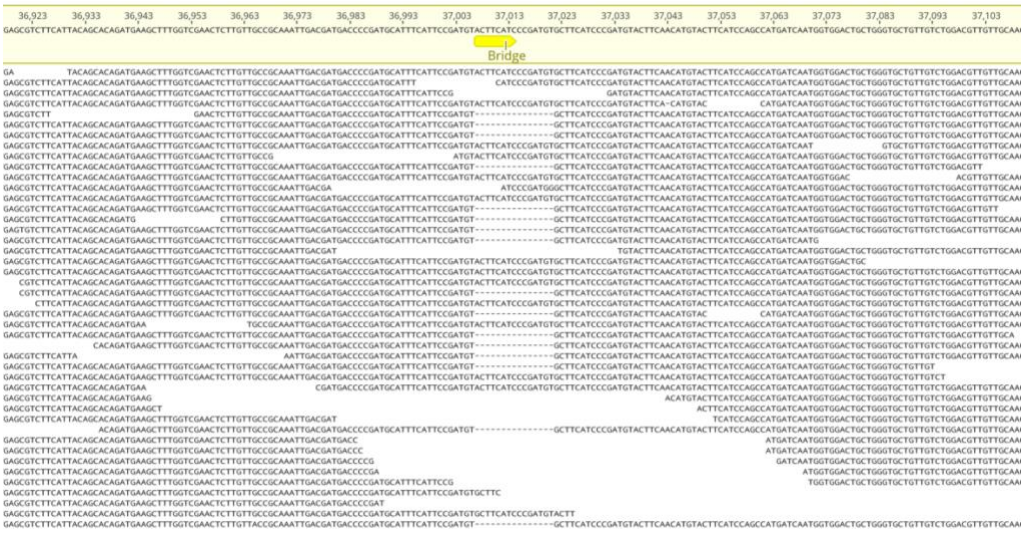


Figure 4.1 – Read mapping to the assembly breakpoint of the Myc fragment reconstructions. Sequence reads were mapped to the fragment reconstruction sequence (top – light-yellow background) to analyse the bridged region (yellow arrow). Reads show support for two arrangements, one placing the bridge within in a 15bp INDEL.

Interestingly, another contig (19 kbps) from the cosmid clone assembly mapped to this new extended **Myc** contig, abutting with the mapping region of the Myc_frag_1 annotated cosmid clone fragment (Figure 4.2). These two sequences were presumably derived from the same original cosmid molecule that had failed to assemble into a single contig in our cosmid sequencing experiment (Chapter 3). Although this additional fragment had initially been identified as encoding PKS and NRPS biosynthetic genes, we had not previously been able to conclusively link it to **Myc** biosynthesis without this further context.

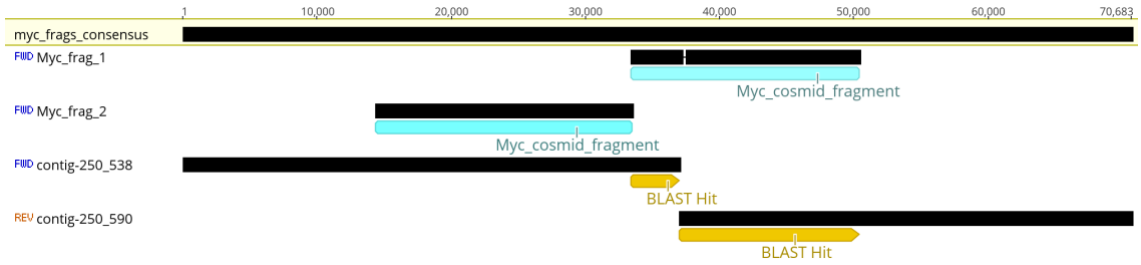


Figure 4.2 – Reconstruction of the mycalamide BGC (Myc). The constituent sequence fragments mapped to the reconstructed Myc BGC consensus (myc_frags_consensus) to show the layout of the reconstruction. Two contigs from the MH_PAT_IDBA-UD_10 Gbp_PE-250-bp_Illumina assembly, contig-250_538 and contig-250_538, were blastn hits (yellow annotations) for the **Myc** candidate query Myc_frag_1. Myc_frag_1 and Myc_frag_2 (blue annotations) are contigs from the cosmid clone assembly. Narrowing of the black bar of Myc_frag_1 highlights the position of the (8bp) gap between the two blastn hits.

The reconstructed **Myc** scaffold was annotated using antiSMASH and the overall architecture of the BGC was compared to that of the onnamide[43] and pederin BGCs (BGC0001108/BGC0001105), two molecular relatives of mycalamide, to check for completeness. The BGC sequence appeared to be nearly complete and was annotated with large biosynthetic genes that spanned over the sequence joins of the new extension (Figure 4.3). Additional biosynthetic genes were also detected in the extensions which maintained gene synteny with the onnamide gene cluster.

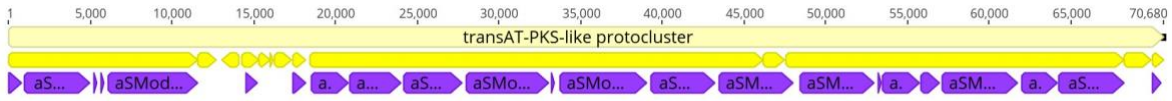


Figure 4.3 – antiSMASH annotation of the reconstructed Myc BGC. Large biosynthetic genes (bright yellow) are annotated by antiSMASH with positions of underlying biosynthetic modules (purple) shown. The numeric sequence positions are shown above (1-70,680 nt). The annotated genes span over the previous edges of the initial **Myc** fragment at ~35k and ~50K. The entire sequence length is annotated as a *trans*-AT PKS, suggesting the cluster may continue beyond the boundaries of this contig.

The presence and order of the biosynthetic modules within the detected genes were plausible for the biosynthesis of mycalamide through a *trans*-AT PKS NRPS hybrid mechanism, the details of which are discussed in depth in a later chapter.

As a plausible candidate for the **Myc** BGC had been recovered from the lower range of acceptably sized contigs of the assembly, other BGCs from this organism may also be captured on comparably sized contigs with enough context and information to assign them to their reciprocal secondary metabolite products. To ascertain if additional BGCs were also assembled to a useful degree, the *M. hentscheli* metagenome was comprehensively mined for natural product biosynthetic loci.

4.2.4. BGC mining of the short-read *M. hentscheli* metagenome assembly

The *M. hentscheli* metagenome short-read assembly of contigs >10 kbp was submitted for genome mining with antiSMASH to detect and classify a broad range of BGCs. The results were then manually scrutinised to find potential hits for the target BGCs.

The original processing of this data was done using antiSMASH-standalone version 3 installed on a private cloud-computing service (AWS). Due to availability constraints, results from the improved antiSMASH version 6.0.1 web-server was used for the following discussion with no significant impact on the main findings.

In total, 124.2 Mb of sequence data in 4,988 contigs were mined. From this, antiSMASH classified 90 BGC regions (the highest level of antiSMASH annotation logic)[308] in 83 of the contigs (Table 4-2).

BGC type annotation	Number	BGC type annotation	Number
terpene	14	betalactone	4
trans-AT-PKS-like	12	RRE-containing	4
RiPP-like	11	other	3
T1-PKS	8	phosphonate	2
hserlactone	6	PKS-like	2
NRPS	6	LAP	2
T3-PKS	5	arylpolyene	1
ectoine	4	ranthipeptide	1
NRPS-like	4	proteusin	1

Table 4-2 – BGC Annotation types enumerated for short-read *M. hentscheli* metagenome assembly. Count of BGC types detected by antiSMASH in the short-read *M. hentscheli* metagenome assembly. A total of 90 BGCs were classified in contigs >10 kbp from the assembly. RiPP; ribosomally synthesized and post-translationally modified peptides, hserlactone; homoserine lactone, T3-PKS; type III PKS, RRE; RiPP recognition element, LAP; Linear azol(in)e-containing peptides.

Many of the detected BGC regions were of the types predicted to produce the target molecules, namely *trans*-AT-PKS and NRPS, with some of these regions annotated as chemical hybrids of these two types. The detected BGCs were distributed over a range of contig lengths, with the largest contig submitted (3.5 Mb) having four BGC loci detected, the largest BGC region annotated was 52.2 kb and the shortest classified contig was 10,002 bps. This meant that many of the detected BGCs were potentially split over several contigs of this moderately fragmented assembly, while other BGCs captured on the lager contigs were conceivably complete and flanked by their genomic context.

Examination of the target-type classifications (i.e., *trans*-AT-PKS/NRPS) identified two contigs of interest that were plausible candidates for fragments of the pateamine BGC (**Pat**). These two contigs were both in the lower length range at 26.6 kbp and 39.8 kbp respectively. The structural features that we determined to be particularly useful in identifying the **Pat** BGC include the presence of β-methylations, a thiazole moiety, an *N*,*N*-dimethyl amine, and a rare dilactone-containing macrocycle backbone. Diagnostic sequence features consistent with the incorporation of a thiazole moiety (heterocyclisation, cysteine-adenylation, and oxidation domains)[164,229,309,310] were used here to identify a putative thiazole-incorporating module that was split between these two contigs. The annotated BGC regions for both contigs extended to the contig edges, suggestive of an incomplete BGC fragment derived from a larger cluster. The shorter fragment (26.6 kbp) contained two consecutive NRPS **A** domains, an oxidoreductase domain (**Ox**), and three consecutive *trans*-AT-PKS KS modules. The **A** domains resided at the extreme 5` edge of the contig and were followed by an usual pattern of three consecutive peptidyl-carrier-protein (**PCP**) sites. The substrate prediction consensus for the **A** domains were serine (Ser) and cystine (Cys). A Cys-activating **A**-domain and **Ox** are required in a thiazole incorporating module. The thiazole incorporating Cys-activating A-domain is usually preceded by two consecutive condensation domain annotations.[311] The second contig (39.8 kbp) was terminated with two such condensation domains. This longer contig also included a glycine-activating NRPS module with an associated N-methyltransferase that was a feasible origin for the dimethyl-amino moiety of pateamine A. Moreover, the pattern of methylating, reducing, and dehydrating modules present in this larger contig was consistent with biosynthesis of the pateamine A backbone. The detailed analysis of a proposed pateamine A biosynthesis will be discussed in a later chapter.

The remaining BGC classified contigs were carefully examined and it was noted that this metagenome appeared to be a rich source of novel biosynthesis with many BGC loci detected

that were evolutionary distant from known BGC sequences. The five largest biosynthetic regions (52.2 – 43.9 kbp) were all annotated as type-I PKS or NRPS containing sequences. However, in the current state of fragmentation, none of the remaining target-type classified contigs could be attributed to any known molecules, including peloruside (**Pel**), without further genetic context. Based on these results, we concluded that the organism(s) accommodating the **Myc** and **Pat** BGCs were sufficiently abundant to facilitate the sequencing depth required for a modest assembly. As we presumed that these BGCs may originate from a single organism, single producers were usually observed in sponge halobionts,[222] the **Pel** BGC may also be encoded in the same genome but too excessively fragmented for reliable detection here. To improve the genetic context and facilitate further identification and reconstruction of these BGCs, we endeavoured to isolate individual genomes from the assembly using the techniques of metagenomic binning.

4.2.5. Metagenome binning of the short-read *M. hentscheli* assembly

Our antiSMASH analysis of the metagenomic assembly allowed us to identify numerous BGC fragments, some of which were diagnostic of the **Myc** and **Pat** clusters, however, based on these data alone, we were not able to connect the disparate fragments, or assign them to individual symbionts. We reasoned that missing fragments from the identified target BGCs could potentially be assigned based on genomic context within the microbiome. To this end, the assembly data was analysed using the concepts of metagenomic binning in order to assign contigs to putative genomes. A full analysis of the microbiome population including an ensemble binning method is carried out in a later chapter.

Binning approaches

Metagenomic binning aims to cluster the contigs derived from the discrete species in a metagenomic assembly into bins to give "metagenomic assembled genomes" (MAGs).[302] That is to say, the near-complete genomes[312] of species are extracted from the mixture of metagenomic contigs. When no reference genomes are available to guide the extraction of contigs from the mixture it is called unsupervised binning and it is a challenging task.[313] The broad topic of this ongoing research[314] is excellently reviewed in depth in several publications[315–317], I wish to touch on it briefly here to give context to the following sections. There are broadly two approaches to unsupervised binning, the first based on sequence-abundance and the second is based on sequence-composition.[317,318] Sequence-abundance based binning relies on the differences and changes in sequencing depth between the species

in a sample.[319] The apparent sequencing depth of an organism is controlled by two factors, the abundance of the organism in the sample and the comparative DNA extraction efficiency from that organism. The abundance of organisms will vary over space and time, as such different physical samples will present with different abundance profiles for the organisms within. The DNA extraction efficiency with a particular method will vary between species, mainly to differences in the organism cell wall.[320] Thus, sequence-abundance methods can leverage information across multiple samples or extraction methods with a technique termed differential binning.[262] The technique relies on the principle that the sequencing depth of contigs from the same genomes should have highly correlated covariance across samples or extraction methods.[317] Sequence-composition based binning utilises features inherent to the contig sequences to bin related contigs together, such as GC content and oligonucleotide usage patterns. The core idea is the observation, that in DNA sequences originating from the same taxon, these features are conserved across the genomes.[321,322] Tetranucleotide (oligonucleotides of length four) frequency is a common "genomic signature" used as a clustering feature in many automated binning tools.[323–326] This high dimensional representation is usually dimensionally reduced and clustered by the use of machine-learning approaches, such as k-means clustering or self-organising maps.[317] The statistical power of these approaches improve with sequence length and struggles with regions of genome heterogeneity, a particular issue for BGC loci as these can diverge from the parent genome.[315] Binning tools can use both abundance and composition methods to improve resolution.[318]

4.2.6. Assembly plotting

To visualise and explore the structure of the relationships of the contigs in the assembly, the approach used for differential coverage binning by Abertsen (2013)[262] was adopted and modified. Sequencing coverage (read-coverage), GC-content, sequence length, superphylum-level taxonomy and tetranucleotide frequency information were calculated for each contig in the assembly (See Methods). This information was then plotted as read-coverage over GC-content, and data points were annotated with the additional information (Figure 4.4). Once these data were plotted, clusters of related contigs on the plot plane representing potential genome bins, could be manually extracted for further investigation. The tetranucleotide frequency profiles of these extracted contigs were used to generate PCA plots to validate and refine the relationship of the contigs in the selected bin. This did confirm that contigs with additional biosynthetic loci detected by antiSMASH were included in the genome bins containing the previously identified target BGC fragments (**Pat** and **Myc**).

Attempts to reassemble the extracted bins or scaffold these contigs together using the paired-end read mapping data did not yield any improvements, especially in regards to the fragmentation of the target BGCs.

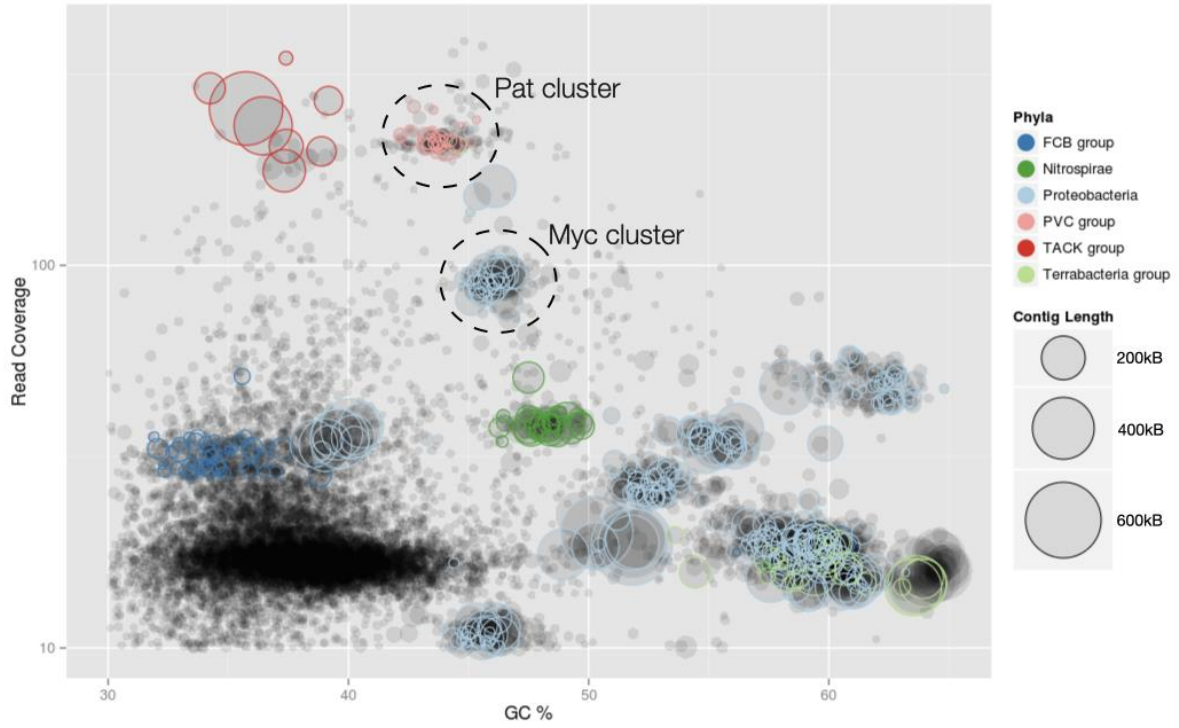


Figure 4.4 – Sequence composition and coverage plot of metagenomic contigs from a short-read *M. hentscheli* assembly. Contigs plotted by average read-coverage over GC content. The length of the contig (bp) are represented by the size of the data point. Contigs that could be taxonomically classified at the superphylum clade level are coloured by classification. FCB group; (Fibrobacteres, Chlorobi, and Bacteroidetes), PVC group; (Planctomycetes, Verrucomicrobia, and Chlamydiae), TACK-group; (Thaumarchaeota, Aigarchaeota, Crenarchaeota and Korarchaeota). Many of the contigs in the dense cluster at the lower left-hand corner of the plane were classified as sponge or other marine eukaryote. The clusters that contained the **Pat** and **Myc** BGC fragments are circled by dashed lines and labelled "Pat cluster" and "Myc cluster" respectively. This plot was derived from an earlier assembly of the data set. Not all assemblies were plotted due to the computational expense of read mapping and taxonomic classification required to generate plots for each calculation of an assembly. The y-axis (read-coverage) is log scaled.

The major insight from this analysis was the apparent separation and phylogenetic difference of the contig clusters (bins) that the BGC fragments of **Pat** and **Myc** belonged to. This suggested that these secondary metabolites were produced by different organisms of high evolutionary divergence. It also showed these two organisms were indeed at relative high-abundance compared to the majority of the microbial population. The only genome bin with comparable or higher read-coverage was classified as belonging to the TACK (Thaumarchaeota, Aigarchaeota, Crenarchaeota and Korarchaeota) group, a taxonomic group of archaea, and appeared to be defined by a relativity contiguous assembly. Several iterations of manual binning and refinement based on the plotted contig clustering structure above was able to confidently produce about 11 distinct genome bins. Genome completeness

was assessed manually using a database of conserved single-copy essential genes[327] as well as the automated CheckM tool.[328] However, this left many contigs unassigned and was subject to poor repeatability and subjective interpretation as it was not automated, nor did it use the most sophisticated techniques that were being developed for use in automated binning software. The automated binning software that were available either needed multiple samples to operate or produced sets of bins that were dubious upon manual inspection based on genome completeness and taxonomic profiling of the binned contigs. The high quantity sponge host derived contigs in the assembly perhaps influenced the performance of the appointed binning software.

The above findings of high-abundance target organism represented as fragmented MAGs supported the case for further investment to improve the metagenomic assembly of the *M. hentscheli* holobiont. The lab had previously attained good results using single-molecule sequencing produced by the PacBio RS-II instrument from a third-party provider for sequencing isolated cosmid clones from sponge and soil metagenomic libraries (not discussed here). Based on the established capacities and analysis experience gained working with PacBio sequence data during the clone sequencing efforts, we commissioned the generation of PacBio long-read sequence data for the *M. hentscheli* metagenome with the intent of improving the metagenome assembly. It had been reported earlier that year (January 2017), that several attempts to improve sponge metagenome assemblies in this way was unsuccessful, however we deemed that the relatively modest investment in sequencing was a justified risk.

"…we have made several attempts to close gaps in the metagenome by rounds of PacBio and Illumina sequencing. Unfortunately, these attempts did not significantly reduce the number of contigs" – Lacknera *et al* 2017

4.3. Long-read sequencing of *M. hentscheli*

In pursuit of improving the promising BGC mining results from the short-read metagenomic assembly discussed above, we proposed to leverage the HMW DNA extraction techniques we had established already to facilitate the generation of long-read sequence data. This long-read data would be used to improve the overall continuity of the metagenomic assembly with the hopes of providing more complete BGC sequences. Due to the repetitive genetic architecture of the BGC of the target types under investigation in this study, they are difficult to resolve completely with short-read sequence data.[214] The sequencing platform chosen to

generate the long-read data was the PacBio Sequel system as it produced the highest yield of long-read data of any instrument at the time.

4.3.1. *M. hentscheli* PacBio long-read sequencing

Much larger amounts of eDNA (>10 ug) were required for PacBio long-read sequencing compared to the short-read sequencing campaign (<1 ug). Necessarily, the DNA also needed to be of high molecular weight (HMW) to yield long sequencing reads as the reads are generated from single molecules. The sequencing method also specifies high purity and concentration of the sample. To meet these requirement, approximate 3.5 g of frozen sponge-tissue powder was subject to the established extraction and size selection process used to generate the DNA sample used to build the *M. hentscheli* cosmid clone library. This sample was additionally treated with RNase-A and subsequently repurified with carboxyl-coated solid phase reversible immobilization magnetic beads. The final sample was sufficiently concentrated and of HMW (Figure 4.5).

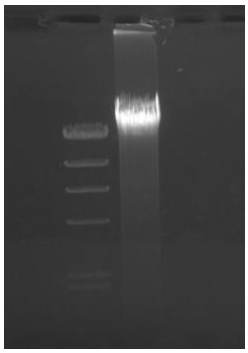


Figure 4.5 – PacBio sequencing DNA sample. 5ul of HMW Agarose gel electrophoresis of *M. hentscheli* DNA sample used for PacBio sequencing compared to λ DNA-HindIII digest marker.

The sample was submitted to the third party sequencing provider Macrogen (Korea) to prepare the SMRTbell >20 kbp template for sequencing on the PacBio Sequel System and sequenced on a single M1 SMRTcell. The submitted sample passed third party quality control prior to sequencing. Sequencing results are tabulated below (Table 4-3).

PacBio sequencing metrics	
Total number of reads:	529,530
Total data yield:	4.86 Gbp
Total Read N50:	15.6 kbp
Average read length:	9.2 kbp
Longest read:	72.5 kbp
Shortest read:	50 bp
Number of reads >10 kbp:	207,119
Data yield for reads >10 kbp:	3.60 Gbp

Table 4-3 – PacBio sequencing data results for *M. hentscheli*. Metrics calculated from all subreads extracted from the polymerase reads.

The single SMRTcell sequencing run returned a modest yield of just under 5 Gbp of sequencing data with a generous proportion of this data distributed in reads above 10 kbp in length. The read length distribution of the full data set is plotted in Figure 4.6.

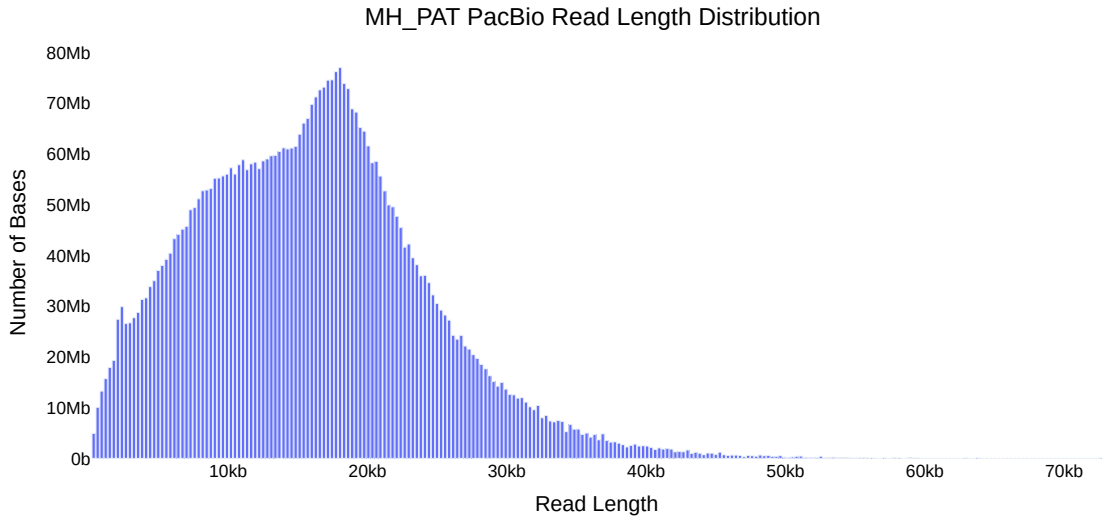


Figure 4.6 – PacBio weighted read length distribution histogram. Histogram weighted by total bases per bin (250 bins). Read length distribution matches the expected results for the library preparation method used, centring around 17-20 kbp.

The sequencing run had returned within the expected parameters for the library preparation and SMRTcell used. The data yield and read lengths of this sequencing data set were deemed acceptable for proceeding with computing the next iteration of metagenomic assembly for *M. hentscheli*. This data set was deposited in the SRA under the accession number: SRX6940356.

4.3.2. Hybrid *M. hentscheli* metagenomic assembly

With two high quality sequence datasets generated from the same biological sample by complementary sequencing technologies in hand, we aimed to conduct a hybrid assembly approach for improving upon the initial short-read metagenome assembly. The use of long-reads, specifically PacBio data, to improve metagenomic assemblies had been recently reported by Frank et al., (2016).[329] Immediately prior to our acquisition of the *M. hentscheli* long-read data, a marine sponge metagenome analysis using both Illumina and PacBio data was reported, by Slaby et al., (2017),[301] and showed improvements in assembly metrics for a hybrid assembly strategy. Both studies used 8 PacBio SMRTcells to generate the required sequence depth. Around this time, advances on the algorithmic challenge of hybrid assembly using Illumina and error-prone long-reads were also being published, as long-read technology was becoming more attenable and the benefits of the hybrid assembly approach were attractive.[330–332] Nether of the aforementioned hybrid assembly metagenomic studies reported employing the improved algorithms. These novel algorithms were not explicitly reported to be aimed at solving metagenomic assemblies which present distinct challenges.[265] One of the recently published hybrid assembly programmes, hybridSPAdes (SPAdes v3.6.2), was applied here to calculate a hybrid assembly of the two *M. hentscheli* data sets. Unfortunately, calculation of an assembly with this strategy was intractable with any available computational resource, including a memory optimised AWS cloud compute node (instance type x1.32xlarge) with >1.9 Tb of RAM. We also tried comparably primitive methods of hybrid assembly of these data. This involved stand alone pre-assembly error-correction of the long-reads with the Illumina data followed by overlap-layout-consensus (OLC) assembly.[333] These attempts were also very computationally intensive and were counterproductive for our task. Another contemporary hybrid assembly strategy, executed by the MaSuRCA assembler (v3.2.8), was able to complete a hybrid assembly of the hybrid data set. This assembler requires some level of coverage by the long-reads to generate assembly of the (meta)genomic region, but is able to process very large genomes within practical memory requirements.[330] This trade-off implies very low abundant species may be missed where the long-read sequencing coverage is insufficient.

The hybrid assembly (`MH_PAT`_masurca_250_PB) results are summarised below (Table 4-4). The assembly is deposited with NCBI under the accession: GCA_012263305.1

Assembly: MH_PAT_masurca_250_PB			
Assembly metrics		Number of contigs by length	
Total length:	442.3 Mbp	Total :	6393
N50:	111,968 bp	>= 10 kbp:	6183
GC (%):	40.83	>= 50 kbp:	2507
Largest contig:	4.7 Mbp	>= 100 kbp:	1054
Shortest contig:	3.2 kb	>= 500 kbp:	45

Table 4-4 – Assembly statistics for MH_PAT_masurca_250_PB assembly. Assembly metrics as assessed by Quast (v5.0.2) for all contigs >1000 bp.

Vitaly, this hybrid assembly showed vast improvements over the short-read only assembly. The continuity, as judged by the increase in N50 (~32X increase), the lower number of contigs (~20X decrease) and increase in number of large contigs, was markedly superior, while maintaining the very similar overall metagenome size (100.2% size of previous short-read assembly) and GC content.

The likelihood of improvements to the continuity and completeness of BGCs in this assembly also seemed plausible, especially for those of higher coverage. Again, the metagenome assembly was plotted as sequencing coverage over GC content with taxonomic and contig size data, in a similar fashion to Figure 4.4 above, to help contextualise the BGC annotations within the metagenome (Figure 4.7).

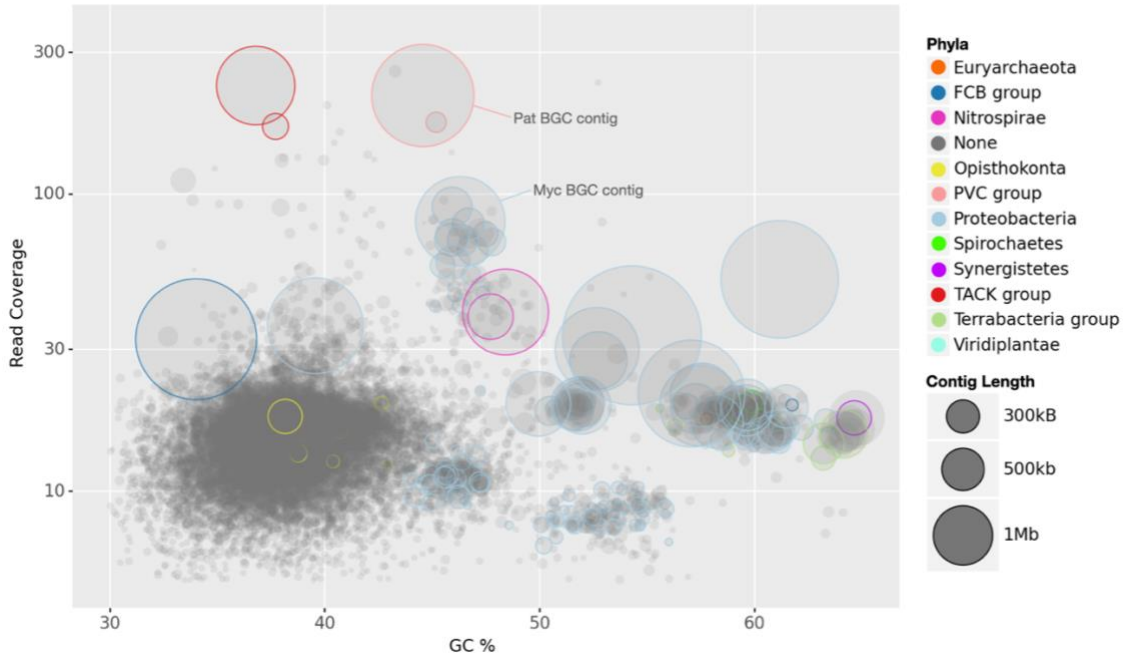


Figure 4.7 - Sequence composition and coverage plot of *M. hentscheli* hybrid assembly. Contigs plotted by average read-coverage over GC content. The length of the contigs (bp) are represented by the size of the data point. Contigs that could be taxonomically classified at the superphylum clade level are coloured by classification. FCB group; (Fibrobacteres, Chlorobi, and Bacteroidetes), PVC group; (Planctomycetes, Verrucomicrobia, and Chlamydiae), TACK-group; (Thaumarchaeota, Aigarchaeota, Crenarchaeota and Korarchaeota). The contigs that contain the **Pat** and **Myc** BGC are labelled. The y-axis (read-coverage) is log scaled. A portion of the taxonomically coloured points are buried beneath other points on the plot. This can be easily observed with the Opisthokonta points (yellow) and is an unfortunate trade-off of using a static plot for this task. The displayed view was selected as balanced overview of all the information, in practice, these plots were used to explore the data dynamically.

The continuity improvements can be clearly observed when comparing the sequence composition plots (Figure 4.4, Figure 4.7) of the two assemblies. Of note, several groups of taxonomically similar contigs that clustered together in the in the initial short-read assembly plot were now represented as single, or a few, large contigs with the same taxonomic assignment. This suggested that the hybrid assembly had succeeded in resolving previously highly fragmented bacterial genomes into near fully assembled genomes within the metagenome. The initial clusters were shown to be composed of related contigs derived from a single molecule, this was confirmed by extracting and mapping the initial clusters to the newly assembled contigs. This supported the use of this simple clustering method to gauge the relationship between contigs. The hybrid assembly also facilitated the assignment of several new phylotypes to the contigs, from six superphyla in the initial assembly to eleven in the hybrid assembly, while maintaining all previous phylotypes. The central finding allowed by the hybrid assembly, and observable on the composition plot (Figure 4.7), was the positioning of the **Pat** and **Myc** BGC fragments now on large contigs, both >2Mb. These

contigs were immediately extracted for BGC loci annotation to appraise any improvement to these fragmented targets.

4.4. Annotation of Pat and Myc BGCs contigs

The contigs form the MH_PAT_masurca_250_PB hybrid assembly identified as containing the previously annotated **Pat** and **Myc** BGC fragments were isolated from the metagenome for annotation by mapping (BLASTN) the known BGC fragment sequences and extracting the hit contigs from the referenced assembly. The known sequences of both BGC fragments mapped completely to internal regions of two large and distinct contigs, of size 3.1 MB for **Pat** and 2.6 Mb for **Myc**. The entire contigs were then annotated with antiSMASH to examine the BGCs located on these contigs. AntiSMASH applies an annotation framework progressively to the underlying sequence by building up upon several layers of biosynthetic logic. The general annotation hierarchy follows down from the top-level "biosynthetic region" annotation, which envelopes one or several related "biosynthetic gene" annotations to distinguish the "biosynthetic region" type. In the case of polyketide and NRPS biosynthetic genes, these "biosynthetic gene" annotations are naturally sub-divided by the annotations of the underlying "modules", which in turn are composed of "domain" annotations. Finally, the "domain" annotations are built up from the annotations of the elemental protein-family ("pfam") domains.[202] This annotation scheme is outlined here to aid in the following discussion and will also offer useful context for later chapters.

4.4.1. Annotation of a complete Pateamine BGC

Seven biosynthetic regions were detected within the extracted **Pat** contig which were all annotated as NRPS or (*trans*-AT) PKS types. Two of the largest BGC regions (166.8 kb and 83.3 kb) showed biosynthetic arrangements directly corresponding to the previously annotated **Pat** fragments. These annotated biosynthetic regions were separated on the contig by a genomic interval of 16.5 kb although the distance between biosynthetic genes was larger. Together, these two regions (**Pat** region-1 and **Pat** region-2) were considered to make up the entire pateamine BGC, which spanned, including the inter-region space, a continuous 266.7 kb of the contig. On further inspection of the annotations, these two biosynthetic regions appeared to share a large repeat of overlapping biosynthetic module annotation architecture (Figure 4.8). This repeat region was within large *trans*-AT-PKS genes present in each of the biosynthetic regions.

A)

Region	Type	From	To
Region 1	transAT-PKS-like, NRPS, T3PKS	516,981	683,781
Region 2	NRPS-like, transAT-PKS-like	700,369	783,687
Region 3	PKS-like	1,147,309	1,188,559
Region 4	transAT-PKS-like	1,328,823	1,375,143
Region 5	transAT-PKS-like	1,428,810	1,519,014
Region 6	transAT-PKS-like	2,727,692	2,768,750
Region 7	T1PKS	2,964,744	3,008,652

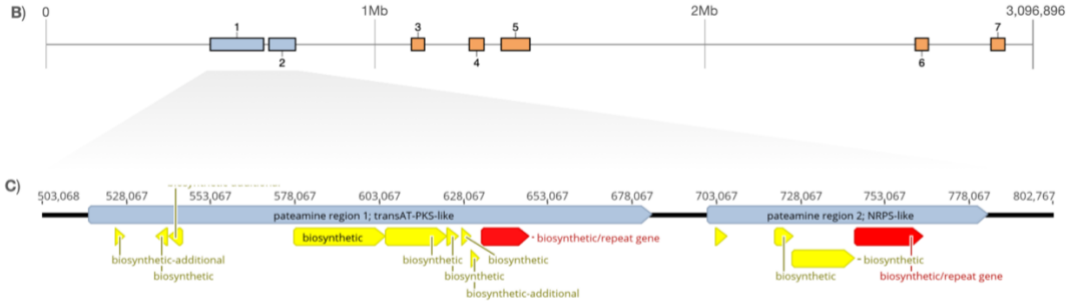


Figure 4.8 – Biosynthetic gene cluster annotation of Pat contig. Overview of the annotation output from antiSMASH for the contig identified as containing the putative **Pat** BGC. **A**) Table of all seven biosynthetic regions detected on the contig with the regions type and contig position (adapted from antiSMASH output). Blue highlighting of the *trans*-AT PKS/NRPS hybrid genes of the **Pat** regions, orange highlighting of other PKS genes. **B**) Schematic of the entire ~3 Mb contig with positions of the detected biosynthetic regions mapped. **C) Pat** regions with biosynthetic genes >2.5 kb annotated (yellow). Red annotations show the genes with the detected repeat region.

A comparison of these repetitive *trans*-AT PKS genes found in each of the **Pat** regions by nucleotide alignment (MAFFT L-INS-I v7.450) showed that the genes were of different length, 13.7 kbp for the gene in **Pat** region-1 and 20.2 kbp for the gene in **Pat** region-2, and that they aligned perfectly for 9.6 kbp at the 5` end of the gene (Figure 4.9). There was a sharp transition to sequence diversion in the alignment in the middle of an annotated PKS module within the gene. Interestingly, the overlying biosynthetic architecture throughout the divergent sequences were conserved, suggesting an evolutionarily distant replication event with a section of the repeat under strong purifying selection.

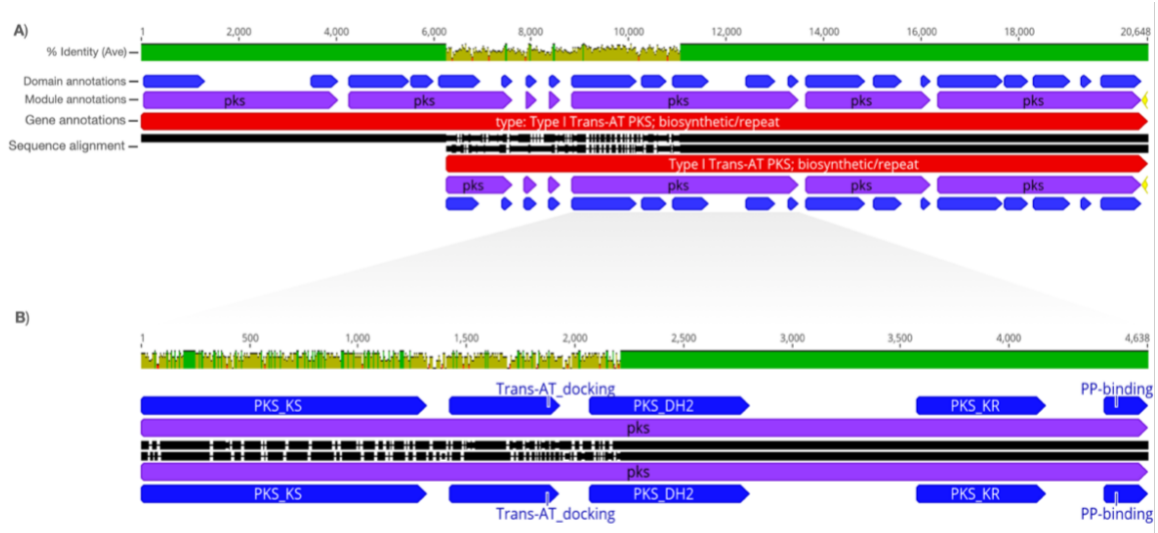


Figure 4.9 – Nucleotide alignment of the Pat BGC repeated *trans*-AT PKS gene. Gene and PKS module nucleotide sequences extracted and pairwise aligned with MAFFT. **A**) Detail view of the gene (red) alignment. Level of alignment identity is depicted by fluctuations in green bar above alignment. Module (purple) and domain (blue) annotations also shown. Sequence alignment depicted by black bars with indels shown by breaks in bar. The alignment show perfect conservation for the 5` end of the gene. This conservation is interrupted within a PKS module. **B**) Extraction and realignment of the nucleotide sequence of the PKS module spanning the conserved/divergent transition.

Comparison of the underlying sequences of the two main pateamine biosynthetic regions showed that there was in fact another larger repeat region shared between them (Figure 4.10). The additional repeat also displayed the pattern of including partially perfect conservation of biosynthetic genes. These repeat structures would be exceedingly difficult to resolve without the use of the long read hybrid assembly.

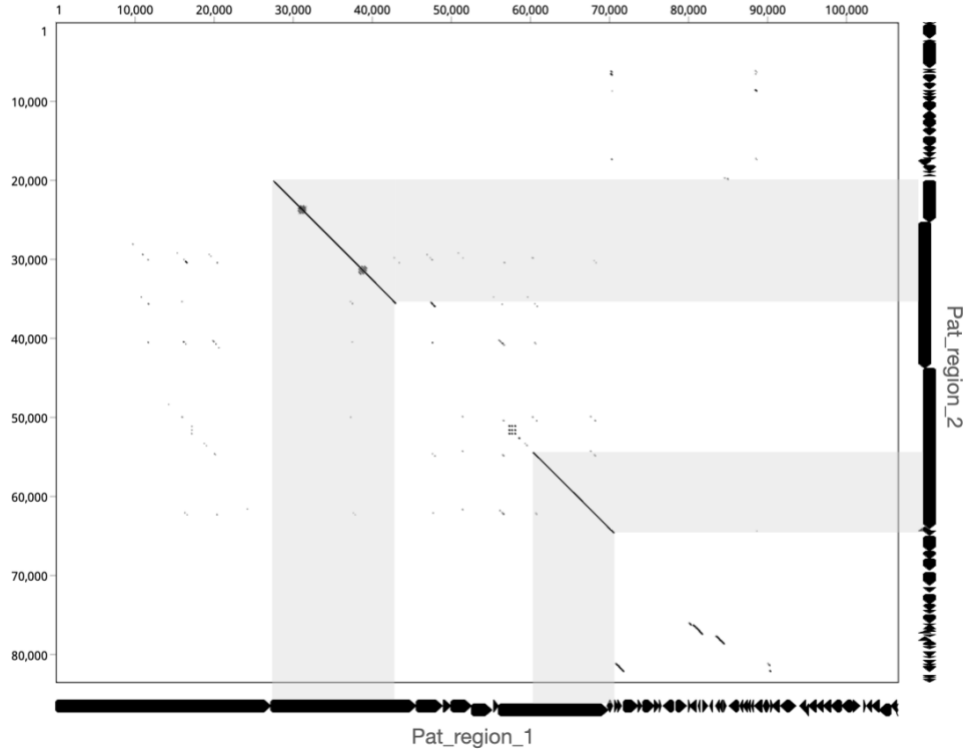
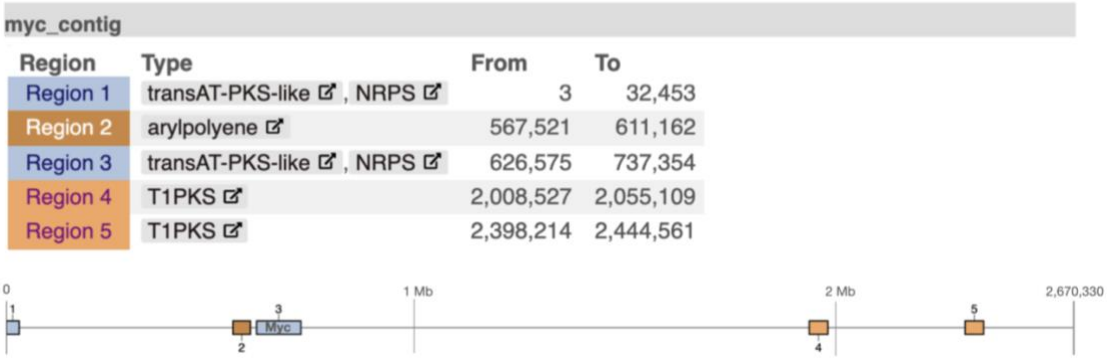


Figure 4.10 – Dot-plot of Pat biosynthetic regions. Comparison of the two main regions of **Pat** highlighting the large duplications of the pathway sequence. Boundaries of the repeat regions lay within ORFs, showing partial gene conservation. Length in nucleotides and ORF annotations shown on axis.

4.4.2. Annotation of a complete Mycalamide BGC.

The large contig identified as containing the putative **Myc** BGC sequence was similarity annotated with antiSMASH to examine any additions or changes afforded by the hybrid assembly. This contig housed five biosynthetic regions, all (*trans-AT*) type I PKS like, except one aryl polyene BGC. Again, the **Myc** cluster could be identified by its identity with the previous elucidated putative **Myc** BGC sequence. However, akin to the **Pat** contig, there were two biosynthetic regions detected on this contig that bore resemblance to the known **Myc** sequence. One of these was a complete uninterrupted **Myc** BGC in the interior of the contig and was 110.8 kbp in length, ~40 kbp longer than previously recovered, and included a transcriptional regulator gene. The other was an incomplete BGC (32.4 kbp), cut off at the contig margin (Figure 4.11). The two were split by a genomic interval of over 500 kbp. The large biosynthetic repeat regions here are likely responsible for an incomplete assembly of the genome of intertest.



myc_contig

Region	Type	From	To
Region 1	transAT-PKS-like, NRPS	3	32,453
Region 2	arylpolyene	567,521	611,162
Region 3	transAT-PKS-like, NRPS	626,575	737,354
Region 4	T1PKS	2,008,527	2,055,109
Region 5	T1PKS	2,398,214	2,444,561

Figure 4.11 – Biosynthetic gene cluster annotation of Myc contig. Overview of the BGC annotations laid-out on the **Myc** containing contig with type and contig position tabulated. The **Myc** BGC is "Region 3" and the incomplete repeat is "Region 1". Regions are color coded type. Blue; *trans*-AT PKS/NRPS, Brown; arylpolyene, Orange; type-I PKS.

Nucleotide alignment of the two **Myc** like biosynthetic sequences on this contig showed a high-level of sequence identity (99.9%), only broken by one large 1.6 kbp indel and a small constellation of single-nucleotide polymorphisms or small indels. Although the full extent of the replication could not be determined from this single chromosome, as one repeat was truncated by a break in the assembly, this BGC may be partially or completely replicated within the organism's genome. The observation of multiple copies of *trans*-AT PKS type BGCs is emerging as a common theme, especially in symbiont genomes, and can effect gene-dosage to increase productions or perhaps allow for alternative biosynthetic pathway processing.[33,334]

4.5. Sequence mining the hybrid *M. hentscheli* metagenome assembly.

Two of the three main target BGCs sequences had been discovered complete in the hybrid assembly. This was aided by having partial fragmented BGC sequences at hand from results of our earlier metagenomic sequencing and library experiments. However, no such sequence data was yet available for the remaining target BGC, peloruside (**Pel**). Rigorous chemical structure-based retrobiosynthetic prediction of the remaining BGC detected in the **Pat** and **Myc** contigs did not yield any candidates for a **Pel** BGC. To this end, we undertook an unbiased strategy towards identifying additional BGCs in the hybrid assembly, with the aim of discovering the **Pel** BGC, the assembled contigs (>10 kb) were submitted for BGC annotation by antiSMASH with default setting. As the majority of this assembly data is found in contigs >10 kb in length, much more of the data can be effectively processed with genome

mining tools. 99.5% of this hybrid assembly was able to be effectively mined, compared to ~25% of the short-read only assembly.

In total, 90 biosynthetic regions were detected in the submitted contigs, including those already annotated on the **Pat** and **Myc** contigs. A summary of the overall findings are tabulated below (Table 4-5).

BGC type annotation	Number	BGC type annotation	Number
Terpene	14	other	4
trans-AT-PKS	13	PKS	2
NRPS	13	arylpolyene	2
RiPP	11	LAP	2
T1PKS	8	redox-cofactor	1
hserlactone	7	ranthipeptide	1
T3PKS	6	proteusin	1
ectoine	5	phosphonate	1
RRE-containing	4	lassopeptide	1
betalactone	4	hglE-KS	1

Table 4-5 – BGC region types enumerated for hybrid assembly of *M. hentscheli*. Counts of BGC region types detected by antiSMASH in the *M. hentscheli* hybrid assembly for contigs >10 kbp. Chemical-hybrid regions were split prior to enumeration for clarity, so total count of region types (101) exceeds the number of detected regions (90).

As peloruside lacks obvious structural moieties that indicate its PKS type, all of the possible candidates for a **Pel** BGC in the collection of detected BGC regions (i.e., types; *trans*-AT-PKS, NRPS, T1PKS and PKS) were examined by manual structure-based retrobiosynthetic analysis. Unfortunately, none of these BGCs regions were deemed to be a plausible **Pel** BGC or fragments thereof. The lack of a **Pel** BGC candidate in these assemblies was consistent with the H^1 NMR measured chemotype (peloruside negative) of this sponge sample. This deficiency prompted a supplementary campaign, enrolling several additional peloruside positive *M. hentscheli* sponge samples, aimed specially at the recovery of the **Pel** BGC by repeating our success so far with this sample in an appropriate chemotype.

4.5.1. Additional BGCs

Several additional and apparently complete or nearly complete BGCs, of types including PKSs, NRPSs and RiPPs, were revealed in the unbiased genome mining that could not be assigned to any known *M. hentscheli* metabolites, increasing the catalogue of orphan BGCs recovered from this sponge. This suggested the biosynthetic potential was higher than what

had been previously expected by chemical analysis alone and supported the case for an unbiased metagenomic BGC discovery approach. The orphan BGCs were distributed among phylogenetically distinct species (based on analysis discussed in chapter 7). Even within the putative genomes containing the target **Myc** and **Pat** BGCs, additional BGCs were identified, suggesting these organisms can produce additional compounds. These particular gene clusters were not examined in-depth here other than to rule out their role as a potential **Pel** BGC.

One "off-target" BGC found elsewhere in the metagenome and examined further was a polytheonamide like RiPP BGC. This BGC was interesting to us as a polytheonamide like molecule has not been detected in this sponge species. Compounds of this family usually occur in the unrelated sponge *Theonella swinhoei* and have been attributed to the uncultivated bacterial symbiont '*Candidatus* Entotheonella factor'.[222] Due to the linkage to this well characterised and biosynthetically rich symbiotic system we justified efforts to further characterise this novel RiPP BGC which was subsequently named **Gan**. The analysis of this gene cluster is detailed in chapter 6.

4.6. *M. hentscheli* direct metagenomic sequencing summary

Analysis of a *M. hentscheli* metagenomic large insert clone library implied that one of the target BGC, **Myc**, was in notably high abundance. This observation provoked us to attempt a bold untargeted direct metagenomic sequencing strategy to recover the target BGC's full sequences, which were expected to reside within the same genome. An initial short-read, sequencing only highly fragmented assembly, was useful in recovering fragments of two target BGCs (**Myc** and **Pat**) and confirmed they were both at high relative abundance. However, these BGC sequences were poorly resolved and appeared to be in phylogenetically distinct genomes. Many additional BGC fragments were detected within the assembly, but we were unable to assign these to any of our three targets, including **Pel**. To reduce the assembly fragmentation, in the hopes of expanding and bridging gaps in the recovered BGC sequences and recovering the final target, we opted to generate long-read sequencing data. This data facilitated a much-improved hybrid assembly which generated many Mb scale contigs, including contigs with the complete **Myc** and **Pat** BGCs encoded within them. These assembled contigs also revealed that these BGCs exhibit a repetitive pattern in the genomes of the producing organisms, and these organisms may produce yet unknown additional secondary metabolites. Much additional biosynthetic potential beyond the known

metabolites was uncovered by the direct metagenomic approach. However, the final target (**Pel**) remained elusive.

The following chapters will cover the extended efforts undertaken in pursing the **Pel** BGC and an in-depth analysis into the biosynthetic mechanism of the recovered BGCs.

5 Peloruside biosynthetic gene cluster exploration

5.1 Introduction

The declared aim of this study was to discover the BGCs sequences of the three main known secondary metabolites of *M. hentscheli:* mycalamide, pateamine and peloruside.
In the last chapter (4), we had succeeded in identifying the sequences of two (**Myc** and **Pat**) of the three target BGCs from a hybrid metagenomic assembly. The sample of *M. hentscheli* use for generating this assembly (`MH_PAT`) was found to be clearly chemotype positive for the two respective compounds by NMR. Thus, the absence of a clear candidate for peloruside biosynthesis was not entirely unexpected from these results.
However, subsequent examination of ethanoic extracts the sample supplied to collaborating natural product chemists from the lab group of Dr Rob Keyzers (VUW), namely Sarah Andreassend and Joe Bracegirdle, using liquid chromatography-tandem mass spectrometry (LC-MS/MS) revealed that peloruside was also present in this sample (Figure 5.1).

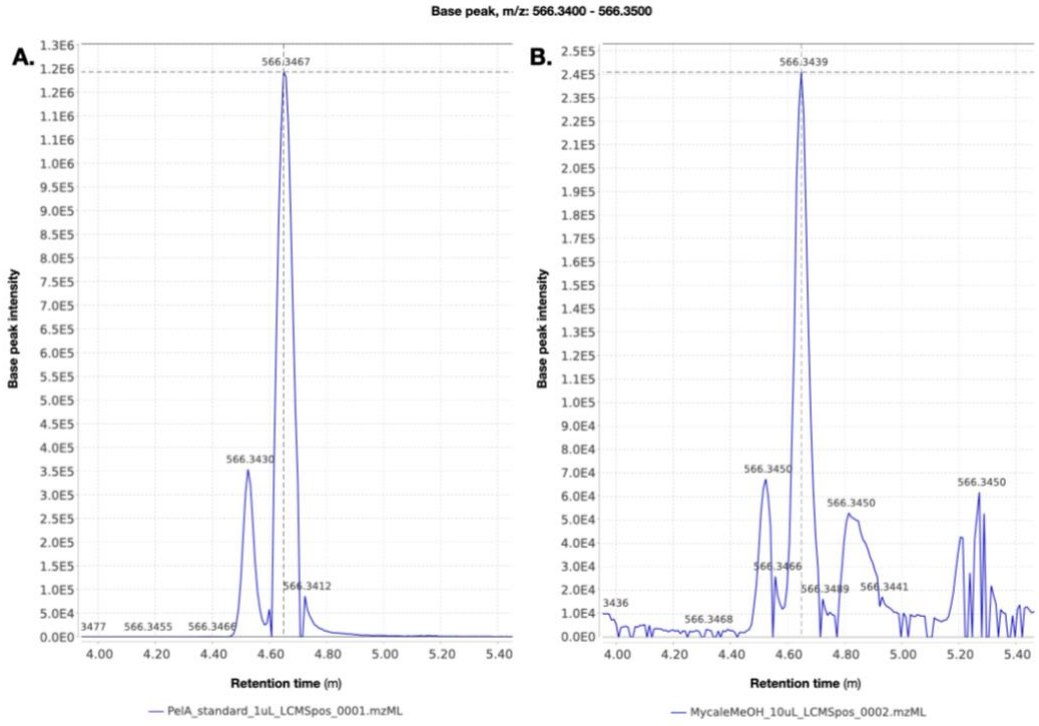


Figure 5.1 – Peloruside detected by LC-MS/MS in `MH_PAT`. (A) LC-MS-extracted chromatogram of peloruside standard. (B) LC-MS-extracted chromatogram of crude methanolic extract from sample PAT showing compound with congruent mass and retention time.

We had an expectation that a single microorganism may be responsible for producing all three of the compounds and encode all three BGCs, based on the experimental precedents set by previous marine sponge secondary metabolite studies.[222,277] However, we had not seen any evidence to support this scenario from the initial investigations of the metagenomic assemblies and considered the possibility that the microbe harbouring **Pel** was distinct and of low-abundance in the studied sample. We reasoned that applying a similar direct metagenomic sequencing approach to a *M. hentscheli* sample that was more strongly chemotype positive for peloruside may aid in yielding the remaining BGC and finally complete the primary objective of the study. The recovery of **Pel** was deemed important, due to the potential this compound had shown as a microtubule stabilizing agent[335–337] and the ongoing research centred around its clinical use, synthesis and sustainable produciton.[243,244,338–340] To this end, we sought access to additional peloruside positive samples to progress this work, and committed to allocating significant recourses and time to pursuing the **Pel** BGC.

5.2 Peloruside positive sponge samples

The known chemical profiles of the additional *M. hentscheli* samples discussed below were previously disclosed in section 3.2.2 (figure 3.2), along with the initial sample (`MH_PAT`) for comparison. These additional samples were not in fact accessed until the metagenomic sequencing and analysis of `MH_PAT` was complete. For this phase of the study, six total additional voucher samples (Named: `MH_PEL`, `S1`, `S2`, `S3`, `S4`, `S5`) were recovered from archival storage. These samples were selected from hundreds of stored samples for being labelled as, or measuring positive for, peloruside and yielding adequate DNA. Multiple samples were recruited for three reasons. 1) To permit redundancy and allow for failure of any of the technical steps leading up to the final analysis, 2) to permit attempts at co-assembly and covariant abundance (differential coverage) binning techniques and 3) to examine the temporal and spatial variability in the microbiome of *M. hentscheli*. The sample `S5` was peloruside negative and was also included to allow for controls and to contribute to covariant binning. Where possible, the final chemotype measurements were made by observing chemical shifts in H^1 NMR spectra at known positions correlating with the presence of peloruside (figure 3.2) from cyclic loaded polystyrene-divinylbenzene column enriched methanoic sponge extracts, using an established protocol.[282,341] The `S1`, `S2` and `S3` samples were found to be peloruside positive (P^+) with this method, displaying all expected

chemical shift signals associated with peloruside, while none of these shifts were present in MH_PAT or S5 (figure 3.2.2). MH_PEL was included based on initial peloruside screening and archival records, but this could not be definitively confirmed as the material was entirely consumed by the DNA extraction process.

5.3 DNA extraction and sequencing

Roughly 1-2g of tissue from each sponge was ground in ice cold "spin buffer". The sample was then palleted and resuspended in "sponge lysis buffer" (see Methods section for details). After several rounds of phenol-chloroform extraction, the raw nucleic-acid extract was treated with RNase A, then the DNA was purified and concentrated on lab prepared carboxyl-coated SPRI magnetic beads. This process appeared to generate enough DNA suitable for the short-read sequencing required here, as the required yield was not as demanding as that of long-read sequencing technologies. The overall yields and molecular weights for the samples were slightly variable, as judged by gel electrophoresis, but all samples produced a useful amount of intact and pure HMW DNA. One of the samples, MH_pel, was also taken forward for large-insert clone library construction to facilitate the preservation of this precious DNA for perpetuity and to potentially capture the **Pel** BGC on cosmid clones. The extracted and purified DNA from the selected samples was submitted for short-read sequencing. An additional portion of MH_PAT DNA was also included for sequencing in this run to increase the coverage of this sample and to act as a positive control. Illumina TruSeq libraries (2 x 150 PE) with an average insert size of ~500 bp were prepared and sequenced on the HiSeq 4000 platform. Library preparation and data acquisition were carried out by a third-party sequencing provider, Genewiz (Suzhou, China). Although all submitted samples passed the initial quality control conducted by the third-party provider, the sequencing library preparation failed for MH_PEL, S1, S2, S3, S4, S5 on the first attempt. Sequencing library construction of the MH_PAT sample did not fail on this attempt. The failed samples were repurified and resubmitted, which finally resulted in successful sequencing library construction on the second attempt.

5.3.1. Additional sponge sample sequencing results.

Tabulated below are sequencing yield, quality and GC-composition results for the 2 x 150 PE sequencing of samples S1, S2, S3, S5, MH_PEL and MH_PAT (Table 5-1). The sample S4 failed to produce any sequencing data and was dropped from further analysis. We were

particularly interested in increasing the coverage for the previously sequenced sample MH_PAT, to explore the less abundant species in this sample. As such, the sequencing run was loaded with 2.5 times the amount of sequencing library from MH_PAT, compared to the others, to increase the sequencing output of this sample.

Data set	Total reads	Total data (Gbp)	>Q30 (%)	GC (%)
S1	45,286,922	13.58	93.6	40.59
S2	51,250,044	15.37	93.5	40.08
S3	50,541,963	15.16	93.3	45.33
S5	38,253,156	11.48	93.8	41.12
MH_PEL	184,220,767	55.27	92.8	51.96
MH_PAT	73,816,310	22.15	91.9	40.28

Table 5-1 – Summary of sequencing results for 150-bp paired-end sequencing of sponge sample. Data summary including total reads, total data in Gbps, percent of data with quality score over Q30 and the GC percentage of the data. Sample MH_PEL has much higher GC content than the others, as well as a much higher yield.

The total yields and quality of the sequencing data returned appeared to be acceptable, with all samples at >10 Gbp output, and >90% of the data was of high quality (>Q30). It was noted that the yield of MH_PAT was not the expected 2.5-fold increase, although the yield was higher than most of the samples. It was noted that the yield of MH_PEL was ~ 4-5 times higher than what might be expected. A sample mix-up with MH_PAT was later ruled out buy read-mapping to previous assemblies. Additionally, the GC content of MH_PEL was significancy higher that the other samples. The data was independently examined for quality assurance using fastQC,[342] and all samples were seen to be of consistently high quality and fit for assembly with only the standard quality filtering and trimming operations required.

5.4 MH_PEL metagenomic cosmid library attempt.

An attempt at producing a large-insert cosmid metagenomic library from MH_PEL DNA, while the ongoing chemotyping and short-read sequencing was underway, resulted in a cosmid library with an insert efficiency of only ~20%. Thus, the entire remaining sample was consumed for DNA extraction to produce as many clones as possible from the available material. This was deemed to be an appropriate course of action, as the remaining sample was too scarce to meet the requirements for long-read sequencing. We reasoned that a cosmid library could potentially be used to produce the required DNA for long-read sequencing or

otherwise be screened for **Pel** based on any positive results from the complementary short-read sequencing of the sample that was being conducted in parallel.

A library was eventually produced with a practical number of insert containing cosmids (~100,000) after eight round of cosmid packaging. After library propagation to expand available material, the inserts were successfully excised from the library using a CRISPR/Cas-9 system targeting the cosmid sequence flanking the inserts. These inserts were intended to be used as input material for long-read sequencing, however this work was never carried out as alternative approaches were prioritised.

5.5 Metagenome assemblies

5.5.1. Data pre-processing

The sequencing data for all samples was trimmed for quality and removal of sequencing adaptor contamination with BBDuk from the BBtools package (v38.11).[257] The sequence data decontamination process also included the removal reads aligning to PhiX sequencing library preparation control, common laboratory strains and reagent contaminants (e.g., *E. coli*, *Ralstonia pickettii* and *Acinetobacter calcoaceticus*) and human sequences, using a database of low-complexity and ribosomal masked references.[343] Following the quality filtering and decontamination operations, reads were error-corrected and, where possible, overlapping paired-reads were merged into a single continuous read, using BBMerge[344] and auxiliary tools from the BBtools package. The read-merging step was performed on these data sets as recent updates to the SPAdes assembler (v 3.12) now allowed the merged-read type as input as it can potentially improve assembly.[344,345] The previous short-read sequencing data (PE 2 x 250bp) from `MH_PAT` was also processed and combined with the new additional sequence data (PE 2 x 150bp) to give a combined short-read data set `MH_PAT_sr`, with a total size of 33.15 Gb.

5.5.2. Initial MH_PEL assemblies.

Individual metagenomic assemblies were calculated for each newly sequenced sample with the SPAdes assembler (v3.12) invoking the `--meta` flag for metagenomic assemblies. Most likely due to the shorter read length (150 bp) of the underlying data, expectedly, assemblies from these reads were highly fragmented. The additional data collected for `MH_PAT` was destined for hybrid assembly so was not include in this phase of analysis.

As the primary focus at this stage of investigation was examination on the MH_PEL sample, because of its expected Pel$^+$ chemotype, an entire work-up of this assembly was prioritised. This first involved several rounds of binning the SPAdes assembled metagenomic contigs (>1000 bp) using multiple automated binning tools, including CONCOCT,[318] metaBAT(2),[326,346] MaxBin2,[324] MyCC,[347] and the manual Mmgenome[348] process. The aim of this step was to attempt to cluster together any potential fragments of the **Pel** BGC within a single MAG, to allow for the identification and reconstruction of the BGC, as fragments of the BGC in isolation would unlikely be sufficient to allow confident identification. The binning results from these efforts, with respect to genome completeness, were noticeably poor, for example, MaxBin2 produced 38 MAGs of which only 3 were >80% complete as assessed by CheckM. Each of the resulting MAGs from each round of binning were then investigated for the presence of BGC fragments, by analysis with antiSMASH (v4.2.0), that might constitute a plausible **Pel** BGC. All remaining non-binned contigs were also inspected to ensure full coverage of the assembly. This failed to produce any plausible **Pel** candidates and was subsequently repeated with contigs produced by an alternative assembler, IDBA-UD, with the same outcome.

Not only did these combined efforts not present any plausible **Pel** candidate, but it also revealed that the microbiome and BGC assemblage of the sample was highly divergent from the accompanying samples (S1, S2, S3, S5, MH_PAT). Although the MH_PEL assemblies did contain a partial and identifiable fragment of the known **Myc** BGC, many other BGC sequences, identified taxa and large swathes sequence common to the accompanying samples were missing from this assembly. Based on these results, the sample was eliminated from further research in this study, as it appeared unlikely to contribute the goal of discovering the **Pel** BGC and would likely confound any comparative or compounded analysis due to the stark difference in the metagenomic make up. This also meant that the metagenomic cosmid library work for this sample was discontinued to free up resources for other arms of the study.

5.5.3. Co-assembly

In isolation none of the newly sequenced samples (S1, S2, S3, S5 and MH_pel) produced any definitive or suspected **Pel** sequence fragments after extensive scrutiny of the antiSMASH results of these assemblies. There were, however, numerous short unassigned (*trans*-AT) type-I PKS/NRPS fragments detected that could potentially, in concert, comprise

the **Pel** BGC, but were being overlooked due to the lack of context and linkage to other fragments. It was also possible that this target BGC was of such low abundance that it was not being assembled effectively from the data generated for each sample, causing this BGC to be partially or completely missing from the assemblies. To alleviate these potential deficits of individual data set assemblies, i.e., unlinked target BGC fragments and insufficient sequencing coverage for assembly the target BGC, we opted to generate a co-assembly of the sequencing data sets `S1`, `S2`, `S3`, `S5` and `MH_PAT_sr`. These data sets appeared to share an underlying metagenomic core population (discussed in chapter 7), making them suitable for covariant based binning methods, which required a co-assembly, and together would constitute a total sequencing read data set size of ~87.5 Gbp. This would result in a more complex assembly with the addition of all the axillary and incidental species in each sample being included, and a likely increased strain micro-diversity between samples.[349,350] However, it was supposed that the low-abundance core species might be assembled more contiguously due to the increase in sequencing depth of the combined data sets, and covariant binning might provide a higher resolution for low abundance MAGs. If the **Pel** BGC indeed resided in one of these low abundance core species, we hoped a co-assembly might be fruitful in revealing it by leveraging the increased coverage and MAG binning advantages that may be afforded by this method.

The pre-processed short-read data sets (5.5.1) were co-assembled with SPAdes (v3.13.0) by provision of a YAML config file to the assembler. This resulted in an 840.9 Mb assembly of contigs >1000 bp, almost twice as large as the previous largest *M. hentscheli* metagenome (hybrid) assembly in this study. The assembly, `MH_co-assembly`, statistics are summarised below (Table 5-2).

Assembly: `MH_co-assembly` SPAdes (v3.13.0)			
Assembly metrics (> 1000 bp)		Number of contigs (> 1000 bp)	
Total length:	840.9 Mbp	Total :	217332
N50:	4788 bp	>= 10 kbp:	12847
GC (%):	41.88	>= 50 kbp:	755
Largest contig:	4.69 Mbp	>= 100 kbp:	237
Shortest contig:	1000bp	>= 500 kbp:	12

Table 5-2 – Assembly metrics for MH_co-assembly. Assembly metrics and contig counts calculated by Quast (v5.0.2) for contigs > 1000 for the co-assembly of data sets `S1`, `S2`, `S3`, `S5` and `MH_PAT_sr`.

The assembly was larger than any of the previous short-read or hybrid assemblies, and the fragmentation was comparable to the assembly of the 2 x 250 PE only data set of MH_PAT, showing a marginal increase in longer contigs and N50. However, most of the additional data in this assembly was contributed by the shorter contigs (< 50 kbp).

5.5.3.1. MH_co-assembly BGC mining

The first analysis of this new assembly was the annotation of the BGCs. All contigs > 2.5 kb from this co-assembly were submitted to antiSMASH (v 4.2.0) for BGC annotation prior to conducting the computationally expensive multi-sample read-mapping required for the proposed covariant binning analysis. Smaller contigs than those previously used in antiSMASH analysis (2.5 kbp vs. 10kbp) were considered here out of determination to find even small contigs of a potentially highly fragmented **Pel** BGC that may be reconciled with metagenomic binning. AntiSMASH returned a total of 307 BGC annotations on 293 unique contigs. 84 of these annotations were in the target (*trans*-AT)-PKS/NRPS type categories (Table 5-3).

MH_co-assembly target annotation type counts	
t1pks: 23	transatpks: 18,
nrps: 30	transatpks-nrps: 6,
t1pks-otherks: 4	t1pks-nrps: 3

Table 5-3 – AntiSMASH target annotation type counts for MH_co-assembly. Annotation type labels as shown in antiSMASH (4.2.0) output.[351] t1pks: Type I PKS, nrps: NRPS, t1pks-otherks: Type I PKS-other KS, transatpks: *trans*-AT-PKS, transatpks-nrps: *trans*-AT-PKS/NRPS hybrid, t1pks-nrps: Type I PKS/ NRPS hybrid.

Manual retro-biosynthetic analysis was used to examine all the target annotations listed. The familiar **Myc** and **Pat** BGCs were present with the same level of fragmentation as seen in the original MH_PAT 2 x 250 PE assembly. In addition to these, was a *trans*-AT PKS/NRPS labelled BGC that was a plausible candidate for the **Pel** BGC. This BGC spanned almost the entire length of a 55.6 kbp contig and appeared to contain all biosynthetic modules and the terminal thioesterase domain required to produce the peloruside backbone, suggesting a substantially complete BGC. The full analysis is descried in chapter 6.

5.5.4. Pel BGC origin

After the discovery of the **Pel** BGC candidate we were interested to know from which sample(s) it was derived as it had not been previously detected in any individual sample assembly. To ascertain this, we mapped a subset of 20 million reads to the isolated **Pel** BGC contig. This revealed that the only read-sets yielding any notion of consistent coverage across the **Pel** contig were those from MH_PAT, both the 2 x 150 bp and 2 x 250 bp sets, suggesting this was the only sample that the contig was derived from.

These results raised two vexing questions that warranted some consideration. Firstly, why was the apparent **Pel** BGC profile of the assemblies in stark contrast with the chemotyping results? Secondly, why was the **Pel** BGC not detected in previous, presumably better, assemblies of the same sample? Addressing the first point, we postulated one scenario that may explain the counter-intuitive observation. An exceeding low relative abundance, and possible genomic degradation, of the peloruside producing organism in the samples that were chemotyped as peloruside positive, coupled with a lack of detectable peloruside bio-synthesis in the MH_PAT sample. Conducting **Pel** specific PCR experiments on the chemotype positive samples may have supported this position, but we could not be conclusive with the evidence at hand. This remains an open question and further investigations were not prioritised at this time. As for the failure to assemble the **Pel** BGC in previous attempts, this was primarily explained by the improvements to the constantly developing assembly software made between this and the earlier assemblies of the same sample. The specific improvements were made between the 3.11.0 and 3.12.0 version releases of the SPAdes assembler.

As attested to by the change log from the 3.12.0 version release.[345]

- NEW: Support for merged paired-end reads.
- CHANGE: Improvements in metaSPAdes results.
- CHANGE: Overall performance improvements.

With the **Pel** candidate sequence in hand, reanalysis of the MH_PAT data-sets showed that there was no coverage of this sequence within the long-read data set and that the increase in short-read coverage alone was not enough to allow for the assembly of the contig containing **Pel** BGC. Unfortunately, a hybrid assembly using the SPAdes assembler and all available data from the MH_PAT sample was intractable using the computational resources available with all attempts leading to the software crashing.

5.6 Summary

The purpose of this chapter was to describe the supplemental research efforts required to discover the **Pel** BGC and justify the alternative mode of its emergence which otherwise may have appeared cryptic. In contrast to **Myc**, **Pat** and many additional BCGs, this BGC target was not apparent in the initial metagenomic libraries or assemblies we produced. To broaden our search, we enrolled additional samples into the study as well as generating additional sequencing depth for the primary sample. The new samples were sequenced, and we also endeavoured to build a traditional metagenome clone library with one of the samples for perpetuity. The success of this library building, and sequencing efforts were limited towards achieving our end. However, the additional samples did facilitate a comparative microbiome analysis. Ultimately, the reanalysis of sequencing data of the primary sample with updated methods resulted in the discovery of the target BGC. This was necessary due to the low abundance of the producing organism and the resulting low genome sequence coverage making assembly challenging.

6 *M. hentscheli* symbiont secondary metabolite biosynthesis

6.1 Introduction

In this chapter we will analyse and discuss the mechanisms of the bacterial symbiont secondary metabolite biosynthesis by **Myc**, **Pat**, **Pel** and **Gan** (Section 6.5) based on the BGC sequences discovered in the metagenomic assembly of *M. hentscheli.*

De novo assembly of deep metagenomic shotgun sequence data using multiple strategies, including a hybrid approach incorporating only a modest amount of long-read data, produced the ostensibly complete BGCs for three target bacterial secondary metabolites specified at the outset of this study. Our metagenomic assemblies also contained a number of additional BGC which could not be assigned to any of the known secondary metabolites isolated from *M. hentscheli* and currently remain uncharacterised. One of these additional orphan BGCs was of particular interest as it resembled the RiPP BGC of the known sponge metabolite polytheonamide isolated from *T. swinhoei*.[228] This BGC was likely capable of producing a structurally related molecule and thus was amenable to *in-silico* characterisation. The following sections will describe these four BGCs, outlaying the biosynthetic mechanism we proposed for the three target secondary metabolites and propose a predicted final structure of the metabolite encoded by the **Gan** based on the BGC architecture and similarity to characterised related BGCs.

6.2 Mycalamide biosynthetic model

6.2.1. Myc gene cluster description

The mycalamides are structurally related to the pederin[124] family of polyketide secondary metabolites. Previous known members of this family are enzymatically produced by *trans*-AT PKSs found in a wide range of host-symbiotic systems and free-living bacteria. The structural similarities of the pederin-family of molecules are mirrored by a shared sequence homology and PKS domain architecture in the BGCs that produce them.[85] This resemblance is likewise evident when comparing the published **Myc** BGC to the hybrid *trans*-AT PKS–NRPS BGCs of pederin (**Ped**) and onnamide[95] (**Onn**), the pederin-type compounds that are most structurally similar to mycalamide (Figure 6.1). The naming convention for the genes found in **Myc** has been broadly conveyed from that used in **Ped** owing to the very close gene synteny between these two BGC.

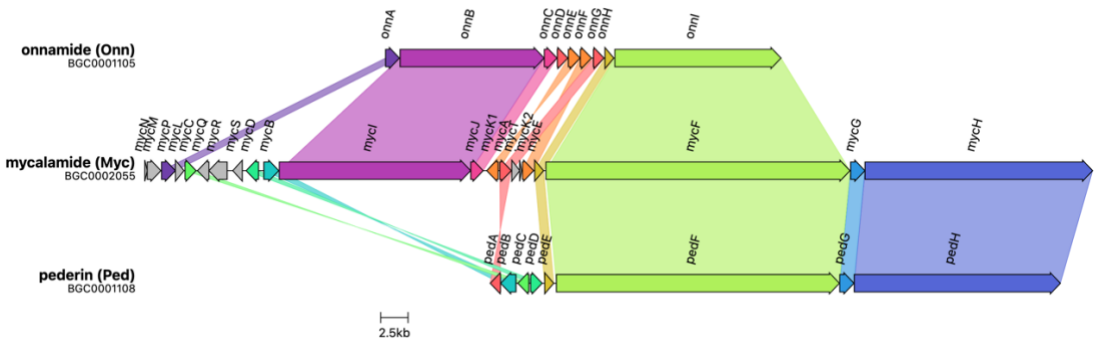


Figure 6.1 – Homologous BGCs from the pederin family of compounds. Comparison of the **Onn** (top), **Myc** (middle) and **Ped** (bottom) gene clusters retrieved from the MiBIG database. MiBIG accession numbers are displayed below the BGC labels. Translated gene corelations above a homology cut-off of 0.3 are shown with matching colours. The architecture of **Myc** closely resembles the hybrid *trans*-AT PKS–NRPS assembly lines of the pederin and onnamide. The corresponding homologue of *onnB*/*mycI* in **Ped** (*pedI*) is absent from the shown gene cluster as it resides at a disconnected genomic region. The **Onn** homologue of *pedH*/*mycH* is absent from the **Onn** gene cluster available in the MiBIG repository. The homology of *mycG* and *pedG* is shared privately between **Myc** and **Ped**.

The sequence of our proposed **Myc** BGC was initially discovered by recognising the close homology of *mycG* with *pedG* from the published **Ped** BGC. On further investigation, the PKS genes of **Myc** had an almost identical module arrangement to those of onnamide (Figure 6.2). The similarity of **Myc** to its counterpart BGCs in the pederin-family facilitates the proposal of a biosynthetic mechanism based on these previously characterised gene clusters. The majority of the proposed **Myc** cluster was found on a single contig which contained all but one of the 12 modules that matched the expected configuration of protein architecture for the biosynthesis of mycalamide. The BGC region on this contig also contained a PKS-NRPS *pedH* homologue (*mycH*) which, like pederin, has no structural representation in

mycalamide. Directly upstream from *mycH* is another positionally conserved **Ped** homologue *mycG*, this gene belongs to a family of PKS-associated monooxygenases, and in **Ped** functions as oxygen-inserting Baeyer–Villigerase.[352] The oxygenated polyketide terminus of both pederin and mycalamide suggest hydrolytic cleavage of the ester-moiety introduced by the action of this monooxygenase, explaining the absence of the region of the molecule produced by the PKS modules in the gene mycH.

The structural backbone of mycalamide is produced by three other *trans*-AT PKS/NRPS genes *mycI1*, *mycI2* and *mycF*. One of these, *mycI2*, was found at a second biosynthetic region and was predicted to encode the initiation module of the megasynthases and the requisite *trans*-AT domain(s) and β-branching enzymes. This second region was located on a contig "binned" in the same *Gammaproteobacterial* designated MAG as the contig of the main biosynthetic region. The *O*-methyltransferase (OMT) and oxidoreductase (OX) genes required for the final tailoring of the mature molecule are situated in a cluster of accessory biosynthetic genes between *mycI1* and *mycF*. All of the gene required for the beta-branching can also be observed within the cluster.

6.2.2. Myc biosynthetic mechanism

The proposed PKS/NRPS biosynthetic assembly line to produce mycalamide is modelled in the figure below (Figure 6.2).

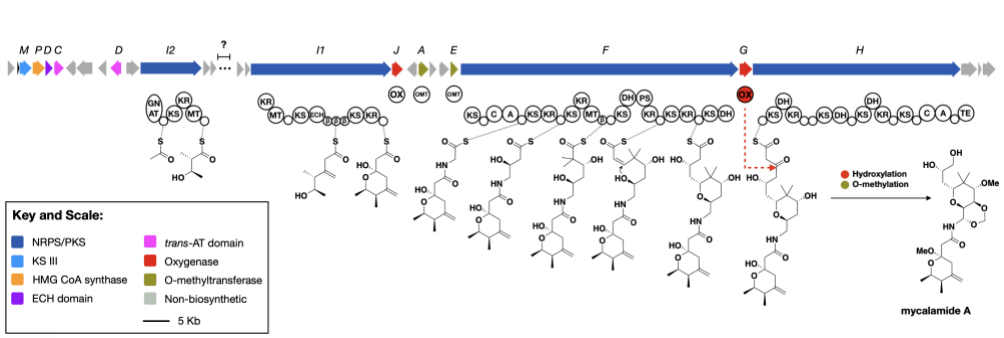


Figure 6.2 – Biosynthetic model for mycalamide. The complete mycalamide biosynthetic gene cluster (**Myc**) above the proposed biosynthetic model for the mycalamides. The red dashed line indicates the putative site of oxidative chain termination catalysed by the pathway-encoded MycG. The gene labelling of this diagram matches the convention used in the main text. Gaps between domains denote the protein boundaries. The ellipsis denotes an unknown chromosomal gap between the two fragments. Biosynthetic intermediates are shown tethered to the ACP/PCP domains (small circles). Domain abbreviations are as follows: KS, ketosynthase; KR, ketoreductase; DH, dehydratase; MT, *C*-methyltransferase; GNAT, GCN5-related *N*-acetyltransferase; OMT, *O*-methyltransferase; OX, oxidoreductase; TE, thioesterase; C, condensation; A, adenylation.

6.3 Pateamine biosynthetic model

6.3.1. Pat gene cluster description

The hybrid assembly of `MH_PAT` afforded the discovery and annotation of a complete BGC for the production of pateamine. The pathway consists of 13 acting *trans*-AT PKS/NRPS biosynthetic modules. The majority of the pathway was spread over two megasynthase biosynthetic loci, on the same 3.1 Mb contig, separated by 40 kb of primary metabolic genes. Analysis of the underlying sequence data of these two regions revealed there were two large (15 kb and 10 kb) partial duplications of the pathway shared between the two megasynthase loci. We also identified a biosynthetic cassette at a distal locus, 320 kb from the megasynthases loci on the same contig, that encoded the hydroxymethyl-glutamate coenzyme A (HMG-CoA) synthase and ECH 1/2 enzymes necessary for the expected β-methyl incorporation, as well as two *trans*-AT domains and polyketide chain initiation machinery.

The genome of the pateamine producing organism was resolved as a complete MAG consisting of one large and one small contig, with a total size of 3.1 Mb, and the producing organism was subsequently assigned the name '*Candidatus* Patea custodiens'. See Chapter 7 for the bases of these naming conventions.

5.1.1. Pat biosynthetic mechanism

Based on the configuration of the domains in the biosynthetic modules found across the two loci and the complement of additional biosynthetic genes a biosynthetic model was formulated. In the model we propose, the cyclic-diester structure of pateamine is produced by first linking two separate polyketide chains (chains A and B) to give a linear ester and then completed by macro-lactonization concatenated chains (Figure 6.4). This model is supported by predictions of the substrate specificity of the participating KS domains based on domain phylogeny.[353]

6.3.1.1. Polyketide chain-A biosynthesis

The first chain of pateamine is initiated with *N*,*N*-dimethylglycine, correspondingly the first module (**1**) contains both a glycine-activating adenylation (A) domain and an *N*-methyltransferase domain to generate this functionality. Modules **2** and **3** are both dehydrating PKS modules appropriate for the two consecutive α,β-olefinic chain extensions. Module **2** also contains a *C*-methyltransferase to incorporate the α-methyl moiety along with its extension. Module **4** is consistent with providing a β-methyl α,β-olefinic extension. This

module has three acyl carrier protein (ACP) domains, all with *in-trans* β-branching enzyme recruitment signatures. The fifth module (**5**) consists of a dehydratase (DH) and a ketoreductase (KR) domain. However, we expect that the DH domain of this module is catalytically inactive and **5** is acting as a β-hydroxy incorporation module. A multiple protein alignment showed an H -> D substitution, eliminating the catalytic histidine residue of the DH domain (Figure 6.3).[125] The predicted substrate specificity of the downstream KS domain also supports this notion.

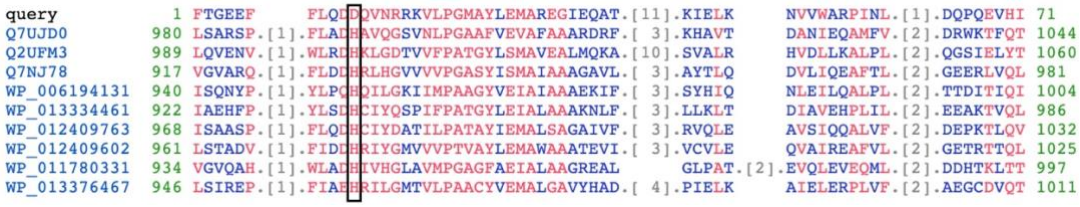


Figure 6.3 – Multiple sequence alignment of Pat Module 5 DH catalytic domain. This is a query centric alignment with selected DH domains from various PKS systems. The query protein sequence is a region of the DH domain from Module **5** of the **Pat** BGC covering the position of the nominal catalytic histidine. Accession numbers are shown left of the sequences. The black box shows the normally conserved catalytic histidine (H) residue is substituted to aspartic acid (D) in the query.

Module **6** contained the diagnostic domains that were invaluable for the discovery of **Pat**, the domain arrangement expected for the incorporation of the observed thiazole moiety, i.e., the heterocyclisation, adenylation, and oxidation domains. Each of these domains are duplicated in the module, which also contains three copies of a peptidyl carrier protein (PCP) domain. The domains comprising this module are split between the terminals of two large biosynthetic genes within the cluster. Module **7** contains a PKS enoylreductase (ER) and three β-branching ACP domains, consistent with a fully reduced extension and a β-methyl branch addition. Module **8** contains a pyridoxal 5′-phosphate (PLP) dependant aminotransferase[354] domain to dictate the inclusion of the β-amino concluding chain-A of pateamine.

6.3.1.1. Polyketide chain-B biosynthesis

The biosynthesis of the shorter chain-B is carried out by modules **10**-**13** of **Pat**. Module **10** appears ketoreductive but is lacking the usual ACP domain of a PKS module. We propose that this module acts to reduce ACP-bound β-ketobutyrate which is used to initiate chain-B. The distal biosynthetic cluster, containing several accessory biosynthetic genes associated with **Pat**, includes a predicted type-III KS homolog, a free-standing ACP domain and a PPTase. We reason these accessory gene work in concert to produce the ACP-bound β-ketobutyrate from malonyl-ACP and acetyl-CoA. Module **11** contains a quartet of β-branching ACP domains, consistent with its expected role as a β-methylating module.

Module **12** is comprised of a reductive and a dehydrating domain, and likely installs a *Z*-configured α,β-olefinic moiety to complete the production of chain-B.

6.3.1.2. Chain transfer, esterification and macrolactonisation

Module **9** was not involved in the co-linear extension of chain-A, instead, this module has non-extending domain arrangement (ACP_1-C-ACP_2) that has been attributed with the linear esterification of two polyketide chains. In the malleilactone/burkholderic-acid biosynthetic pathway, a module with this arrangement tethers a polyketide chain to each ACP for subsequent C-domain catalysed condensation, producing a liner ester.[355–357] In pateamine biosynthesis, we propose that this condensation reaction operates by the transfer of a complete chain-B from module **12** to ACP_2 of module **9** (Figure 6.4, i), the terminal hydroxyl group of chain-B condensates with chain-A from ACP_1 forming the ester in a reaction that is catalysed by the intervening C-domain (Figure 6.4, ii). The linear ester is then passed back down to the final module (**13**) (Figure 6.4, iii) where it is released by thioesterase-catalysed macrolactonisation (Figure 6.4, iv) to give the final macrodiolide.

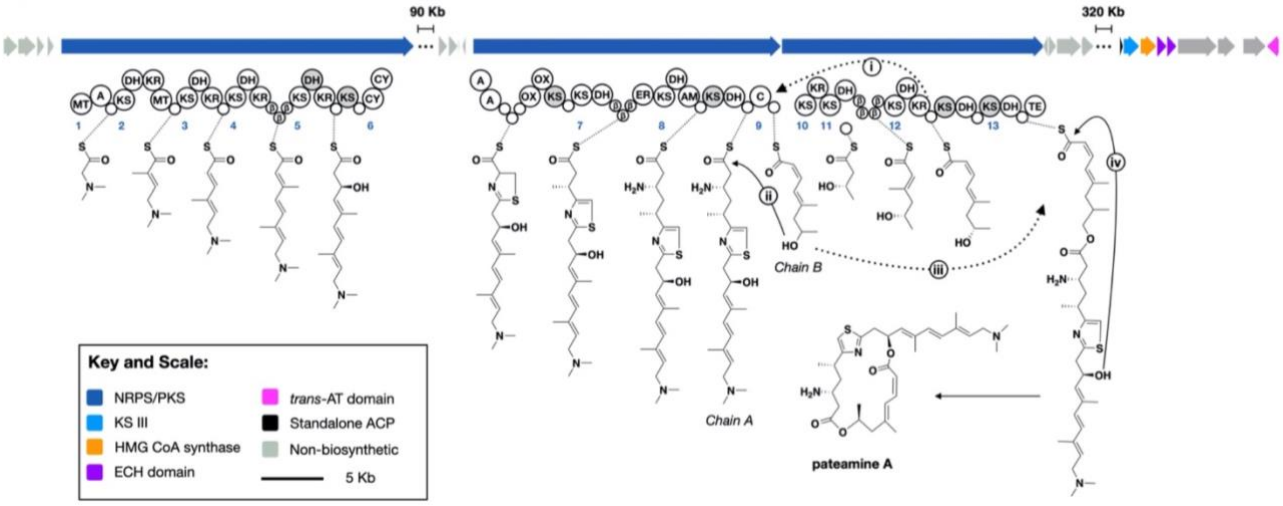


Figure 6.4 – Biosynthetic model for Pateamine. The complete pateamine biosynthetic gene cluster (**Pat**) above the proposed biosynthetic model for pateamine. Dashed arrows indicate chain transfer events; solid arrows indicate esterification events. Modules and events are numbered matching the description in the main text. Ellipses denote measured chromosomal gaps between cluster fragments. Biosynthetic intermediates are shown tethered to the ACP/PCP domains (small circles). Inactive domains are grey. Domain abbreviations are as follows: KS, ketosynthase; KR, ketoreductase; DH, dehydratase; MT, *C*-methyltransferase; TE, thioesterase; C, condensation; A, adenylation.

6.4 Peloruside biosynthetic model

In a revised short-read only assembly of the sample `MH_PAT`, we identified a 55.6-kb contig annotated with biosynthetic features that were an excellent match for producing the structure

of peloruside. This contig was composed entirely of the presumed **Pel** BGC sequence and could not be assigned to any MAG or scaffolded to any other contigs in the assemblies, as such, additional information about the producing organism and some biosynthetic components of **Pel** are currently lacking. However, the BGC sequence on this single contig was complete enough to infer a plausible biosynthetic model for peloruside.

6.4.1. Pel gene cluster description

The recovered peloruside biosynthetic gene cluster contains two large *trans*-AT PKS genes, *pelC* and *pelD*, 23.9 kb and 25.0 kb respectively, which occupy most of the contig. The remaining genes annotated on the contig also belong to the cluster, these are *pelA* a standalone ACP, *pelB* a standalone acyl-CoA ligase, *pelE* a 2-oxoglutarate (2OG)-Fe(II)-dependent oxygenase, and *pelF* an *O*-methyltransferase. Both PKS genes contain methyl-transferase domains at appropriate positions to install the expected α-methyl complexes of peloruside. The first PKS gene in the pathway also contains two auxiliary thioesterase (**TE**) like domains in a non-canonical arrangement within modules **2** and **3** of the pathway, obscuring the functional assignment of these modules. Both PKS genes also appear to carry multiple inactive KS domains (Figure 6.5). There are no *trans*-acting AT domains found in the recovered cluster, and these are likely to be located elsewhere in the genome of the producing organism which is yet to be elucidated.

6.4.2. Pel biosynthetic mechanism

In our proposed model for the synthesis of peloruside, the free-standing acyl-CoA ligase domain encoded in *pelB* activates a 2-methyl-butanoic acid starter unit to the corresponding CoA thioester, which in turn is transferred to the freestanding ACP of *pelA* to initiate the chain. Module **1** of *pelC* contains KR and MT domains and is predicted to extend an α-methyl-β-hydroxyl. Rust *et al* (2020)[250] proposed this starter unit may be the result of the first KS domain acting twice, however this could not be confirmed by biochemical experiments using the purified enzymes. This module lacks the expected DH domain which we propose is performed by a DH domain of the up-stream module **5** which also carries out a β-hydroxy extension. The intervening and highly unusual modules **2**, **3** and **4** were assigned as non-elongating in our model. However, a similar domain arrangement has been observed to introduce *Z*-configured double bonds in the PKSs of spliceostatin[358] and thailanstatin[359] and may be responsible for the expected elimination reaction rather than the DH domain of

5, a model that was first proposed in **Pel** by Rust *et al* (2020)[250]. . Determining which of these two possible routes the pathway follows will likely require biochemical investigations. Module **6** and module **7** together carry out a β-hydroxy, γ-methoxy chain extension. Module **7** is split over the two PKS genes and in addition to the *O*-methyltransferase also contains a *C*-methyltransferase domain which may act iteratively to install the α-*gem*-dimethyl of peloruside. The phylogeny of the **KS** domain in the following module supports this hypothesis (Figure 6.6). Module **8** is expected to perform a non-reductive extension, the resulting carbonyl will persist to act as the electrophile in the spontaneous cyclisation of the incoming downstream extensions to produce the pyranose moiety of peloruside. Module **9** is the compound of a reductive extension module and a non-extending *O*-methyltransferase containing module. This module construct results in the expected β-methoxy extension. Modules **10** and **11** are both reductive modules and incorporate successive β-hydroxy extensions.

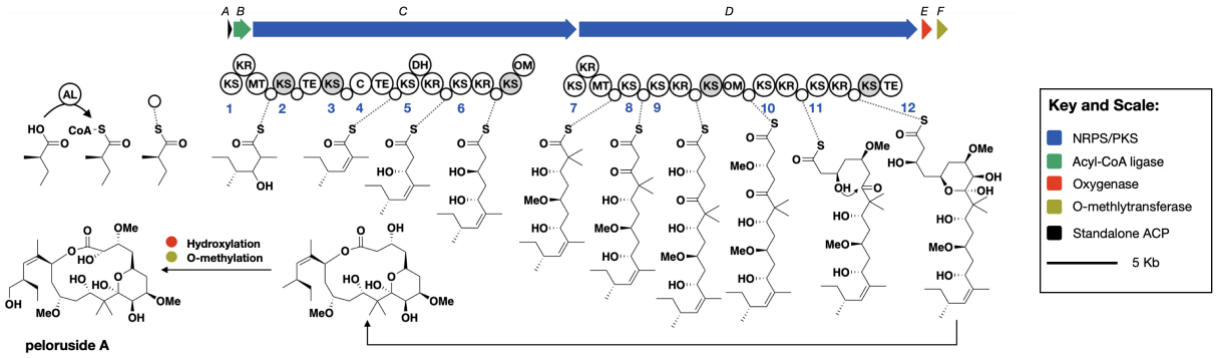


Figure 6.5 - Peloruside biosynthetic model. The peloruside biosynthetic gene cluster (**Pel**) and module domain arrangement above the proposed biosynthetic model for peloruside. Module numbering matches the description of the main test text. Biosynthetic intermediates are shown tethered to the ACP (small circles). Inactive domains are grey. Domain abbreviations are as follows: KS, ketosynthase; KR, ketoreductase; DH, dehydratase; MT, *C*-methyltransferase; OM, *O*-methyltransferase; AL, Acyl-CoA ligase; TE, thioesterase; C, condensation.

The hydroxy of the first extension preforms the nucleophilic attack on the carbonyl of module **8** to close the pyranose ring and the second is *O*-methylated by the standalone *O*-methyltransferase of *pelF*. The chain is then released by intramolecular cyclisation catalysed by the TE domain of the non-extending module **12**. The final molecule undergoes further oxidative tailoring by the 2OG-Fe(II)-dependent oxygenase *pelE*.

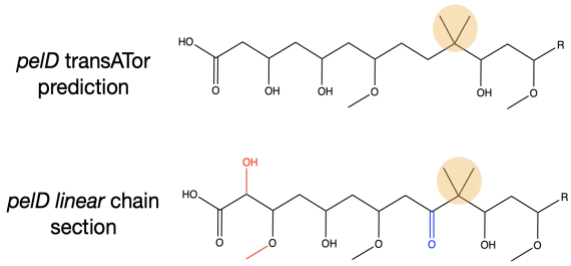


Figure 6.6 - *pelD* transATor structure prediction. Polyketide chain prediction of PKS *pelD* of **Pel** compared to the corresponding linearised portion of peloruside. Highlighted is the predicted and expected dimethyl introduced by module **7**. Blue structure indicates divergence of the transATor prediction. Red structure indicates the expected molecular tailoring post chain-extension.

6.5 A new polytheonamide-like gene cluster

The metagenome assembles of *M. hentscheli* produced in this study contained a number of orphan BGCs that did not appear to match reported secondary metabolites from this sponge. Some of the large and conceivably intact BGCs may warrant further study to guide the discovery of novel metabolites or assign molecules to BGCs.[248] One of these orphan BGC, subsequently assigned abbreviation "**Gan**", was considered for further investigation here. The **Gan** BGC is a 25.0 kb RiPP that appeared to specify a molecule structurally related to polytheonamide.[227] This BGC was found within a 7.0 Mb MAG which contains an additional 9 putative BGCs. The extracted 16S rRNA sequence from this MAG positions the producer strain of this RiPP as a new genus, within the family *Nitrosococcaceae*, which falls under the *Gammaproteobacteria* phylum, the name '*Candidatus* Caria hoplita' has been assigned to this species (Chapter 7). The polytheonamides are complex peptides that involves up to 50 posttranslational modifications[230] during maturation of a 48 residue core-peptide encoded by the precursor-encoding gene *poyA* of the polytheonamide RiPP BGC (**Poy**). These cytotoxic secondary metabolites are produced in the sponge *T. swinhoei* by the super-producer symbiont '*Candidatus* Entotheonella factor'.[222] The **Gan** BGC architecture and gene content is very similar to that of **Poy,** with homologues for five of the six RiPP maturation enzymes of **Poy** appearing to be closely syntenic in the **Gan** BGC (Figure 6.7), in including a PoyD like epimerase which in important for dictating the overall structure of polytheonamide.[227,360,361]

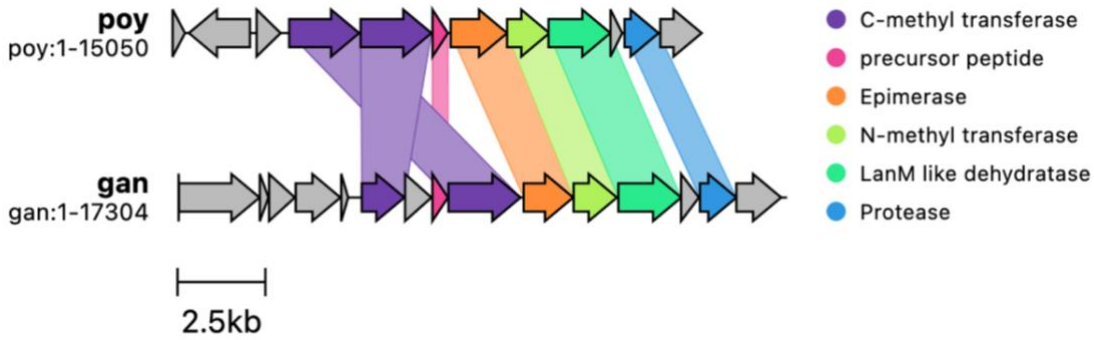


Figure 6.7 – Comparison of the Poy and Gan BGC gene arrangement and homology. The five maturation genes with homology between the two clusters are the *C*-methyltransferases (2), the epimerase, the *N*-methyltransferase and the dehydratase. The architectural similarity of the two clusters is apparent. Also shown is the homology of the precursor-peptide, and the homologous protease, which cleaves the leader sequence from the final molecule. The most 5` gene of the **poy** cluster is a hydroxylase maturation gene for which there is no homologue in **Gan.**

The leader sequence of the precursor peptide encoded by *ganA*, like *poyA*, shares homology with the alpha subunit of nitrile hydratases, and the first 20 amino acids of both core peptides align closely with each other. However, there are also key differences that indicate **Gan** encodes a structurally distinct but functionally related compound. The sequence alignment of the core peptides diverges completely after residue 20 where the peptide sequence encoded by *ganA* proceeds with the conspicuous hexapeptide motif GANANA, repeated three times in succession, from which the BGC bears its name-sake. Accordingly, the expected orphan natural product produced by **Gan** has been assigned gananamide.

The spacing of achiral residues (glycines) in the core and the presence of the PoyD-like epimerase, suggests that, like polytheonamide, the product of **Gan** likely possesses D-configured or achiral residues at every second position. Analysing the sequence of the core peptide and taking into account the content of tailoring enzymes encode in **Gan**, we predicted (with some uncertainty) a possible final structure of the encoded metabolite (Figure 6.8).

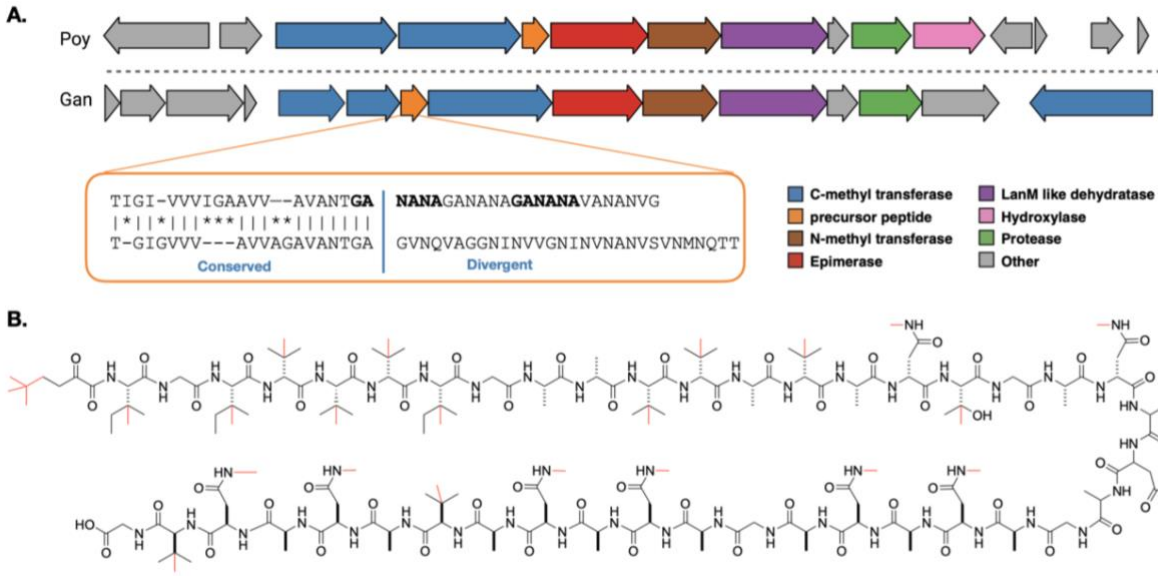


Figure 6.8 The Gan BGC and the predicted structure of the gananamide metabolite. (**A**) Comparison of the **Poy** BGC and the full **Gan** BGC from the genome of '*Candidatus* Caria hoplita'. Genes are coloured by predicted function. The inset orange box indicates regions of conservation and divergence between the predicted precursor peptides found in each pathway. The GANANA motif repeat is highlighted. (**B**) The predicted structure of gananamide, the secondary metabolite produce of **Gan** RiPP BGC. The red structure show all possible methylations in the predicted structure, only a subset of these are likely to occur in the final molecule.

6.6 Summary

The primary focus of this study was to access the genetic elements responsible for the production of three bioactive secondary metabolites mycalamide, pateamine and peloruside from the New Zealand marine sponge *M. hentscheli*. This was ultimately achieved by untargeted direct metagenomic sequencing of eDNA isolated from a sample of the producing sponge and analysis of the resulting metagenomic assemblies. This facilitated the discovery of complete BGCs for both mycalamide and pateamine, and the very nearly complete BGC of peloruside. The untargeted discovery approach also surfaced many other apparently intact orphan BGCs which may be responsible for yet to be characterised natural products.

One of these orphan BGCs appeared to encode for the production of a molecule similar to the RiPP polytheonamide which has been isolated form an entirely geographically and taxonomically distinct sponge. The similarity of these BGCs and the nature of RiPP biosynthesis allowed up to make some reasonable assumptions about the structure of the product of the orphan BGC, a molecule which was subsequently named gananamide.

The similarities to the terrestrial beetle derived molecule pederin aided in the discovery and assignment of a *trans*-AT PKS/NRPS BGC to the production of mycalamide. The similarity of these BGC even extends to the inclusion of a PKS gene in both clusters that does not contribute any structure to the final molecules. This is indicative of a tight evolutionary

relationship between these BGCs despite the markedly different symbiotic environment origins.

As known molecular counterparts for the assignment of BGCs to pateamine and peloruside were not available, the BGCs for these two key secondary metabolites were discovered by a retro-biosynthetic analysis approach. The structure of the polyketide and non-ribosomal peptide backbones of the molecules were analysed to conceive plausible modular pathways for distinctive sections and features of the molecules.[247] Then, the assortment of BGCs recovered from the metagenome assemblies were analysed at the domain level to identify candidate BGCs for the production of the target metabolites based on pairings with the expected retro-biosynthesis. Candidate BGC were further analysed to reconstruct the entire proposed biosynthetic pathways. This analysis revealed that each of the assigned **Pat** and **Pel** pathways displayed non-canonical biosynthetic arraignments that is somewhat typical of these classes of BGC often born form bacterial symbiont genomes. Most notably within the **Pat** pathway was the extension of two distinct polymers which are joined to produce the final macrolide molecule. This pathway is encoded by a BGC that disturbed across a large region of the producer's genome with a curious partially repeating pattern. In our model of the **Pel** pathway, the extension modules act essentially co-linear, however several of the early modules contain premature TE domains and were designated as non-extending. These modules may play a role in specifying the local stereospecific reduction seen in peloruside at this position of the molecule.

Of the BGCs discussed in this chapter only **Pel** could not be placed into a wider genomic context. The stand-alone AT domain was the only crucial element expected to be missing from this gene cluster. Often these domains are often found at distal regions of the genome separate from the body of the BGC. All of the expected extension modules and tailoring enzymes to produce peloruside were present on the removed contig giving a full and complete BGC. Efforts using advances in long-read sequencing to connect the **Pel** BGC to a complete symbiont genome are ongoing.

The full and complete BGCs described here will be crucial for ushering in heterological expression studies of these potential valuable molecules and facilitate biochemical studies to further our understanding of secondary metabolite biosynthesis in microbial symbionts of marine organisms.

7 *M. hentscheli* microbiome binning and taxonomy

7.1 Introduction

In this chapter we discuss the 'binning' of contigs from the metagenome assemblies of *M. hentscheli* produced in this study to resolve high quality metagenome assembled genomes (MAGs) of individual members of the sponge associated microbiome. This was done to complement the discovery of the target BGCs by recovering the entire genomes of the producing organisms and gain insight into the diversity of the sponge microbiome and its functional potential. Access to the complete genomes of the symbionts was essential to detect distal biosynthetic elements of BGCs that were distributed across organisms' genomes. Genomic based primary-metabolic analysis of the secondary-metabolite producing organism is also valuable for guiding *in vitro* culturing conditions as a path to sustainable production of natural products.[362] The high-quality MAGs we recovered were used as references to compare the core and auxiliary microbiomes across several different chemotypes of *M. hentscheli*. A taxonomic analysis of the recovered MAGs necessitated the proposal of several new candidate bacterial genera to accommodate the classification of the producing organisms. This analysis also showed that the biosynthetic potential of *M. hentscheli* is distributed across many taxonomically distinct producing organisms and revealed relationships between symbiosis and chemotype.

7.1.1. Binning of sponge metagenomes

The process of metagenomic binning attempts to cluster together the contigs of a metagenomic assembly that originate from individual microbial genomes. Sponge biomass can be comprised of up to 35% microbial symbionts[292] with a community diversity ranging from just a few operational taxonomic units (OTUs)[363] to several thousand genetically distinct symbionts.[364,365] Binning contigs from complex ecosystems required advanced methods usually taking multiple factors such as sequence composition, coverage and taxonomy into account.[366] An additional challenge with binning sponge derived metagenomes is the unavoidable inclusion of contigs in the assembly from the sponge host genome and incidental species from the surrounding marine environment.[264,367] We attempted to mitigate these challenges by using an ensemble binning approach which

included a binning pipeline that separates host genome sequences prior to contig clustering and then selecting only the highly complete and pure MAGs to proceed with downstream analysis.[264,368,369]

7.1.2. Ensemble binning approach

The binning approach applied to the metagenomic datasets produced in this study implements multiple binning methods for each sample in concert followed by a MAG dereplication and quality assessment step to return a non-redundant set of MAGs from the output of all the methods across all samples. The use of co-assembly of multiple samples was found to produce inferior quality MAGs so this method was excluded from the analysis.[369] The automated binning pipelines used to extract MAGs from the metagenome assembly of each sample assembly were MaxBin 2.0,[324] CONCOCT,[318] MetaBAT 2,[326] and Autometa.[264] The execution of these binning programs are detailed in the Methods chapter. Multiple binning methods were used in order to maximise the recovery of possible MAGs and produce the best quality MAGs by leveraging the strengths of the different binning programs.[366,369] These four programs were selected based on the appropriateness of the sample type, data type and assemblies.

Following the binning of contigs into MAGs for each sample with each binning method, the total outputs were combined. A pairwise comparison of all binned MAGs was then performed to group highly-similar or identical MAGs from the various output streams together. The grouping similarity threshold is set to approximate a species level representation of 99% average nucleotide identity. The genome comparison and grouping tool dRep[369] was used for this task. From these groups of similar MAGs, the MAG returning the highest aggregate score based on continuity, completeness and contamination was returned as the representative MAG for that group. Only the MAGs that met a quality threshold of >85% complete and <15% contamination, as reported by CheckM,[328] were considered as high quality and taken forward for this and further analysis. Although efforts were made to collate as many bins a possible, we reasoned that MAGs below the chosen quality threshold did not warrant reporting and conceded that some biodiversity would be overlooked.

7.2 Binning Results

7.2.1. Raw binning

A single assembly for samples S1, S2, S3 and S5, and both the hybrid and short-read only assembly of sample MH_PAT were used as input for the ensemble binning process. These were a hybrid assembly, MH_PAT_all (MaSuRCA v3.2.8), using all available data sets for this sample, and the short-read only assembly, MH_PAT_sr (SPAdes v 3.13), with the remaining samples using the SPAdes (v 3.12) assemblies reported on in previous chapters. A total of 1321 redundant MAG bins were produced from all of these assemblies after running the four independent binning methods with the assemblies of each sample. Only the two binning methods, Autometa and MetaBAT 2, were used with the MH_PAT_sr assembly. The result of each binning and assembly combination is tabulated below (Table 7-1).

	MH_Pat all	**MH_Pat sr**	**s1**	**s2**	**s3**	**s5**	Total
Autometa	22	22	13	15	20	4	96
MetaBAT 2	65	37	66	88	94	56	406
CONCOCT	95	-	154	153	142	126	670
MaxBin 2.0	25	-	38	36	38	12	149
Total	207	59	271	292	294	198	**1321**

Table 7-1 – Number of raw bins for each sample. Contigs from the assembly of each sample were binned with four different automated binning methods (left column). The number of bins for each method per sample is shown in the main table. The total number of bins for each method and each sample are in the totals column and row.

7.2.2. MAG quality filtering

Although Autometa consistently produced the lowest number of bins per sample, these sets of bins had the highest average quality scores (CheckM: completeness and contamination). The inverse was true for CONCOCT which produced many bins with a reported 0% completeness. The variation in the number and quality of bins from each method highlights the differences in sensitivity and stringency between methods. With a completeness and contamination threshold set at >85% and <15% respectively, only 165 (~12.5%) of the total MAGs produced by these rounds of binning would proceed to dereplication. Although bins falling below this threshold may represent true partial MAGs, we did not want to speculate on potential errors regarding the complement of biosynthetic pathways of an organism.

7.2.3. MAG dereplication

The 165 set of MAGs passing the quality threshold were dereplicated with dRep (v2.2.3), this resulted in 26 high quality non-redundant MAGs from across all samples. Each assembly contributed at least one MAG to this final set, and all binning methods expect MaxBin 2.0 were represented, demonstrating the utility of employing multiple methods in an ensemble approach. The assembly and binning method source of each MAG in the dereplicated set is tabulated below Table 7-2.

	MH_Pat all	MH_Pat sr	s1	s2	s3	s5	Total
Autometa	9	4	0	1	2	1	17
MetaBAT 2	2	2	1	1	2	0	8
CONCOCT	0	0	0	1	0	0	1
Total	11	6	1	3	4	1	**26**

Table 7-2 - Origins of final MAG bins. The assembly and binning method origin of each of the final bins in the dereplicated set of MAGs. The total counts for each assembly and binning method are also shown.

The binning method, Autometa, that consistently supplied the fewest bins per assembly for the total input set has contributed them most bins to the final set based on these bins scoring the highest quality metrics. Disregarding the exclusion of MaxBin 2.0, again the inverse is true for CONCOT which has only contributed a single bin to the final set.

7.3 Recovered MAGs

Each of the 26 recovered MAGs were analysed to determine assembly metrics, taxonomy and assignment of target BGCs to the producer genomes. The levels of MAG completeness were determined using CheckM as this is an easily comparable standard that has been adopted by the field.[302]

7.3.1. MAG assembly statistics

The assembly quality and statistics for each MAG is tabulated below. The MAGs provisional names were constructed from the assembly and binning method of origin and appended with a consecutive number for each replicate combination. The MAGs containing the BGC significant to this study were subsequently assigned candidate binomial names to distinguish their importance. The most complete genome recovered in this set was reported as 98.9% complete while the lowest was 88.2%, the average completeness of all 26 recovered MAGs

was 94.4%. The fragmentation of the recovered MAGs ranged from a single contig to 903 contigs for the most fragmented MAG.

Bin name	Completeness	Contamination	MAG size (Mb)	Contigs	N50 (kb)	GC %
s2_metabat2_1	98.91	0.54	2.47	67	67.4	55.15
MH-Pat-sr_metabat2_1	98.91	0	4.11	25	225.5	51.98
s3_autometa_2	98.67	0.6	4.49	214	32.7	52.65
MH-Pat-all_metabat2_1	98.54	0.97	1.92	1	1922.1	36.87
MH-Pat-all_autometa_6	98.48	14.63	5.79	21	493.9	46.41
MH-Pat-all_autometa_7	98.12	0.45	3.60	1	3603.2	55.35
MH-Pat-all_autometa_3	97.93	0	4.02	2	4011.1	61.2
MH-Pat-all_autometa_1	96.92	1.23	4.31	1	4308.0	34
MH-Pat-sr_autometa_2	96.7	12.69	9.36	188	71.4	59.44
MH-Pat-all_autometa_8	96.11	0.68	3.14	2	3096.9	44.59
MH-Pat-all_autometa_5	95.55	1.82	2.76	3	2124.4	48.29
s3_metabat2_2	95.01	1.39	3.69	203	26.8	54.01
MH-Pat-all_autometa_4	94.89	1.95	2.65	1	2650.0	39.55
MH-Pat-sr_autometa_1	94.26	0.74	4.53	4	1278.2	45.78
MH-Pat-all_autometa_9	94.13	1.3	6.33	24	527.2	59.32
MH-Pat-all_metabat2_2	94.12	0.84	1.16	1	1162.5	49.89
s2_autometa_1	93.96	1.71	4.81	9	1042.5	64.26
MH-Pat-sr_autometa_3	93.89	5.11	5.82	21	4647.9	50.77
s5_autometa_1	93.38	6.01	6.86	291	39.4	57.61
MH-Pat-sr_autometa_4	91.53	3.73	4.19	17	632.7	46.06
MH-Pat-sr_metabat2_2	90.7	1.22	3.36	339	13.2	45.9
MH-Pat-all_autometa_2	89.57	0.43	3.08	25	152.8	60.6
s1_metabat2_1	88.98	4.23	2.23	296	14.0	62.43
s2_concoct_1	88.6	5.14	2.98	907	4.4	61.7
s3_autometa_1	88.53	3.08	1.65	243	8.2	43.6
s3_metabat2_1	88.17	2.07	1.66	96	32.7	62.65

Table 7-3 – MAG assembly and quality statistics. Completeness and contamination percentage as reported by CheckM. MAGs are ordered by level of completeness.

7.3.2. MAG Taxonomic classifications

Each of the MAGs were examined with two complementary methods to determine the phylogenetic placement. The initial method compared the 16S rRNA genes extracted from each MAG to the SILVA[370] database (release 132) and inferred the taxonomy using the lowest common ancestor (LCA) algorithm. This database contains over six million rRNA gene sequences covering a very wide distribution of taxa including sequences from uncultivated environmental samples. The second method of taxonomic placement employed the Genome taxonomy database toolkit (GTDB-Tk v 0.3.2)[371] to generate whole-genome-based taxonomic classifications. This method can provide a more robust classification, as it relies on 120 bacterial or 122 archaeal marker genes and domain specific reference trees to infer the taxonomic placement. However, this is a database (GTDB R04-RS89) of fully sequenced reference genomes and only represents 23,458 bacterial species, this excludes the inclusion of many environmentally derived species. As the MAGs that we had recovered were likely from uncultivated species or genera, a precise taxonomic placement was not possible. We were able to affiliate a minimum of phylum level taxonomies for all MAGs with nine MAGs placed at the genus level. The results of the taxonomic profiling, based on 16S rRNA gene identity to known taxa, suggests that many of the MAGs were from entirely uncategorised genera (Figure 7.1). In two cases, a 16S rRNA gene could not be detected in the recovered MAG. This may be due to the target sequence not being correctly binned into the MAG, as they often fail to assemble,[297] or failure to detect a highly divergent 16S rRNA sequence.[372] The taxonomic agreement between the two methods was highly comparable and most differences were attributed to the non-standardisation in the naming of the underlying taxonomies. An incentive to taxonomically classify these MAGs was to detect any presence of the previously reported sponge symbiont "super producer" phylum 'Tectomicrobia',[227] which includes the polytheonamide producer genus 'Entotheonella'. None of the MAGs nor any of the contigs from the assemblies were placed near this phylum.

7.4 MAG abundance profiles and comparison of samples.

With the high-quality MAGs recovered and taxonomically classified, we were interested in the abundance profile of these MAGs across the samples to ascertain which were members of the core microbiome (shared between all samples) and which were axillary in the pan-metagenome (not shared between all samples). Many sponge species have microbiomes conserved between specimens even when specimens were geographically distributed,[363] we

wanted to understand if *M. hentscheli* also displayed this pattern and if the producing organisms appeared to be members of the core microbiome. We reasoned that differences in the microbiome profiles may be responsible for the variations in chemotypes observed in *M. hentscheli*.

To assess the abundance profile of the MAGs in each sample, each short-read set was independently mapped to each sample's assembly in an all-versus-all manner. The read mapping information for the contigs that comprised each MAG were then used to calculate the average depth of coverage for the MAGs. Only MAGs with a breadth of coverage threshold[373] of 1.5-fold coverage over >80% of the sequence length from a sample were considered as detected in a sample. This allowed for some spurious read-mapping within the complex reference while avoiding false positive detection. The average coverage was then normalised by the total size of the read, set to help with comparisons between samples. A quantitative comparison between samples is somewhat problematic as the ratio of host to microbiome sequencing reads cannot be controlled for between the samples, however this method was sufficient for intra-sample comparison. The calculated relative abundance of each detected MAG is shown by the clustered heatmap of Figure 7.1 alongside the 16S rRNA taxonomic rankings.

MAG name	16S %	Kingdom	Phylum	Class	Order	Family	Genus	BGC
MH-Pat-all_autometa_3	94.7	Bacteria	Proteobacteria	Alphaproteobacteria	Rhodobacterales	Rhodobacteraceae	-	
MH-Pat-all_autometa_9	95.0	Bacteria	Proteobacteria	Alphaproteobacteria	Rhodospirillales	Magnetospiraceae	Magnetospira	
MH-Pat-all_autometa_2	97.3	Bacteria	Proteobacteria	Alphaproteobacteria	-	-	-	
MH-Pat-all_autometa_7	96.0	Bacteria	Proteobacteria	Alphaproteobacteria	Rhodobacterales	Rhodobacteraceae	-	
MH-Pat-sr_metabat2_1	91.9	Bacteria	Proteobacteria	Alphaproteobacteria	Rhodospirillales	Magnetospiraceae	-	
s3_autometa_2	98.3	Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobiales	Rhizobiaceae	Pseudahrensia	
s3_metabat2_2	95.8	Bacteria	Proteobacteria	Alphaproteobacteria	Rhodobacterales	Rhodobacteraceae	-	
MH-Pat-all_autometa_4	91.4	Bacteria	Proteobacteria	Gammaproteobacteria	-	-	-	
MH-Pat-all_autometa_6	94.1	Bacteria	Proteobacteria	Gammaproteobacteria	UBA10353	-	-	Myc
MH-Pat-sr_autometa_1	95.9	Bacteria	Proteobacteria	Gammaproteobacteria	Oceanospirillales	Pseudohongiellaceae	Pseudohongiella	
MH-Pat-sr_autometa_3	94.1	Bacteria	Proteobacteria	Gammaproteobacteria	KI89A clade	-	-	
MH-Pat-sr_autometa_4	95.4	Bacteria	Proteobacteria	Gammaproteobacteria	-	-	-	
s3_autometa_1	85.2	Bacteria	Proteobacteria	Gammaproteobacteria	Legionellales	Legionellaceae	Legionella	
s5_autometa_1	94.6	Bacteria	Proteobacteria	Gammaproteobacteria	Nitrosococcales	Nitrosococcaceae	-	Gan
s2_autometa_1	89.8	Bacteria	Acidobacteria	Subgroup 26	-	-	-	
s2_concoct_1	95.2	Bacteria	Actinobacteria	Acidimicrobiia	Microtrichales	Ilumatobacteraceae	Ilumatobacter	
s3_metabat2_1	99.5	Bacteria	Actinobacteria	Actinobacteria	PeM15	-	-	
MH-Pat-all_autometa_1	93.6	Bacteria	Bacteroidetes	Bacteroidia	Flavobacteriales	Flavobacteriaceae	Polaribacter 4	
s2_metabat2_1	99.6	Bacteria	Cyanobacteria	Oxyphotobacteria	Synechococcales	Cyanobiaceae	Synechococcus	
MH-Pat-all_metabat2_2	85.6	Bacteria	Dadabacteria	Dadabacteriia	Dadabacteriales	-	-	
MH-Pat-all_autometa_8	90.4	Bacteria	Kiritimatiellaeota	Kiritimatiellae	WCHB1-41	-	-	Pat
MH-Pat-all_autometa_5	92.1	Bacteria	Nitrospirae	Nitrospira	Nitrospirales	Nitrospiraceae	Nitrospira	
MH-Pat-sr_autometa_2	89.1	Bacteria	PAUC34f	-	-	-	-	
MH-Pat-all_metabat2_1	98.9	Archaea	Thaumarchaeota	Nitrososphaeria	Nitrosopumilales	Nitrosopumilaceae	Nitrosopumilus	
MH-Pat-sr_metabat2_2	-	-	-	-	-	-	-	
s1_metabat2_1	-	-	-	-	-	-	-	

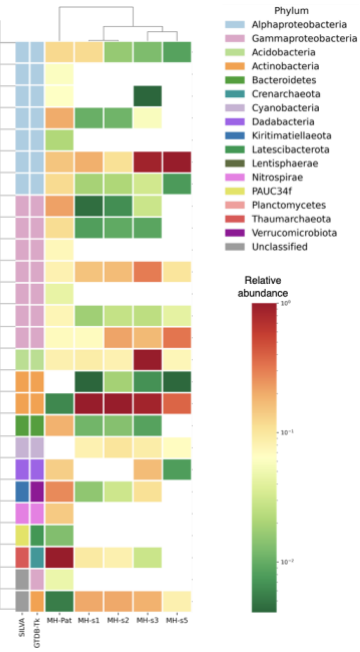


Figure 7.1 – MAG abundance and taxonomic ranking. Taxonomic placement of the 16S rRNA genes extracted from the MAGs is tabulated. Two MAGS did not contain a detectable 16S rRNA gene. The percent identity of the 16S rRNA genes to the closest sequence in the SILVA database is shown in the "**16S%"** column. The presence of target BGCs for **Myc**, **Gan** and **Pat** are shown in the **BGC** column. The phylum level taxonomy for both the SILVA and GTDB methods are color-coded to the right of the abundance heatmap and the columns of the heatmap are labelled by sample. Proteobacteria are split by class (Alpha and Gamma). The hierarchical clustering of samples is shown above the heatmap. Relative abundance is shown as a log normalised scale. Empty (white) tiles in the heatmap depict the absence of a species in the sample at the detection threshold.

7.4.1. Core microbiome

Nine of the 26 MAGs were detected as present in all (five) sponges sampled, with an additional eight being detected in at least four of the samples. This may be an underestimation of core species distribution, as the sponge sample (MH-s5) that was deficient in the most members of the core community and had the lowest overall number of detected MAGs, also had the lowest sequencing coverage of the samples. All but two of the MAGs were detected in the MH-Pat sample which had the highest coverage of the samples. This suggests the diversity of a sample's metagenome when measured in this way is a function of sequencing depth and the core metagenome may be more closely conserved across the species than the data displayed by abundance heatmap shows. In previous amplicon based studies of this sponge, a significant portion (>35%) of the microbial community was common across a range of geographic locations for a larger cohort of sponges.[245]

The MAG encoding the **Pat** BGC was present with various abundance in four of the five samples and missing only from MH-s5, the same pattern was also seen for the MAG encoding the **Myc** BGC. The MAG encoding the **Gan** BGC was detected in all sponge samples at relatively high abundance. The contig encoding the **Pel** BGCs was not included in any of the MAGs constructed in this study with the applied filtering scheme.

7.4.1.1. Novel taxonomic groups

Many of the recovered MAGs appeared to be distant to any described taxa making classification to the species or even genus level difficult. The full length 16S rRNA identities to the closest known neighbour in the SILVA database ranged from 85.6% to 99.6% (Figure 7.1). Although no single value of 16S rRNA identity has been defined to objectively delineate genus or species, in the absence of supporting data such as morphology or physiology, the canonically used species clustering threshold is >97% with recent publications suggesting an updated value of >99% for full length 16S rRNA sequences.[374] Following this convention, at least 20 of the recovered MAGs would represent a newly discovered species. A conservative estimate for a genus identity cut-off has been cited at <95%.[375] This would place each of the MAGs encoding the **Myc**, **Pat** and **Gan** BGCs in new candidate genera. To facilitate further collaborative discussion and analysis, specific nomenclature was proposed for these three MAGs.

7.4.2. Biosynthetic MAGs proposed nomenclature

7.4.2.1. "*Candidatus* Entomycale ignis"

The lowest level taxonomic ranking of the 16S rRNA gene sequence isolated from the MAG encoding the **Myc** BGC was placement within the order of the 'UBA10353 marine group'. This was corroborated by the GTBD genome-based phylogeny. This taxon was represented in the SILVA database by uncultivated clones and a metagenomic sequence and although this taxon type has been detected in sponges[376] it was not a known producer of the pederin family of secondary metabolites. For the mycalamide producing biosynthetic organism represented by this 98.5% complete 21 contig MAG, we proposed the name '*Candidatus* Entomycale ignis'. The species name 'ignis' (Latin: fire) refers to the intense skin blistering that arises from exposure to the mycalamides. The full MAG sequence of this organism is available under the accession number: GCA_012103415.1

7.4.2.2. "*Candidatus* Patea custodiens"

The 16S rRNA gene of the pateamine producer places the organism in the phylum *Kiritimatiellaeota* with 90.4% sequence identity to the closest relative. This phylum is a recent branching of *Verrucomicrobia* and falls under the PVC (*Planctomycetes, Verrucomicrobia, Chlamydiae*) superphylum.[377] The affiliation to *Kiritimatiellaeota* was confirmed by GTDB whole genome phylogeny which identified the sole cultivated member of *Kiritimatiellaeota* as the closest neighbour. As only one cultivated member of *Kiritimatiellaeota* phylum was known, the addition of this MAG represented the second representative full genome of this taxa.[378] The 3.14 Mb MAG was measured to be 96.1% complete and constructed by only two contigs.

For this organism, we proposed the name '*Candidatus* Patea custodiens'. The genus name Patea reflecting the historical indigenous population of the area in which the original pateamine- producing sponge was collected and the species name custodiens (from the Latin word custos for guardian) referring to the protective role of the metabolites produced by the organism. The full MAG sequence of this organism is available under the accession number: GCA_012103575.1

7.4.2.3. "*Candidatus* Caria hoplita"

'*Candidatus* Caria hoplita' is the proposed name for the gananamide producer. The 16S rRNA gene sequence recovered from the MAG encoding the **Gan** BGC was classified at the lowest level to the family *Nitrosococcaceae* with the closest sequence identity of 94.6%. The GTDB whole genome phylogeny was consistent at a higher taxonomic level with placement

of the MAG in the class Gammaproteobacteria. Inspection of the GTDB reference tree shows the closest neighbouring genome as a *Nitrospira*-like organism derived from a metagenome. This MAG was 93.4% complete and comprised of 291 contigs. The most complete version of this MAG was recovered from a short-read only assembly resulting in the higher level of fragmentation. A less fragmented version of the MAG was also recovered form he hybrid assembly with very comparable MAG quality metrics. The naming of this organism is based on the "hoplites", the ancient Greek soldiers that helped defend the region of Caria during the battle of Mycale in 479 BC. The full MAG sequence of this organism is available under the accession number: GCA_012103455.1

7.4.3. Phylum-level microbiome composition

The recovered MAGs comprising the *M. hentscheli* pan-microbiome present in each sponge sample likely only represents some fraction of the total microbiome of a sample due to the stringent conditions we used to quality filter the bins. To supplement the comparison of the MAG abundance profiles and attempt to capture more of the true sponge diversity we compared the abundance of all 16S rRNA genes detected across all samples at the phylum level. To achieve this we calculated the abundance of each detected non-redundant (clustered at >97% identity) 16S rRNA gene in each sample by read mapping (Figure 7.2). The identity of the 16S rRNA genes were again determined by LCA analysis within the SILVA database. Perhaps surprisingly, the number of 16S rRNA genes recovered from any single sample in this manner was not more that the total number of MAGs recovered, supportive of the relatively modest metagenome diversity suggest by the number of recovered MAGs. An in-depth amplicon sequencing approach may confirm this but was not a part of this studies objetive.[379]

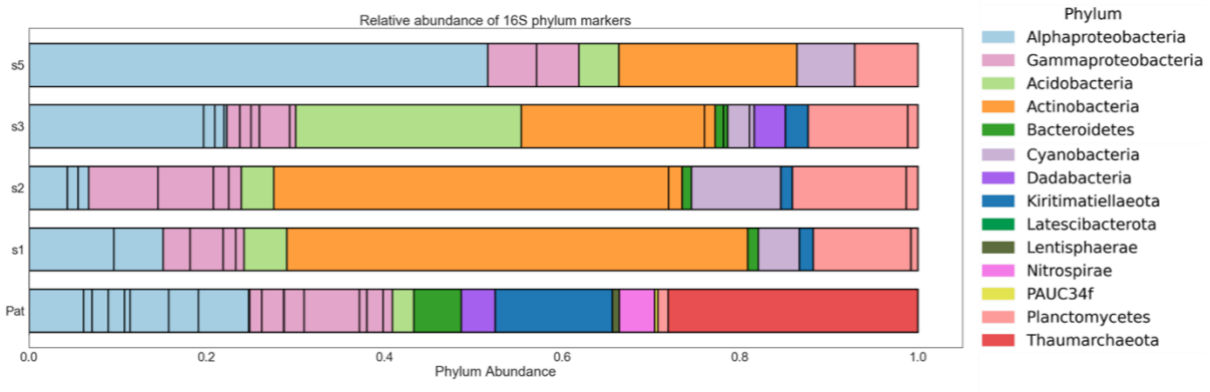


Figure 7.2 – Phylum-level microbiome composition for the five *M. hentscheli* specimens. Taxonomy deduced by extracting 16S sequences directly from metagenome assemblies. Abundance values were derived from coverage of the corresponding contig in the assembly. Black bars within the same coloured block denote multiple species within the same phylum. Proteobacteria are split by class (Alpha and Gamma).

The comparison highlights the broad similarity between the samples at this taxonomic level and reiterates the general pattern shown by the MAG abundance analysis. Of note is the absence of the *Cyanobacterial* and *Actinobacterial* members in the MH-Pat samples. Otherwise many of the singularly occurring phylum are shared among most or all samples. This analysis again depicts a well conserved microbiome of modest diversity for the *M. hentscheli* samples used in this study.

7.5 Summary

The process of binning contigs form metagenomic assemblies into MAGs is commonplace in metagenomic studies and especially useful for resolving fragmented assemblies.[297] We applied this process with the intention of convening BGCs that were potentially distributed across distinct contigs. In the case of the **Pat** BGC this was not necessary as the assembly of the producing organism was very contiguous. The discovery of the full **Myc** BGC did require binning to piece together distal elements that were split over contigs. Further assembly of the **Pel** BGC and its genetic context was one of the primary drivers for executing the ensemble binning process, however it was not possible for us to assign the **Pel** BGC to any plausible MAG. The ensemble binning process we employed was purposefully conservative to avoid the inaccurate assignment of the **Pel** BGC to an unreliable MAG and resulted in a collection high quality MAGs that may be useful for studying the ecology of secondary-metabolite rich sponges. This also facilitate the examination of many of the additional BGCs

present in these specimens that were not covered in this study and provides a resource to the wider community for exploration of novel biosynthetic chemistry.

With the set of high quality MAGs and metagenomic sequence data for a variety of sponge samples in hand, we performed a cursory investigation into the microbiomes diversity and distribution. We found many members of the microbiome, including those responsible for producing the target secondary metabolites, to be genetically distant from any known organism. The microbiome appeared to be partially conserved across the samples. Of note was the presence of producing organism in sponge samples that did not reflect the matching chemotype. For example, the sponge samples `S1`, `S2` and `S3` all harboured the pateamine producing orgasm and the requisite BGC however we did not detect this metabolite in these samples. This is consistent with the cryptic nature of the spatial and temporal variability of the toxic secondary metabolites discovered in this sponge.[84,245]

8 Summary, conclusions and future directions

8.1 Research motivation

The existence of highly potent bioactive compounds in extracts of the marine sponge *Mycale hentscheli* has been recognized for decades.[283] The complex structures and therapeutic potential of these natural products continues to elicit research interest. However, important research programmes to unlock the full therapeutic potential have been hampered by a lack of access to these compounds form their natural supply, and extremely difficult synthesis. One '*missing piece of the puzzle*' from our understanding has been the biosynthetic origin and mechanism of these valuable natural products. Gaining insights into the genetic determinants underlying the biosynthesis are crucial to a synthetic biology pathway to producing these compounds while also contributing to our broader understanding of complex secondary metabolism.

Based on structural elements of three specific compounds, pateamine peloruside and mycalamide, we hypothesised that their natural origins were a microbial symbiont of *M. hentscheli* and they would be encoded by type-I polyketide synthases (T1PKS) biosynthetic gene clusters (BGC) in the genome of this symbiont.

The aim of this research was to discover and attempt to resolve these hypothetical biosynthetic gene clusters. We proposed to apply a metagenomic screening technique that had proven successful in recovering BGCs from symbiont and terrestrial bacteria for fermentation based production. This technique would establish a library of large metagenomic fragments that could be explored to identify and isolate fragments with sequences pertaining to the target BGCs. If required, multiple overlapping fragments could be stitched together to build up full BGCs. The sequences of recovered BGCs could then be analysed to resolve the biosynthetic mechanisms resulting in the production of the natural products, and build a foundation for fermentation based production.

This study would involve two main research fronts, one based on molecular biology to process and manipulate the samples metagenomic DNA, and the other employing

bioinformatics to analyse the required sequence data to define and annotate relevant biosynthetic genes.

8.2 Advancing field

Sponges and their associated microorganisms remain to be a target of enquiry due to the continued discovery of chemical and biological diversity and the biosynthetic/biotechnology potential which can be accessed by metagenomic methods.[380–382] During the course of this research, advances have been made in the field of marine/sponge metagenomics and drug discovery with respect to the discoveries made, and to the methods and resources being applied.[383]

For example, the source the macrolide palmerolide A was discovered recently. This potential chemotherapeutic agent targeting melanoma has structural resemblance to pateamine A, and like pateamine in this study, this molecule was originally isolated from the host of the producing microorganism, in this case an Antarctic ascidian.[384] Motivated to understand the biosynthetic origin, the researchers (Murry, et al.) hypothesized, based on the molecular structure, that the molecule was produced by a host associated microorganism and proposed a metagenomic approach to identify the biosynthetic gene cluster. There were many other similarities in the biology discovered and methods used by Murry et al. to those presented in this thesis. The palmerolide A BGC was a *trans*-AT PKS-NRPS type with multiple copies encoded in the bacterial chromosome, comparable to the **Pat** BGC. The palmerolide A producing bacteria discovered was taxonomically classified as *Verrucomicrobia*, in the same phylum as the pateamine producer. The research was based on a hybrid-assembly and contig binning approach, not dissimilar to the methods presented in this thesis. These findings exemplify the validity of the methods and motivations adopted in this thesis.

The putative biosynthetic pathways of other drug-like molecules have been discovered in recent sponge metagenomic studies, such as lasonolide A[385], using similar approaches as ours.

As this field relies on the intersection and cooperation of several distinct disciplines, such as, at least, natural-product isolation and chemical analysis, molecular- and micro-biology, second- and third-generation sequencing technologies, and bioinformatics and computer science,[386] it is difficult to pinpoint specific cases of advancements that capture all of these aspects.

One of the powerful characteristics of the current approach to natural-product discovery is the feedback of data from new discoveries into the databases that are used to train and validate the computational and network analysis based methods used to assign function to sequence data.[387] In this way, new discoveries are able to be re-encoded into the various models and databases used by these tools to improve their sensitivity and accuracy.[386] This has facilitated the design and building of machine learning and artificial intelligence tools for efficiently performing tasks in the natural-product discovery pipeline, such as structural elucidation and dereplication.[388,389] The discovery of BGCs of known molecules can be systematically incorporated into the growing body of knowledge used to make such discoveries.

General advances in other supporting technologies also play a role in the advance of natural-product discovery and research. The democratisation of cloud computing and the use of containerisation for managing scientific workflows will ease the requirements of both experienced and aspiring labs to contribute to the field.[390,391] Access to these technologies was crucial to the founding members of the "Owen" lab-group to participate in the field at a level historically only obtained by established academics and private organisations.

Another pivotal aspect on the advance of the field is the ongoing improvements to sequencing technology especially regarding single molecule long read sequencing. One particularly exciting development that is beginning to impact metagenomic sequencing studies is the introduction of adaptive sampling on the ONT platform.[392] Adaptive sampling allows for the programmed exclusion or enrichment of target DNA sequences by the sequencer at the time of sequencing. This is achieved by observing the sequence of the first few hundred bases as they are proceeding through the pore and rejecting any off target molecules from further sequencing by reversing the polarity of the current across the sequencing pore.[393] In the context of metagenomics, this can be used to decrease the sequencing data that is derived from the host or other known unwanted organisms and increase the coverage of low abundances species without any special sample preparation or increase in overall sequencing depth and cost. [394,395]

The heterologous expression of BGCs in lab amenable strain for production of useful products at a large scale is one of the ultimate goals and promises of metagenomic studies and the field at large. This aspect of the field continues to advance with increasingly complex molecules being successfully produced in fermentation reactors by refactoring BGCs to be active in heterologous hosts. One notable recent example saw the expression of a large complex BGC occur across taxonomic kingdoms, the production of the anti-cancer drug

vinblastine from the plant *Catharanthus roseus* in a genetically reprogramed yeast strain.[396] The production of 1g of vinblastine requires 500 kg of dried *C. roseus* leaves which can now be replaced by continuous fermentation of the engineered yeast. This effort exemplifies the advances made in heterologous expression technology and the proof of its value, this may influence the adoption of other useful BGCs to be considered as candidates for expression studies.

Sponge metagenomic and microbiome studies still remain a challenge for researches with many aspects that can be improved. These challenges usually arise from the difficulty of generating a suitable quantity and quality of sequencing data from sponges due to the complexity of the sample.[397] This limits the ability to generate complete assemblies of microbes and access to the low abundance species. Both of these were issues that hampered progress in this study. To overcome these limitations, researches are turning to more sophisticated methods to such as single-cell genomics, where individual microbial cells are isolated using microfluidics and the single genome copy is amplified to allow sequencing. This technique can unequivocally link the BGCs to the cellular source of uncultivated producers and was instrumental in discovery of the source of the previously orphaned antifungal compound aurantoside.[398]

8.3 Key findings

8.3.1. Metagenomic clone library survey.

The initial stage of this research was the isolation of high molecular weight DNA and the construction of cosmid metagenome libraries from the primary sponge sample. After some optimisations, a cosmid library of approximately 400,000 unique clones was produced. As well as being pivotal to this study, this library will serve as an ongoing resource and metagenomic sequence bank for future studies of the sponge's microbiome.

As part of the 'due diligence' conducted before finalising the library building phase and undertaking intensive library screening, we selected 48 random clones for sequence analysis. This revealed not only that the symbiont microbial genomes had been captured by the library but also an unexpected high abundance of BGC like sequences of the target type were present within this random sampling. Included in this was a candidate for a partial mycalamide BGC.

8.3.2. Direct metagenomic sequencing.

Due to the apparent higher than expected presence of BGC like sequences in the constructed metagenomic library, we proposed to supplement our discovery efforts with a direct metagenomic sequencing approach. We hypothesised that the target organism was adequately abundant to allow a direct sequencing approach to produce useful sequencing depth of the target organisms from a modestly sized data set.

We generated and assembled a pilot short-read data set which produced promising results including partial sequences for two of the three target BGGs, those of mycalamide and pateamine. This was followed by generation of additional data including long-read sequencing and a hybrid assembly approach which vastly improved the continuity of the assemblies and completed the two partial BGCs. Although many additional BGCs were also observed in the improved assembly, none of these were candidates for the remaining target BGC of peloruside. One of the additional BGCs, for a predicted molecule that was named gananamide, was of interested to us. This BGC was a partial homologue for an already characterised bacterial secondary metabolite gene cluster that was not known to be associated with *M. hentscheli*, but had been previously discovered in the biosynthetically rich bacterial genus *Entotheonella.* We can conclude that the use of direct metagenomic sequencing with 'recent' advances in sequencing technology, assembly and bioinformatics is a productive and economic way to discover BGCs from complex environments.

8.3.3. Multiple bacterial producers.

The direct metagenomic assembly approach afforded some additional insights into the biosynthetic systems responsible for the production of the bioactive compounds isolated from *M. hentscheli*. Firstly, in accordance with our hypothesis, production of the secondary metabolites were in fact encoded by bacterial T1PKS BGCs, confirming their symbiont origin. Additionally, the genomic context of the BGCs made it immediately apparent that the secondary metabolites were produced by distinct microorganisms. This was in contrast with our initial expectations, and previous studies of sponge secondary metabolite biosynthetic systems[222], that a single microorganism was responsible for producing these compounds. We were able to recover complete metagenome assembled genomes (MAGs) for some of the organisms that were found to carry the BGCs and these were not represented by any known bacterial genera. This may explain why culture-based efforts to access the compounds has failed, as the currently available culturing technology was not appropriate for these enigmatic organisms. The production by distinct microorganism also explained why the BGC for

peloruside had eluded our discovery efforts thus far. We now expect this BGC to be encoded in the genome of an organism that was either absent from this sponge sample analyses or at very low abundance.

Three of the producing species that we identified are the founding members of newly defined bacterial genera *Entomycale*, *Patea*, and *Caria*. The presence of both onnamide and polytheonamide-like gene clusters directly parallels the situation in the unleaded sponge *Theonella swinhoei*; however, in *M. hentscheli* these two clusters are hosted by symbionts that are phylogenetically distant from the *T. swinhoei* producer (*Entotheonella factor*). The observation of similar BGCs, playing similar roles in such distantly related microbial species, emphasizes the extreme horizontal migration of this cluster and contributes to a growing body of work that suggests that acquisition of BGCs encoding compounds with potential defensive properties might drive the formation of stable long-term associations between hosts and their microbial symbionts.

8.3.4. Peloruside BGC recovery.

Although the recovery of many full and partial, target and off-target BGCs was considered to be a definitive success for the project and a breakthrough in understanding the biosynthesis of *M. hentscheli* derived secondary metabolites, we had still not met one of our primary goals, the recovery of the peloruside BGC. To this end we enlisted additional samples and increased the sequencing depth to cover the possibilities that the peloruside producing organism was absent or below the sensitivity threshold of our approach in the primary sample. We also attempted to produce a metagenomic clone library from one of the additional sponge samples, suspected to be peloruside positive based on preliminary chemotyping results, however this resource intensive process was of limited success.

The sequence data from each additional sample was analysed independently and no peloruside BGC candidates were detected. When we combined all the available data sets to produce a co-assembly a candidate peloruside BGC was detected on a low coverage short contig. This BGC was ultimately detected in data derived from the primary sample, however was only present when assembled with an updated assembly software release. Unlike the other BGCs focused on in this study, the peloruside BGC was not resolved within the producing organisms genome and was at very low relative abundance. Efforts were made to link the peloruside BGC contig to a producing organism by various binning methods and manual data analysis with no success. Further attempts to extend this sequence were made that were not covered in the main body of this work but are worth mentioning briefly. The

metagenomic clone library of the same sample was interrogated by PCR and found to contain clones covering the BGC sequence, unfortunately these were not able to be isolated after multiple attempts, for unknown reasons. Pools of clones from the same library that tested positive for fragments of the BGC were sequenced directly and did not result in useful extension of the BGC. The peloruside BGC remains isolated from its producing organism at this point.

8.3.5. Biosynthetic mechanisms

Our approach enabled us to identify complete BGCs for each of the cytotoxic polyketides previously isolated from *M. hentscheli*, without any prior knowledge of the localization, morphology, or phylogeny of producing organisms. A biosynthetic mechanism for each of these compounds was proposed. These proposed mechanisms include non-canonical features and architecture which is typical for symbiont *trans*-AT PKSs. Of note was the dual PKS chain coupling of the pateamine BGC.

Functional assignments in this work are based on comparison to other known biosynthetic systems; however, it will be necessary in this, and other cases, for biochemical characterization to be undertaken in order to confirm putative functional assignments.

8.3.6. Core microbiome

As the project evolved to include multiple sponge samples we were able to repurpose the metagenomic sequencing data to conduct a comparative analysis of the microbiome across the samples.

This analysis revealed that the identity of the species present in each of the *M. hentscheli* samples was remarkably stable. Of the 26 high-quality MAGS assembled across the pan-metagenome, 17 were seen in at least four of the five specimens examined, an observation that suggests that *M. hentscheli* stably maintains a defined microbiome that contributes cooperatively to the secondary metabolic output of the holobiont. This feature is typical of the microbiomes of Mycale sponges.[399] Another striking feature was the relative simplicity of individual microbiomes that were consistently dominated by just a few (<4) high-abundance species with the remaining species presenting at lower variable levels.

Manual examination of the BGCs detected in the putative "*Candidatus* Caria hoplita," "*Candidatus* Patea custodiens," and "*Candidatus* Entomycale ignis" MAGs, across each sample in which they were found, revealed that in each case the BGCs for mycalamide,

pateamine, and the putative polytheonamide-like RiPP were present in the expected MAG. This further supported our assignment of these strains as producers of their respective metabolites. Interestingly, the presence of a producing organism did not directly correlate with detection of the expected metabolite. This suggests that although *M. hentscheli* maintains a stable cohort of microbes, in some cases this chemical potential is latent and awaiting an appropriate environmental cue.

8.4 Future directions

To continue to derive benefits from the data and key insights produced by this study several research initiatives could be considered. Firstly, the heterologous expression of the recovered biosynthetic gene clusters to produce the secondary metabolites could be attempted. This would be an ambitious project and would require construction of the large BGC, and finding or engineering an appropriate host for expression. Another approach to achieve this may be in analysing the full genomes of the producing organisms to guide direct cultivation and employ the now known sequence to aid with screening. An unresolved intention of the project was the completion of the genome of the peloruside producer, this may now or very soon be a tractable goal with the continued advances in long-read sequencing and assembly. The project also uncovered a number of orphan BGCs, some of which may warrant much more attention than was allocated in this work. Some of these BGCs were of the type that we were focusing on in this study (*trans*-AT PKS) and likely fully or very nearly completely assembled. These BGCs may have utility in the isolation of novel compounds from *M. hentscheli* by guiding metabolite isolation and structural elucidation based on BGC features, or be used as templates for heterologous expression. One example of this was the presence of an azumamide like gene cluster in the primary assembly. As a pilot investigation, organic extracts of *M. hentscheli* were fractionated and tested for expected histone deacetylases (HDAC) inhibitor activity.[400] Active fractions were then analysed for the presence of expected amino acid residues based on the BGC substrate predictions. This research was not completed in the time frame of this study.

The success of direct metagenomic sequencing and bioinformatic analysis as a reasonable approach to access the BGCs from a sponge microbiome opened new research directions for the supporting lab. Several projects based on this approach were initiated during the course of this study which included activities such as sampling over 1000 sponges from a local

collection held by a CRI and developing inhouse sequencing library construction methods. I look forward to the outputs of this ongoing work.

This work also uncovered new bacterial and archeal diversity; this data could be expanded on to further understand sponge ecology and the important symbiotic relationship with the hosted microbiome. This field of research was outside the direct expertise of the supporting lab but found utility with willing collaborators.

8.5 Data availability

The sequencing data produced for and used in this project has been deposited in online repositories and the respective links can be found throughout the main body of this work in the relevant sections describing the data. All relevant data for the project can also be accessed under the NCBI BioProjects IDs: PRJNA515312 and PRJNA576275.

A table of accession numbers linking to all deposited MAG sequences produced in this study is provided in Appendix A.

8.6 Concluding remarks

The fundamental objective of this study was to discover three target BGCs from the microbiome of the endemic marine sponge *M. hentscheli* using metagenomic techniques.

Eventually this goal was completely achieved by building upon observations at various stages of the research and adapting the methodology to progress the findings along the way. Foundational to the success of this research was the availability of high quality sponge material and the prior work that had been done to characterise the bioactive molecules of interest. It was a privilege to contribute to this important body of work. The high quality sponge tissue provided allowed for the extraction of ample high quality HMW DNA. This was also crucial to the success of this project and the importance of the care taken at this stage of the research cannot be over stated. The ability to isolate HMW DNA allowed us to utilise several complementary metagenomic and sequencing approaches. Continued advance in long-read sequencing and sequencing analysis techniques may allow for even deeper insights to be gained from the isolated DNA. It may now be feasible to link the peloruside BGC to the producing organisms genome. Even during the course of this study, forward strides in bioinformatic software allowed for new discoveries to be made. Advances continue to be made across the sequencing technologies and computational methods that can be applied to metagenomic drug discovery, maintaining a working knowledge of these systems

and the associated data management and visualisation tools will be crucial for the success of future research programmes.

9 References

1. Walsh, C. T. & Fischbach, M. A. Natural products version 2.0: Connecting genes to molecules. *Journal of the American Chemical Society* vol. 132 2469–2493 (2010).
2. Foulston, L. Genome mining and prospects for antibiotic discovery. *Current Opinion in Microbiology* vol. 51 1–8 (2019).
3. Lewis, K. Platforms for antibiotic discovery. *Nature Reviews Drug Discovery* vol. 12 371–387 (2013).
4. Wright, G. D. Opportunities for natural products in 21st century antibiotic discovery. *Natural Product Reports* vol. 34 694–701 (2017).
5. Fritz, S. *et al.* Full-length title: NRPPUR database search and in vitro analysis identify an NRPS-PKS biosynthetic gene cluster with a potential antibiotic effect. *BMC Bioinformatics* **19**, (2018).
6. Stevenson, L. J., Owen, J. G. & Ackerley, D. F. Metagenome Driven Discovery of Nonribosomal Peptides. *ACS Chem. Biol.* **14**, 2115–2126 (2019).
7. Brady, S. F. Construction of soil environmental DNA cosmid libraries and screening for clones that produce biologically active small molecules. (2007) doi:10.1038/nprot.2007.195.
8. Quince, C., Walker, A. W., Simpson, J. T., Loman, N. J. & Segata, N. Shotgun metagenomics, from sampling to analysis. *Nature Biotechnology* vol. 35 833–844 (2017).
9. Baltz, R. H. Marcel Faber Roundtable: Is our antibiotic pipeline unproductive because of starvation, constipation or lack of inspiration? in *Journal of Industrial Microbiology and Biotechnology* vol. 33 507–513 (Springer, 2006).
10. Trivella, D. B. B. & de Felicio, R. The Tripod for Bacterial Natural Product Discovery: Genome Mining, Silent Pathway Induction, and Mass Spectrometry-Based Molecular Networking. *mSystems* **3**, (2018).
11. Clardy, J., Fischbach, M. A. & Walsh, C. T. New antibiotics from bacterial natural products. *Nature Biotechnology* vol. 24 1541–1550 (2006).
12. Lam, Y. C. & Crawford, J. M. Discovering antibiotics from the global microbiome. *Nat. Microbiol.* **3**, 392–393 (2018).
13. Newman, D. J. & Cragg, G. M. Natural Products as Sources of New Drugs over the

Last 25 Years ⊥. (2006) doi:10.1021/np068054v.

14. Newman, D. J. & Cragg, G. M. Natural Products as Sources of New Drugs from 1981 to 2014. *J. Nat. Prod.* **79**, 629–661 (2016).
15. Peláez, F. The historical delivery of antibiotics from microbial natural products - Can history repeat? *Biochem. Pharmacol.* **71**, 981–990 (2006).
16. Butler, M. S., Blaskovich, M. A. T., Owen, J. G. & Cooper, M. A. Old dogs and new tricks in antimicrobial discovery. *Current Opinion in Microbiology* vol. 33 25–34 (2016).
17. Lewis, K. Antibiotics: Recover the lost art of drug discovery. *Nature* vol. 485 439–440 (2012).
18. Durand, G. A., Raoult, D. & Dubourg, G. Antibiotic discovery: history, methods and perspectives. *International Journal of Antimicrobial Agents* vol. 53 371–382 (2019).
19. Genilloud, O. *et al.* Current approaches to exploit actinomycetes as a source of novel natural products. *Journal of Industrial Microbiology and Biotechnology* vol. 38 375–389 (2011).
20. Datta, S., Rajnish, K. N., Samuel, M. S., Pugazlendhi, A. & Selvarajan, E. Metagenomic applications in microbial diversity, bioremediation, pollution monitoring, enzyme and drug discovery. A review. *Environmental Chemistry Letters* vol. 18 1229–1241 (2020).
21. Pereira, F. Metagenomics: A gateway to drug discovery. in *Advances in Biological Science Research: A Practical Approach* 453–468 (Elsevier, 2019). doi:10.1016/B978-0-12-817497-5.00028-8.
22. Wrighton, K. H. Antibacterial drugs: Discovering antibiotics through soil metagenomics. *Nature Reviews Drug Discovery* vol. 17 240 (2018).
23. Kwan, J. C. The Who, Why, and How of Small-Molecule Production in Invertebrate Microbiomes: Basic Insights Fueling Drug Discovery. *mSystems* **3**, e00186-17 (2018).
24. Staley, J. T. & Konopka, A. Measurement of in situ activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats. *Annual review of microbiology* vol. 39 321–346 (1985).
25. Harwani, D. The Great Plate Count Anomaly and the Unculturable Bacteria Cryptic genes View project Bioinformatics View project. *Artic. Int. J. Sci. Res.* (2012) doi:10.15373/22778179/SEP2013/122.
26. Amann, R. I., Ludwig, W. & Schleifer, K. H. Phylogenetic identification and in situ detection of individual microbial cells without cultivation. *Microbiol. Mol. Biol. Rev.*

59, (1995).

27. Balvočiute, M. & Huson, D. H. SILVA, RDP, Greengenes, NCBI and OTT - how do these taxonomies compare? *BMC Genomics* **18**, 114 (2017).
28. Glöckner, F. O. *et al.* 25 years of serving the community with ribosomal RNA gene reference databases and tools. *J. Biotechnol.* **261**, 169–176 (2017).
29. Karimi, E. *et al.* Metagenomic binning reveals versatile nutrient cycling and distinct adaptive features in alphaproteobacterial symbionts of marine sponges. *FEMS Microbiol. Ecol.* **94**, (2018).
30. Mori, T. *et al.* Single-bacterial genomics validates rich and varied specialized metabolism of uncultivated Entotheonella sponge symbionts. *Proc. Natl. Acad. Sci. U. S. A.* **115**, 1718–1723 (2018).
31. Nakabachi, A. *et al.* Defensive Bacteriome Symbiont with a Drastically Reduced Genome. *Curr. Biol.* **23**, 1478–1484 (2013).
32. McCutcheon, J. P. & Moran, N. A. Extreme genome reduction in symbiotic bacteria. *Nat. Rev. Microbiol.* **10**, 13–26 (2012).
33. Lopera, J., Miller, I. J., McPhail, K. L. & Kwan, J. C. Increased Biosynthetic Gene Dosage in a Genome-Reduced Defensive Bacterial Symbiont. *mSystems* **2**, e00096-17 (2017).
34. Waterworth, S. C. *et al.* Horizontal gene transfer to a defensive symbiont with a reduced genome in a multipartite beetle microbiome. *MBio* **11**, (2020).
35. Wiegand, S. *et al.* Cultivation and functional characterization of 79 planctomycetes uncovers their unique biology. *Nat. Microbiol. 2019* 1–15 (2019) doi:10.1038/s41564-019-0588-1.
36. Walsh, C. Antibiotics: actions, origins, resistance. *Antibiot. actions, Orig. Resist.* (2003).
37. Piel, J. *Metabolites from symbiotic bacteria*. *Natural Product Reports* vol. 26 338–362 (The Royal Society of Chemistry, 2009).
38. Budzikiewicz, H. Secondary metabolites from fluorescent pseudomonads. *FEMS Microbiol. Lett.* **104**, 209–228 (1993).
39. Kramer, J., Özkaya, Ö. & Kümmerli, R. Bacterial siderophores in community and host interactions. *Nature Reviews Microbiology* vol. 18 152–163 (2020).
40. Shank, E. A. & Kolter, R. New developments in microbial interspecies signaling. *Current Opinion in Microbiology* vol. 12 205–214 (2009).
41. Höfer, I. *et al.* Insights into the biosynthesis of hormaomycin, an exceptionally

complex bacterial signaling metabolite. *Chem. Biol.* **18**, 381–391 (2011).

42. Tyc, O., Song, C., Dickschat, J. S., Vos, M. & Garbeva, P. The Ecological Role of Volatile and Soluble Secondary Metabolites Produced by Soil Bacteria. *Trends in Microbiology* vol. 25 280–292 (2017).
43. Piel, J. *et al.* Antitumor polyketide biosynthesis by an uncultivated bacterial symbiont of the marine sponge Theonella swinhoei. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 16222–16227 (2004).
44. Zan, J. *et al.* A microbial factory for defensive kahalalides in a tripartite marine symbiosis. *Science (80-.).* **364**, (2019).
45. Nakabachi, A. *et al.* Defensive bacteriome symbiont with a drastically reduced genome. *Curr. Biol.* **23**, 1478–1484 (2013).
46. Wright, G. D. Unlocking the potential of natural products in drug discovery. *Microb. Biotechnol.* **12**, 55–57 (2019).
47. Wright, G. D. Opportunities for natural products in 21 st century antibiotic discovery. *Nat. Prod. Rep.* **34**, 694–701 (2017).
48. Scott, T. A. & Piel, J. *The hidden enzymology of bacterial natural product biosynthesis*. *Nature Reviews Chemistry* vol. 3 404–425 (Nature Publishing Group, 2019).
49. Milshteyn, A., Schneider, J. S. & Brady, S. F. Mining the metabiome: Identifying novel natural products from microbial communities. *Chemistry and Biology* vol. 21 1211–1223 (2014).
50. Demain, A. L. Importance of microbial natural products and the need to revitalize their discovery. *J. Ind. Microbiol. Biotechnol.* **41**, 185–201 (2014).
51. Bérdy, J. Thoughts and facts about antibiotics: Where we are now and where we are heading. *J. Antibiot. (Tokyo).* **65**, 385–395 (2012).
52. Bérdy, J. Bioactive Microbial Metabolites. *J. Antibiot. (Tokyo).* **58**, 1–26 (2005).
53. Watve, M. G., Tickoo, R., Jog, M. M. & Bhole, B. D. How many antibiotics are produced by the genus Streptomyces? *Arch. Microbiol.* **176**, 386–390 (2001).
54. Katz, L. & Baltz, R. H. Natural product discovery: past, present, and future. *Journal of Industrial Microbiology and Biotechnology* vol. 43 155–176 (2016).
55. Medema, M. H. *et al.* Minimum Information about a Biosynthetic Gene cluster. *Nature Chemical Biology* vol. 11 625–631 (2015).
56. Martin, J. F. & Liras, P. Organization and expression of genes involved in the biosynthesis of antibiotics and other secondary metabolites. *Annual Review of*

Microbiology vol. 43 173–206 (1989).

57. Hong, H., Demangel, C., Pidot, S. J., Leadlay, P. F. & Stinear, T. Mycolactones: Immunosuppressive and cytotoxic polyketides produced by aquatic mycobacteria. *Natural Product Reports* vol. 25 447–454 (2008).
58. Fischbach, M. A. & Walsh, C. T. Assembly-Line Enzymology for Polyketide and Nonribosomal Peptide Antibiotics: Logic, Machinery, and Mechanisms. (2006) doi:10.1021/cr0503097.
59. Walsh, C. T. Insights into the chemical logic and enzymatic machinery of NRPS assembly lines. *Nat. Prod. Rep.* **33**, 127–135 (2016).
60. Arnison, P. G. *et al.* Ribosomally synthesized and post-translationally modified peptide natural products: overview and recommendations for a universal nomenclature. doi:10.1039/c2np20085f.
61. Owen, J. G. *et al.* Multiplexed metagenome mining using short DNA sequence tags facilitates targeted discovery of epoxyketone proteasome inhibitors. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 4221–4226 (2015).
62. Huo, L. *et al.* Heterologous expression of bacterial natural product biosynthetic pathways. *Nat. Prod. Rep.* **36**, 1412–1436 (2019).
63. Brady, S. F., Chao, C. J., Handelsman, J. & Clardy, J. Cloning and Heterologous Expression of a Natural Product Biosynthetic Gene Cluster from eDNA. *Org. Lett.* **3**, 1981–1983 (2001).
64. Hannigan, G. D. *et al.* A deep learning genome-mining strategy for biosynthetic gene cluster prediction. *Nucleic Acids Res.* **47**, e110–e110 (2019).
65. Nah, H. J., Pyeon, H. R., Kang, S. H., Choi, S. S. & Kim, E. S. Cloning and heterologous expression of a large-sized natural product biosynthetic gene cluster in Streptomyces species. *Front. Microbiol.* **8**, 394 (2017).
66. Kim, J. H. *et al.* Cloning large natural product gene clusters from the environment: Piecing environmental DNA gene clusters back together with TAR. *Biopolymers* **93**, 833–844 (2010).
67. Brady, S. F., Chao, C. J. & Clardy, J. New natural product families from an environmental DNA (eDNA) gene cluster. *J. Am. Chem. Soc.* **124**, 9968–9969 (2002).
68. Brady, S. F. & Clardy, J. Cloning and Heterologous Expression of Isocyanide Biosynthetic Genes from Environmental DNA. *Angew. Chemie Int. Ed.* **44**, 7063–7065 (2005).
69. Rutledge, P. J. & Challis, G. L. Discovery of microbial natural products by activation

of silent biosynthetic gene clusters. *Nature Reviews Microbiology* vol. 13 509–523 (2015).

70. Xu, F. *et al.* A genetics-free method for high-throughput discovery of cryptic microbial metabolites. *Nat. Chem. Biol.* **15**, 161–168 (2019).
71. Wang, B., Guo, F., Dong, S. H. & Zhao, H. Activation of silent biosynthetic gene clusters using transcription factor decoys. *Nat. Chem. Biol.* **15**, 111–114 (2019).
72. Ochi, K. & Hosaka, T. New strategies for drug discovery: Activation of silent or weakly expressed microbial gene clusters. *Applied Microbiology and Biotechnology* vol. 97 87–98 (2013).
73. Mao, D., Okada, B. K., Wu, Y., Xu, F. & Seyedsayamdost, M. R. Recent advances in activating silent biosynthetic gene clusters in bacteria. *Current Opinion in Microbiology* vol. 45 156–163 (2018).
74. Myronovskyi, M. *et al.* Generation of a cluster-free Streptomyces albus chassis strains for improved heterologous expression of secondary metabolite clusters. *Metab. Eng.* **49**, 316–324 (2018).
75. Wang, G. *et al.* CRAGE enables rapid activation of biosynthetic gene clusters in undomesticated bacteria. *Nat. Microbiol.* **4**, 2498–2510 (2019).
76. Sekurova, O. N., Schneider, O. & Zotchev, S. B. Novel bioactive natural products from bacteria via bioprospecting, genome mining and metabolic engineering. *Microb. Biotechnol.* **12**, 828–844 (2019).
77. Kim, H., Ji, C. H., Je, H. W., Kim, J. P. & Kang, H. S. MpCRISTAR: Multiple Plasmid Approach for CRISPR/Cas9 and TAR-Mediated Multiplexed Refactoring of Natural Product Biosynthetic Gene Clusters. *ACS Synth. Biol.* **9**, 175–180 (2020).
78. Kang, H.-S., Charlop-Powers, Z. & Brady, S. F. Multiplexed CRISPR/Cas9-and TAR-Mediated Promoter Engineering of Natural Product Biosynthetic Gene Clusters in Yeast. (2016) doi:10.1021/acssynbio.6b00080.
79. Bauman, K. D. *et al.* Refactoring the Cryptic Streptophenazine Biosynthetic Gene Cluster Unites Phenazine, Polyketide, and Nonribosomal Peptide Biochemistry. *Cell Chem. Biol.* **26**, 724-736.e7 (2019).
80. Montiel, D., Kang, H.-S. S., Chang, F.-Y. Y., Charlop-Powers, Z. & Brady, S. F. Yeast homologous recombination-based promoter engineering for the activation of silent natural product biosynthetic gene clusters. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 8953–8958 (2015).
81. Kallifidas, D., Kang, H. S. & Brady, S. F. Tetarimycin A, an MRSA-active antibiotic

identified through induced expression of environmental DNA gene clusters. *J. Am. Chem. Soc.* **134**, 19552–19555 (2012).

82. Yamanaka, K. *et al.* Direct cloning and refactoring of a silent lipopeptide biosynthetic gene cluster yields the antibiotic taromycin A. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 1957–1962 (2014).
83. Tomm, H. A., Ucciferri, L. & Ross, A. C. Advances in microbial culturing conditions to activate silent biosynthetic gene clusters for novel metabolite production. *Journal of Industrial Microbiology and Biotechnology* vol. 46 1381–1400 (2019).
84. Page, M., West, L., Northcote, P., Battershill, C. & Kelly, M. Spatial and Temporal Variability of Cytotoxic Metabolites in Populations of the New Zealand Sponge Mycale hentscheli. *J. Chem. Ecol.* **31**, 1161–1174 (2005).
85. Helfrich, E. J. N. N. N. & Piel, J. Biosynthesis of polyketides by trans-AT polyketide synthases. *Nat. Prod. Rep.* **33**, 231–316 (2016).
86. Bhushan, A., Peters, E. E. & Piel, J. *Entotheonella Bacteria as Source of Sponge-Derived Natural Products: Opportunities for Biotechnological Production*. *Progress in molecular and subcellular biology* vol. 55 291–314 (Springer, Cham, 2017).
87. Hopwood, D. A. *Genetic Contributions to Understanding Polyketide Synthases*. https://pubs.acs.org/sharingguidelines (1997).
88. Moore, B. S. & Hopke, J. N. Discovery of a New Bacterial Polyketide Biosynthetic Pathway. *ChemBioChem* **2**, 35–38 (2001).
89. Staunton, J. The Extraordinary Enzymes Involved in Erythromycin Biosynthesis. *Angew. Chemie Int. Ed. English* **30**, 1302–1306 (1991).
90. Cane, D. E., Prabhakaran, P. C., Tan, W. & Ott, W. R. Macrolide biosynthesis. 6 Mechanism of polyketide chain elongation. *Tetrahedron Lett.* **32**, 5457–5460 (1991).
91. Rawlings, B. J. Type I polyketide biosynthesis in bacteria (Part B). *Natural Product Reports* vol. 18 190–227 (2001).
92. Keatinge-Clay, A. T. The structures of type i polyketide synthases. *Natural Product Reports* vol. 29 1050–1073 (2012).
93. Dutta, S. *et al.* Structure of a modular polyketide synthase. *Nature* **510**, 512–517 (2014).
94. Lin, S., Huang, T. & Shen, B. Tailoring enzymes acting on carrier protein-tethered substrates in natural product biosynthesis. in *Methods in Enzymology* vol. 516 321–343 (2012).
95. Moore, B. S. Biosynthesis of marine natural products: Macroorganisms (Part B).

Natural Product Reports vol. 23 615–629 (2006).

96. Moss, S. J., Martin, C. J. & Wilkinson, B. Loss of co-linearity by modular polyketide synthases: A mechanism for the evolution of chemical diversity. *Natural Product Reports* vol. 21 575–593 (2004).
97. Staunton, J. & Wilkinson, B. *Biosynthesis of Erythromycin and Rapamycin*. (1997).
98. Ray, L. & Moore, B. S. Recent advances in the biosynthesis of unusual polyketide synthase substrates. *Nat. Prod. Rep.* **33**, 150–161 (2016).
99. Herbst, D. A. *et al.* The structural organization of substrate loading in iterative polyketide synthases. *Nat. Chem. Biol.* **14**, 474–479 (2018).
100. Moore, B. S. & Hertweck, C. Biosynthesis and attachment of novel bacterial polyketide synthase starter units. *Nat. Prod. Rep.* **19**, 70–99 (2002).
101. Sheehan, L. S. *et al.* Engineering of the spinosyn PKS: Directing starter unit incorporation. *J. Nat. Prod.* **69**, 1702–1710 (2006).
102. Chan, Y. A., Podevels, A. M., Kevany, B. M. & Thomas, M. G. *Biosynthesis of polyketide synthase extender units*. *Natural Product Reports* vol. 26 90–114 (Royal Society of Chemistry, 2009).
103. Hertweck, C. The biosynthetic logic of polyketide diversity. *Angewandte Chemie - International Edition* vol. 48 4688–4716 (2009).
104. Walsh, C. T. The chemical versatility of natural-product assembly lines. *Acc. Chem. Res.* **41**, 4–10 (2008).
105. Shao, L., Zi, J., Zeng, J. & Zhan, J. Identification of the herboxidiene biosynthetic gene cluster in Streptomyces chromofuscus ATCC 49982. *Appl. Environ. Microbiol.* **78**, 2034–2038 (2012).
106. Horsman, M. E., Hari, T. P. A. A. & Boddy, C. N. Polyketide synthase and non-ribosomal peptide synthetase thioesterase selectivity: Logic gate or a victim of fate? *Natural Product Reports* vol. 33 183–202 (2016).
107. Koch, A. A. *et al.* A Single Active Site Mutation in the Pikromycin Thioesterase Generates a More Effective Macrocyclization Catalyst. *J. Am. Chem. Soc.* **139**, 13456–13465 (2017).
108. Du, L. & Lou, L. PKS and NRPS release mechanisms. *Natural Product Reports* vol. 27 255–278 (2010).
109. Hantke, V., Skellam, E. J. & Cox, R. J. Evidence for enzyme catalysed intramolecular [4+2] Diels-Alder cyclization during the biosynthesis of pyrichalasin H. *Chem. Commun.* **56**, 2925–2928 (2020).

110. Pang, B., Wang, M. & Liu, W. Cyclization of polyketides and non-ribosomal peptides on and off their assembly lines. *Nat. Prod. Rep.* **33**, 162–173 (2016).
111. Hubbard, B. K. & Walsh, C. T. Vancomycin Assembly: Nature's Way. *Angew. Chemie Int. Ed.* **42**, 730–765 (2003).
112. Xu, J., Wan, E., Kim, C. J., Floss, H. G. & Mahmud, T. Identification of tailoring genes involved in the modification of the polyketide backbone of rifamycin B by Amycolatopsis mediterranei S699. *Microbiology* **151**, 2515–2528 (2005).
113. Sundaram, S. & Hertweck, C. On-line enzymatic tailoring of polyketides and peptides in thiotemplate systems. *Current Opinion in Chemical Biology* vol. 31 82–94 (2016).
114. Robbins, T., Kapilivsky, J., Cane, D. E. & Khosla, C. Roles of Conserved Active Site Residues in the Ketosynthase Domain of an Assembly Line Polyketide Synthase. *Biochemistry* **55**, 4476–4484 (2016).
115. Molnár, I. *et al.* The biosynthetic gene cluster for the microtubule-stabilizing agents epothilones A and B from Sorangium cellulosum So ce90. *Chem. Biol.* **7**, 97–109 (2000).
116. Zheng, J. & Keatinge-Clay, A. T. The status of type i polyketide synthase ketoreductases. *MedChemComm* vol. 4 34–40 (2013).
117. Struck, A. W., Thompson, M. L., Wong, L. S. & Micklefield, J. S-Adenosyl-Methionine-Dependent Methyltransferases: Highly Versatile Enzymes in Biocatalysis, Biosynthesis and Other Biotechnological Applications. *ChemBioChem* vol. 13 2642–2655 (2012).
118. Skiba, M. A. *et al.* Domain Organization and Active Site Architecture of a Polyketide Synthase C-methyltransferase. *ACS Chem. Biol.* **11**, 3319–3327 (2016).
119. Stevens, D. C., Wagner, D. T., Manion, H. R., Alexander, B. K. & Keatinge-Clay, A. T. Methyltransferases excised from trans-AT polyketide synthases operate on N-acetylcysteamine-bound substrates. *J. Antibiot. (Tokyo).* **69**, 567–570 (2016).
120. Walker, P. D., Weir, A. N. M., Willis, C. L. & Crump, M. P. Polyketide β-branching: diversity, mechanism and selectivity. *Nat. Prod. Rep.* (2020) doi:10.1039/d0np00045k.
121. Piel, J. Biosynthesis of polyketides by trans-AT polyketide synthases. *Natural Product Reports* vol. 27 996–1047 (2010).
122. Calderone, C. T., Kowtoniuk, W. E., Kelleher, N. L., Walsh, C. T. & Dorrestein, P. C. Convergence of isoprene and polyketide biosynthetic machinery: Isoprenyl-S-carrier proteins in the pksX pathway of Bacillus subtilis. *Proc. Natl. Acad. Sci. U. S.*

A. **103**, 8977–8982 (2006).

123. Skiba, M. A. *et al.* Structural Basis of Polyketide Synthase O-Methylation. *ACS Chem. Biol.* **13**, 3221–3228 (2018).
124. Piel, J. A polyketide synthase-peptide synthetase gene cluster from an uncultured bacterial symbiont of Paederus beetles. *Proc. Natl. Acad. Sci. U. S. A.* **99**, 14002–7 (2002).
125. Nguyen, T. *et al.* Exploiting the mosaic structure of trans-acyltransferase polyketide synthases for natural product discovery and pathway dissection. *Nat. Biotechnol.* **26**, 225–233 (2008).
126. Staunton, J. & Weissman, K. J. Polyketide biosynthesis: A millennium review. *Natural Product Reports* vol. 18 380–416 (2001).
127. Maloney, F. P., Gerwick, L., Gerwick, W. H., Sherman, D. H. & Smith, J. L. Anatomy of the β-branching enzyme of polyketide biosynthesis and its interaction with an acyl-ACP substrate. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 10316–10321 (2016).
128. Edrada, R. A. *et al.* Swinhoeiamide A, a new highly active calyculin derivative from the marine sponge Theonella swinhoei. *J. Nat. Prod.* **65**, 1168–1172 (2002).
129. Kosol, S., Jenner, M., Lewandowski, J. R. & Challis, G. L. Protein-protein interactions in trans-AT polyketide synthases. *Natural Product Reports* vol. 35 1097–1109 (2018).
130. Calderone, C. T. Isoprenoid-like alkylations in polyketide biosynthesis. *Natural Product Reports* vol. 25 845–853 (2008).
131. Stein, T. Bacillus subtilis antibiotics: Structures, syntheses and specific functions. *Molecular Microbiology* vol. 56 845–857 (2005).
132. Hofemeister, J. *et al.* Genetic analysis of the biosynthesis of non-ribosomal peptide- and polyketide-like antibiotics, iron uptake and biofilm formation by Bacillus subtilis A1/3. *Mol. Genet. Genomics* **272**, 363–378 (2004).
133. Stein, T. Bacillus subtilis antibiotics: structures, syntheses and specific functions. *Mol. Microbiol.* **56**, 845–857 (2005).
134. Poust, S. *et al.* Divergent Mechanistic Routes for the Formation of *gem* -Dimethyl Groups in the Biosynthesis of Complex Polyketides. *Angew. Chemie Int. Ed.* **54**, 2370–2373 (2015).
135. Meinke, J. L. *et al.* Structural and Functional Studies of a *gem* -Dimethylating Methyltransferase from a *trans* -Acyltransferase Assembly Line. *ACS Chem. Biol.* **13**, 3306–3314 (2018).

136. Ansari, M. Z., Yadav, G., Gokhale, R. S. & Mohanty, D. NRPS-PKS: A knowledge-based resource for analysis of NRPS-PKS megasynthases. *Nucleic Acids Res.* **32**, W405–W413 (2004).
137. Zhang, J. J., Tang, X., Huan, T., Ross, A. C. & Moore, B. S. Pass-back chain extension expands multimodular assembly line biosynthesis. *Nat. Chem. Biol.* **16**, 42–49 (2020).
138. Miyanaga, A., Kudo, F. & Eguchi, T. Protein-protein interactions in polyketide synthase-nonribosomal peptide synthetase hybrid assembly lines. *Natural Product Reports* vol. 35 1185–1209 (2018).
139. Caboche, S. N. *et al.* Diversity of monomers in nonribosomal peptides: Towards the prediction of origin and biological activity. *J. Bacteriol.* **192**, 5143–5150 (2010).
140. Michael A. Fischbach†, ‡ and & Christopher T. Walsh*, †. Assembly-Line Enzymology for Polyketide and Nonribosomal Peptide Antibiotics: Logic, Machinery, and Mechanisms. (2006) doi:10.1021/CR0503097.
141. Mootz, H. D., Schwarzer, D. & Marahiel, M. A. Ways of Assembling Complex Natural Products on Modular Nonribosomal Peptide Synthetases A list of abbreviations can be found at the end of the text. *ChemBioChem* **3**, 490 (2002).
142. Stachelhaus, T., Mootz, H. D. & Marahiel, M. A. The specificity-conferring code of adenylation domains in nonribosomal peptide synthetases. *Chem. Biol.* **6**, 493–505 (1999).
143. Marahiel, M. A., Stachelhaus, T. & Mootz, H. D. Modular peptide synthetases involved in nonribosomal peptide synthesis. *Chem. Rev.* **97**, 2651–2673 (1997).
144. Conti, E., Stachelhaus, T., Marahiel, M. A. & Brick, P. Structural basis for the activation of phenylalanine in the non-ribosomal biosynthesis of gramicidin S. *EMBO J.* **16**, 4174–4183 (1997).
145. Cane, D. E. & Walsh, C. T. The parallel and convergent universes of polyketide synthases and nonribosomal peptide synthetases. *Chem. Biol.* **6**, R319–R325 (1999).
146. Ehmann, D. E., Shaw-Reid, C. A., Losey, H. C. & Walsh, C. T. The EntF and EntE adenylation domains of Escherichia coli enterobactin synthetase: Sequestration and selectivity in acyl-AMP transfers to thiolation domain cosubstrates. *Proc. Natl. Acad. Sci. U. S. A.* **97**, 2509–2514 (2000).
147. Stachelhaus, T., Mootz, H. D., Bergendah, V. & Marahiel, M. A. Peptide bond formation in nonribosomal peptide biosynthesis: Catalytic role of the condensation domain. *J. Biol. Chem.* **273**, 22773–22781 (1998).
148. Rausch, C., Hoof, I., Weber, T., Wohlleben, W. & Huson, D. H. Phylogenetic analysis

of condensation domains in NRPS sheds light on their functional evolution. *BMC Evol. Biol.* **7**, 1–15 (2007).

149. Keating, T. A. & Walsh, C. T. Initiation, elongation, and termination strategies in polyketide and polypeptide antibiotic biosynthesis. *Current Opinion in Chemical Biology* vol. 3 598–606 (1999).
150. Reimer, J. M., Aloise, M. N., Harrison, P. M., Martin Schmeing, T. & Schmeing, T. M. Synthetic cycle of the initiation module of a formylating nonribosomal peptide synthetase. *Nature* **529**, 239–242 (2016).
151. Tanovic, A., Samel, S. A., Essen, L. O. & Marahiel, M. A. Crystal structure of the termination module of a nonribosomal peptide synthetase. *Science (80-.).* **321**, 659–663 (2008).
152. Kohli, R. M., Trauger, J. W., Schwarzer, D., Marahiel, M. A. & Walsh, C. T. Generality of peptide cyclization catalyzed by isolated thioesterase domains of nonribosomal peptide synthetases. *Biochemistry* **40**, 7099–7108 (2001).
153. Walsh, C. T. *et al. Tailoring enzymes that modify nonribosomal peptides during and after chain elongation on NRPS assembly lines*. *Current Opinion in Chemical Biology* vol. 5 525–534 (Elsevier Ltd, 2001).
154. Chevrette, M. G., Aicheler, F., Kohlbacher, O., Currie, C. R. & Medema, M. H. SANDPUMA: Ensemble predictions of nonribosomal peptide chemistry reveal biosynthetic diversity across Actinobacteria. *Bioinformatics* **33**, 3202–3210 (2017).
155. Konz, D. & Marahiel, M. A. How do peptide synthetases generate structural diversity? *Chem. Biol.* **6**, R39–R48 (1999).
156. Challis, G. L., Ravel, J. & Townsend, C. A. Predictive, structure-based model of amino acid recognition by nonribosomal peptide synthetase adenylation domains. *Chem. Biol.* **7**, 211–224 (2000).
157. Stachelhaus, T. & Walsh, C. T. Mutational analysis of the epimerization domain in the initiation module PheATE of gramicidin S synthetase. *Biochemistry* **39**, 5775–5787 (2000).
158. Payne, J. A. E. E., Schoppet, M., Hansen, M. H. & Cryle, M. J. Diversity of nature's assembly lines-recent discoveries in non-ribosomal peptide synthesis. *Mol. Biosyst.* **13**, 9–22 (2017).
159. Li, T.-L. *et al.* Biosynthetic Gene Cluster of the Glycopeptide Antibiotic Teicoplanin. *Chem. Biol.* **11**, 107–119 (2004).
160. Walsh, C., Freel Meyers, C. L. & Losey, H. C. Antibiotic glycosyltransferases:

Antibiotic maturation and prospects for reprogramming. in *Journal of Medicinal Chemistry* vol. 46 3425–3436 (American Chemical Society , 2003).

161. Radkov, A. D. & Moe, L. A. Bacterial synthesis of D-amino acids. *Applied Microbiology and Biotechnology* vol. 98 5363–5374 (2014).
162. Hubbard, B. K., Thomas, M. G. & Walsh, C. T. Biosynthesis of L-p-hydroxyphenylglycine, a non-proteinogenic amino acid constituent of peptide antibiotics. *Chem. Biol.* **7**, 931–942 (2000).
163. Samel, S. A., Marahiel, M. A. & Essen, L. O. How to tailor non-ribosomal peptide products-new clues about the structures and mechanisms of modifying enzymes. *Mol. Biosyst.* **4**, 387–393 (2008).
164. Schneider, ‡ Tanya L *et al.* Oxidase domains in epothilone and bleomycin biosynthesis: Thiazoline to thiazole oxidation during chain elongation. *Biochemistry* **42**, 9722–9730 (2003).
165. Süssmuth, R. D. & Mainz, A. Nonribosomal Peptide Synthesis—Principles and Prospects. *Angewandte Chemie - International Edition* vol. 56 3770–3821 (2017).
166. Walker, K. D., Klettke, K., Akiyama, T. & Croteau, R. Cloning, heterologous expression, and characterization of a phenylalanine aminomutase involved in taxol biosynthesis. *J. Biol. Chem.* **279**, 53947–53954 (2004).
167. Maddah, F. El, Nazir, M. & König, G. M. The rare amino acid building block 3-(3-furyl)-Alanine in the formation of non-ribosomal peptides. *Natural Product Communications* vol. 12 147–150 (2017).
168. Dowling, D. P. *et al.* Structural elements of an NRPS cyclization domain and its intermodule docking domain. doi:10.1073/pnas.1608615113.
169. Roy, R. S., Gehring, A. M., Milne, J. C., Belshaw, P. J. & Walsh, C. T. Thiazole and oxazole peptides: Biosynthesis and molecular machinery. *Nat. Prod. Rep.* **16**, 249–263 (1999).
170. Kashyap, A. *et al.* Review on Synthetic Chemistry and Antibacterial Importance of Thiazole Derivatives. *Curr. Drug Discov. Technol.* **15**, 214–228 (2018).
171. Zhang, W. *et al.* Family-wide structural characterization and genomic comparisons decode the diversity-oriented biosynthesis of thalassospiramides by marine proteobacteria. *J. Biol. Chem.* **291**, 27228–27238 (2016).
172. Medema, M. H. & Fischbach, M. A. *Computational approaches to natural product discovery*. *Nature Chemical Biology* vol. 11 639–648 (Nature Publishing Group, 2015).

173. Ziemert, N., Alanjary, M. & Weber, T. The evolution of genome mining in microbes-a review. *Natural Product Reports* vol. 33 988–1005 (2016).
174. Ricart, E. *et al.* RBAN: Retro-biosynthetic analysis of nonribosomal peptides. *J. Cheminform.* **11**, 1–14 (2019).
175. Romek, K. M. *et al.* A retro-biosynthetic approach to the prediction of biosynthetic pathways from position-specific isotope analysis as shown for tramadol. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 8296–8301 (2015).
176. Chen, Y. *et al.* Characterization of the Chemical Space of Known and Readily Obtainable Natural Products. *J. Chem. Inf. Model.* **58**, 1518–1532 (2018).
177. Nivina, A., Yuet, K. P., Hsu, J. & Khosla, C. Evolution and Diversity of Assembly-Line Polyketide Synthases Focus Review. (2019) doi:10.1021/acs.chemrev.9b00525.
178. Miller, S. J. & Clardy, J. Natural products: Beyond grind and find. *Nat. Chem.* **1**, 261–263 (2009).
179. Bachmann, B. O., Van Lanen, S. G., Baltz, R. H., Lanen, S. G. Van & Baltz, R. H. Microbial genome mining for accelerated natural products discovery: Is a renaissance in the making? *Journal of Industrial Microbiology and Biotechnology* vol. 41 175–184 (2014).
180. Costa, M. S., Clark, C. M., Ómarsdóttir, S., Sanchez, L. M. & Murphy, B. T. Minimizing Taxonomic and Natural Product Redundancy in Microbial Libraries Using MALDI-TOF MS and the Bioinformatics Pipeline IDBac. *J. Nat. Prod.* **82**, 2167–2173 (2019).
181. Piddock, L. J. V. V. The crisis of no new antibiotics-what is the way forward? *The Lancet Infectious Diseases* vol. 12 249–253 (2012).
182. Genilloud, O. Natural products discovery and potential for new antibiotics. *Current Opinion in Microbiology* vol. 51 81–87 (2019).
183. Bentley, S. D. *et al.* Complete genome sequence of the model actinomycete Streptomyces coelicolor A3(2). *Nature* **417**, 141–147 (2002).
184. Ikeda, H. *et al.* Complete genome sequence and comparative analysis of the industrial microorganism Streptomyces avermitilis. *Nat. Biotechnol.* **21**, 526–531 (2003).
185. Challis, G. L. & Ravel, J. Coelichelin, a new peptide siderophore encoded by the *Streptomyces coelicolor* genome: structure prediction from the sequence of its non-ribosomal peptide synthetase. *FEMS Microbiol. Lett.* **187**, 111–114 (2000).
186. Omura, S. *et al.* Genome sequence of an industrial microorganism Streptomyces avermitilis: Deducing the ability of producing secondary metabolites. *Proc. Natl.*

Acad. Sci. U. S. A. **98**, 12215–12220 (2001).

187. Baltz, R. H. Natural product drug discovery in the genomic era: realities, conjectures, misconceptions, and opportunities. *J. Ind. Microbiol. Biotechnol.* **46**, 281–299 (2019).
188. Harvey, A. L., Edrada-Ebel, R. & Quinn, R. J. The re-emergence of natural products for drug discovery in the genomics era. *Nature Reviews Drug Discovery* vol. 14 111–129 (2015).
189. Katz, M., Hover, B. M. & Brady, S. F. Culture-independent discovery of natural products from soil metagenomes. *J. Ind. Microbiol. Biotechnol.* **43**, 129–141 (2016).
190. Banegas-Luna, A. J. *et al.* Advances in distributed computing with modern drug discovery. *Expert Opin. Drug Discov.* **14**, 9–22 (2019).
191. Prihoda, D. *et al.* The application potential of machine learning and genomics for understanding natural product diversity, chemistry, and therapeutic translatability. *Nat. Prod. Rep.* (2020) doi:10.1039/d0np00055h.
192. Kalantar, K. L. *et al.* IDseq-An open source cloud-based pipeline and analysis service for metagenomic pathogen detection and monitoring. *Gigascience* **9**, 1–14 (2020).
193. Kautsar, S. A. *et al.* MIBiG 2.0: A repository for biosynthetic gene clusters of known function. *Nucleic Acids Res.* **48**, D454–D458 (2020).
194. Medema, M. H. *et al.* antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. doi:10.1093/nar/gkr466.
195. Blin, K. *et al.* antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline. *Nucleic Acids Res.* **47**, 81–87 (2019).
196. Medema, M. H. *et al.* AntiSMASH: Rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. *Nucleic Acids Res.* **39**, (2011).
197. Delcher, A. L., Bratke, K. A., Powers, E. C. & Salzberg, S. L. Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics* **23**, 673–679 (2007).
198. Li, L., Stoeckert, C. J. & Roos, D. S. OrthoMCL: Identification of ortholog groups for eukaryotic genomes. *Genome Res.* **13**, 2178–2189 (2003).
199. Durbin, Richard and Eddy, Sean R and Krogh, Anders and Mitchison, G. *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. (Cambridge university press, 1998).
200. Forney, G. D. The Viterbi Algorithm. *Proc. IEEE* **61**, 268–278 (1973).
201. Jurafsky, D. & Martin, J. Hidden Markov Models - Stanford. *Speech Lang. Process.*

21 (2017) doi:10.1016/S0959-440X(96)80056-X.

202. Finn, R. D. *et al.* The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.* **44**, 279–285 (2016).
203. Haft, D. H., Selengut, J. D. & White, O. The TIGRFAMs database of protein families. doi:10.1093/nar/gkg128.
204. Schultz, J., Milpetz, F., Bork, P. & Ponting, C. P. SMART, a simple modular architecture research tool: Identification of signaling domains. *Proc. Natl. Acad. Sci. U. S. A.* **95**, 5857–5864 (1998).
205. Letunic, I. *et al. Recent improvements to the SMART domain-based sequence annotation resource*. *Nucleic Acids Research* vol. 30 242–244 (2002).
206. De Jong, A., Van Heel, A. J., Kok, J. & Kuipers, O. P. BAGEL2: mining for bacteriocins in genomic data. doi:10.1093/nar/gkq365.
207. Van Heel, A. J., De Jong, A., Montalbá N-Ló Pez, M., Kok, J. & Kuipers, O. P. BAGEL3: automated identification of genes encoding bacteriocins and (non-)bactericidal posttranslationally modified peptides. doi:10.1093/nar/gkt391.
208. *HMMER*.
209. Finn, R. D., Clements, J. & Eddy, S. R. HMMER web server: interactive sequence similarity searching. doi:10.1093/nar/gkr367.
210. Letunic, I., Doerks, T. & Bork, P. SMART 6: recent updates and new developments. *Nucleic Acids Res.* **37**, 229–232 (2008).
211. Anand, S. *et al.* SBSPKS: Structure based sequence analysis of polyketide synthases. *Nucleic Acids Res.* **38**, (2010).
212. Yadav, G., Gokhale, R. S. & Mohanty, D. Towards Prediction of Metabolic Products of Polyketide Synthases: An In Silico Analysis. *PLoS Comput. Biol.* **5**, e1000351 (2009).
213. Kenshole, E., Herisse, M., Michael, M. & Pidot, S. J. Natural product discovery through microbial genome mining. *Current Opinion in Chemical Biology* vol. 60 47–54 (2021).
214. Latorre-Pérez, A., Villalba-Bermell, P., Pascual, J. & Vilanova, C. Assembly methods for nanopore-based metagenomic sequencing: a comparative study. **10**, 1–14 (2020).
215. Bhushan, A., Egli, P. J., Peters, E. E., Freeman, M. F. & Piel, J. Genome mining- and synthetic biology-enabled production of hypermodified peptides. *Nat. Chem.* **11**, 931–939 (2019).
216. Ke, J. & Yoshikuni, Y. Multi-chassis engineering for heterologous production of

microbial natural products. *Current Opinion in Biotechnology* vol. 62 88–97 (2020).

217. Barzkar, N., Jahromi, S. T., Poorsaheli, H. B. & Vianello, F. Metabolites from marine microorganisms, micro, and macroalgae: Immense scope for pharmacology. *Marine Drugs* vol. 17 464 (2019).
218. Stien, D. Marine Microbial Diversity as a Source of Bioactive Natural Products. *Mar. Drugs* **18**, 215 (2020).
219. Li, R. *et al.* Natural Product Reports Marine natural products †. **37**, 139–294 (2020).
220. Carroll, A. R. *et al.* Marine natural products. *Nat. Prod. Rep.* **36**, 122–173 (2019).
221. Fukuhara, K. *et al.* Colony-wise Analysis of a Theonella swinhoei Marine Sponge with a Yellow Interior Permitted the Isolation of Theonellamide I. *J. Nat. Prod* **81**, 2595–2599 (2018).
222. Wilson, M. C. *et al.* An environmental bacterial taxon with a large and distinct metabolic repertoire. *Nature* **506**, 58–62 (2014).
223. Wegerski, C. J., Hammond, J., Tenney, K., Matainaho, T. & Crews, P. A Serendipitous Discovery of Isomotuporin-Containing Sponge Populations of Theonella swinhoei. *J. Nat. Prod* **70**, 38 (2007).
224. Hentschel, U., Piel, J., Degnan, S. M. & Taylor, M. W. *Genomic insights into the marine sponge microbiome*. *Nature Reviews Microbiology* vol. 10 641–654 (Nature Publishing Group, 2012).
225. Lackner, G., Peters, E. E., Helfrich, E. J. N. & Piel, J. Insights into the lifestyle of uncultured bacterial natural product factories associated with marine sponges. *Proc. Natl. Acad. Sci. U. S. A.* **114**, E347–E356 (2017).
226. Hamada, T., Matsunaga, S., Yano, G. & Fusetani, N. Polytheonamides A and B, Highly Cytotoxic, Linear Polypeptides with Unprecedented Structural Features, from the Marine Sponge, Theonella swinhoei. (2005) doi:10.1021/ja045749e.
227. Freeman, M. F., Vagstad, A. L. & Piel, J. Polytheonamide biosynthesis showcasing the metabolic potential of sponge-associated uncultivated 'Entotheonella' bacteria. *Curr. Opin. Chem. Biol.* **31**, 8–14 (2016).
228. Freeman, M. F. *et al.* Metagenome mining reveals polytheonamides as posttranslationally modified ribosomal peptides. **338**, 387–390 (2012).
229. Julien, B. *et al.* Isolation and characterization of the epothilone biosynthetic gene cluster from Sorangium cellulosum. *Gene* **249**, 153–160 (2000).
230. Freeman, M. F., Helf, M. J., Bhushan, A., Morinaka, B. I. & Piel, J. Seven enzymes create extraordinary molecular complexity in an uncultivated bacterium. *Nat. Chem.*

9, 387–395 (2017).

231. Rinke, C. *et al.* Insights into the phylogeny and coding potential of microbial dark matter. *Nature* **499**, 431–437 (2013).
232. Brück, W. M., Sennett, S. H., Pomponi, S. A., Willenz, P. & McCarthy, P. J. Identification of the bacterial symbiont Entotheonella sp. in the mesohyl of the marine sponge Discodermia sp. *ISME J.* **2**, 335–339 (2008).
233. Nakashima, Y., Egami, Y., Kimura, M., Wakimoto, T. & Abe, I. Metagenomic Analysis of the Sponge Discodermia Reveals the Production of the Cyanobacterial Natural Product Kasumigamide by 'Entotheonella'. *PLoS One* **11**, e0164468 (2016).
234. Gunasekera, S. P., Gunasekera, M., Longley, R. E. & Schulte, G. K. Discodermolide: A New Bioactive Polyhydroxylated Lactone from The Marine Sponge Discodermia Dissoluta. *J. Org. Chem.* **55**, 4912–4915 (1990).
235. Kato, Y. *et al.* Calyculin A, a novel antitumor metabolite from the marine sponge Discodermia calyx. *J. Am. Chem. Soc.* **108**, 2780–2781 (1986).
236. Fagerholm, A., Habrant, D. & Koskinen, A. M. Calyculins and Related Marine Natural Products as Serine- Threonine Protein Phosphatase PP1 and PP2A Inhibitors and Total Syntheses of Calyculin A, B, and C. *Mar. Drugs* **8**, 122–172 (2010).
237. Hentschel, U., Usher, K. M. & Taylor, M. W. Marine sponges as microbial fermenters. *FEMS Microbiol. Ecol.* **55**, 167–177 (2006).
238. Hochmuth, T. & Piel, J. Polyketide synthases of bacterial symbionts in sponges - Evolution-based applications in natural products research. *Phytochemistry* vol. 70 1841–1849 (2009).
239. Hill, R. T. *Microbes from Marine Sponges: A Treasure Trove of Biodiversity for Natural Products Discovery*. *Microbial Diversity and Bioprospecting* 177–190 (ASM Press, 2014). doi:10.1128/9781555817770.ch18.
240. Unson, M. D., Holland, N. D. & Faulkner, D. J. A brominated secondary metabolite synthesized by the cyanobacterial symbiont of a marine sponge and accumulation of the crystalline metabolite in the sponge tissue. *Mar. Biol.* **119**, 1–11 (1994).
241. Munro, M. H. G. G. *et al.* The discovery and development of marine compounds with pharmaceutical potential. *Prog. Ind. Microbiol.* **35**, 15–25 (1999).
242. Northcote, P. T., Blunt, J. W. & Munro, M. H. G. G. Pateamine: a potent cytotoxin from the New Zealand Marine sponge, mycale sp. *Tetrahedron Lett.* **32**, 6411–6414 (1991).
243. Page, M. J., Northcote, P. T., Webb, V. L., Mackey, S. & Handley, S. J. Aquaculture

trials for the production of biologically active metabolites in the New Zealand sponge Mycale hentscheli (Demospongiae: Poecilosclerida). *Aquaculture* **250**, 256–269 (2005).

244. Page, M. J., Handley, S. J., Northcote, P. T., Cairney, D. & Willan, R. C. Successes and pitfalls of the aquaculture of the sponge Mycale hentscheli. *Aquaculture* **312**, 52–61 (2011).
245. Anderson, S. A., Northcote, P. T. & Page, M. J. Spatial and temporal variability of the bacterial community in different chemotypes of the New Zealand marine sponge Mycale hentscheli. *FEMS Microbiol. Ecol.* **72**, 328–342 (2010).
246. Richter, A., Kocienski, P., Raubo, P. & Davies, D. E. The in vitro biological activities of synthetic 18-O-methyl mycalamide B, 10-epi-18-O-methyl mycalamide B and pederin. *Anticancer. Drug Des.* **12**, 217–227 (1997).
247. Fisch, K. M. *et al.* Polyketide assembly lines of uncultivated sponge symbionts from structure-based gene targeting. *Nat. Chem. Biol.* **5**, 494–501 (2009).
248. Dejong, C. A. *et al.* Polyketide and nonribosomal peptide retro-biosynthesis and global gene cluster matching. *Nat. Chem. Biol.* **12**, 1007–1014 (2016).
249. Storey, M. A. *et al.* Metagenomic exploration of the marine sponge mycale hentscheli uncovers multiple polyketide-producing bacterial symbionts. *MBio* **11**, (2020).
250. Rust, M. *et al.* A multiproducer microbiome generates chemical diversity in the marine sponge Mycale hentscheli. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 9508–9518 (2020).
251. Hohn, B. In Vitro Packaging of λ and Cosmid DNA. *Methods Enzymol.* **68**, 299–309 (1979).
252. Gunther, E. J., Murray, N. E. & Glazer, P. M. High efficiency, restriction-deficient in vitro packaging extracts for bacteriophage lambda DNA using a new E.coli lysogen. *Nucleic Acids Res.* **21**, 3903–3904 (1993).
253. Gurgui, C. & Piel, J. Metagenomic Approaches to Identify and Isolate Bioactive Natural Products from Microbiota of Marine Sponges. in *Methods in molecular biology (Clifton, N.J.)* vol. 668 247–264 (Humana Press, Totowa, NJ, 2010).
254. Winn, R. N. & Norris, M. B. Analysis of mutations in l transgenic medaka using the cII mutation assay. *Tech. Aquat. Toxicol.* **2**, (2005).
255. Jiang, H., Lei, R., Ding, S.-W. & Zhu, S. Skewer: a fast and accurate adapter trimmer for next-generation sequencing paired-end reads. *BMC Bioinformatics* **15**, 182 (2014).
256. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina

sequence data. *Bioinformatics* **30**, 2114–2120 (2014).

257. Bushnell, B. BBTools software package. *URL http//sourceforge. net/projects/bbmap* (2014).
258. Peng, Y., Leung, H. C. M., Yiu, S. M. & Chin, F. Y. L. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* **28**, 1420–1428 (2012).
259. Bankevich, A. *et al.* SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *J. Comput. Biol.* **19**, 455–477 (2012).
260. Zimin, A. V. *et al.* The MaSuRCA genome assembler. *Bioinformatics* **29**, 2669–2677 (2013).
261. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
262. Albertsen, M. *et al.* Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. **31**, 533–538 (2013).
263. Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119 (2010).
264. Miller, I. I. J. *et al.* Autometa: Automated extraction of microbial genomes from individual shotgun metagenomes. *Nucleic Acids Res.* **47**, e57–e57 (2019).
265. Nurk, S., Meleshko, D., Korobeynikov, A. & Pevzner, P. A. metaSPAdes: a new versatile metagenomic assembler. *Genome Res.* **27**, 824–834 (2017).
266. Yunzi Luo *et al.* Engineered biosynthesis of natural products in heterologous hosts. *Chem. Soc. Rev.* **44**, 5265–5290 (2015).
267. Zhang, J. J. *et al. Genetic platforms for heterologous expression of microbial natural products*. *Natural Product Reports* vol. 36 1313–1332 (Royal Society of Chemistry, 2019).
268. Ngara, T. R. & Zhang, H. Recent Advances in Function-based Metagenomic Screening. *Genomics, Proteomics and Bioinformatics* vol. 16 405–415 (2018).
269. Baltz, R. H. Molecular beacons to identify gifted microbes for genome mining. *J. Antibiot. (Tokyo).* **70**, 639–646 (2017).
270. Charlop-Powers, Z., Banik, J. J., Owen, J. G., Craig, J. W. & Brady, S. F. Selective enrichment of environmental DNA libraries for genes encoding nonribosomal peptides and polyketides by phosphopantetheine transferase- dependent complementation of siderophore biosynthesis. *ACS Chem. Biol.* **8**, 138–143 (2013).
271. Bunet, R. *et al.* A Single Sfp-Type Phosphopantetheinyl Transferase Plays a Major

Role in the Biosynthesis of PKS and NRPS Derived Metabolites in Streptomyces ambofaciens ATCC23877. *PLoS One* **9**, e87607 (2014).

272. Owen, J. G., Robins, K. J., Parachin, N. S. & Ackerley, D. F. A functional screen for recovery of 4′-phosphopantetheinyl transferase and associated natural product biosynthesis genes from metagenome libraries. *Environ. Microbiol.* **14**, 1198–1209 (2012).
273. Craig, J. W., Chang, F. Y., Kim, J. H., Obiajulu, S. C. & Brady, S. F. Expanding small-molecule functional metagenomics through parallel screening of broad-host-range cosmid environmental DNA libraries in diverse proteobacteria. *Appl. Environ. Microbiol.* **76**, 1633–1641 (2010).
274. McMahon, M. D., Guan, C., Handelsman, J. & Thomas, M. G. Metagenomic analysis of Streptomyces lividans reveals host-dependent functional expression. *Appl. Environ. Microbiol.* **78**, 3622–3629 (2012).
275. Lohman, J. R. *et al.* Structural and evolutionary relationships of 'AT-less' type I polyketide synthase ketosynthases. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 12693–12698 (2015).
276. Telenius, H. *et al.* Degenerate oligonucleotide-primed PCR: General amplification of target DNA by a single degenerate primer. *Genomics* **13**, 718–725 (1992).
277. Schirmer, A. *et al.* Metagenomic analysis reveals diverse polyketide synthase gene clusters in microorganisms associated with the marine sponge Discodermia dissoluta. *Appl. Environ. Microbiol.* **71**, 4840–4849 (2005).
278. Macher, J.-N. *et al.* Comparison of environmental DNA and bulk-sample metabarcoding using highly degenerate cytochrome *c* oxidase I primers. *Mol. Ecol. Resour.* **18**, 1456–1468 (2018).
279. Hover, B. M. *et al.* Culture-independent discovery of the malacidins as calcium-dependent antibiotics with activity against multidrug-resistant Gram-positive pathogens. *Nat. Microbiol.* **3**, 415–422 (2018).
280. Reddy, B. V. B. *et al.* Natural product biosynthetic gene diversity in geographically distinct soil microbiomes. *Appl. Environ. Microbiol.* **78**, 3744–3752 (2012).
281. Jenner, M. *et al.* Acyl-Chain Elongation Drives Ketosynthase Substrate Selectivity in *trans* -Acyltransferase Polyketide Synthases. *Angew. Chemie* **127**, 1837–1841 (2015).
282. West, L. M. *et al.* Peloruside A: a potent cytotoxic macrolide isolated from the new zealand marine sponge Mycale sp. *J. Org. Chem.* **65**, 445–449 (2000).
283. Perry, N. B. *et al.* Mycalamide A, an antiviral compound from a New Zealand sponge

of the genus Mycale. *J. Am. Chem. Soc.* **110**, 4850–4851 (1988).

284. Von Schelling, H. & Schelling, H. Von. Coupon Collecting for Unequal Probabilities. *Am. Math. Mon.* **61**, 306–311 (1954).
285. Okamura, Y. *et al.* Isolation and Characterization of a GDSL Esterase from the Metagenome of a Marine Sponge-associated Bacteria. doi:10.1007/s10126-009-9226-x.
286. Charlop-Powers, Z., Milshteyn, A. & Brady, S. F. Metagenomic small molecule discovery methods. *Curr. Opin. Microbiol.* **19**, 70–75 (2014).
287. Charlop-Powers, Z. *et al.* Global biogeographic sampling of bacterial secondary metabolism. *Elife* **2015**, (2015).
288. Williams, S. T. & Cross, T. Chapter XI Actinomycetes. *Methods Microbiol.* **4**, 295–334 (1971).
289. Wilson, M. C. & Piel, J. Metagenomic approaches for exploiting uncultivated bacteria as a resource for novel biosynthetic enzymology. *Chemistry and Biology* vol. 20 636–647 (2013).
290. Lam, K. N., Hall, M. W., Engel, K., Vey, G. & Cheng, J. Evaluation of a Pooled Strategy for High-Throughput Sequencing of Cosmid Clones from Metagenomic Libraries. *PLoS One* **9**, 98968 (2014).
291. Hrvatin, S. & Piel, J. Rapid isolation of rare clones from highly complex DNA libraries by PCR analysis of liquid gel pools. *J. Microbiol. Methods* **68**, 434–436 (2007).
292. Webster, N. S. & Thomas, T. The sponge hologenome. *MBio* **7**, 135–151 (2016).
293. Helf, M. J., Jud, A. & Piel, J. Enzyme from an Uncultivated Sponge Bacterium Catalyzes S-Methylation in a Ribosomal Peptide. *ChemBioChem* **18**, 444–450 (2017).
294. Oulas, A. *et al.* Metagenomics: Tools and Insights for Analyzing Next-Generation Sequencing Data Derived from Biodiversity Studies: *https://doi.org/10.4137/BBI.S12462* **9**, 75–88 (2015).
295. Tyson, G. W. *et al.* Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nat. 2003 4286978* **428**, 37–43 (2004).
296. Martín, H. G. *et al.* Metagenomic analysis of two enhanced biological phosphorus removal (EBPR) sludge communities. *Nat. Biotechnol. 2006 2410* **24**, 1263–1269 (2006).
297. Parks, D. H. *et al.* Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. (2017) doi:10.1038/s41564-017-0012-7.

298. Luo, C., Tsementzi, D., Kyrpides, N. C. & Konstantinidis, K. T. Individual genome assembly from complex community short-read metagenomic datasets. *ISME J. 2012 64* **6**, 898–901 (2011).
299. Rodrigue, S. *et al.* Whole Genome Amplification and De novo Assembly of Single Bacterial Cells. *PLoS One* **4**, e6864 (2009).
300. Burgsdorf, I. *et al.* Lifestyle evolution in cyanobacterial symbionts of sponges. *MBio* **6**, 391–406 (2015).
301. Slaby, B. M., Hackl, T., Horn, H., Bayer, K. & Hentschel, U. Metagenomic binning of a marine sponge microbiome reveals unity in defense but metabolic specialization. **11**, 2465–2478 (2017).
302. Bowers, R. M. *et al.* Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat. Biotechnol. 2017 358* **35**, 725–731 (2017).
303. Gao, Z. M. *et al.* Symbiotic Adaptation Drives Genome Streamlining of the Cyanobacterial Sponge Symbiont "Candidatus Synechococcus spongiarum". *MBio* **5**, (2014).
304. Donia, M. S., Fricke, W. F., Ravel, J. & Schmidt, E. W. Variation in Tropical Reef Symbiont Metagenomes Defined by Secondary Metabolism. *PLoS One* **6**, (2011).
305. Kwan, J. C. *et al.* Genome streamlining and chemical defense in a coral reef symbiosis. *Proc. Natl. Acad. Sci.* **109**, 20655–20660 (2012).
306. Roumpeka, D. D., Wallace, R. J., Escalettes, F., Fotheringham, I. & Watson, M. A Review of Bioinformatics Tools for Bio-Prospecting from Metagenomic Sequence Data. *Front. Genet.* **0**, 23 (2017).
307. Markowitz, V. M. *et al.* IMG/M: the integrated metagenome data management and comparative analysis system. *Nucleic Acids Res.* **40**, D123–D129 (2012).
308. *Understanding the output - antiSMASH Documentation*.
309. Gehring, A. M. *et al.* The nonribosomal peptide synthetase HMWP2 forms a thiazoline ring during biogenesis of yersiniabactin, an iron-chelating virulence factor of Yersinia pestis. *Biochemistry* **37**, 11637–11650 (1998).
310. Du, L. *et al.* An oxidation domain in the BlmIII non-ribosomal peptide synthetase probably catalyzing thiazole formation in the biosynthesis of the anti-tumor drug bleomycin in *Streptomyces verticillus* ATCC15003. *FEMS Microbiol. Lett.* **189**, 171–175 (2000).
311. Bloudoff, K., Fage, C. D., Marahiel, M. A. & Schmeing, T. M. Structural and

mutational analysis of the nonribosomal peptide synthetase heterocyclization domain provides insight into catalysis. *Proc. Natl. Acad. Sci.* **114**, 95–100 (2017).

312. Meziti, A. *et al.* The Reliability of Metagenome-Assembled Genomes (MAGs) in Representing Natural Populations: Insights from Comparing MAGs against Isolate Genomes Derived from the Same Fecal Sample. *Appl. Environ. Microbiol.* **87**, 1–15 (2021).
313. Sczyrba, A. *et al.* Critical Assessment of Metagenome Interpretation - A benchmark of metagenomics software. *Nat. Methods* **14**, 1063–1071 (2017).
314. Nissen, J. N. *et al.* Improved metagenome binning and assembly using deep variational autoencoders. *Nat. Biotechnol. 2021 395* **39**, 555–560 (2021).
315. Mande, S. S., Mohammed, M. H. & Ghosh, T. S. Classification of metagenomic sequences: methods and challenges. *Brief. Bioinform.* **13**, 669–681 (2012).
316. Breitwieser, F. P., Lu, J. & Salzberg, S. L. A review of methods and databases for metagenomic classification and assembly. *Brief. Bioinform.* **20**, 1125–1139 (2018).
317. Sedlar, K., Kupkova, K. & Provaznik, I. Bioinformatics strategies for taxonomy independent binning and visualization of sequences in shotgun metagenomics. *Comput. Struct. Biotechnol. J.* **15**, 48–55 (2017).
318. Alneberg, J. *et al.* Binning metagenomic contigs by coverage and composition. *Nat. Methods 2014 1111* **11**, 1144–1146 (2014).
319. Wu, Y.-W. & Ye, Y. A Novel Abundance-Based Algorithm for Binning Metagenomic Sequences Using l-tuples. *https://home.liebertpub.com/cmb* **18**, 523–534 (2011).
320. Li, X. *et al.* Efficiency of chemical versus mechanical disruption methods of DNAextraction for the identification of oral Gram-positive and Gram-negativebacteria. *J. Int. Med. Res.* **48**, 1–12 (2020).
321. Land, M. *et al.* Insights from 20 years of bacterial genome sequencing. *Funct. Integr. Genomics 2015 152* **15**, 141–161 (2015).
322. PA, N., RW, C. & OA, O. Tetranucleotide frequencies in microbial genomes. *Electrophoresis* **19**, 528–535 (1998).
323. Mohammed, M. H. *et al.* INDUS - a composition-based approach for rapid and accurate taxonomic classification of metagenomic sequences. *BMC Genomics 2011 123* **12**, 1–14 (2011).
324. Wu, Y.-W., Simmons, B. A. & Singer, S. W. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics* **32**, 605–607 (2016).

325. Teeling, H., Waldmann, J., Lombardot, T., Bauer, M. & Glöckner, F. O. TETRA: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences. *BMC Bioinforma. 2004 51* **5**, 1–7 (2004).

326. Kang, D. D. *et al.* MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* **7**, e7359 (2019).

327. Dupont, C. L. *et al.* Genomic insights to SAR86, an abundant and uncultivated marine bacterial lineage. *ISME J. 2012 66* **6**, 1186–1199 (2011).

328. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043–55 (2015).

329. Frank, J. A. *et al.* Improved metagenome assemblies and taxonomic binning using long-read circular consensus sequence data. *Sci. Reports 2016 61* **6**, 1–10 (2016).

330. Zimin, A. V *et al.* Hybrid assembly of the large and highly repetitive genome of Aegilops tauschii, a progenitor of bread wheat, with the MaSuRCA mega-reads algorithm. *Genome Res.* **27**, 787–792 (2017).

331. Antipov, D., Korobeynikov, A., McLean, J. S. & Pevzner, P. A. hybridSPAdes: an algorithm for hybrid assembly of short and long reads. *Bioinformatics* **32**, 1009–1015 (2016).

332. Wick, R. R., Judd, L. M., Gorrie, C. L. & Holt, K. E. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. *PLOS Comput. Biol.* **13**, e1005595 (2017).

333. Koren, S. *et al.* Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nat. Biotechnol.* **30**, 693–700 (2012).

334. Avalon, N. E. *et al.* Bioinformatic and mechanistic analysis of the palmerolide PKS-NRPS biosynthetic pathway from the microbiome of an Antarctic ascidian. *bioRxiv* 2021.04.05.438531 (2021) doi:10.1101/2021.04.05.438531.

335. A, K., PT, N. & JH, M. Peloruside A: a lead non-taxoid-site microtubule-stabilizing agent with potential activity against cancer, neurodegeneration, and autoimmune disease. *Nat. Prod. Rep.* **33**, 549–561 (2016).

336. Hood, K. A. *et al.* Peloruside A, a novel antimitotic agent with paclitaxel-like microtubule-stabilizing activity. *Cancer Res.* **62**, 3356–3360 (2002).

337. Cao, Y. N. *et al.* Recent advances in microtubule-stabilizing agents. *Eur. J. Med. Chem.* **143**, 806–828 (2018).

338. Ganguly, A., Cabral, F., Yang, H. & Patel, K. D. Peloruside A is a microtubule-stabilizing agent with exceptional anti-migratory properties in human endothelial cells. *Oncoscience* **2**, 585 (2015).
339. Chan, A., Singh, A. J., Northcote, P. T. & Miller, J. H. Peloruside A, a microtubule-stabilizing agent, induces aneuploidy in ovarian cancer cells. *Investig. New Drugs 2016 344* **34**, 424–438 (2016).
340. Kennington, S. C. D., Romo, J. M., Romea, P. & Urpí, F. Stereoselective Synthesis of the C9–C19 Fragment of Peloruside A. *Org. Lett.* **18**, 3018–3021 (2016).
341. Keyzers, R. A. & Alexander, R. The Isolation of Biologically Active Secondary Metabolites from New Zealand Marine Organisms. (2003).
342. *Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data*.
343. Clum, A. *et al.* DOE JGI Metagenome Workflow. *mSystems* **6**, (2021).
344. Bushnell, B., Rood, J. & Singer, E. BBMerge – Accurate paired shotgun read merging via overlap. *PLoS One* **12**, e0185056 (2017).
345. *Release SPAdes 3.12.0 · ablab/spades*.
346. Kang, D. D., Froula, J., Egan, R. & Wang, Z. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ* **3**, e1165 (2015).
347. Lin, H.-H. & Liao, Y.-C. Accurate binning of metagenomic contigs via automated clustering sequences using information of genomic signatures and marker genes. *Sci. Reports 2016 61* **6**, 1–8 (2016).
348. Karst, S. M., Kirkegaard, R. H. & Albertsen, M. mmgenome: a toolbox for reproducible genome extraction from metagenomes. *bioRxiv* 059121 (2016) doi:10.1101/059121.
349. Sharon, I. *et al.* Accurate, multi-kb reads resolve complex populations and detect rare microorganisms. *Genome Res.* **25**, 534–543 (2015).
350. Nelson, W. C., Maezato, Y., Wu, Y. W., Romine, M. F. & Lindemann, S. R. Identification and resolution of microdiversity through metagenomic sequencing of parallel consortia. *Appl. Environ. Microbiol.* **82**, 255–267 (2016).
351. *Glossary - antiSMASH Documentation*.
352. Meoded, R. A. *et al.* A Polyketide Synthase Component for Oxygen Insertion into Polyketide Backbones. *Angew. Chemie Int. Ed.* **57**, 11644–11648 (2018).
353. Helfrich, E. J. N. N. N. *et al.* Automated structure prediction of trans-acyltransferase

polyketide synthase products. *Nat. Chem. Biol.* **15**, 813–821 (2019).

354. Yi-Ling Du & S. Ryan, K. Pyridoxal phosphate-dependent reactions in the biosynthesis of natural products. *Nat. Prod. Rep.* **36**, 430–457 (2019).
355. Lin, S. *et al. A free-standing condensation enzyme catalyzing ester bond formation in C-1027 biosynthesis*. *Proceedings of the National Academy of Sciences* vol. 106 www.pnas.org/cgi/content/full/ (2009).
356. Biggins, J. B., Ternei, M. A. & Brady, S. F. Malleilactone, a Polyketide Synthase-Derived Virulence Factor Encoded by the Cryptic Secondary Metabolome of Burkholderia pseudomallei Group Pathogens. *J. Am. Chem. Soc.* **134**, 13192–13195 (2012).
357. Franke, J., Ishida, K. & Hertweck, C. Genomics-Driven Discovery of Burkholderic Acid, a Noncanonical, Cryptic Polyketide from Human Pathogenic Burkholderia Species. *Angew. Chemie* **124**, 11779–11783 (2012).
358. Eustáquio, A. S., Janso, J. E., Ratnayake, A. S., O'donnell, C. J. & Koehn, F. E. Spliceostatin hemiketal biosynthesis in Burkholderia spp. is catalyzed by an iron/α-ketoglutarate-dependent dioxygenase. doi:10.1073/pnas.1408300111.
359. Liu, X. *et al.* Genomics-Guided Discovery of Thailanstatins A, B, and C As Pre-mRNA Splicing Inhibitors and Antiproliferative Agents from Burkholderia thailandensis MSMB43. *J. Nat. Prod.* **76**, 685–693 (2013).
360. Morinaka, B. I. *et al.* Radical S-Adenosyl Methionine Epimerases: Regioselective Introduction of Diverse D-Amino Acid Patterns into Peptide Natural Products. *Angew. Chemie Int. Ed.* **53**, 8503–8507 (2014).
361. Morinaka, B. I., Verest, M., Freeman, M. F., Gugger, M. & Piel, J. An Orthogonal D2O-Based Induction System that Provides Insights into d-Amino Acid Pattern Formation by Radical S-Adenosylmethionine Peptide Epimerases. *Angew. Chemie - Int. Ed.* **56**, 762–766 (2017).
362. Garza, D. R. & Dutilh, B. E. From cultured to uncultured genome sequences: metagenomics and modeling microbial ecosystems. *Cell. Mol. Life Sci. 2015 7222* **72**, 4287–4308 (2015).
363. Schmitt, S. *et al.* Assessing the complex sponge microbiota: core, variable and species-specific bacterial communities in marine sponges. *ISME J. 2012 63* **6**, 564–576 (2011).
364. Reveillaud, J. *et al.* Host-specificity among abundant and rare taxa in the sponge microbiome. *ISME J. 2014 86* **8**, 1198–1209 (2014).

365. Thomas, T. *et al.* Diversity, structure and convergent evolution of the global sponge microbiome. *Nat. Commun.* **7**, 11870 (2016).

366. Sieber, C. M. K. K. *et al.* Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. *Nat. Microbiol. 2018 37* **3**, 836–843 (2018).

367. Pereira-Marques, J. *et al.* Impact of host DNA and sequencing depth on the taxonomic resolution of whole metagenome sequencing for microbiome analysis. *Front. Microbiol.* **10**, 1277 (2019).

368. Yue, Y. *et al.* Evaluating metagenomics tools for genome binning with real metagenomic datasets and CAMI datasets. *BMC Bioinforma. 2020 211* **21**, 1–15 (2020).

369. Olm, M. R., Brown, C. T., Brooks, B. & Banfield, J. F. dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. **11**, 2864–2868 (2017).

370. Balvočiūtė, M., Huson, D. H., Balvočiute, M. & Huson, D. H. No Title. *BMC Genomics* **18**, 114 (2017).

371. Chaumeil PA, Mussig AJ, Hugenholtz P, P. D. GTDB-Tk: A toolkit to classify genomes with the Genome Taxonomy Database. <in prep>. (2019).

372. *GitHub - tseemann/barrnap: Bacterial ribosomal RNA predictor*.

373. Sims, D., Sudbery, I., Ilott, N. E., Heger, A. & Ponting, C. P. Sequencing depth and coverage: key considerations in genomic analyses. *Nat. Rev. Genet. 2014 152* **15**, 121–132 (2014).

374. Edgar, R. C. Updating the 97% identity threshold for 16S ribosomal RNA OTUs. *Bioinformatics* **34**, 2371–2375 (2018).

375. Beye, M., Fahsi, N., Raoult, D. & Fournier, P. E. Careful use of 16S rRNA gene sequence similarity values for the identification of Mycobacterium species. *New Microbes New Infect.* **22**, 24 (2018).

376. Venter, J. C. *et al.* Environmental Genome Shotgun Sequencing of the Sargasso Sea. *Science (80-.).* **304**, 66–74 (2004).

377. Rivas-Marín, E. & Devos, D. P. The Paradigms They Are a-Changin': past, present and future of PVC bacteria research. *Antonie van Leeuwenhoek 2017 1116* **111**, 785–799 (2017).

378. Sackett, J. D. *et al.* Four Draft Single-Cell Genome Sequences of Novel, Nearly Identical Kiritimatiellaeota Strains Isolated from the Continental Deep Subsurface. *Microbiol. Resour. Announc.* **8**, e01249-18 (2019).

379. Yang, Q., Franco, C. M. M. M. & Zhang, W. Uncovering the hidden marine sponge microbiome by applying a multi-primer approach. *Sci. Reports 2019 91* **9**, 1–13 (2019).
380. Taufa, T., Subramani, R., Northcote, P. T. & Keyzers, R. A. Natural Products from Tongan Marine Organisms. *Molecules* **26**, (2021).
381. Caso, A. *et al.* Exploring Chemical Diversity of Phorbas Sponges as a Source of Novel Lead Compounds in Drug Discovery. *Mar. Drugs* **19**, (2021).
382. Dharamshi, J. E. *et al.* Genomic diversity and biosynthetic capabilities of sponge-associated chlamydiae. *ISME J. 2022* 1–16 (2022) doi:10.1038/s41396-022-01305-9.
383. Suet, T. *et al.* Recent Advances of Marine Sponge-Associated Microorganisms as a Source of Commercially Viable Natural Products. *Mar. Biotechnol.* doi:10.1007/s10126-022-10130-2.
384. Murray, A. E. *et al.* Discovery of an Antarctic Ascidian-Associated Uncultivated Verrucomicrobia with Antimelanoma Palmerolide Biosynthetic Potential . *mSphere* **6**, (2021).
385. Uppal, S. *et al.* Uncovering Lasonolide A Biosynthesis Using Genome-Resolved Metagenomics. *MBio* (2022) doi:10.1128/MBIO.01524-22.
386. Hemmerling, F. & Piel, J. Strategies to access biosynthetic novelty in bacterial genomes for drug discovery. *Nat. Rev. Drug Discov. 2022 215* **21**, 359–378 (2022).
387. Lyu, C. *et al.* CMNPD: a comprehensive marine natural products database towards facilitating drug discovery from the ocean. *Nucleic Acids Res.* **49**, D509–D515 (2021).
388. Sahayasheela, V. J. *et al.* Artificial intelligence in microbial natural product drug discovery: current and emerging role. *Nat. Prod. Rep.* (2022) doi:10.1039/D2NP00035K.
389. Saldívar-González, F. I., Aldas-Bulos, V. D., Medina-Franco, J. L. & Plisson, F. Natural product drug discovery in the artificial intelligence era. *Chem. Sci.* **13**, 1526–1546 (2022).
390. Spjuth, O., Frid, J. & Hellander, A. The machine learning life cycle and the cloud: implications for drug discovery. *https://doi.org/10.1080/17460441.2021.1932812* **16**, 1071–1079 (2021).
391. Frye, L., Bhat, S., Akinsanya, K. & Abel, R. From computer-aided drug discovery to computer-driven drug discovery. *Drug Discov. Today Technol.* **39**, 111–117 (2021).
392. Martin, S. *et al.* Nanopore adaptive sampling: a tool for enrichment of low abundance species in metagenomic samples. *Genome Biol.* **23**, 1–27 (2022).

393. Payne, A. *et al.* Readfish enables targeted nanopore sequencing of gigabase-sized genomes. *Nat. Biotechnol. 2020 394* **39**, 442–450 (2020).
394. Cheng, H. *et al.* An ultra-sensitive bacterial pathogen and antimicrobial resistance diagnosis workflow using Oxford Nanopore adaptive sampling sequencing method. *medRxiv* 2022.07.03.22277093 (2022) doi:10.1101/2022.07.03.22277093.
395. Payne, A. *et al.* Barcode aware adaptive sampling for GridION and PromethION Oxford Nanopore sequencers. *bioRxiv* 2021.12.01.470722 (2022) doi:10.1101/2021.12.01.470722.
396. Zhang, J. *et al.* A microbial supply chain for production of the anti-cancer drug vinblastine. *Nat. 2022 6097926* **609**, 341–347 (2022).
397. Strehlow, B. W., Schuster, A., Francis, W. R. & Canfield, D. E. Metagenomic data for Halichondria panicea from Illumina and nanopore sequencing and preliminary genome assemblies for the sponge and two microbial symbionts. *BMC Res. Notes* **15**, 1–4 (2022).
398. Kogawa, M. *et al.* Single-cell metabolite detection and genomics reveals uncultivated talented producer. *PNAS Nexus* **1**, 1–13 (2022).
399. Cárdenas, C. A. *et al.* High similarity in the microbiota of cold-water sponges of the Genus *Mycale* from two different geographical areas. *PeerJ* **6**, e4935 (2018).
400. Nakao, Y. *et al.* Azumamides A–E: Histone Deacetylase Inhibitory Cyclic Tetrapeptides from the Marine SpongeMycale izuensis. *Angew. Chemie Int. Ed.* **45**, 7553–7557 (2006).