Metagenomic exploration of natural product biosynthesis in marine microbial communities

by

Vincent Nowak

A thesis submitted to the Victoria University of Wellington in partial fulfillment of the requirements for the degree of Master of Science

Victoria University of Wellington

2019

Abstract

Marine environments are home to a staggering diversity of microbial life, the majority of which has yet to be cultivated in a laboratory setting. These microbes are the base of the food chain that supports marine life and key drivers of the planets biogeochemical cycles. Marine microbes also produce biologically active natural products, yet surprisingly little is known about the diversity and ecological roles of microbial natural products in marine ecosystems. Beyond their possible ecological roles, natural products from marine microbes are of immense interest as a source of new drug leads. The aim of this thesis was to develop experimental and bioinformatic approaches for the metagenomic exploration of marine microbes. Initial work was conducted on planktonic marine microbes and sea ice microbes with the goal of developing a robust method for constructing large insert (30-45 kb) metagenomic libraries from these challenging low-biomass samples. A series of optimisation experiments were conducted to this end, with some success, however this goal proved intractable given the sample availability and time frame of this thesis. Instead, direct shotgun sequencing, metagenome binning and biosynthetic pathway analysis was used to examine the microbiomes of six Tongan marine sponges. This collection included a sample of the marine sponge *Cacospongia mycofijiensis*, which contained high levels of the cytotoxic polyketides (-)-zampanolide, latrunculin A and laulimalide A. A large insert (30-45 kb) metagenomic library was ultimately constructed from this sponge sample. The newly sequenced Tongan sponge metagenomes were also compared to a collection of marine sponge metagenomes from New Zealand as well as publicly available metagenomes obtained from Mediterranean marine sponges. In addition to developing robust and cost effective experimental and bioinformatic techniques for marine sponge microbiome analysis, key findings from this work were: (1) Host taxonomy and large-scale oceanographic location both appear to be important drivers of microbial community composition (2) Tongan marine sponge microbiomes are more similar to Mediterranean sponge microbiomes than New Zealand marine sponge microbiomes (3) Tongan marine sponge microbiomes are rich in natural product biosynthesis, particularly in ribosomally synthesized and post-translationally modified peptides with potential antimicrobial activity that may play an important defensive role in the context of this symbiosis.

Acknowledgements

First and foremost I would to thank my supervisor Jeremy Owen for taking me on as a Masters student, providing me with an interesting and relevant research topic and offering help and support throughout the past two years. I have learned a lot of skills from Jeremy and have certainly enjoyed the inspiration he has had to offer. One of my favorite quotes would be “There is nothing that irritates me more in life, than weak coffee”. I certainly sympathize with that sentiment.

I would also like to thank my secondary supervisor, Ken Ryan, who has inspired me to study microbial organisms and fascinated me for the Antarctic continent early on in undergraduate lectures.

Furthermore, I would like to thank all the past and new members of the Owen and Ackerley lab groups for creating an awesome working environment as well as fun social activities. Whoever said, scientists can’t be fun.

Last but not least, I would like to thank my partner Lauren, my family and my friends for being such a big part of my life and supporting me in my endeavor of becoming a scientist.

Contents

List of Figures and Tables

Abbreviations

A-domain	Adenylation domain
ACP	Acyl carrier protein
ANI	Average nucleotide identity
AT-domain	Acyltransferase domain
BGC	Biosynthetic gene cluster
C-domain	Condensation domain
DH-domain	Dehydration domain
eDNA	Environmental DNA
ER-domain	Enoyl reductase domain
GCF	Gene cluster family
gDNA	Genomic DNA
GTDB	Genome Taxonomy Database
HMA	High microbial abundance
HMW	High molecular weight
KR-domain	Ketoreductase domain
KS-domain	Ketosynthase domain
LMA	Low microbial abundance
LMW	Low molecular weight
MAG	Metagenome assembled genome
MT-domain	Methyltransferase domain
NRPS	Non-ribosomal peptide synthetase
ORF	Open reading frame
NZ	New Zealand
PCP	Peptidyl carrier protein
PKS	Polyketide synthase
PP	Phosphopantetheine
PPTase	Phosphopantetheinyl transferase
RiPP	Ribosomally synthesized and post-translationally modified peptide
TE-domain	Thioesterase domain
TMTC	Too many to count

Chapter 1

Introduction

1.1 Marine natural products

With increasing antibiotic resistance and a predicted increase in the number of mortalities attributed to antibiotic resistance from 700,000 to 10 million by 2050, there is an urgent need for new antibiotics [1,2]. Natural products have historically been an invaluable source of antibiotics and other antimicrobials [3-5] and the marine environment is a rich source of natural products that is widely considered underexplored [6-8]. One of the most prolific sources of marine natural products has been marine invertebrates [9-12] and in particular sponges [7,10,11]. There are countless marine natural products that display high therapeutic potential [10,13-15].

1.1.1 Natural products from marine bacteria

Bacteria produce a vast range of bioactive molecules [2,16-19]. And in the case of marine bacteria, many of the natural products identified to date have been from culturable members of the Actinomycetales order, with the genera *Streptomyces* [20-22] and *Salinospora* [23,24] being particularly prolific. Other prominent bacterial taxa include Cyanobacteria [25,26], Myxobacteria [27,28] and members of the recently identified uncultivated microbial genus Entotheonellaeota [29,30]. Marine bacteria that produce natural products are ubiquitous in the world's oceans, with confirmed producers from seawater, sediments and symbioses with a wide variety of marine organisms [31-35]

1.1.2 Sponge natural products

Marine sponges have traditionally been a source of chemically diverse natural products, with approximately 200 new compounds discovered per year [6,10,11,36]. These compounds account for almost half of all marine natural products discovered so far and include a number of clinically employed agents [6,7,37-39]. Some of the taxonomic groups of sponges that appear prominent in recent natural product literature include the Theonellidae [30,40-42] and Irciniidae [43] families as well as the genera *Mycale*, *Biemna* and *Clatharia* [44] and *Amphimedon* [45]. Chemical extraction of whole animals, coupled with bioactivity, or analytical chemistry guided fractionation has been used for decades to discover functionally and structurally diverse natural products [10,46]. While these efforts have been extremely fruitful, they are not without their limitations. Specific problems associated with using chemical methods of natural product discovery include low extraction yields, compounds lost during isolation and difficulty of finding a cheap

and high-yielding direct synthesis pathway to support sustainable supply of clinically relevant candidates [10,46,47].

1.1.3 Bacterial symbionts as sponge natural product producers

It is now widely accepted that the majority of bioactive compounds isolated from sponges are likely produced by sponge-associated microbes rather than the animals themselves [37,41,48-51]. This presents a new possible mode for discovery and supply, in which microbial biosynthetic pathways are identified and used to produce new molecules [30,52-54]. The Piel lab was the first to conclusively link an invertebrate marine natural product, onnamide A (Figure 1, 170), to a bacterial producer, a *Pseudomonas* sp. found in the marine sponge *Theonella swinhoei* [41]. This work was inspired by previous work from the same group that identified a symbiotic *Pseudomonas* sp. as the producer of a related compound, pederin (Figure 1, 169) in the beetle *Paederus fuscipes* [55]. Pederin and onnamide A are part of a large family of structurally related cytotoxic polyketides, that include the potent mycalamides (Figure 1, 171) isolated from the New Zealand marine sponge *Mycale hentscheli* [41,56]. Difficulty in sustainably producing pederin-like compounds has hindered activity testing and thus also their possible development into clinical applications [57-59]. While to this date, the majority of sponge natural products have not been linked to bacterial producers, this is a key step in being able to produce desired secondary metabolites in a sustainable and scalable fashion [33,60,61]. Entotheonellaeota bacteria are a prominent example of sponge-associated natural product producers and have been shown to produce pederin-like polyketides, post-translationally modified ribosomal peptides, non-ribosomal peptides and more in *T. swinhoei* [29,33,62-64]. These filamentous bacteria, which have been likened to soil Actinomycetes in terms of their secondary metabolite richness and diversity, belong to the proposed Candidate phylum Tectomicrobia, members of which appear to be present in other marine sponges [29,30,63,64].

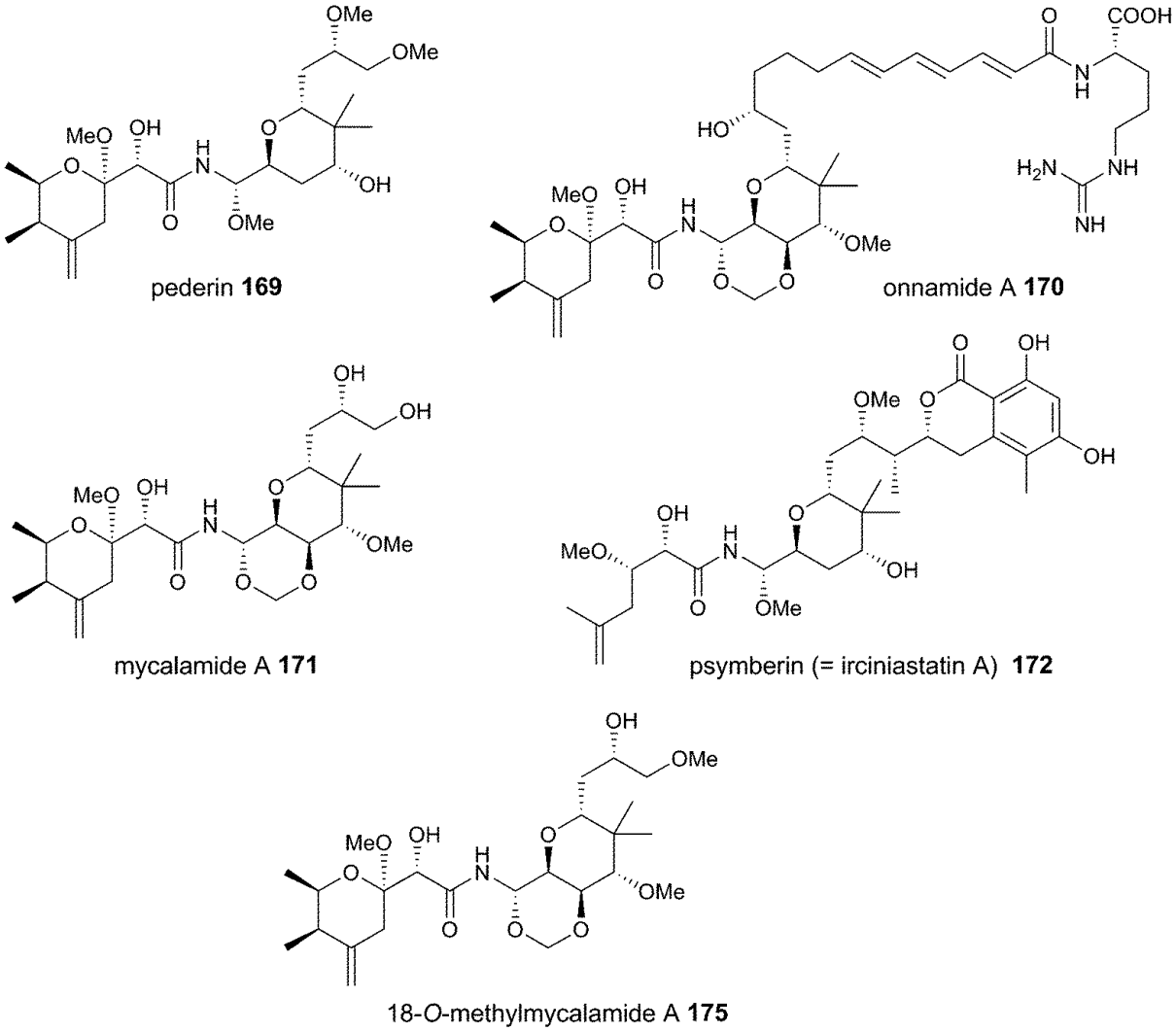


Figure 1: Pederin-like compounds Reprinted from [56] with permission from The Royal Society of Chemistry (License number 4622180191500). Shows the chemical structures of some pederin-like polyketides including pederin (169), onnamide A (170), mycalamide A (171), psymberin (172) and 18-O-methylmycalamide A (175).

Extensive studies have been carried out on Entotheonellaeota bacteria in the hope of one day being able to culture them [29,64]. Metaproteomic data suggests that methanol may be their primary carbon source, whilst metagenomic data suggests they are able to break down various complex organic molecules including acids, alcohols, polysaccharides and aromatics [29]. Classical virulence and communication factors could not be detected in these bacteria and it has been suggested that newly identified secondary metabolite gene clusters may be involved in communication and maintenance of symbiosis [29]. Metabolic function and compounds involved in host-symbiont communication are crucial considerations to make when attempting to culture symbiotic bacteria [29,30,64,65].

Direct culturing has also proven very difficult for less studied symbiotic bacteria but a few success stories do exist and these have yielded compounds with a range of bioactivities, e.g. antibacterial, cytotoxic, antifungal, anti-malarial [10,52,62,66-70]. Aquaculture of marine sponges themselves for chemical extraction of natural products has also proven non-viable due to environmental factors affecting sponge growth and generally low yields [10,71,72]. This variability in secondary metabolite production using culture-based approaches is likely caused by environmental factors, e.g. nutrient levels, or culture conditions affecting bacterial growth and gene expression [10,29,65].

1.2 Bacterial natural product biosynthesis

Bacterial natural products are typically produced by machinery encoded in biosynthetic gene clusters (BGCs). BGCs are collections of co-localised genes encoding most or all of the biosynthetic enzymes, regulatory genes, resistance elements and transporters required for compound production [73-75]. Sizes of up to 150 kb have been reported for some of these BGCs [76]. In the case of non-ribosomal peptide synthetases (NRPSs) and polyketide synthases (PKSs), BGCs are often colinear with their products, i.e. the order of genes in a BGC matches the order of subunits incorporated into the final product [75,77,78]. There are many classes of bacterial natural products, however NRPSs, PKSs and ribosomally synthesized and post-translationally modified peptides (RiPPs) have proven to be particularly bioactive [3,6,10,56] and their BGCs are frequently identified in abundance from sponge-associated bacteria as well as other marine bacteria [7,12,24,29,41,48,52,79].

1.2.1 Non-ribosomal peptide synthetases and polyketide synthases

Non-ribosomal peptides (NRPs) and polyketides (PKs), are produced by large modular enzymes (NRPSs, PKSs respectively) that construct molecules in a modular fashion from monomer building blocks [77,78,80,81]. NRPSs assemble small peptide molecules from individual proteinogenic, modified or non-proteinogenic amino acid monomers using at least an adenylation (A) domain and a peptidyl carrier protein (PCP, also called a thiolation domain) in each module (Figure 2; [78,80]). The A-domain recognizes the substrate and tethers the substrate to the 4'-phosphopantetheine (PP) prosthetic group of the downstream PCP-domain [78,80]. Phosphopantetheinyl

transferases (PPTases) are present in NRPSs and PKSs and are the enzymes responsible for attaching the PP-prosthetic group to the PCP or acyl carrier protein (ACP) respectively [80,81]. A peptide bond is then formed by the C-domain between the growing peptide chain tethered to the PCP immediately upstream and the substrate tethered to the PCP [78,80]. Elongating modules may contain added enzymatic domains, e.g. epimerization, methyltransferase, glycosylation, oxidation, cyclisation, etc., which modify the core monomers [78,80]. The final module then ends with a catalytic domain, usually a thioesterase (TE) domain, that causes the release of the final product, often via macrocyclisation between an internal nucleophilic group and the terminal enzyme-linked thioester moiety [78,80].

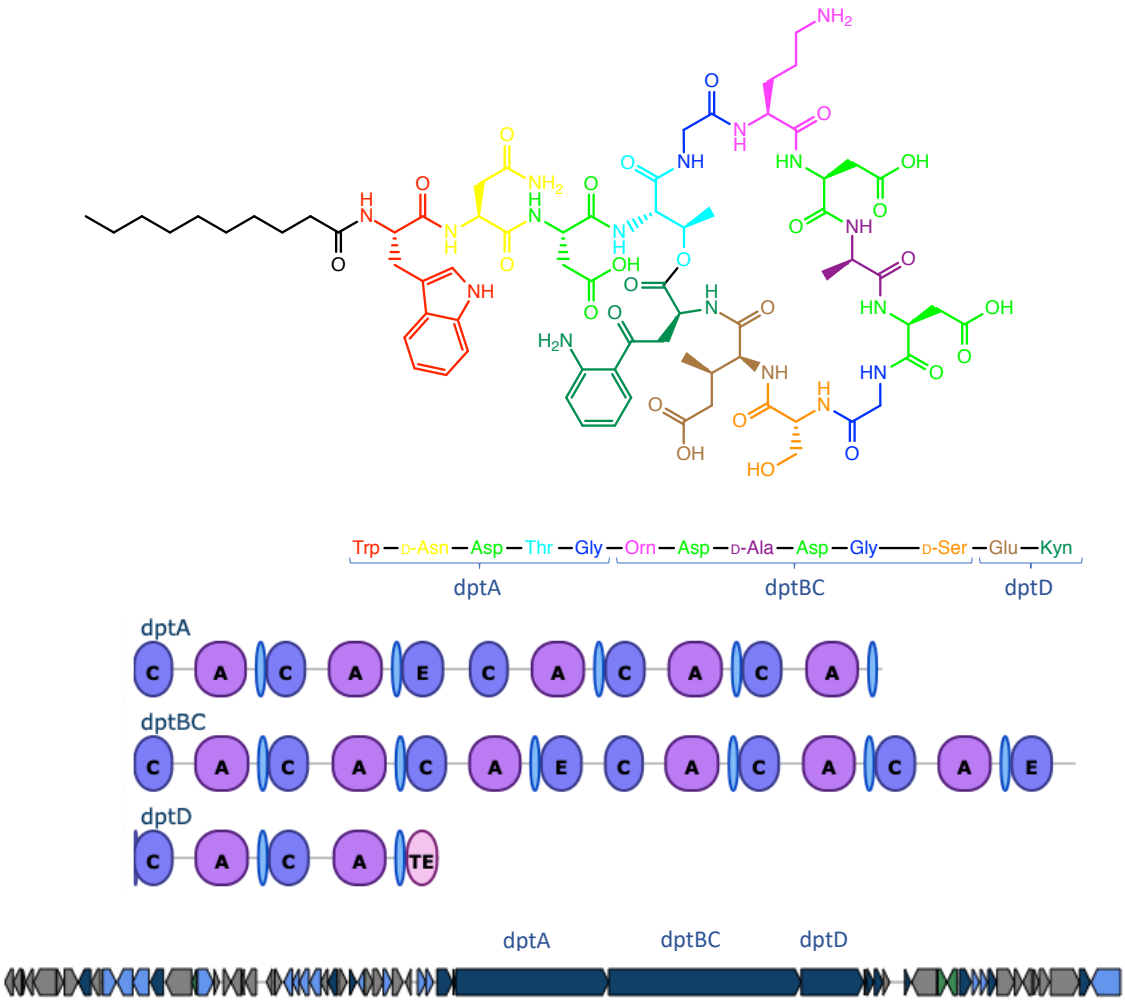


Figure 2: Daptomycin biosynthesis Daptomycin is a so called “last-resort” antibiotic and NRP natural product used to treat drug-resistant bacterial infections (Miao et al. 2005, Pader & Edwards 2017). Figure was drawn and colored with ChemDraw according to AA-sequence. NRPS domains and corresponding BGC organization from MiBIG (BGC0000336). C = Condensation domain. A = Adenylation domain. E = Epimerisation domain. TE = Thioesterase domain. Blue oval = PCP-domain.

PKSs are usually distinguished into three types (I - III) based on their size, domain organization and function. In Type I PKSs the acyltransferase (AT) domain recognizes acyl monomers and, similar to NRPSs, tethers it to the PP-prosthetic group on the ACP [77,81]. The ketosynthase (KS) domain then catalyses a Claisen condensation between the substrate and the growing polyketide chain [77,81]. Several modifying domains, e.g. ketoreductase (KR), dehydration (DH), methyltransferase (MT) and more, may be present altering the structure of the usually simple acyl monomers [77,81]. A distinction has to be made between *cis*-AT and *trans*-AT Type I modular PKSs with the former containing an AT-domain in each module and the latter having an AT-domain as well as potentially other domains acting in trans to several modules [56,77,81]. These *trans*-AT PKSs generally speaking do not follow the collinearity typically observed in "normal" cis-AT PKSs but may account for up to 38% of all bacterial PKSs including the pederin-like polyketides mentioned above [56,82].

Type II PKSs are iterative rather than modular and have a characteristic heterodimeric KS-domain [83]. This heterodimeric KS-domain is composed of a catalytic KS domain (KSα) carrying out the Claisen condensation of monomers and a chain length factor (KS β) functioning in determining polyketide chain length [83-85]. It is not well understood how the heterodimeric KS-domain interacts with other domains in the pathway to produce the final PK [83]. Contrary to Type II PKSs, the KS-domain in Type III PKSs is homodimeric but can create numerous complex cyclic and linear polyketide molecules by relaxed substrate specificity, different cyclisation mechanisms and/or variability in iteration number [86,87].

1.2.2 Ribosomally synthesized and post-translationally modified peptides

RiPPs are a very diverse class of natural product, containing numerous subclasses [88-90]. Synthesis of these compounds starts with the production of a precursor peptide consisting of joined leader and core peptides. This precursor peptide is produced by ordinary ribosomal activity but subsequently modified by tailoring enzymes and cleaved to generate the mature natural product [88,91,92]. The leader peptide functions in transport and mediation of the post-translational modification(s) and is usually removed by the action of a peptidase [88,91,92]. The recent development of specialized bioinformatic algorithms has led to an explosion in the number of known RiPP BGCs,

however most of these have yet to be linked to their product molecules by functional studies [91,92].

Lassopeptides are a class of RiPP that are particularly resistant to heat and proteases with a range of identified biological activities making them attractive targets of natural product discovery [92,93]. They are small peptides with an N-terminal macrolactam formed by condensation between the amino and side-chain carboxylate group of glutamine or aspartic acid, which locks the C-terminal tail within the macrocycle and forms the characteristic lasso-fold [88,94]. Lassopeptide BGCs usually contain a lasso cyclase, which is homologous to an asparagine synthetase, and a leader peptidase, which is homologous to a transglutaminase [92].

Thiopeptides are an important subclass of RiPPs as they include compounds with nanomolar antibacterial activity against drug-resistant strains [88,91]. They typically inhibit protein synthesis and several thiopeptides are already in commercial development and production, in particular in the livestock industry [91]. Structurally characteristic is the central six-membered nitrogen-containing ring, usually pyridine but also oxidized to piperidine or dehydropiperidine, which is formed by a trimeric heterocycle synthetase, a split LanB-like dehydratase and a [4+2] cycloaddition enzyme [91].

Lanthipeptides are structurally diverse, widely distributed and have a range of biological activities and potential therapeutic applications [95]. They contain a characteristic lanthionine or methyllanthionine group, which is formed by dehydration of serine or threonine into 2,3-didehydroalanine (Dha) or 2,3-didehydrobutyrine (Dhb) respectively, followed by nucleophilic attack by a cysteine residue to form the thioether linkage [88,95]. Four classes of lanthipeptides exist to date, based on the enzyme(s) responsible for forming the lanthionine or methyllanthionine group [88]. Other noteworthy RiPP subclasses, whose BGCs have been investigated in detail include the bottromycins [96], cyanobactins [97], linaridins [98], microcins [99] and proteusins [62,100].

1.3 Metagenomics

While culturing marine bacteria is a pathway to isolate natural products and identify BGCs, the vast majority (~99%) cannot be grown in the laboratory [3,101-103]. Metagenomics, being the study of environmental DNA (eDNA), allows the study of the unculturable majority of bacteria by sequencing DNA directly isolated from the environment [54,104-106]. When used in combination with synthetic biology approaches such as heterologous expression, metagenomics is a powerful method for the discovery and sustainable supply of new natural products (Figure 3) [27,54,56,79,107]. Applications of metagenomics however, far exceed natural product discovery and include genome recovery, functional studies in symbioses or other contexts, phylogenetic studies and human health studies among many others [29,33,108-113].

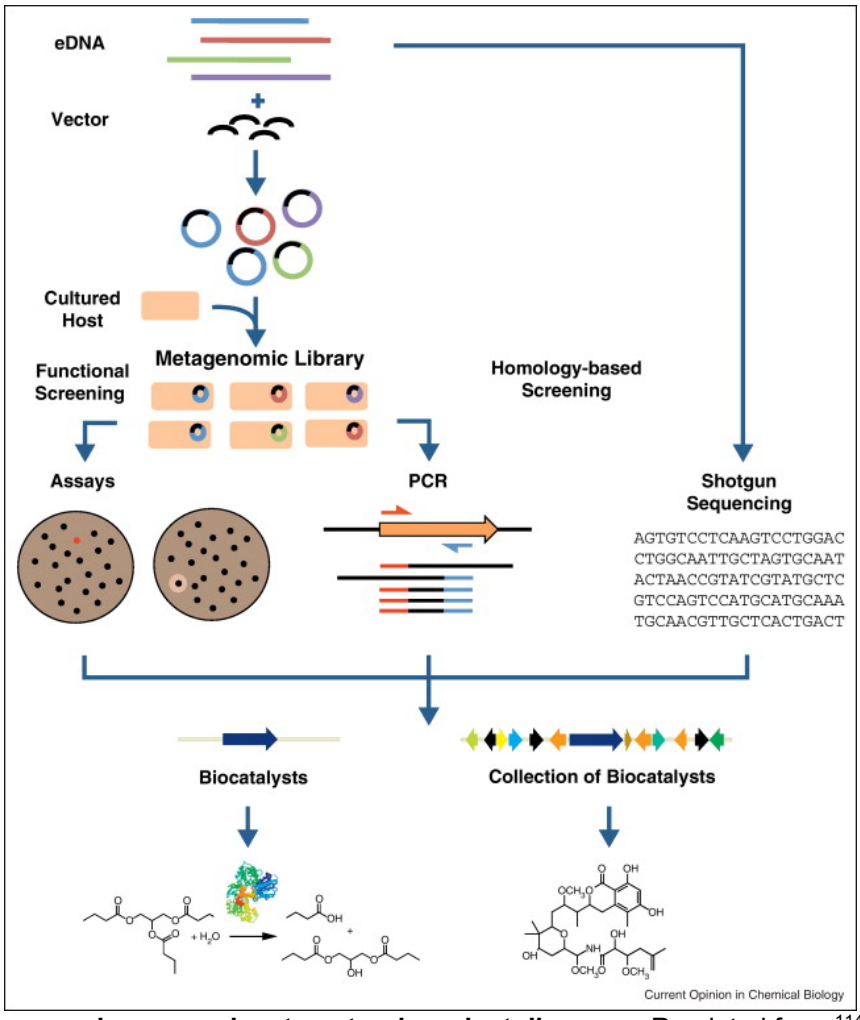


Figure 3: Metagenomic approaches to natural product discovery Reprinted from [114] with permission from Elsevier (License number 4622160783893). Shows the general library construction process on the left, followed by PCR-based screening to identify genes related to natural product/biocatalyst biosynthesis or functional screening to directly identify expression of natural products/biocatalysts. Shotgun sequencing is shown on the right as an alternative route to discover BGCs and natural products.

1.3.1 Metagenomic cosmid libraries

Metagenomic library construction bypasses the challenges of cultivation and direct chemical extracts by cloning environmental DNA into a culturable host (Figure 3), allowing heterologous expression of encoded natural products and the indefinite storage of genomic diversity [50,54,79,105,106]. Cosmid libraries are advantageous over other metagenomic libraries as they allow a large insert size as well as high copy numbers [105,106,108,115]. If the number of unique insert sequences is sufficiently high, it is theoretically possible to redundantly cover an entire metagenome, allowing recovery of even the largest BGCs as sets of overlapping cosmid clones [106,115]. Cosmid cloning uses lambda phage heads, which recognize the cos-site in a vector backbone, selectively package cloned DNA inserts and transfect them into a host cell [116,117]. This allows high efficiency cloning and is selective for large (30-45 kb) DNA inserts [105]. There are numerous cosmid vectors that can be used to construct metagenomic libraries, however one of the most commonly used is pWEB-TNC (Epicenter; Figure 4) because it is readily transfected, compatible with large inserts (30-45 kb) and stably maintained at high copy numbers (5-50). This vector has already been used successfully in numerous metagenomic library construction studies [105,106,118].

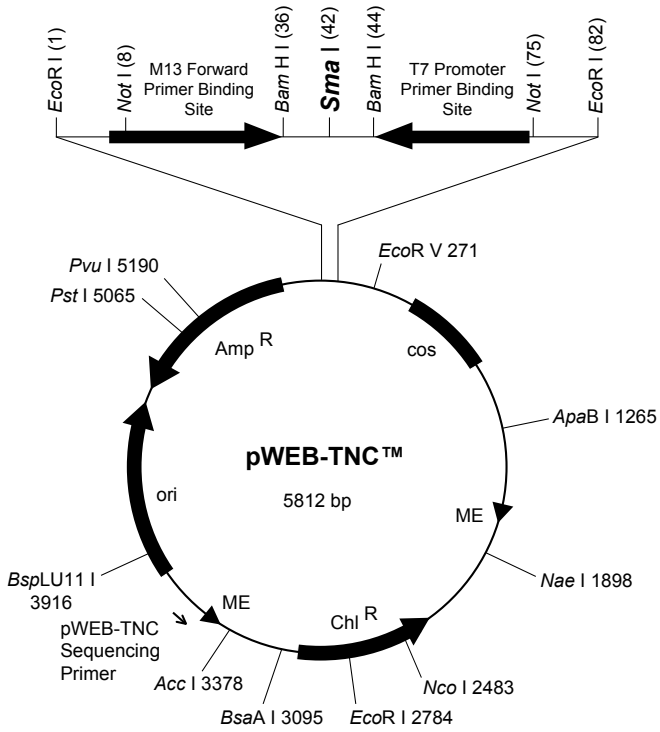


Figure 4: Cosmid map of pWEB-TNC (http://www.epibio.com/docs/default-source/protocols/pweb-tnc-cosmid-cloning-kit.pdf)

Cosmid libraries are often screened for BGCs by PCR or direct functional screening [48,50,54,105,106]. PCR screening involves the use of degenerate primers targeting conserved motifs within key biosynthetic genes, e.g. KS-domains for PKs and A-domains for NRPs, and can be more or less targeted depending on primer choice [18,41,48,55,106,119]. For example in a landmark study by the Piel group, the genome of an uncultivated bacterial symbiont of the *Paederus fuscipes* beetle was captured as a cosmid library that was then screened by degenerate PCR to recover clones containing putative PKSs [55]. Subsequent sequencing and annotation of these clones allowed identification of the BGC responsible for the production of pederin and showed that this natural product is produced by an uncultivated bacterial symbiont of the genus *Pseudomonas* [55]. Activity guided or functional screening is more high throughput and leads to quicker identification of relevant cosmids but is less targeted [105,120,121].

1.3.2 Metagenomic sequencing and genome binning

Since the seminal paper on genome-resolved metagenomics from the Banfield lab[122], the development of sequencing technology and bioinformatic algorithms (in particular short read assembly) have led to an explosion in sequencing projects [103,104,123,124]. Large-scale sequencing, assembly and genome binning studies have demonstrated the feasibility of recovering near-complete or even complete genomes from metagenomes, so called metagenome assembled genomes (MAGs) [125-129].

1.3.3 Metagenomic sequencing studies of marine sponges

Sponges are sessile filter-feeding organisms that harbour microbiomes of variable complexities [37,65,130]. Compared to soils and planktonic ocean communities, these are generally of lesser complexity [37,65,130]. Bacterial biomass varies between different sponges to the extent that numerous papers distinguish them as low microbial abundance (LMA) and high microbial abundance (HMA) sponges [61,65,131]. As many as 72 different bacterial phyla and candidate phyla have been identified from marine sponges by 16S rRNA amplicon-sequencing, with some of the most diverse phyla being Chloroflexi, Proteobacteria, Actinobacteria and the Candidate Phylum “Poribacteria” [61,65,110,111]. Most taxonomic groups previously thought to be exclusive to sponges, including members of the *Nitrospira*, Acidobacteria, Gemmatimonadetes and Poribacteria, have also been detected in seawater but often at significantly lower abundances [130,132,133]. Sponge biology and sponge microbiome studies have mostly

been based on 16S rRNA amplicon diversity [61,65,132]. While 16S rRNA sequencing is an economical method to determine community composition, it is prone to amplification bias and not adequate to quantify species abundances [10,123,132,134]. The filter-feeding lifestyle of sponges and the high amounts of seawater contained in or passing through sponges also make it difficult to discern symbiotic microbes from transient microbes when employing amplicon sequencing [131]. Disparity appears to exist between older 16S rRNA studies suggesting no correlation between sponge phylogeny and microbial community [61] and newer studies using next-generation sequencing suggesting that species-specific microbial communities exist in certain taxa [135,136]. Direct metagenomic shotgun sequencing is an alternative method that provides a less biased view of community structure as it requires relatively high coverage, and thus high microbial abundances, making it less prone to assemble the genomes of transient microbes [137]. With new assembly algorithms it is very tractable to assemble and identify genomes from microbial communities of intermediate complexity, such as sponges (Figure 5; [29,37,132,138]). While some marine sponge metagenomes have been examined and published [29,112,135,137-142], few metagenomic sequencing studies have investigated the secondary metabolite diversity, or attempted to link this to function of the sponge holobiont.

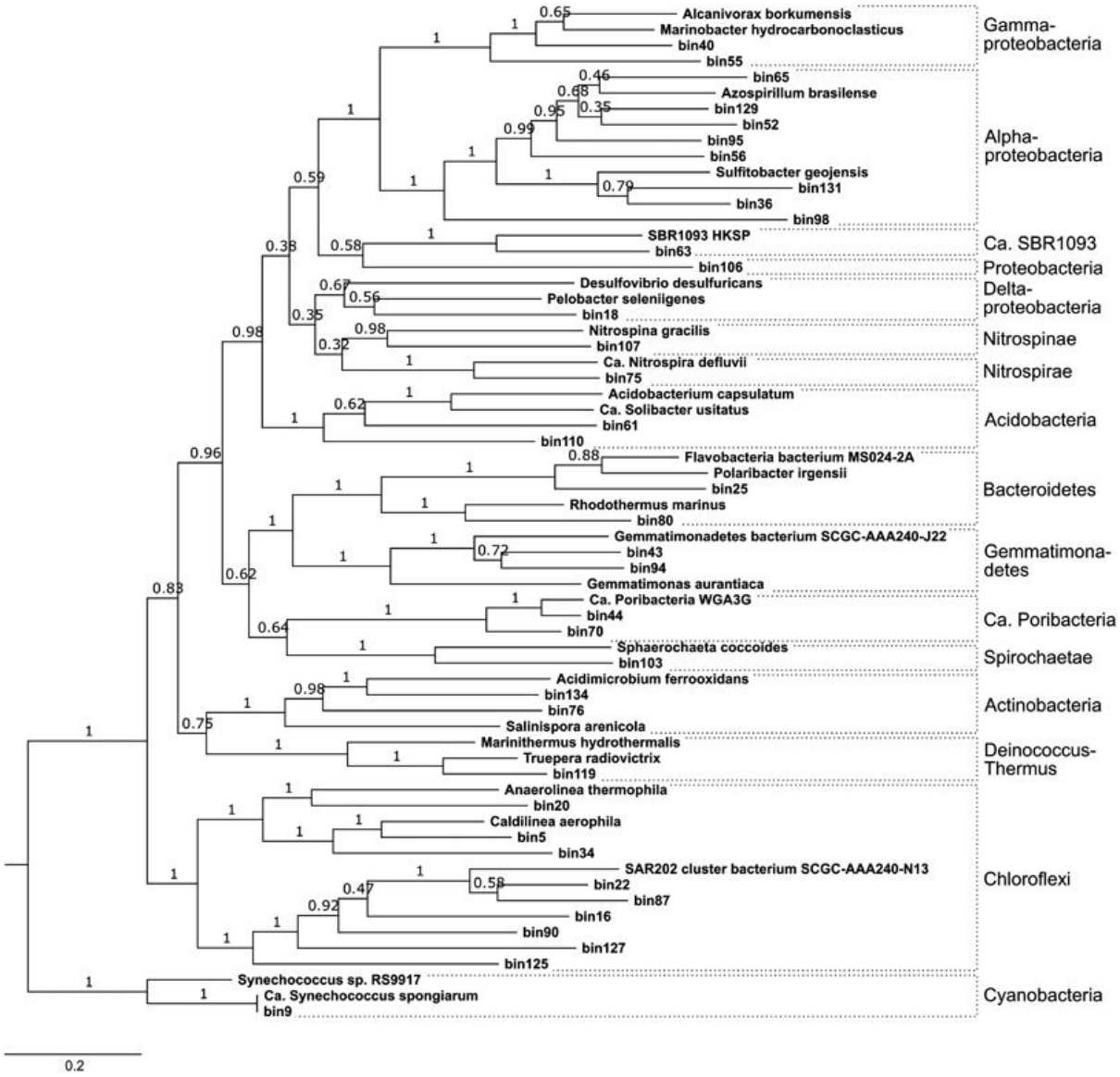


Figure 5: Maximum likelihood phylogenetic tree of 37 MAGs recovered from the metagenome of *Aplysina aerophoba*. Chloroflexi and Proteobacteria were the most abundant and diverse phyla identified in this analysis. Reproduced from [138] with permission under the Creative Commons Attribution 4.0 International License (https://creativecommons.org/licenses/by/4.0/)

1.3.4 Metagenomic cosmid library construction from low biomass environments

Large insert metagenomic cosmid libraries are typically constructed from high biomass environments, as these readily yield large quantities of HMW DNA, e.g. soil [105,115], the human microbiome [108] and marine sponges [33,50]. However, cosmid cloning can be a challenging process and methods for low biomass marine environments are currently lacking, preventing the detailed study of these environments. Seawater is a low biomass environment and planktonic bacteria have been extensively studied in some

parts of the world, in particular in the Northern Hemisphere [143-146], but very little around Antarctica and New Zealand [146]. The Antarctic prokaryote world in particular remains largely under-explored with high prospective bacterial species and chemical diversity [4,8,147-150]. Existing natural product studies are few and have mostly focused on soil, in particular from the McMurdo Dry valleys [151-155] and to a lesser extent the glacial marine environment [156-159]. Sea ice appears to be yet underexplored for natural product chemistry [149] but promises to be very interesting as it is a very heterogenous environment with a variety of stressors and diverse and unusual bacteria [160-165].

1.4 Overview of key bioinformatics used in this thesis

Metagenomic assembly and binning algorithms are essential for attributing certain metabolic functions (including secondary metabolite production) to bacteria within complex uncultivated communities. The assembly of reads from metagenomic samples into contigs followed by binning of these contigs using two or more binning algorithms and consolidation of these bins into one non-redundant set of bins is currently considered the gold standard to recover MAGs [166-171]. Unsurprisingly, a high-quality metagenomic assembly is crucial for accurate binning and consequent MAG recovery [166]. Metagenomic assemblies are extremely challenging due to variability in abundance and thus coverage of different bacteria, repetitive and/or conserved regions of the genomes of different bacterial species as well as microdiversity, i.e. a mixture of different strains in a sample [166,170,172]. Some of the most notable metagenomic assembly algorithms include IDBA_UD [172], Ray Meta [173], MetaVelvet [174], MEGAHIT [175] and metaSPAdes [166].

It should be noted that it is inherently difficult to compare metagenomic assemblers due to a lack of reference metagenomes; thus synthetic metagenomes made up of mixtures of known genomes are often used as an approximation [166]. IDBA was the first assembler to use an iterative de Bruijn graph approach, which iterates from a minimum k-mer to a maximum k-mer using contigs generated from one iteration in the graph construction of the next iteration [176]. This approach was taken over into the metagenomic version IDBA-UD where at each step a variable depth cut-off threshold (based on coverage of neighbouring contigs) is used to remove low depth contigs thought to be of different origin [172]. Paired-end reads are aligned and used to construct

a local assembly, which at high coverage can help resolve branches and coverage gaps [172]. SPAdes [177] was originally developed for single-cell sequencing data, which typically contains extremely variable coverages due to PCR amplification biases, but has also been successfully applied for metagenomic sequencing projects [166]. While it can assemble the metagenomes of simple communities it is not adequate for medium to highly complex communities [166]. The metagenomic version, metaSPAdes, performed better on complex communities in terms of the scaffold lengths constructed (in particular the 1000 longest scaffolds), the number of genes that could be predicted from the data and the number of read pairs that could be aligned with the final contigs (>1 kb) as demonstrated by the assembly of a complex marine and a soil dataset [166]. MetaSPAdes constructs de Bruijn graphs for different k-mer sizes (similar to IDBA-UD) but uses k-bimers, which includes the k-mers from both reads as well as the estimated genomic distance between them, for graph simplification and saves all the modifications made to the graph [166]. Like IDBA-UD it calculates coverage of adjacent edges during graph resolution, to determine coverage ratios and if an edge is defined as weak (≥10-fold lower coverage than its adjacent vertices), it is disconnected rather than removed as in IDBA-UD [166]. In an attempt to address the microdiversity challenge metaSPAdes masks strain variation by identifying characteristic types of edge topologies (so called “filigree edges”), e.g. those that have one high coverage and one low coverage path, and building a consensus assembly, which is then later resolved into different contigs/strains [166]. During read error-correction, estimated genomic distances between k-bimers are adjusted and used for the construction of a paired assembly graph (based on the principle of a paired de-Bruijn graph) from which the final contigs are derived [166]. Based on this algorithmic design and prevalence in recent literature, metaSPAdes therefore appears to be the best metagenomic assembly algorithm at this time [166,170,171,178].

Following metagenomic assembly, BGCs need to be identified using specialized algorithms, such as the commonly used antiSMASH [52,73,179,180]. In antiSMASH, core BGC enzymes are identified using the HMMer algorithm [181] and custom profile Hidden Markov Models (pHMMs), which are multiple sequence alignments with attached probabilities that allow comparison to position-specific scoring matrices created from the query sequence [73,180]. Version 4 of antiSMASH then uses custom cluster rules to identify BGCs from co-localized core enzymes, identified from the pHMM analysis,

and assigns BGCs to one of 45 different secondary metabolite classes. Web server analysis is slow and limited to 1000 contigs at a time but this can be deactivated in standalone installs [73,180]. While structure prediction is possible in antiSMASH [73,180] or using software such as PRISM [182] it is not yet very accurate [79,183,184]. Sequencing studies and the use of bioinformatic algorithms have resulted in the creation of large (>700,000) databases of verified or putative BGCs [185]. This demonstrates that sequencing studies and bioinformatics allow high-throughput and relatively inexpensive natural discovery BGC discovery [73,103,182,186-188], which can be coupled with metabolomics and synthetic biology to study BGC expression [79,189-192] .

1.5 Aims

There is a strong need for interdisciplinary research on how natural products shape ecosystem or interactions between different species. For example, secondary metabolites produced by sponge symbionts and their role in this symbiosis have been shown to act as a chemical defense only in a few select cases [33,65,193]. Linking secondary metabolites to their bacterial producers using metagenomic sequencing allows targeted studies of their biology and function as well as setting the platform for the production of these compounds and discovery of new drug candidates. To this end, the following three aims were formulated for the scope of this Master's thesis.

1) Investigate the feasibility of constructing large insert cosmid libraries for environments that yield low quantities of HMW DNA and optimize protocols for the construction of such libraries.

2) Develop a shotgun sequencing, assembly and genome recovery workflow to allow microbiome analysis at the genome level as well as construct a large insert cosmid library to investigate the secondary metabolism of *Cacospongia mycofijiensis*.

3) Apply the shotgun sequencing, assembly, genome recovery and secondary metabolite analysis workflow to five further Tongan sponges, including *C. mycofijiensis* collected from a different location to compare microbiome composition and secondary metabolism of different marine sponges from the same location as well as the same marine sponge from two different locations.

Chapter 2

Materials and methods

2.1 Sample collection

2.1.1 New Zealand seawater samples

New Zealand seawater samples were collected by Federico Baltar (PhD). Sampling took place during the 2017 season of the Aotearoa New Zealand Ross Ice Shelf Programme at (HWD-2B) (Latitude -80.65767, Longitude 174.46263). Seawater was collected from 3 depths (400 m, 550 m, and 700 m), with each sample consisting of approximately 200 litres collected over a cast of ~2h. A McLane WTS-LV-Bore Hole filter pump fitted with a 142 mm, 0.22 µm filter (Supor membrane filters, Pall Corporation) was used to collect microbes *in situ*. Sampled filters were placed in cryovials and frozen.

2.1.2 Antarctic samples

Collection of Antarctic samples was carried out by Eileen Koh during Antarctic expeditions in 2007, 2008 and 2010 as part of her PhD-studies at Victoria University of Wellington [194]. For the "biomass survey in glycerol" samples from Terra Nova Bay, Antarctica in 2007, a 1 cm * 1 cm *10 cm block of sea ice from the bottom of an ice core was melted in 125 ml autoclaved 0.22 µm-filter-sterilized seawater in a sterilized black plastic box overnight to yield the original sample volume indicated in Appendix A6. 500 µl of these melted sea ice cores was taken and mixed with an equal amount of 50% glycerol to make up the "biomass survey in glycerol" samples. Samples were stored in liquid nitrogen for transportation. This process was repeated for the "biomass survey in glycerol" samples collected in 2008 from Granite Harbour, Antarctica. Seawater samples at Terra Nova Bay in 2007 were collected using a system similar to a Niskin bottle, where a sampling bottle with a bung on each end is lowered into the water on a wire rope and upon reaching a known depth a brass weight is dropped down triggering the bung to close. 500 µl of these samples was taken and mixed with an equal amount of 50% glycerol to yield the "seawater in glycerol" samples.

The site at Terra Nova Bay, Antarctica, where brine samples were collected on the 24/11/08, had an estimated ice thickness of 1.9 m plus approximately 30 cm snow cover. A core was drilled to a depth of 45 cm allowing the brine water to drain into the hole for approximately 30 minute and collecting the sample with the volume indicated in Appendix A6. The same hole was then used to drill to a depth of 65 cm and 85 cm

collecting the drained brine at each depth. 500 µl of these samples was taken and mixed with an equal amount of 50% glycerol to yield the brine samples in glycerol. Salinities observed for the 45 cm, 65 cm and 85 cm samples, were 54%, 44% and 82% respectively. Samples were stored in liquid nitrogen for transportation.

Sample collection for sea ice filters was carried out at Cape Evans, Antarctica in 2010. 5 cm * 5 cm * 10 cm sections of an ice core were melted in approximately 600 ml autoclaved 0.22 µm-filter-sterilized seawater in a sterilized black plastic box overnight. Subsamples from four of these section were combined to give a Top, Middle and Bottom sample with the original sample volume indicated in Appendix A6, which was then filtered through a 0.22 µm MCE filter using an electric pump with pressure set to 50 mmHg. Resulting Top, Middle and Bottom filters were stored in a zip lock plastic bag at -20°C for transportation.

2.1.3 Sponge samples

Sponge samples were collected from Pete's cave and Cathedral cave in 'Eua, Tonga by a team from the School of Chemical and Physical Sciences, Victoria University of Wellington, led by Rob Keyzers (PhD). Pieces of sponge were collected by divers using their hands and knives and were then stored in individual resealable zipper storage bags at -20°C.

2.2 DNA extraction from seawater and sea ice samples

Various DNA extraction protocols were developed and compared for use on seawater and sea ice sample. Each has been given a unique number, which is referred to in the remainder of the text.

2.2.1

Cells were pelleted by centrifugation (4,000 g, 4°C, 15 min). The cell pellet was then resuspended in 1 ml soil lysis buffer (2% SDS [wt/vol], 1% CTAB [wt/vol.], 100 mM Tris-HCl, 100 mM EDTA, 1.5 M NaCl, [pH=8.0]) [106] and incubated (70°C, 2 h) with occasional inversion to mix. After letting the sample cool to approximately 37°C, Proteinase K was added to a working concentration of 200 µg/ml and the sample incubated (37°C, 1 h). Large particulates were pelleted by centrifugation (4,000 g,

room temperature, 30 min) and the supernatant added to a fresh microcentrifuge tube. 0.6x vol room-temperature 100% isopropanol was added to precipitate DNA. DNA was then pelleted by centrifugation (17,000 g, 4°C, 30 min). The DNA pellet was washed with 70% ethanol, taking care not to disturb the pellet. DNA was eluted into 50 µl of TE (10 mM Tris, 1 mM EDTA [pH = 8.0]) by incubating samples at room temperature for 1 hour.

2.2.2

Filter paper was aseptically cut into small pieces, suspended in 3 ml of 100 mM Tris-Cl [pH=8.0] and vortexed for 5 minutes. Lysozyme solution was added to a final concentration of 1 mg/ml followed by incubation (37°C, 30 min). In order to make up the soil lysis buffer (2% SDS [wt/vol], 1% CTAB [wt/vol.], 100 mM Tris-HCl, 100 mM EDTA, 1.5 M NaCl, pH=7.5-8.0) (Owen et al. 2013) missing reagents were added from stock solutions to a final volume of 5ml and the sample incubated (70°C, 2 h). After letting the sample cool to approximately 50°C, Proteinase K was added to a final concentration of 500 µg/ml, $CaCl_2$ to concentration of 1-5 mM and the sample incubated (50°C, 1 h). Cell debris was pelleted by centrifugation (4,000 g, room temperature, 30 min) and the supernatant aliquoted into fresh microcentrifuge tubes. DNA was precipitated by adding 0.1x vol 3 M NaOAc [pH=5.2] and 1x vol cold (-20°C) isopropanol and transferring samples to -20°C freezer for 30 minutes. DNA was then pelleted by centrifugation (15,000 g, 4°C, 20 min). The DNA pellet was washed twice with 70% ethanol, taking care not to disturb the pellet, and air dried for approximately 5 minutes. DNA was eluted in 30 µl of TE (10 mM Tris, 1 mM EDTA [pH=8.0]) by incubating samples at room temperature for 1 hour.

2.2.3

This protocol was modified from Rebecca Cowie's PhD thesis [195]. Filter paper was aseptically cut into small pieces, suspended in 1ml of Enzymatic Lysis Buffer [40 mM EDTA, 50 mM Tris-Cl [pH=7.4], 0.75 M sucrose, 15% Tween 80 and vortexed for 5 minutes or cells were pelleted by centrifugation (4,000 g, 4°C, 15 min) and resuspended in in 1 ml of Enzymatic Lysis Buffer [40 mM EDTA, 50 mM Tris-Cl [pH=7.4], 0.75 M sucrose, 15% Tween 80]. Lysozyme was added to a final concentration of 2 mg/ml and samples incubated (37°C, 1 h). SDS was then added to a final concentration of 1% [wt/vol], Proteinase K to a final concentration of 1 mg/ml

and samples incubated (55°C, 2 h). 1x vol phenol:chloroform (1:1) was then added followed by gentle inversion until an emulsion formed. Phases were separated by centrifugation (12,000 g, room temperature, 3 min) and the upper (aqueous) phase was transferred to a fresh 15 ml falcon tube. This phenol:chloroform wash was repeated once more. 0.1x vol 3 M NaOAc [pH=5.2] and 2x vol 100% isopropanol were then added and samples transferred to -20°C freezer for 2 hours to precipitate DNA. The DNA was then pelleted by centrifugation (15,000 g, 4°C, 30 min) and the DNA pellet washed twice with 70% ethanol, taking care not to disturb pellet. After letting the pellet air dry for approximately 5 minutes, DNA was eluted in 20 µl of TE (10 mM Tris, 1 mM EDTA [pH=8.0]) by incubating samples at room temperature for 1 hour.

2.2.4

This protocol was modified from Streit & Daniel 2017 [196]. 20 µl Proteinase K (20 mg/ml) and 200 µl lysozyme (50 mg/ml) were added to samples suspended in 2.7 ml of seawater lysis buffer (1% CTAB, 1.5 M NaCl, 100 mM sodium phosphate, 100 mM EDTA, 100 mM Tris-HCl, [pH=8.0]) followed by incubation (37°C, 30 min). 3 µl of 100 µg/ml RNase A was then added followed by incubated (37°C, 1 h). Next, 300 µl 20%SDS was added and samples incubated for further lysis (65°C, 2 h). 1x vol phenol:chloroform (1:1) was then added followed by gentle inversion until an emulsion formed. Phases were separated by centrifugation (12,000 g, room temperature, 3 min) and the upper (aqueous) phase was transferred to a fresh 15 ml falcon tube. This phenol:chloroform wash was repeated once more. The lower organic layer was collected from the washes, 500 µl of recovery buffer (1x TE, 0.5 M NaCl) added and the solution mixed by gentle inversion. Phases were again separated by centrifugation (12,000 g, room temperature, 3 min) and the upper (aqueous) phase was transferred to a fresh 15 ml falcon tube. DNA was precipitated by adding 0.7x vol cold (-20°C) isopropanol and pelleted by centrifugation (16,000 g, 4°C, 30 min). The DNA pellet was washed once with 70% ethanol and eluted in 25 µl TE (10 mM Tris, 1 mM EDTA [pH=8.0]) by incubating at room temperature overnight.

2.2.5

Filter paper was aseptically cut into small pieces and suspended in 5ml sponge lysis buffer (8 M Urea, 2% Sodium Lauroyl Sarcosinate (sarkosyl), 1 M NaCl, 50 mM EDTA, 50 mM Tris-Cl, pH=7.5 [197]). The sample was then incubated (65°C, 10 min), gently

inverted three times and incubated further (65°C, 10 min). 1x vol phenol:chloroform (1:1) was then added followed by gentle inversion until an emulsion formed. Phases were separated by centrifugation (12,000 g, room temperature, 3 min) and the upper (aqueous) phase was transferred to a fresh 15 ml falcon tube. This phenol:chloroform wash was repeated once more. The lower organic layer was collected from the washes, 500 µl of recovery buffer (1x TE, 0.5 M NaCl) added and the solution mixed by gentle inversion. Phases were again separated by centrifugation (12,000 g, room temperature, 3 min) and the upper (aqueous) phase was transferred to a fresh 15 ml falcon tube. The sample was split into 700 µl subsamples and 0.1x vol 3 M NaOAc [pH=5.2] as well as 0.6x vol -20°C isopropanol were added to precipitate DNA. Next, DNA was pelleted by centrifugation (17,000 g, room temperature, 20 min) and the DNA pellet washed once with 70% ethanol. After letting the DNA pellet air-dry for 2 minutes, DNA was eluted in 50 µl TE (10 mM Tris, 1 mM EDTA [pH=8.0]) by incubating samples at room temperature overnight.

2.2.6

This protocol was taken from Rebecca Cowie's PhD thesis [195]. Filter paper was aseptically cut into small pieces, suspended in 1 ml of Enzymatic Lysis Buffer [40 mM EDTA, 50 mM Tris-Cl [pH=7.4], 0.75 M sucrose, 15% Tween 80] and vortexed for 5 minutes or cells were pelleted by centrifugation (4,000 g, 4°C, 15 min) and resuspended in in 1 ml of Enzymatic Lysis Buffer [40 mM EDTA, 50 mM Tris-Cl [pH=7.4], 0.75 M sucrose, 15% Tween 80]. Lysozyme was then added to a final concentration of 2 mg/ml and samples were incubated (37°C, overnight). The next morning SDS was added to a final concentration of 1% wt/vol, Proteinase K to a final concentration of 0.5 mg/ml and samples incubated (55 °C, 2 h). 1x vol phenol:chloroform (1:1) was then added followed by gentle inversion until an emulsion formed. Phases were separated by centrifugation (12,000 g, room temperature, 3 min) and the upper (aqueous) phase was transferred to a fresh 15 ml falcon tube. This phenol:chloroform wash was repeated once more. The lower organic layer was collected from the washes, 500 µl of recovery buffer (1x TE, 0.5 M NaCl) added and the solution mixed by gentle inversion. Phases were again separated by centrifugation (12,000 g, room temperature, 3 min) and the upper (aqueous) phase was transferred to a fresh 15 ml falcon tube. 0.25x vol 5 M NH_4OAc [pH=7.0] and 1x vol room-temperature 100 % isopropanol were added and samples transferred to a -20°C

freezer for 3 hours to precipitate DNA. Next, DNA was pelleted by centrifugation (16,000 g, 4°C, 30 min) and the pellet washed once using 70% ethanol. After letting the pellet air dry for approximately 2 minutes, DNA was eluted in 20 µl of TE (10 mM Tris, 1 mM EDTA [pH=8.0]) by incubating samples at room temperature overnight.

2.3 Size selection of HMW DNA

All DNA concentrations were quantified using the NanoPhotometer NP80 (Implen).

Classic: DNA was loaded using 6x purple loading dye (NEB) and run out on a 0.8% agarose TAE gel at 100 V for 1 h. After gel electrophoresis finished, strips wide enough to contain some of the DNA run out on the gel were cut from each side of the gel and stained in EtBr for 2 h. These strips (but not the middle section of the gel) were then visualised under UV-light and one strip was cut above and below the band of HMW DNA respectively. The trimmed strips were then realigned with the middle section of gel and a sterile ruler was used to horizontally cut the gel at the height indicated by the trimmed side strips, yielding a slice from the middle section of the gel that should contain the HMW DNA.

SYBR: DNA was loaded using 6x purple loading dye (NEB) and run out on a 0.8% agarose TAE gel stained with 1x SYBR (Thermo Fisher Scientific) at 100 V for 1 h. After gel electrophoresis finished, the gel was visualised under blue light and the band containing HMW DNA was carefully cut from the gel, trimming excess gel where possible.

2.3.1 – Gel electroelution blocks

Gel electrophoresis was carried out and a slice containing HMW DNA cut from the gel using the Classic or SYBR method described above. Electro-elution was carried out using the CB.S Scientific Electro-eluter/concentrator ECU-040-20 filled with 1x TAE and electro-elution blocks covered with dialysis tubing with MWCO 7,000 (Membra-Cel® MC18x100 CLR, Viskase Companies, USA). Electro-elution was run at 50 V for 16 h overnight. 0.1x vol 3 M NaOAc [pH=5.2]. A long glass pipette was used to aspirate approximately 1 ml TAE solution directly above the dialysis membrane of the electroelution blocks, putatively containing the HMW DNA. 1x vol room temperature

100% isopropanol were then added to precipitate DNA. DNA was pelleted by centrifugation (17,000 g, room temperature, 20 min), the DNA pellet washed using 70% ethanol and eluted in 20 µl of EB (10 mM Tris-Cl, pH 7.5) by incubating samples at room temperature for 3 hours.

2.3.2 – Freeze and squeeze

Gel electrophoresis was carried out and a slice containing HMW DNA cut from the gel using the Classic or SYBR method described above. The excised slice containing HMW DNA was then frozen at -80°C for 1 hour. Said gel slice was then inserted into the top section of a large sterile filter pipette tip with its bottom cut off, which in turn was inserted into a microcentrifuge tube. This setup was then centrifuged (17,000 g, room temperature, 30 sec) and the collected liquid transferred to a fresh microcentrifuge tube. Any remaining gel fragments were pelleted by centrifugation (17,000 g, room temperature, 5 min) and the supernatant transferred to a fresh microcentrifuge tube. 0.1x vol 3 M NaOAc [pH=5.2] and 1x vol room temperature 100% isopropanol were then added to precipitate DNA. DNA was pelleted by centrifugation (17,000 g, room temperature, 20 min), the DNA pellet washed using 70% ethanol and eluted in 20 µl of EB (10 mM Tris-Cl, pH 7.5) by incubating samples at room temperature for 3 hours.

2.3.3 – Gel electroelution clip set up

Gel electrophoresis was carried out and a slice containing HMW DNA cut from the gel using the Classic or SYBR method described above. The gel slice was then inserted into dialysis tubing with MWCO 7,000 (Membra-Cel® MC18x100 CLR, Viskase Companies, USA), which was filed with 2 ml of 1x TAE and sealed using clips making sure no bubbles or air pockets remain. This setup (orientated as if the slice was still in the gel) was placed into an electrophoresis tank filled with 1x TAE and run at 100 V for 1 hour. The liquid phase from inside the dialysis tubing was then transferred into two fresh microcentrifuge tubes. 0.1x vol 3 M NaOAc [pH=5.2] and 1x vol room temperature 100% isopropanol were then added to precipitate DNA. DNA was pelleted by centrifugation (17,000 g, room temperature, 20 min), the DNA pellet washed using 70% ethanol and eluted in 20 µl of EB (10 mM Tris-Cl, pH 7.5) by incubating samples at room temperature for 3 hours.

2.3.4 – Gel digest

Gel electrophoresis was carried out and a slice containing HMW DNA cut from the gel using the Classic or SYBR method described above, but instead of using a 0.8% agarose TAE, a 1.5% low-melting point agarose (SeaPrep - Lonza) TAE gel was used. The following protocol was derived from the ß-agarase I (NEB) and GELase™ (Epibio) user protocols. The gel slice was transferred into a microcentrifuge tube and melted (65°C, 5 min). After ensuring the gel is completely melted by gentle pipetting, the microcentrifuge tube was transferred to 42°C to equilibrate for 5 minutes. Equilibration of the gel slice using the provided reaction buffer was omitted due to a note in the GELase™ (Epibio) manual, stating that the reaction buffer, which is similar to the ß-agarase I reaction buffer, may interfere with lambda phage packaging reactions. ß-agarase I (NEB) was added at 0.75 U/200 µl of melted 1.5% gel and samples incubated (42°C, 2 h). 1x vol NH_4OAc [pH=7.0] and 4x vol ethanol were then added and samples incubated (room temperature, 2 h) to precipitate DNA. DNA was pelleted by centrifugation (17,000 g, room temperature, 20 min), the DNA pellet washed using 70% ethanol and eluted in 20 µl of EB (10 mM Tris-Cl, pH 7.5) by incubating samples at room temperature for 3 hours.

2.4 Miniprep

The buffers and protocol are based on the QIAprep® Miniprep Handbook (Quiagen) but some details have been adjusted. Cells were pelleted by centrifugation (4,000 g, 4°C, 10 min for ≥50 ml culture) and the supernatant discarded. The cell pellet was resuspended in 200 µl of P1 for every 5 ml overnight culture. 200 µl aliquots were transferred into microcentrifuge tubes and 200 µl of P2 was added. Microcentrifuge tubes were mixed by gentle inversion. After 3-5 min of lysis, 300 µl of N3 was added and microcentrifuge tubes mixed gently by inversion. Cell debris was then pelleted by centrifugation (15,000 g, room temperature, 7 min). The supernatant was then transferred to Silica spin columns for DNA (Epoch), which were centrifuged at 15,000 g for 60 seconds. Flow through was discarded, 500 µl of PB added to the spin column and samples were again centrifuged at 15,000 g for 60 seconds. These steps were repeated twice with 500 µl of PE to wash the DNA fixed to the silica membrane. After discarding the last flow through, the column was again centrifuged at 15,000 g for 60 seconds to collect any remaining liquid before being transferred to a microcentrifuge

tube. 70 µl of EB (10 mM Tris-Cl, pH 7.5) was added straight onto the silica membrane and microcentrifuge tubes incubated at room temperature for ~60 minutes. Eluted DNA was then collected by centrifugation (10,000 g, 2 min).

2.5 pWEB-TNC vector preparation

LB Amp_{200} Chl_{25} was inoculated with the pWEB-TNC cell stock (DH5α™) and grown for 16-18 h (37°C, 200 rpm). The pWEB-TNC cosmid was then miniprepped (2.4), quantified and checked for purity (A260/A230 > 1.6 and A260/280 > 1.8) using the NanoPhotometer NP80 (Implen). Aliquots of 40 µg clean pWEB-TNC were then digested overnight using 100 U SmaI (NEB) in a final volume of 500 µl at 25°C. A further 60 U of SmaI (NEB) was added the next morning and incubated (25°C, 2 h). Next, pWEB-TNC was dephosphorylated by adding 8 U of rSAP (NEB) and incubating (37°C, 2 h). Enzymes were then heat-inactivated at 65°C for 5 minutes. pWEB-TNC was then precipitated using 0.7x vol. isopropanol and 0.1x vol 3 M NaOAc [pH=5.2] and pelleted by centrifugation (16,000 g, 4°C, 30 min). The supernatant was discarded and the DNA pellet washed 2x with 70% ethanol. After letting the pellet air dry for approximately 3 minutes, pWEB-TNC was dissolved in 80 µl of EB (10 mM Tris-Cl, pH 7.5) 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 (Implen). Where applicable, pWEB-TNC was further diluted using EB (10 mM Tris-Cl, pH 7.5) to a concentration less than 400 ng/µl. Successful digest of pWEB-TNC was confirmed by running it out on a 1% agarose gel in TAE. Digested pWEB-TNC should be visible as a single band around the 6kb marker and undigested pWEB-TNC as two or three bands due to the circular, coiled and hyper-coiled conformations of the DNA. Ligation efficiency of digested and dephosphorylated pWEB-TNC was quantified by electroporating digested pWEB-TNC, digested pWEB-TNC after a ligation reaction using the Quick Ligation™ Kit (NEB) to the manufacturer's instructions, 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 50 µl EC100™ electrocompetent cells in cold electroporation cuvettes. Electroporation was carried out at 2.5 kV, 2.5 mm for 5.0 msec after which 1 ml of SOC was immediately added by gentle resuspension. Cells were then incubated (37°C, 200 rpm, 30 min) for

recovery. After recovery cells were pelleted by centrifugation (5,000 g, room temperature, 5 min), resuspended in 100 µl of supernatant and plated out on LB Amp_{200} Chl_{25} agar plates using sterile plating beads. 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.

2.6 *Streptomyces albus* gDNA isolation

50 ml of Trypticase Soy Broth (TSB) medium in a 250 ml culturing flasks (Ultra Yield™) was inoculated from a spore suspension of *Streptomyces albus* and grown for 12-18 h (30°C, 200 rpm). The 50 ml culture was divided into two 50 ml falcon tubes and cells were collected by centrifugation (4,000 g, 4°C, 10 min). After carefully discarding all supernatant, cells were resuspended in 12.5 ml SET buffer (75 mM NaCl, 25 mM EDTA [pH=8.0], 20 mM Tris-Cl [pH7.5]) and lysozyme was added to a final concentration of 2 mg/ml. Samples were then incubated (37°C, 30 min) before adding SDS to a final concentration of 1% (wt/vol) and Proteinase K to a final concentration of 0.5 mg/ml followed by incubation (55°C, 2 h), inverting occasionally. After letting the samples cool to approximately 37°C, NaCl was added to a final concentration of 1.25 M. Then, 12.5 ml chloroform was added to each falcon tube and the samples were mixed by gentle shaking or inversion for 30 minutes at room temperature. Samples were centrifuged (10,000 rpm, room temperature, 10 min) to separate the phases and the upper (aqueous) phase was transferred to fresh tubes. To precipitate DNA, 0.6 vol Isopropanol was added, followed by gentle inversion and incubation at room temperature for 5 minutes. DNA was then pelleted by centrifugation (16,000 g, 4°C, 20 min). The supernatant was carefully removed and the DNA pellets gently washed with 70% ethanol. Samples were then centrifuged again (16,000 g, 4°C, 10 min) and all ethanol removed. The DNA pellets were then left to air-dry for 15 minutes and dissolved in 750 µl TE (10 mM Tris, 1 mM EDTA [pH = 8.0]) by incubating samples at 55°C with occasional gentle flicking of the tubes.

2.7 λ-phage extract preparations

Phage extracts were prepared using a modified version of the methods described by Winn and Norris ([117], Appendix I). A full protocol for this is given in Appendix A.1.

2.8 λ-phage packaging of ligated DNA

λ-phage packaging of ligated DNA for cosmid library preparation was carried out using a modified version of the method described by Sean Brady [105]. LB 10 mM $MgSO_4$ was inoculated with EC100™ Δ*entD* (or EC100™) and grown overnight (37°C, 200 rpm). The next day fresh LB 10 mM $MgSO_4$ was inoculated with 0.01x vol of the overnight culture and this day culture incubated (37°C, 200 rpm) until OD_{600} ~ 0.6. Once matured, the day culture was kept on ice until use. 125 ng insert DNA was end-repaired using the NEBNext® End Repair Module to the manufacturer's instructions (unless specified otherwise) and then ligated to 250 ng digested and dephosphorylated pWEB-TNC using the Quick Ligase™ Kit to the manufacturer's instructions in a final volume of 5-25 µl. Aliquots of BHB2688 extracts (45 µl) and NM759 extracts (60 µl) were thawed on ice and mixed to give fully functional λ-phage packaging extracts. 33 µl of the resulting packaging extract was then added to the ligation reaction and samples incubated (30°C, 90 min). A further 33 µl was then added and samples again incubated (30°C, 90). The resulting packaged DNA was diluted using 500 µl of Phage Dilution Buffer (10 mM Tris-Cl [pH 8.3], 100 mM NaCl, 10 mM $MgCl_2$). Excess phage heads were precipitated by addition of 40 µl chloroform and centrifugation (5,000 g, 5 sec). Packaged phage heads were then added to the EC100 Δ*entD* (or EC100) day culture at a 1(packaged phage heads):10(day culture) ratio and incubated (room temperature, 20 min). Samples were then transferred to a shaking incubator (37°C, 200 rpm, 75 min). 100 µl of this culture as well as two dilutions (1:10 and 1:100) were then plated out on LB Amp_{200} Chl_{25} agar plates and incubated (37°C, overnight). The remainder was transferred into cryotubes or 96 well plates (depending on the amount of reactions carried out) as 500 µl aliquots and mixed with an equal volume of filter-sterilized 30% glycerol. Packaging efficiency was evaluated by counting the colonies formed on the dilution agar plates and total expected colonies were calculated as described below.

Total expected colonies (Pfu/µg) = colony count average (colonies neat + (10 * “colonies 1:10”) + (100 * “colonies 1:100”) /3) * 57.6 (since only 10µl of a total volume of 576µl is taken for every 100µl of cells) * F

F = factor accounting for amount of DNA used per packaging reaction, i.e. if 0.25µg of DNA are used, F = 4

2.9 Metagenomic DNA extraction from marine sponges

Sponge tissue was homogenized in minimal amount of spin buffer (100 mM Tris-EDTA, 500 mM NaCl) in sterile food processor or using mortar and pestle. This suspension was then centrifuged (4,000 g, 10 sec) to collect large debris and the supernatant transferred to a fresh falcon tube. Biomass was pelleted by centrifugation (4,000 g, 20 min) and the supernatant used to wash out blender or mortar and pestle. The two centrifugation steps were repeated and the supernatant then discarded. Resulting cell pellet(s) were resuspended in 2.5 ml sponge lysis buffer (8 M urea, 1 M NaCl, 2% sodium lauroyl sarcosinate, 50 mM EDTA, 50 mM Tris-Cl, [pH=7.5] [197]) and incubated for lysis (55°C, 10 min), gently inverted and incubated further (55°C, 10 min). 1x vol phenol:chloroform (1:1) was then added followed by gentle inversion until an emulsion formed. Phases were separated by centrifugation (12,000 g, room temperature, 3 min) and the upper (aqueous) phase was transferred to a fresh 15 ml falcon tube. This phenol:chloroform wash was repeated once more. The lower organic layer was collected from the washes, 500 µl of recovery buffer (1x TE, 0.5 M NaCl) added and the solution mixed by gentle inversion. Phases were again separated by centrifugation (12,000 g, room temperature, 3 min) and the upper (aqueous) phase was transferred to a fresh 15 ml falcon tube. Sample was split into 700 µl subsamples and 0.1x vol 3 M NaOAc [pH=5.2] as well as 0.6x vol cold (-20°C) isopropanol were added to precipitate DNA. Samples were then centrifuged (16,000 g, 4°C, 30 minutes) to pellet DNA. The pellet was washed twice with 70% ethanol, taking care not to disturb the pellet, and air dried for 3 minutes. DNA was eluted in 30 µl of TE (10 mM Tris, 1 mM EDTA [pH=8.0]) by incubating samples at room temperature overnight.

2.10 Retransformation of CS_1 and CS_T

Separate overnight cultures of the master pools CS_1 and CS_T, which were originally constructed in standard EC100 cells, were grown (37°C, 200 rpm) in LB Amp_{200} Chl_{25} 0.4% glucose [wt/vol]. DNA was miniprepped (2.4) and size-selected using the SYBR method (2.2) and ß-agarase I (2.2.4). 50 ng of size-selected cosmid plus insert DNA (vol < 5 µl) was then electroporated into 100 µl EC100^{TM} Δ*entD* electrocompetent cell stock with 2.5 kV, 2.5 mm for ~5.0 msec. 1 ml of SOC (0.5% (w/v) yeast extract, 2% (w/v) tryptone, 10 mM NaCl, 2.5 mM KCl, 20 mM $MgSO_4$, 20 mM glucose) was immediately added and mixed with cells by gentle pipetting. Cells were transferred into a shaking incubator for recovery (37°C, 200 rpm, 30 min). Centrifugation (2,000 g, room temperature, 5 min) was carried out to pellet cells, which were then resuspended in 100 µl of supernatant. Resuspended cells were then plated on LB Amp_{200} Chl_{25} agar plates and incubated (37°C, overnight). An initial test electroporation yielded ~12,000 clones. CS_1 and CS_T were retransformed at 2x coverage (384,000 and 120,000 clones respectively) to account for possible variations in clone number and diversity (between the original and retransformed library) created during culturing and electroporation. Consequently, 32 and 10 electroporations were carried out for CS_1 and CS_T respectively, which were grown as an overnight culture in LB Amp_{200} Chl_{25} and stored as glycerol stocks at -80°C the next morning rather than being plated on agar plates.

2.11 PPTase enrichment functional screen

This protocol was adapted from Charlop-Powers et al. 2013 [120]. Screening agar (M9, 1 g/L casamino acids, 0.4% glucose [wt/vol], 20 mM $MgSO_4$, 100 µM $CaCl_2$, 100 µM 2,2-dipyridyl, 10 µM thiamine-HCl, Amp_{200} Chl_{25}) was prepared by mixing 5x M9 salts with melted agar, mixing all other sterile components with MQ (50 ml final volume) and then adding this mixture to the melted agar. 3 mL LB Amp_{200} Chl_{25} overnight cultures were inoculated with the master pool of each CS_library plates respectively. The next morning 100 µl overnight culture was used to inoculate 3 ml of LB Amp_{200} Chl_{25}. After 3 hours, 2 ml of the day culture were transferred into a microcentrifuge tube and cells pelleted by centrifugation (5,000 g, room temperature, 5 min). The cell pellet was then resuspended in 1.5 ml 10% glycerol to wash cells from any media. This pelleting and washing sequence was repeated twice more. At the final wash step cells were

resuspended in 2 ml 10% glycerol and OD_{600} measured. Cells/ml was calculated from the optical density using the formula below and a dilution factor determined to achieve 10x coverage of each CS_library plate. Two 200 µl screening pools were created, one containing the retransformed CS_1 and CS_T plates and one containing CS_2 – CS_5.

OD_{600} of 1.0 = 1.24 * 10^8 cells/mL.

The two 200 µl screening pools were then plated on separate selection medium agar plates using sterile plating beads and incubated (37°C, <13 h). The small and translucent colonies (often surrounded by satellite cells; ideally 200-500 patches per plate) were swiftly resuspended in 3 ml 10% glycerol using a sterile bent glass rod to scrape all the colonies off the agar surface and the liquid was quickly transferred into microcentrifuge tubes. 500 µl of this cell suspension was stored at -80°C as a glycerol stock. Remaining cells were pelleted by centrifugation (5,000 g, room temperature, 5 min) and thoroughly washed three times using 1 ml 10% glycerol as described above. At the final wash step cells were resuspended in 2 ml 10% glycerol and OD_{600} measured. Using the formula above, a dilution factor was determined to achieve 50,000 cells/ml for each screening pool. 200 µl (~10,000 cells) of each screening pool was then plated out on separate selection medium agar plates using sterile plating beads and incubated (37°C, >13 h). Small and translucent colonies (ideally 70-100 per plate) were picked without touching any surrounding satellite cells and re-streaked onto selection medium agar plates using sterile toothpicks. These plates were then incubated (37°C, overnight), single colonies picked and cultured in 10 ml LB Amp_{200} Chl_{25} (37°C, 200 rpm, overnight). 500 µl of each overnight culture was stored as a glycerol stock and the remainder miniprepped (2.4).

2.12 Preparation of crude metagenomic DNA from marine sponges for Illumina sequencing

1 µl of RNAse stock solution (10 mg/ml) was added to a microcentrifuge containing the DNA sample (<100 µl volume) in TE (10 mM Tris, 1 mM EDTA [pH=8.0]) and incubated (37°C, 1 h). Next, 1 µl of Proteinase K stock solution (10 mg/ml) was added and the sample incubated (55°C, 1 h). 0.6x vol of magnetic bead solution (Carboxyl

modified Sera-Mag™ SpeedBeads (FisherScientific #09-981-123) prepared using protocol from Rohland & Reich 2012) was then added and the sample incubated (room temperature, 5 min). The sample was then transferred to a magnetic bead rack (built in-house). Once solution had cleared completely, the supernatant was aspirated and the bead pellet washed with 1 ml 70% ethanol without disturbing the pellet. The washing step was repeated once more and all ethanol was carefully removed. 30 µl EB (10 mM Tris-Cl, pH 7.5) was then added, the microcentrifuge tube removed from the magnetic rack and incubated (50°C, 1h). The sample was then transferred back to the magnetic rack until solution clears completely. Finally, the supernatant was transferred to a fresh microcentrifuge tube, making sure no magnetic beads are carried over.

2.13 Nanopore sequencing of *C. mycofijiensis*

Nanopore sequencing was carried out in-house by Matt Storey (PhD-student) following the "One-pot ligation protocol for Oxford Nanopore" (Josh Quick, Loman Labs, University of Birmingham) using the MinION Mk1B (R9.4.1 Flow Cell) with the Ligation Sequencing Kit 1D R9 Version (SQK_LSK108). Due to availability and expense of Oxford Nanopore sequencing technologies this sequencing run contained both, metagenomic DNA from *C. mycofijiensis* and DNA isolated from the microbiome of *Mycale hentscheli*, a New Zealand sponge of interest to the lab group. Sequencing data was then separated using the bioinformatic workflow described in 2.14.2.

2.14 Bioinformatic analyses

This section contains a brief description of the general bioinformatic workflow used; for details on exact commands used, see Appendix and the GitHub repository (https://github.com/MaxMeta/Vincent_Masters.git).

2.14.1 Read trimming

The script **`adap_ID.sh`** (Appendix A2) was run on raw reads to identify any adapter sequences. These were then appended to the **`TruSeq2-PE.fa`** file, provided with Trimmomatic v.0.36 [198] for thorough trimming. Trimming was then carried out using Trimmomatic v.0.36 with the following arguments: **`PE`**

ILLUMINACLIP:/path/to/adapter_file:2:30:10:4:4:/true TRAILING:9 SLIDINGWINDOW:4:15 MINLEN:36

2.14.2 Base-calling and sequence alignment of Nanopore reads

Base-calling was carried out using Albacore v.2.2.6 (Oxford Nanopore) with default settings. Passed reads were then aligned to the PE_150 assembly using BBMap's v.38.26 **bbmapskimmer.sh** with **maxlen=1000** and aligned reads extracted using a script based on samtools v.0.1.19 (Appendix A3 for details).

2.14.3 Assembly of trimmed reads

Paired-end Illumina short reads were assembled using metaSPAdes from [166] SPAdes v.3.12.0 with the parameters **–meta -k 21,33,55,77,99,127 --pe-1 /path/to/reads_1 --pe-2 /path/to/reads_2** sepcified. Paired—end Illumina short reads plus the extracted Nanopore reads from above (*C.mycofijiensis* only) were assembled using SPAdes v.3.12.0 with the parameter **--meta -k 21,33,55,77,99,127--pe-1 /path/to/reads_1 --pe-2 /path/to/reads_2 --nanopore /path/to/Nanopore_reads**. PE250 Illumina reads with PE150 as trusted contigs (*C.mycofijiensis* only) were assembled using SPAdes v.3.12.0 (hybrid mode) with the parameters **--pe-1 /path/to/reads_1 --pe-2 /path/to/reads_2 --trusted-contigs /path/to/PE150_contigs**

2.14.4 Binning

Contigs obtained from the assembly were binned using the metaWRAP v.1.0.5 [171] binning module with the parameters **--metabat2 --maxbin2 --a /path/to/assembly_contigs /path/to/reads_1 /path/to/reads_2**. These were also independently binned using Autometa (commit c6e398e) with the included **make_taxonomy_table.py -l 1000, run_autometa.py --length_cutoff 1000** and by **cluster_process.py** (default settings) python scripts run in the given sequence.

2.14.5 Bin dereplication, comparison and taxonomic identification

The resulting bins were then dereplicated using the dRep v.2.2.3 [167] **`dereplicate`** module with default settings. Bins were also compared using the dRep v.2.2.3 **`compare`** module with default settings. Taxonomic identification of bins was obtained by referencing the marker lineage identified by checkM [199] run as part of dRep v.2.2.3 to the Genome Taxonomy Database (GTDB; Appendix A25).

2.14.6 16S rRNA sequence comparison of sponge microbiomes

Full code and results are available on the Git-Hub. In short; all bacterial rRNA sequences were extracted from the metagenomic assemblies using barrnap v.0.9 (Torsten Seemann 2015) with default settings. 16S rRNA sequences longer than 800bp were then extracted from these reads and dereplicated by clustering them using CD-HIT (Fu et al. 2012) with a 0.99 similarity cutoff

```
$ cd-hit-est -i <input> -o <output> -c 0.99 -n 10 -d 0 -M 1000 -T 1 -sc 1
```

Dereplicated 16S rRNA sequences were then standardized, clustered based on cosine distance and visualized using seaborn v.0.8.1. This clustering was carried out with and without coverage information. Coverage information was extracted from the FASTA headers of the respective contigs the 16S rRNA sequence originated from.

```
>>> seaborn.clustermap(df, cmap='RdYlGn_r', linewidths=1, figsize=(5,25), mask=df==0, method = "average", metric="cosine", standard_scale=1)
```

Dataframes were then created based on the taxonomy of 16S rRNA sequences identified by the SILVA ACT online tool (SINA 1.2.11, SILVA release 132) and coverage information summarized, e.g. per class (Appendix A23).

2.14.7 Taxonomic identification of sponges

Eukaryotic rRNA sequences were extracted from the metagenomic assemblies using barrnap v.0.9 (Torsten Seemann 2015) with **`--kingdom euk`** specified. These sequences were then analyzed using the SILVA ACT online tool (SINA 1.2.11, SILVA release 132) with default settings for “basic alignment parameters” and “search and classify”.

2.14.8 Identification of secondary metabolite gene clusters

Contigs resulting from the metagenomic assembly were length-filtered using the **`fasta_len_filter.sh`** script with a length filter of 5000 and then analyzed using antiSMASH v.4.1.0 **`with --full-hmmer --clusterblast --subclusterblast --knownclusterblast --smcogs`**. antiSMASH v.4.2.0 [180] with the same options as above was used for analysis of the dereplicated MAGs.

2.14.9 Secondary metabolite BGC clustering

BiG-SCAPE v.0.0.0 [200] was run with **`–mode glocal –mibig –cutoffs 0.3 0.5 0.7`** in order to analyze and visualize relatedness of BGCs to each other and known BGCs deposited in the MiBIG database.

2.14.10 Construction of coverage over GC content plots of sponge metagenome assemblies

Data for these plots was calculated and concocted using UNIX commands and R v.3.4.3 (2017-11-30) – “Kite-Eating Tree” with the **`calc.gc.pl`** script as well as a general workflow derived from Albertsen et al. 2013 [201]. Plots were then created using the ggplot2 library (Appendix A5 for details).

2.14.11 Primers

T7_promoter: 5’-TAATACGACTCACTATAGGGAGA-3’
M13F: 5’-CGCCAGGGTTTTCCCAGTCACGAC-3’

Chapter 3

Optimizing cosmid library construction for environments yielding low quantities of HMW DNA

Microbial communities in Antarctic and New Zealand seawater are largely unexplored [146] and the Antarctic prokaryote world is thought to have particularly high bacterial and chemical diversity [4,8,147-150]. Construction of cosmid libraries from these environments would theoretically allow redundant coverage and preservation of metagenomes, [50,54,105,106,115] and would be an immensely useful resource for both ecological and drug discovery studies of these inherently challenging environments.

There are several steps involved in the construction of a large insert cosmid library [105], of which DNA isolation (Figure 6, 1) is the first and one of the most crucial. Prior to cloning, DNA must be subjected to preparative agarose gel electrophoresis (Figure 6, 2). This step serves to both remove inhibitors, and select the correct size range for cloning [33,105]. In order to prevent the loss of precious NZ seawater and Antarctic samples during protocol optimization, these steps were trialed using genomic DNA (gDNA) from a common laboratory strain (*Streptomyces albus*). This non-precious gDNA was also used to optimize packaging efficiency (Figure 6, 6). Both the pWEB-TNC (Epicenter) vector (Figure 6, 2) and the λ bacteriophages (Figure 6, 3) were produced in house and thus had to be quality controlled and benchmarked.

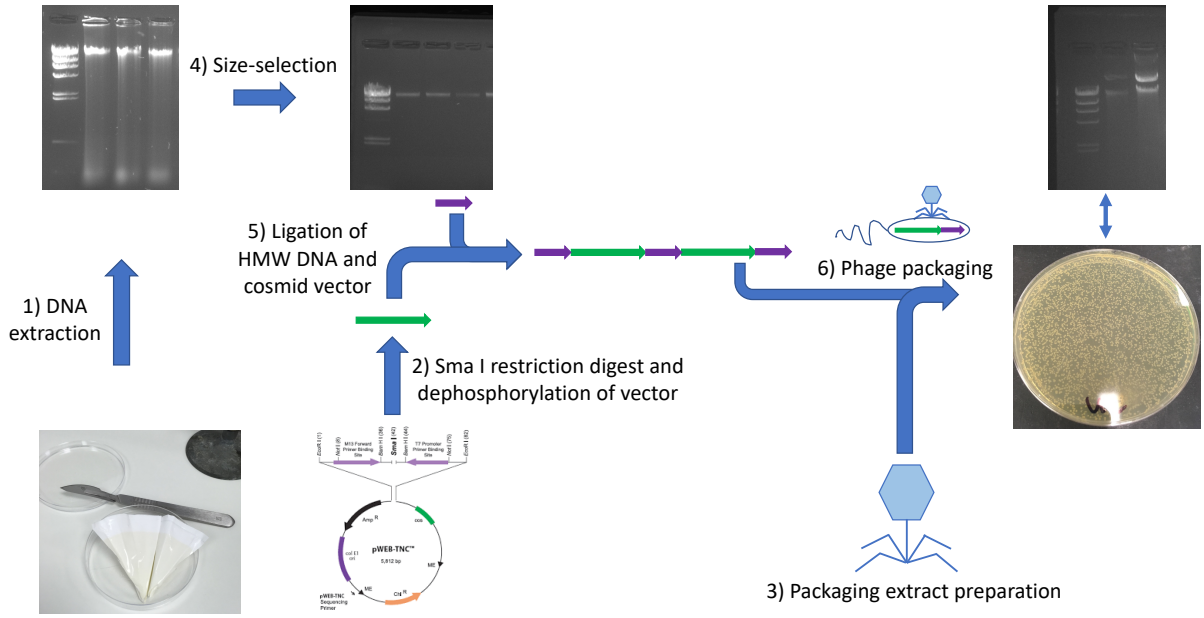


Figure 6: Cosmid library construction workflow. All gels contain lambda DNA/HindIII Marker in Lane 1. (1) DNA is extracted from an environmental sample and visualized using preparative gel electrophoresis to check for HMW DNA. (2) pWEB-TNC vector is prepared by SmaI restriction digest and dephosphorylation to allow ligation to environmental DNA (eDNA). (3) Phage packaging extracts are prepared and tested for efficiency to allow the selective packaging of DNA+vector constructs into the *E. coli* host. (4) eDNA is size selected for HMW DNA and visualized using preparative gel electrophoresis to verify size and quality of size-selected DNA. (5) Digested and dephosphorylated pWEB-TNC vector and size-selected DNA are ligated, often forming concatemers. (6) DNA+vector constructs are transfected into the *E. coli* host using the phage packaging extracts, which selectively package one construct per bacterial cell. Inserts are then verified by miniprep plasmid extraction of a cultured colony and preparative gel electrophoresis of the plasmid DNA.

3.1 Extraction of high molecular weight DNA

Six different DNA extraction protocols were trialed here in order to find an optimal protocol for the isolation of HMW DNA from both New Zealand seawater samples and Antarctic samples. These involved either enzymatic or chemical lysis, as well as different detergents and buffer compositions. Variation in DNA precipitation and elution were also compared. Details for each of the methods trialed can be found in 2.2.1 - 2.2.6 and detailed results are presented in Appendix A6. Overall, method 2.2.6 was the most suited for Antarctic samples with the most notable difference to other protocols being the overnight lysozyme digest (2.2.6). However, most suited for New Zealand seawater samples was Method 2.2.5, which is based on chemical lysis rather than enzymatic lysis and consistently yielded HMW DNA. Both 2.2.5 and 2.2.6 involved the recovery of nucleic acids from the collected organic phenol-chloroform layer, which resulted in additional samples containing DNA at approximately half the

concentration of the original sample. These protocols also included the final elution of DNA overnight rather than for a few hours (2.2.5, 2.2.6).

3.2 Optimizing electrophoresis and DNA size-selection

For each test condition, 20 µg of *S. albus* gDNA was used as starting material and the final yield quantified as a percentage of the original 20 µg. Standard DNA clean up columns could not be used as they shear the large pieces of DNA desired here (30-45 kb). Another critical consideration was the avoidance of UV light as this is well known to damage DNA and reduce cloning efficiency. Overall, eight trials were conducted (Table 1). These used four separate methods for agarose size selection and elution (2.3.1-2.3.4), in combination with two separate visualization methods that avoid UV light (2.3). The visualization method was not expected to affect DNA yield, however it was possible that the use of SYBR™ dye might impact downstream cloning efficiency. Of the four methods, 2.3.4 emerged as the best choice, yielding the largest amount of DNA in both trials (Table 1).

Method number (visualization technique)	Concentration (ng/ul)	Yield
2.3.1 (Classic)	9.6	40%
2.3.1 (SYBR)	11.7	50%
2.3.1 (Classic) repeat	14.8	57%
2.3.1 (SYBR) repeat	17.3	64%
2.3.2 (Classic)	1.3	3%
2.3.2 (SYBR)	6.1	3%
2.3.3 (Classic)	1.1	5%
2.3.3 (SYBR)	2.0	20%
2.3.4 (Classic)	20.5	58%
2.3.4 (SYBR)	30.2	70%

Table 1: Concentration and yield of HMW DNA samples after different size selection methods Note that yield is compared to original quantity of DNA, which includes LMW DNA and non-DNA contaminants if present.

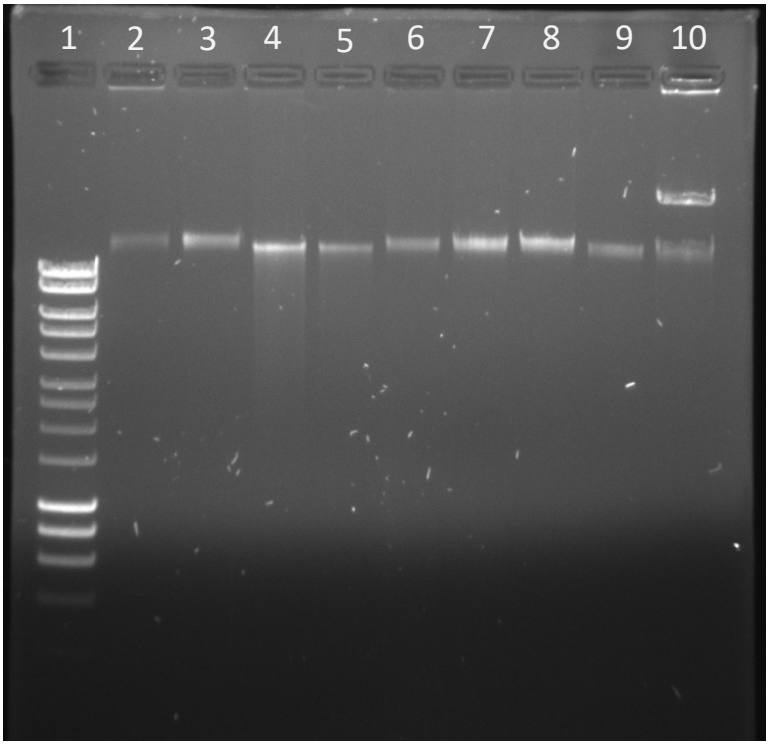


Figure 7: Qualitative comparison of DNA recovered using the size selection Methods 2.3.1-2.3.3
Lane 1 - Hyperladder 1kb; Lane 2 - Method 2.3.1 (Classic); Lane 3 - Method 2.3.1 (SYBR); Lane 4 - Method 2.3.2 (Classic); Lane 5 - Method 2.3.2 (SYBR); Lane 6 - Method 2.3.3 (Classic); Lane 7 - Method 2.2.3 (SYBR); Lane 8 - Method 2.3.1 (Classic) repeat; Lane 9 - Method 2.3.1 (SYBR) repeat; Lane 10 - pWEB-TNC + insert.

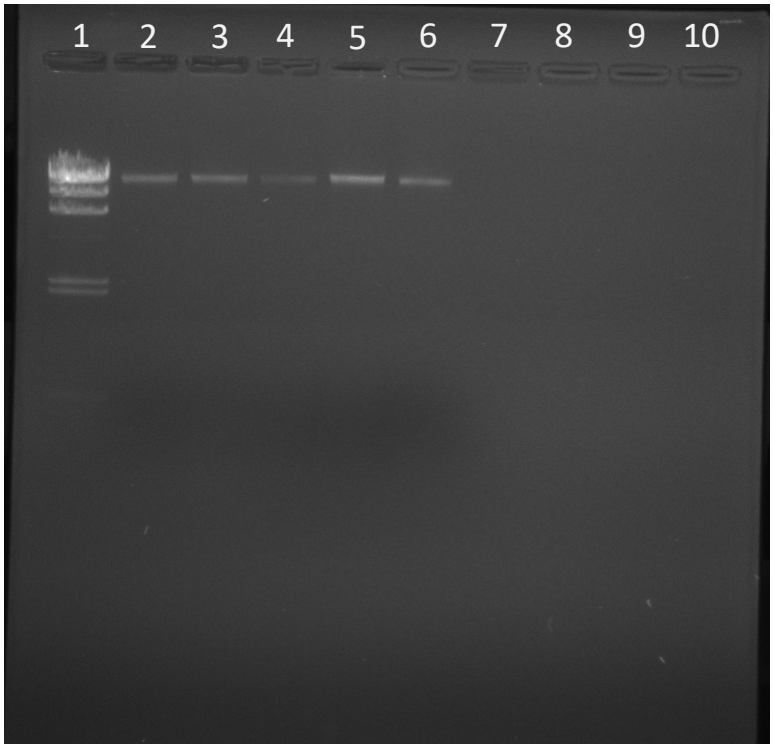


Figure 8: Qualitative comparison of DNA recovered using the size selection Method 2.3.4
Lane 1 - Lambda HindIII ladder; Lane 2 - Method 2.3.4 (Classic); Lane 3 - Method 2.3.4 (Classic); Lane 4 - Method 2.3.4 (SYBR); Lane 5 - Method 2.3.4 (SYBR); Lane 6 - Method 2.3.4 (Classic); Lane 7-10 - empty.

3.3 Testing the efficiency of in-house λ-phage packaging extracts

λ bacteriophages were prepared using a modified version of the methods described by Winn and Norris (2.7, Appendix A1). The current version of the protocol was extensively optimized by Dr. Mark Calcott (Victoria University of Wellington). Reagent quality and even brand emerged as being crucial for success. This protocol employs two engineered *E. coli* strains (BHB2688 and NM759), the extracts from which are combined to produce mature phage heads capable of packaging cosmid cloned DNA. Packaging extracts were tested in EC100^{TM} as well as EC100^{TM} Δ*entD* (Table 2). Commercially provided λ phage control DNA (Epibio) was packaged to measure of maximum packaging efficiency and digested and dephosphorylated pWEB-TNC as a negative control to measure relative efficiency of unwanted vector cloning (Table 2).

Test DNA packaged (0.25 µg)	Cell line	Colonies neat	Colonies 1:10	Colonies 1:100	Expected pfus/µg
Control Ligated Lambda DNA	EC100	TMTC	TMTC	~10,000	$>2*10^8$
Control Ligated Lambda DNA	EC100 Δ*entD*	TMTC	TMTC	~10,000	$>2*10^8$
pWEB-TNC (circular)	EC100	36	0	0	8,294
pWEB-TNC (circular)	EC100 Δ*entD*	19	0	0	4,378
pWEB-TNC (digested and dephosphorylated)	EC100	16	0	0	3,686
pWEB-TNC (digested and dephosphorylated)	EC100 Δ*entD*	4	0	0	922

Table 2: Results from the efficiency testing of in-house packaging extracts "Colonies neat" are the result of 100 µl of packaged DNA plated out before overnight incubation, with respective 1:10 and 1:100 dilutions of those 100 µl. TMTC = Too many to count.

As a final confirmation of successful cloning, putative cosmid DNA was isolated by miniprepping (2.4) six control reactions for each of two strains tested. These were visualised using agarose gel electrophoresis and the DNA from all minipreps was visible above a 23.1 kb marker indicating successful large insert cloning. The results indicate that the λ phage packaging extracts were of high efficiency and would be suitable for downstream cloning applications using precious metagenomic samples.

3.4 Testing end-repair, ligation and λ-phage packaging efficiency of a cloned insert

As a final test of the suitability of prepared reagents and cloning protocols, end repair, ligation, packaging and transfection were conducted using *S. albus* gDNA, and six different vector stocks (Appendix A7). This experiment was designed to determine:

(1) Which of the prepared vector stocks were of highest quality and should therefore be employed with the precious metagenome samples
(2) Whether the end repair and ligation protocols were functioning as expected
(3) Which size selection and gel visualization method(s) give the highest downstream cloning efficiency

The results showed that there was variation between the different vector stocks, which can likely be attributed to the level of digestion and dephosphorylation achieved from the respective minipreps, and Vector stock 1 (V1) yielded the most pfus/µg DNA by a considerable margin (Figure 9). Furthermore the size-selection method also affected the pfus/µg DNA with GELase (2.3.4) using SYBR to visualize the gel yielding notably more colonies (Figure 10). This method was least prone to sample loss, and best able to be applied to limited sample amounts. This is because the DNA of interest is directly visualised, rather than being inferred as in the Classic method. Also, the exclusion of dialysis electroelution in favour of GELase digestion and recovery reduces the chance that sample will be lost due to membrane breach or improper sealing. Vector stock 1 and GELase (2.3.4) SYBR were thus chosen as the most optimal for library construction from low-yielding environments.

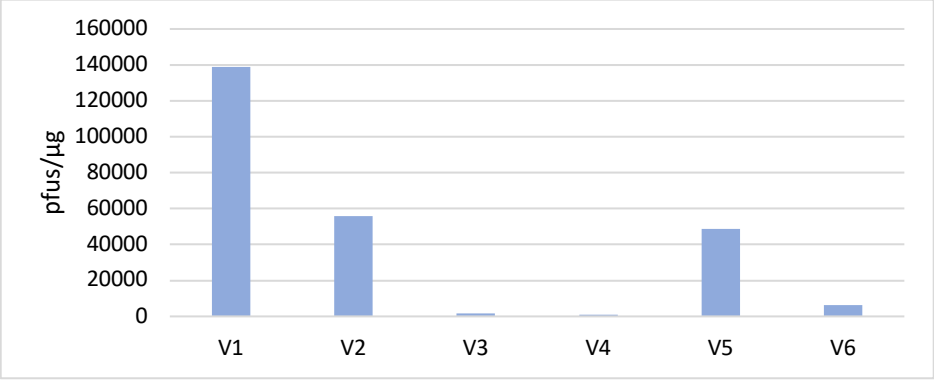


Figure 9: Average pfus/µg obtained from optimisation packaging reactions summarized by pWEB-TNC vector stock (V1-6) (Appendix A7)

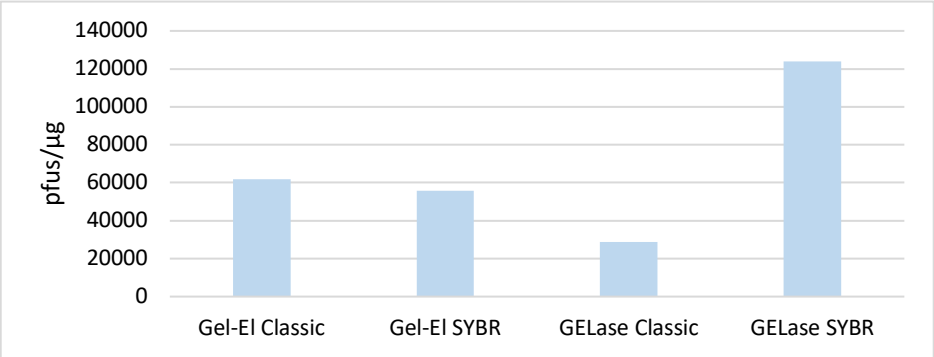


Figure 10: Average pfus/ µg obtained from optimisation packaging reactions summarized by DNA size-selection methodology (Appendix A7).

3.5 Construction of a cosmid library from Antarctic samples

Due to the extremely limited biomass of Antarctic samples [164,202], all DNA extracted from Antarctic samples in 3.1 was combined amounting to approximately 1 µg of extracted nucleic acids. This material was then size-selected using Method 2.3.4 resulting in HMW DNA samples with concentrations ranging from 10-15 ng/µl and a total of ~450 ng. Nine separate ligation and packaging reactions of 50 ng size-selected Antarctic DNA and 100 ng of digested and dephosphorylated pWEB-TNC vector were carried out and transfected. The resulting Antarctic library contained approximately 2,000 clones. 24 of these clones were miniprepped and visualized by agarose gel electrophoresis, revealing 12 clones that appeared to contain full insert. This indicates that approximately half of the 2,000 clones in the library are likely to contain insert, which would correspond to a packaging efficiency of 2222 pfu/µg.

3.6 Construction of a cosmid library from NZ seawater

DNA was extracted from three separate NZ seawater samples following protocol 2.2.5 and size-selected using protocol 2.3.1 to yield ligation competent insert DNA. The concentration of HMW DNA samples obtained after size selection was very low and centrifugal concentration was attempted in order to rectify this. However, this typically resulted in a significant amount of sample loss. A total of 21 different packaging reactions were carried using a total of 1.46 µg of HMW DNA for library construction (Appendix A8). This resulted in 28,659 clones, of which ~1/3 appeared to have insert based on subsequent miniprep (2.4) and visualization. Taking the number of clones with insert (9,457) results in a packaging efficiency of 6464 pfu/µg.

As a final check for library quality, 28 clones from 10 different packaging reactions were assessed by Sanger sequencing from conserved vector priming sites. This analysis allowed sequence from a small portion (< 1kb) of the ~30 kb inserts to be obtained. Comparative blast analysis was then used to determine whether the inserts were likely to be of marine bacterial origin, which would indicate successful cloning of marine metagenome inserts rather than laboratory contaminants, or vector only concatenation. As shown in Table 3 below, these clones did appear to contain a diverse collection of bacterial inserts that were likely to be of marine bacterial origin, indicating that the cloning effort was successful, albeit with low efficiency.

Well	Description	Accession	Pairwise identity	Bit-score	E-value	Sequence length
A02	hypothetical protein [Oceanicoccus sp. KOV_DT_Chl]	WP_101757164	86.2%	200.68	3.62e-61	109
A07	thiol peroxidase [Rubripirellula obstinata]	WP_068267256.1	81.9%	303.91	2.27e-101	171
A09	MULTISPECIES: fimbrial chaperone [Proteobacteria]	WP_000465928	99.6%	505.75	6.83e-180	246
D02	hypothetical protein A9Q90_00480 [Gammaproteobacteria bacterium 54_18_T64]	OUS10771	46.6%	289.27	1.51e-93	292
D07	transcription-repair coupling factor [Rubripirellula obstinata]	WP_068261595	61.9%	340.12	9.95e-105	312

Table 3: End-sequencing results for NZ seawater library Sanger-sequenced cosmids were consequently analyzed using Blastx as part of Geneious Prime. Note that 43 of the cosmid analyses returned results related to *E. coli* and 8 did not return any results.

3.7 Discussion

Seawater bacteria appear to produce medically relevant natural products [203,204]. In order to optimize the construction of metagenomic cosmid libraries for these low biomass environments, several modifications were made to the library construction protocol developed by Sean Brady [105], which was originally designed for soil and other high biomass metagenomic cosmid libraries [106,118]. These modifications include the lysis protocol used for DNA extraction, the method of size-selection and the use of in-house vector and packaging extracts. Unfortunately, due to limited sample availability of both New Zealand seawater and Antarctic samples, the resulting libraries were comparatively small (28,659 and 2,000 clones respectively).

Based on the fact that very high packaging efficiency was achieved with control *S. albus* gDNA, it appears that the DNA recovered from these metagenomic samples

was of insufficient quality for efficient cosmid cloning. While these optimization efforts did yield valuable insights and resulted in an optimized protocol, this was not sufficient to allow large scale cloning from the available samples. Given the time constraints on this thesis and the lack of availability of further samples, it was not possible to continue this work as part of this thesis. Focus was shifted to the metagenomic examination of a collection of marine sponges that were available in archived storage and were more likely to yield sufficient DNA for both metagenome shotgun sequencing and cosmid library construction. The optimized protocol described here does however provide a valuable template and is anticipated to be useful in future efforts toward the construction of metagenome libraries from larger samples of planktonic marine bacteria and Antarctic sea ice bacteria.

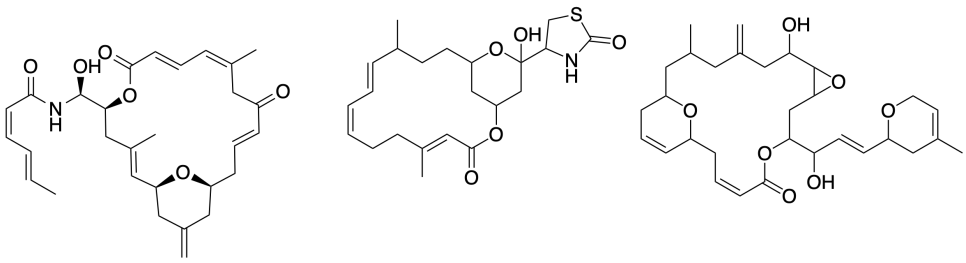
Chapter 4

Metagenomic shotgun sequencing and cosmid library construction using microbial DNA from the marine sponge *Cacospongia mycofijiensis*

The genus *Cacospongia* and specifically *C. mycofijiensis* is particularly rich in actin-binding macrolide compounds and has yet to be investigated using metagenomic methods [205-208]. The compounds (-)-zampanolide, latrunculin A and laulimalide A (Figure 11) are particular potent examples of these and the sample investigated as part of this thesis was shown to contain high levels of each of these compounds by the Keyzers group at Victoria University of Wellington.

(-)-Zampanolide (Figure 11) acts as a microtubule-stabilizing agent and is a promising drug candidate due to its activity against cancer cell lines, which are resistant to established cancer therapeutics [205,208,209]. Laulimalide A (Figure 11) is also a microtubule-stabilizing agent and potent anti-cancer compound against multi-drug resistance cell lines [210,211] while Latrunculin A (Figure 11) is an anti-cancer agent that disrupts the cytoskeleton. As with many other microtubule-stabilizing compounds isolated from marine invertebrates, these cannot be efficiently extracted or synthesized and the availability of source materials is very low, effectively halting their progress into the next stages of clinical development [10,212].

The goal of the research described in this chapter was to develop sequencing, assembly and binning methods for the resolution of individual genomes within the metagenome of the marine sponge *C. mycofijiensis*, ultimately leading to a picture of the microbial community associated with this sponge. It was hoped that analysis of the resulting MAGs using the suite of secondary metabolite tools available via the package antiSMASH, would then allow identification of microbes that were candidates for producing the cytotoxic polyketides associated with this sponge. This work represents the first steps toward providing a sustainable supply of these compounds, either via heterologous expression of their pathways or use of genome sequence data to guide isolation and cultivation of the producing organisms. An additional aim of this chapter was to construct a metagenomic large insert cosmid library from microbial DNA extracted from *C. mycofijiensis*, thereby preserving the genomes of the microbial community for future functional studies.



(-)-Zampanolide Latrunculin A Laulimalide A

Figure 11: Structures of the marine natural products (-)-zampanolide, latrunculin A and laulimalide A.

4.1 Retrobiosynthetic analysis of (-)-zampanolide, latrunculin A and laulimalide A

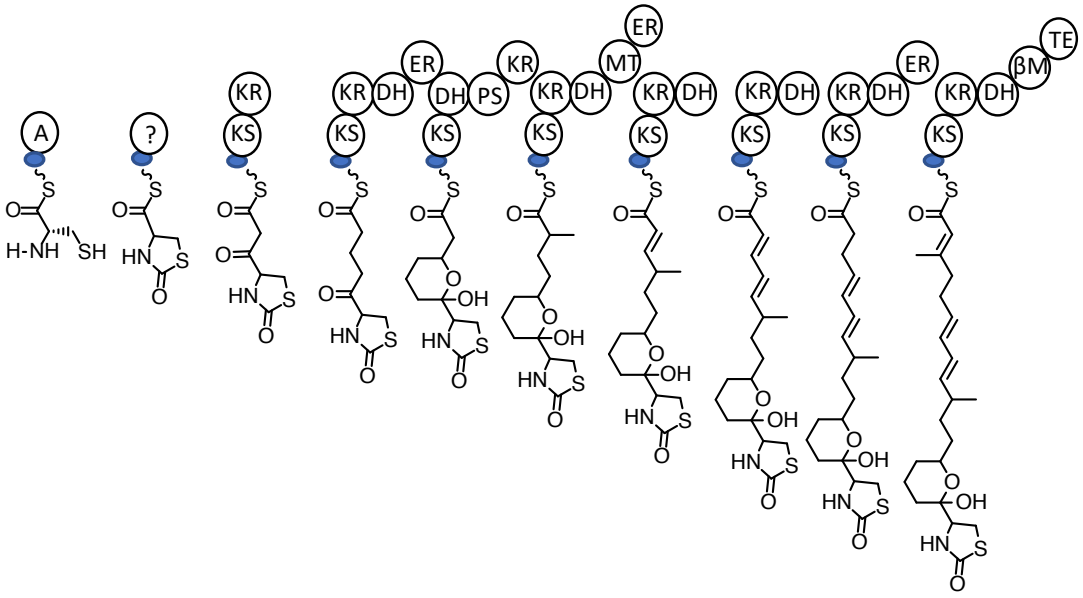
In order to identify the BGCs encoding the production of (-)-zampanolide, latrunculin A and laulimalide A, it was necessary to first conduct a retrobiosynthetic analysis, in which the collection of enzymes responsible for the production of a target compound are inferred from its structure by reversing the biochemical transformations that putatively led to its production [73,127]. This type of analysis has been extensively applied to trans-AT PKs well as other natural product types and relies on identifying substructures within a compound of interest that have been previously experimentally linked to certain enzymatic activities [56,78,82]. Based on the structure of (-)-zampanolide, latrunuclin A and laulimalide A, it is expected that they are produced by symbiotic bacteria, as has been demonstrated in numerous other marine sponge polyketides [4,10,56]. Retrobiosynthetic analysis of (-)-zampanolide, latrunculin A and laulimalide A is outlined in Figure 12 below. It should be noted that this analysis assumes strict adherence to collinearity and canonical biosynthetic rules, assumptions which are often violated, particularly in trans-AT PKSs where several domains may be acting in trans and modules may be inactive, iterative or split [77,81].

When conducting retrobiosynthetic analysis for pathway discovery, identifying rarely occurring functional groups that will yield a characteristic genetic signature can serve as a hook for identification. Characteristic functional groups of laulimalide A are the epoxide group and the 3,4-dihydropyran ring (Figure 12, F), both of which have been previously observed in natural products and have known routes to production. Pyran

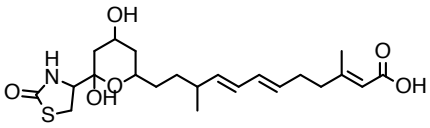
rings are present in many trans-AT PKs and are typically produced by a pyran synthase domain [56] that can be identified from sequence data using algorithms such as antiSMASH. Pyran synthase domains may appear similar to a dehydrogenation domain but have a deletion in the active site [56,213]. The biosynthetic routes to an epoxide group are more varied and several classes of enzymes are able to synthesize this moiety, including the P450 (CYP) family [214,215], luciferase-like monooxygenases [216] as well as flavoenzymes [217,218]. Examples of compounds containing epoxide groups are epothilone A and B [219], hypothemycin [220], fumiquinazoline A [221] and lasalocid [222].

Both latrunculin A and (-)-zampanolide contain amino acids as substituents, suggesting production via a NRPS-PKS hybrid system containing a cysteine-incorporating and a threonine-incorporating adenylation domain respectively. In the case of latrunculin A, the incorporated cysteine is converted to a thiazolidine moiety (Figure 12, C), which suggests the presence of heterocyclisation and oxidation domains in addition to the adenylation domain typical of an AA-incorporation [223]. Additional distinctive aspects of (-)-zampanolide are the uneven number of carbons on either side of the peptide bond, suggesting the presence of a terminal ß-methylation and insertion of threonine rather than serine, as well as two additional methyl groups likely originating from ß-methylation (Figure 12, G and I).

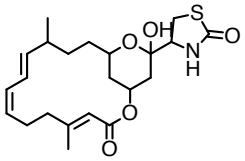
A


A
?
KR
KS
KR DH
ER
DH PS
KR DH
MT
ER
KR DH
KR DH
KR DH
ER
KR DH
βM
TE
H-NH
SH
HN
S
O
OH

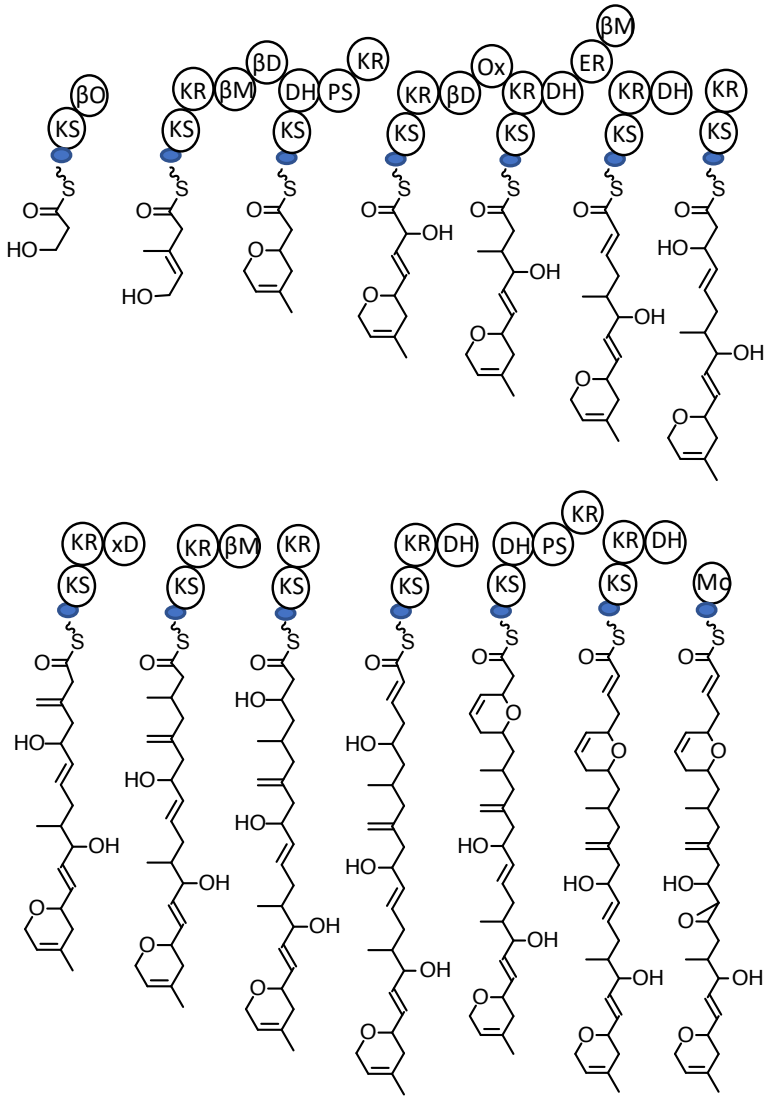
B


OH
H
N
O
S
OH
O
OH

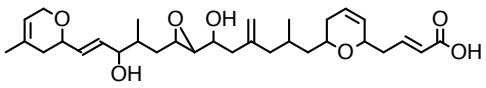
C


OH
S
O
N
H
O
O

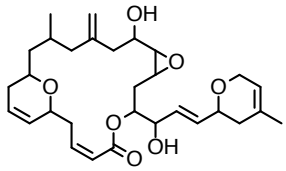
D



E



F



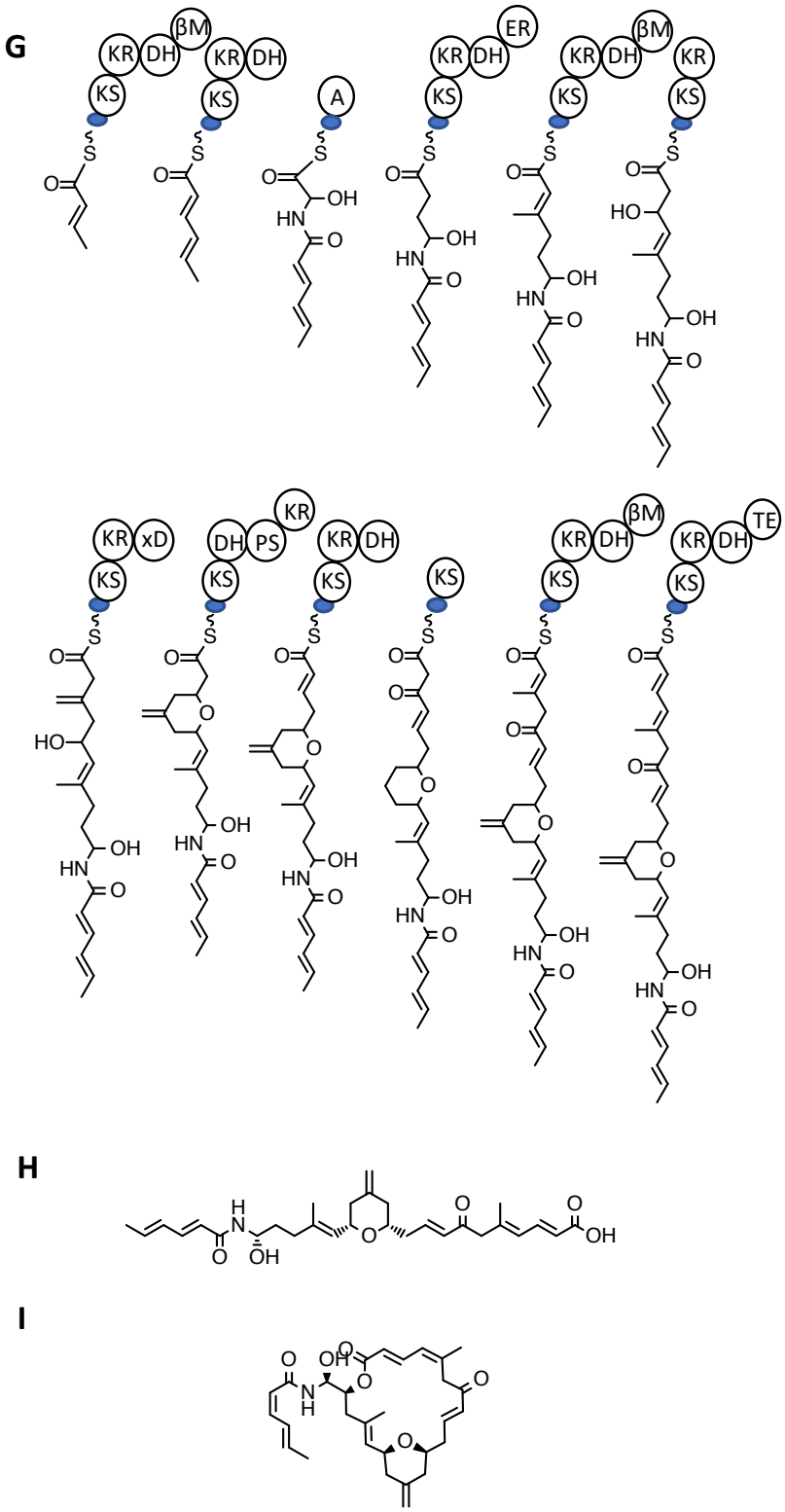


Figure 12: Retrobiosynthesis of Latrunculin A (A-C), Laulimalide A (D-F) and (-)-Zampanolide (G-I) Based on modular trans-AT Type I PKS systems due to the distinct possibility that these pathways may be trans-AT (Wilson et al. 2014, Helfrich & Piel 2016). Domain key: A = Adenylation (NRPS); KS = Ketosynthase; KR = ketoreductase; DH = dehydration; ßD = ßγ-dehydration; ER = enoylreductase; Ox = α-hydroxylation; MT = α-methyltransferase; ßM = ß-methyltransferase; PS = pyransynthase; xD = unknown dehydration ; Mo = mono-oxygenase; TE = thioesterase.

4.2 Direct shogun sequencing and assembly of the microbiome of *C. mycofijiensis*

In order to obtain sequence data for the microbiome of *C. mycofijiensis*, metagenomic DNA was extracted using a protocol heavily modified from Gurgui & Piel [197] (2.9) from *C. mycofijiensis* and cleaned (2.12). Clean metagenomic DNA was then sent to a commercial provider to collect paired-end Illumina reads (2x150 bp and 2x250 bp) while Nanopore long read data were collected in house as described in 2.13 (Appendix A12). By using a variety of read lengths, sequencing technologies and downstream assembly methods, the aim was to derive an optimal protocol for sequencing and assembly of marine sponge microbiomes.

4.2.1 Quality filtering and trimming of reads

Quality analysis of PE150 and PE250 Illumina reads was carried out using FastQC Version 0.11.5 (Babraham Bioinformatics) with default settings. The PE250 reads were flagged as having low sequence quality, irregular GC content, overrepresented sequences, high adapter content (in over 40% of reads) and high k-mer content. This indicates that the sequencing library for this sample was of low quality and contained shorter than expected inserts. Adapter and quality trimming of the PE150 and PE250 Illumina read sets was carried out using Trimmomatic v.0.36 [198] with minimum length set to 36 bp (2.14.1). During this step, almost a quarter of PE250 read pairs were dropped. The relatively lower quality of the PE250 reads might reflect an increased sensitivity of the sequencing chemistry to environmental contaminants and could indicate that PE150 is a better choice for robust shotgun sequencing of metagenomic DNA that may contain inhibitors. Base-calling Nano_reads using Albacore v.2.2.6 (Oxford Nanopore) with default settings yielded 253,097 passed (45-65,289 bp length) and 65,717 failed reads.

4.2.2 Metagenome assembly

Metagenome assembly of Illumina short read data was conducted using metaSPAdes ([166]; 2.14.3). In order to investigate the effect of the low quality PE250 data on the metagenome assembly, assemblies for PE150, PE250 and a combination of both were constructed (Appendix A13). For hybrid assembly using short and long reads, passed Nano_reads were aligned to the assembled contigs from the PE150 only

assembly and aligned reads extracted from the passed Nano_reads using samtools v.0.1.19 resulting in 19,486 reads for hybrid assembly in metaSPAdes (2.14.2). This option first constructs an assembly graph from the PE150 reads using the metaSPAdes module followed by resolution of repeat regions and closure of gaps in the assembly graph using the long reads [166,224]. PE250 was omitted as an option for this hybrid assembly due to the low-quality assembly produced from this read set (Appendix A13). In a final attempt to further improve assembly quality of the *C. mycofijiensis* microbiome a hybridSPAdes was run on the PE250 reads with the PE150 or the PE150_plus_Nano assembly supplied as trusted contigs. These trusted contigs are used for the initial assembly graph construction as well as repeat resolution and gap closure (SPAdes v.3.12.0 Manual).

The assembly of PE150 reads had an N50 of 2,987 and total length of large contigs (≥5000 bp) was 207,685,886 bp (Appendix A13). This assembly was used as a baseline in order to determine whether inclusion of PE250 and/or hybrid assembly using Nano_reads improved assembly statistics. N50 appears the most commonly used metric to assess assembly quality and compare assemblies and is defined as the contig size, which equal to or larger account for 50% of the genome (in this case metagenome). Hybrid assembly of PE150_plus_Nano had the highest N50 (3,008) and is superior to the PE150 assembly in terms of the total length of large contigs (≥5000 bp) (Appendix A13). The most large contigs (≥5000 bp) were obtained from the PE250_onPE150 assembly and the biggest total length of large contigs (≥5000 bp) was achieved by the PE250_on_PE150_plus_Nano assembly. However, both these assemblies have a significantly smaller N50 and can be considered more fragmented based on the distribution of contig sizes, i.e. whilst having almost three times the number of contigs, more than three quarters are below 1000 bp (Appendix A13). A trade-off between reducing fragmentation and increasing the length of larger contigs is apparent. Both the PE150_plus_Nano assembly, as an example of low fragmentation, and the PE250_on_PE150, as an example of prioritising on total length of large contigs, were binned and dereplicated as described in 2.14.4 and 2.14.5. Based on the higher number of dereplicated bins returned (Appendix A11), all analyses from here on were conducted with the PE150_plus_Nano assembly.

4.3 Phylogenetic analysis of *C. mycofijiensis* by extracting ribosomal RNA sequences

Taxonomic identification of *C. mycofijiensis* samples was carried out based on visual inspection at the time of collection. To provide further evidence that this assignment was correct, the barrnap algorithm (Torsten Seemann 2015) was used to isolate eukaryotic ribosomal sequences from the metagenome assemblies (2.14.7). This analysis returned a seemingly complete 18S rRNA gene (1,764 bp) and resulted in the assignment to the Dictyoceratida order (90.27% identity), which is consistent with the full taxonomy of *C. mycofijiensis* as identified in the World Porifera Database (http://www.marinespecies.org/porifera/porifera.php?p=taxdetails&id=165305) and provides support for the original taxonomic assignment of this sample.

4.4 Binning and bin analysis

Numerous binning algorithms exist and as of yet there does not appear to be a universally accepted binning algorithm for application in a particular setting [167,169,225]. Use of multiple algorithms, followed by consolidation or dereplication of the resulting bins has recently been suggested as a way to overcome this problem and combine the strengths of different binning algorithms in order to recover metagenome assembled genomes (MAGs) [167,169,225]. Three binning algorithms were implemented here, these were Metabat2 [178], Maxbin2 [226] and Autometa [225]. Of these algorithms Metabat2 and Maxbin2 are well established and highly cited [126,169,171,225,227], whereas Autometa is a newer algorithm developed by the Kwan group with the specific intent of resolving the microbiome of marine invertebrates.

Metabat2 uses contig abundance probabilities (in the case of multiple samples) and tetranucleotide frequencies to assign contigs to bins or leave them unbinned [178]. It is commonly used as a baseline comparison to other binning methods [169,171,225,227]. Maxbin2 uses tetranucleotide frequencies and coverage to construct bins [226] and is also a commonly used baseline comparison as well as being applied in research [169,171,225,227]. Autometa is a reference-based binning algorithm that uses sequence homology to construct kingdom bins, Prodigal [228] to taxonomically identify contigs and 5-mer frequencies as well as dimension reduction using Barnes-Hut Stochastic Neighbor Embedding (BH-tSNE) as input for *de* novo binning by DBSCAN [225]. It is

specialized for binning of host-associated and highly complex datasets as it, due to genome reduction and adaptation of symbiotic microbes, does not assume all marker genes need to be present for completeness and does not use marker genes to pre-calculate the number of bins [225].

Using multiple binning algorithms results in redundant bins, which need to be clustered and dereplicated. A recently devised tool for achieving this aim is dRep, which conducts two alignment steps to identify and cluster related or identical bins [167]. The first alignment step is fast but imprecise and uses MASH to differentiate a set of genomes up to a level of 90% average nucleotide identity (ANI) forming primary clusters. These primary clusters are then further differentiated up to a level of 99% ANI (default setting) using the more sensitive alignment tool ANIm [167].

4.4.1 Bin size and quality

From the PE150_plus_Nano assembly Metabat2 produced 81 bins, Maxbin2 96 bins and Autometa 101 bins. Dereplication of these three bin sets using dRep with default settings, which only considers genomes of ≥500 kb length, >75% completeness, <25% contamination and <25% strain heterogeneity, resulted in 46 unique MAGs. Genomes can be considered near-complete when ≥90% complete and ≤5% contaminated [126]. Of the 46 unique MAGS from the PE150_plus_Nano assembly, 22 were near-complete genomes (Appendix A15). The MAG with the highest completeness and lowest contamination from the PE150_plus_Nano assembly was constructed by MetaBAT2 and was 97.44% complete and 1.71% contaminated.

4.4.2 Phylogenetic assignment of MAGs

CheckM is integrated in the dRep workflow and was used here to infer rough taxonomic classification based on the marker lineage attributed to each MAG. Referencing this marker lineage to the Genome Taxonomy Database (GTDB) identified the Candidate Phylum “Patescibacteria”, the Protebacteria and Acidobacteria as the most abundant taxonomic groups (Figure 13).

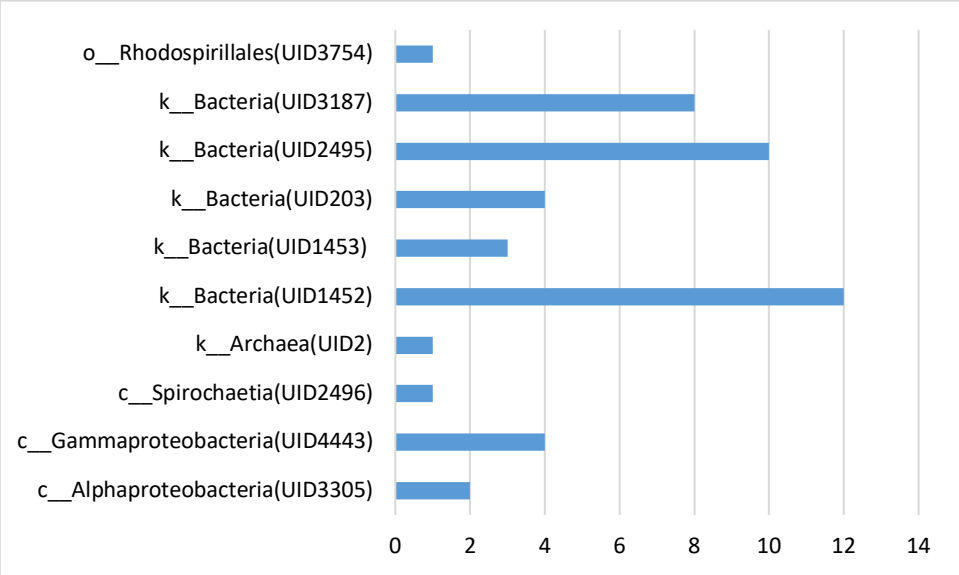


Figure 13: Number of MAGs per marker lineage from the PE150_plus_Nano assembly as identified by dRep (Total number of bins = 46).

Plotting coverage of contigs over the percentage GC content of contigs is a commonly used method to visualize bins, as contigs belonging to the same taxonomic group should have a similar GC content and coverage [201]. Color was added here based on the marker lineage identified by checkM for each of each of the 46 unique MAGs (Figure 14). This plot clearly shows the aggregation of contigs belonging to the same marker lineage in distinct regions of the plot, further validating the correct recovery and identification of the MAGs recovered from the *C. mycofijiensis* microbiome.

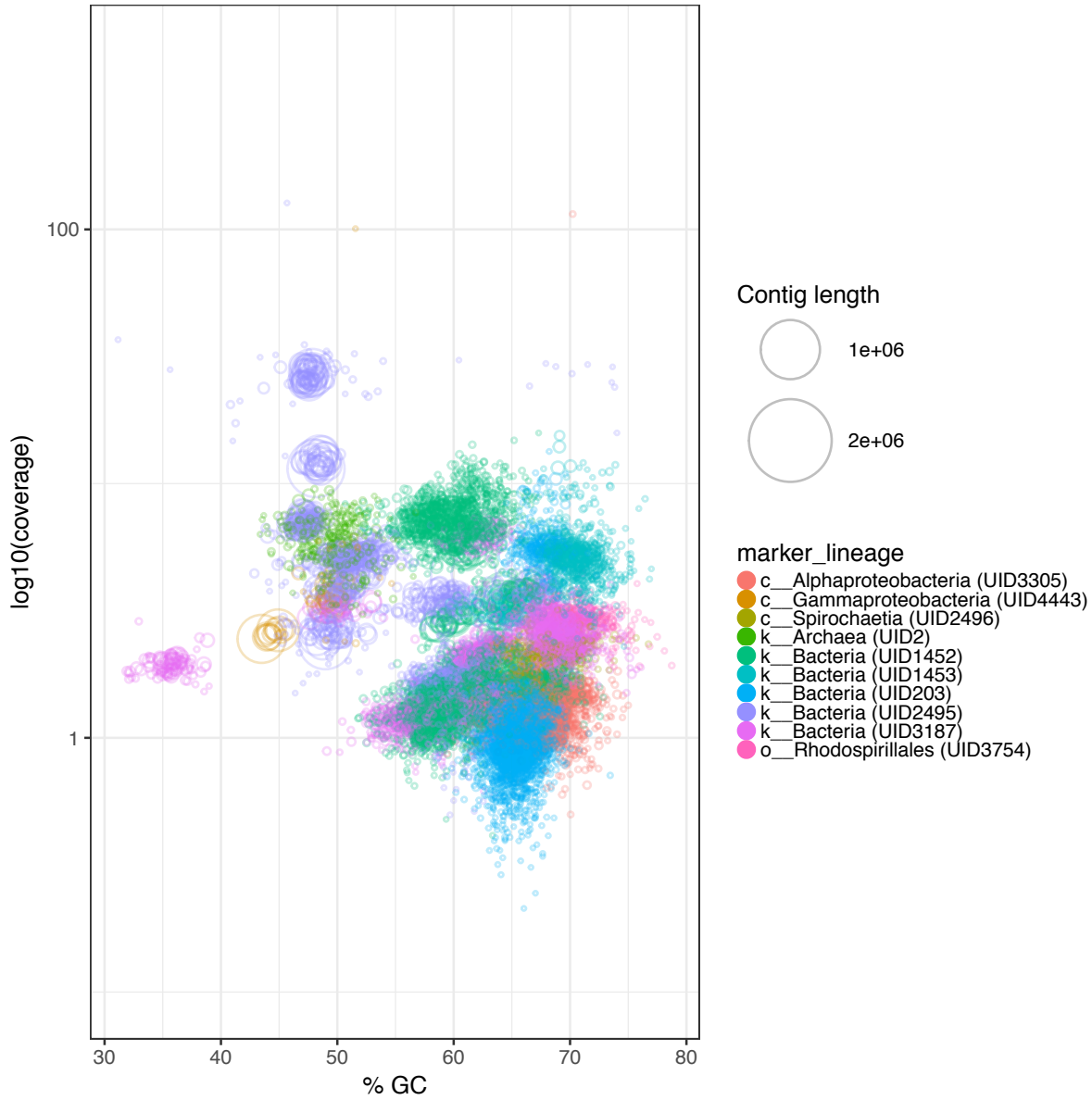


Figure 14: Blobplot of PE150_plus_Nano contigs part of dRep dereplicated bins %GC content of contigs is plotted against log-transformed coverage of contigs (maximum k-mer coverage as calculated by metaSPAdes) and contigs then colored by marker lineage as identified by checkM with size of the circle indicating size of the contig (2.14.10).

4.5 Analysis of secondary metabolism within the *C. mycofijiensis* metagenome

A standalone install of antiSMASH was used to identify and annotate secondary metabolite biosynthetic gene clusters in each of the 46 unique MAGs as well as all length-filtered (≥5 kb) PE150_plus_Nano contigs (2.14.8) derived from the complete assembly. A length-filter of ≥5 kb was chosen because complete secondary metabolite clusters are very unlikely to be shorter than 5kb and the runtime of antiSMASH was

consequently dramatically reduced. Based on the checkM marker lineage identified, Acidobacteria (k__Bacteria(UID3187)) and Proteobacteria were the most prolific producers of secondary metabolites (Figure 15). They encode numerous PKS and RiPP BGCs with the bacteriocin-lanthipeptide and lanthipeptide cluster found unique to the Acidobacteria and the head-to-tail cyclized peptide cluster unique to Proteobacteria (Figure 15). Based on rough taxonomy, these two taxonomic groups are among the three most abundant marker lineages identified (Figure 15). Other unique chemistries include the lassopeptide BGC found in the Rhodospirillales order, as well as a Type II polyketide BGC attributed to the Candidate Phylum "Patescibacteria". Alphaproteobacteria, Gammaproteobacteria and Acidobacteria appear to be the prominent producers in *C. mycofijiensis*. This contrasts with *T. swinhoei*, where Entotheonellaeota have been identified as the prominent producers of secondary metabolites [29,30].

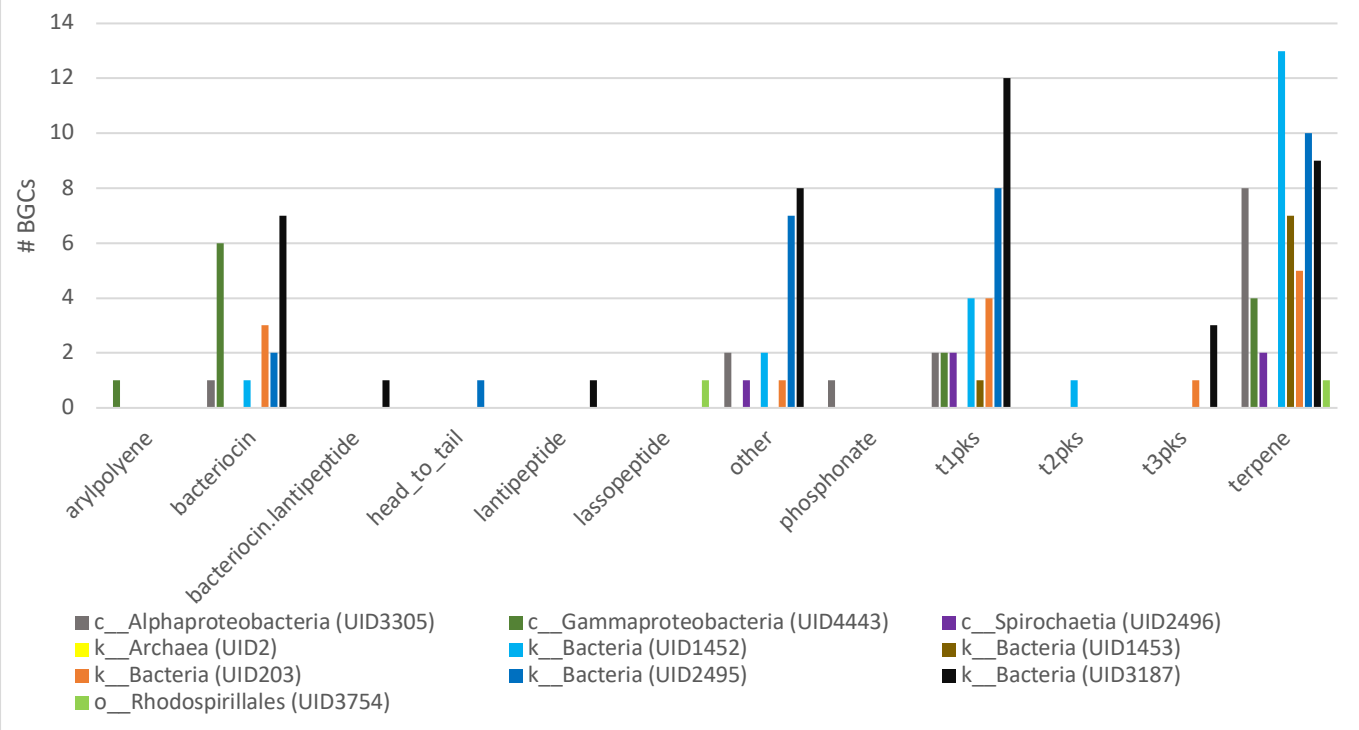


Figure 15: Number of BGCs identified from the PE150_plus_Nano assembly per secondary metabolite class and marker lineage secondary metabolite class was identified by antiSMASH and marker lineage by checkM. T1pks = Type 1 Polyketide synthase; t2pks = Type 2 Polyketide synthase; t3pks = Type 3 Polyketide synthase.

4.5.1 Search for (-)-zampanolide, latrunculin A and laulimalide A candidate BGCs

In order to identify candidate BGCs encoding the compounds of interest latrunculin A, laulimalide A and (-)-zampanolide, the results obtained from antiSMASH were analysed for characteristic aspects of the putative BGC layout elucidated using retrobiosynthesis (Figure 12). Noteworthy is that, while many sponge PKS were identified as trans-AT PKS [56], only cis-AT PKS were identified by antiSMASH from *C. mycofijiensis* (Figure 15).

Focusing first on the epoxide group found in laulimalide A, seven BGCs in the antiSMASH output from the length-filtered assembly were identified as containing a cytochrome P450 gene part of a "secondary metabolism Cluster of Orthologous Group" (SMCOG). Two of these BGCs were classified as Other containing SMCOG1007, three as Terpene, one as Type III PKS and one as Type I PKS containing SMCOG1034. Another BGC contained a luciferase-family protein (SMCOG1251) and was identified as a Type I PKS but none appeared to contain flavoproteins related to secondary metabolism. The BGC containing the luciferase family protein (SMCOG1251; Figure 16 ctg1_1568) stood out because of its high homology (e-value = $1.1e^{-64}$) to luciferase-like monooxygenases identified in BGCs deposited into the MiBIG database. Also, part of the SMCOG1251 was MsnO8 (flavin-dependent monooxygenase) in the mensacaricin BGC from *Streptomyces bottropensis*, which in conjunction with MsnO3 (flavin reductase) was shown to be responsible for the formation of the epoxide moiety during the last step of biosynthesis [216]. Further evidence indicating that this may be the laulimalide A BGC is, that the biggest ORF of the BGC (ctg1_1579, Figure 16) is almost identical (e-value = 0) to the epothilone BGC, more specifically EpoD and EpoF, deposited in MiBIG. Epothilone A and B are structurally very similar to laulimalide A and also contain the characteristic epoxy group [214,219]. Pyran synthase domains are very similar to dehydrogenation domains but have a deletion in the active site [213] and while a pyran synthase domain was not specifically identified, several of the ORFs returned top hits for dehydrogenases that might have been misannotated by antiSMASH.

The candidate laulimalide A BGC was identified from a ~2.76 Gbp contig attributed to a MAG (cluster_DBSCAN_round14_0) recovered from the PE_150_plus_Nano

assembly. Based on the checkM marker lineage, this MAG was a Proteobacterium, 81.25% complete, 0% contaminated, had 0% strain heterogeneity and a size of ~3.31 Gbp (Appendix A15). Although not identified as 100% complete this may well be the whole genome of the bacterium as genome reduction is common in symbiotic bacteria [139,225,229]. It is not contaminated, does not display any strain heterogeneity and is of a size considered to be normal for bacteria. Further evidence that this genome is complete, is given by the high quality assembly of contigs associated with this MAG. This MAG bin consists of only three contigs, which includes the largest contig of the PE_150_plus_Nano assembly (~2.76 Gbp) at a coverage of 4.31, a ~0.54 Gbp contig at a coverage of 4.44 and a 3.68 Kbp contig at a coverage of 0.59.

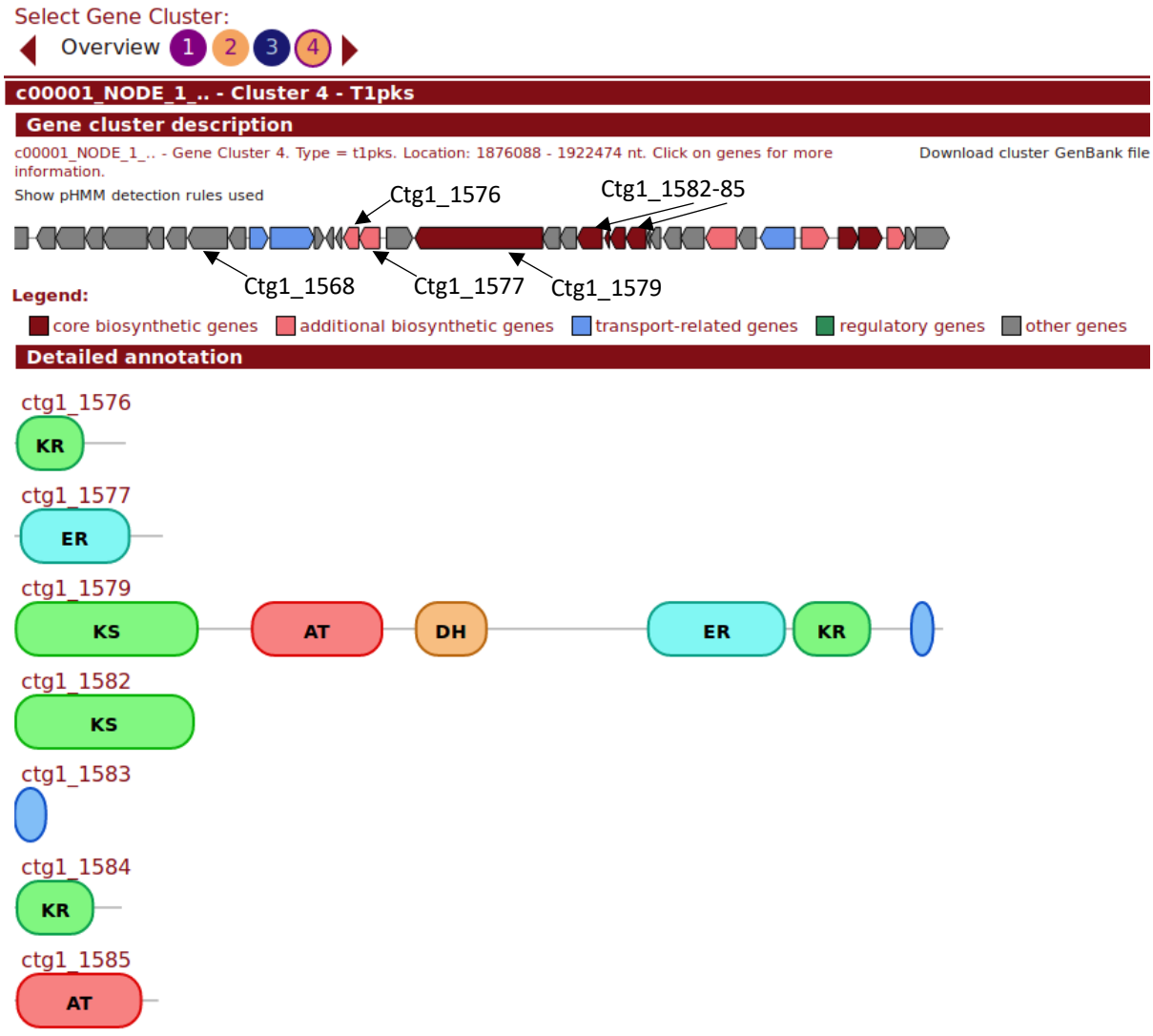


Figure 16: Candidate BGC of laulimalide A as identified by antiSMASH from the MAG cluster_DBSCAN_round14_0 recovered from the PE150_plus_Nano assembly.

No NRPS-PKS hybrid clusters were specifically identified but one Type I PKS BGC stood out as it contained a DH-domain and MT-domain followed by a module containing an ER-domain (Cluster 3, Figure 17), which is what we would expect to find in the latrunculin A BGC (Figure 17). This Type I PKS BGC was attributed to the MAG bin.14.fa.maxbin2 (Appendix A15), which contained another Type I PKS, a Type III PKS and 3 terpene BGCs. The other Type I PKS cluster is most likely incomplete as it lies at the start of a contig (Cluster5, Figure 17). The MAG was 100% complete and 7.89% contaminated with a size of 4,760,373 bp but contained 3.45% strain heterogeneity. While still low, this strain heterogeneity indicates that the metaSPAdes may not have been able to separate this MAG from one or multiple closely related strains [201,230], which in turn may have affected the assembly of the BGC. The candidate BGCs presented here are incomplete but give indicative evidence that the BGCs for latunculin A, laulimalide A and (-)-zampanolide are present in the microbiome of *C. mycofijiensis*. There are various reasons why these BGCs may be incomplete, such as fragmentation of the assembly due to repetitive elements in the BGCs or a lack of coverage of that particular genomic region [231,232].

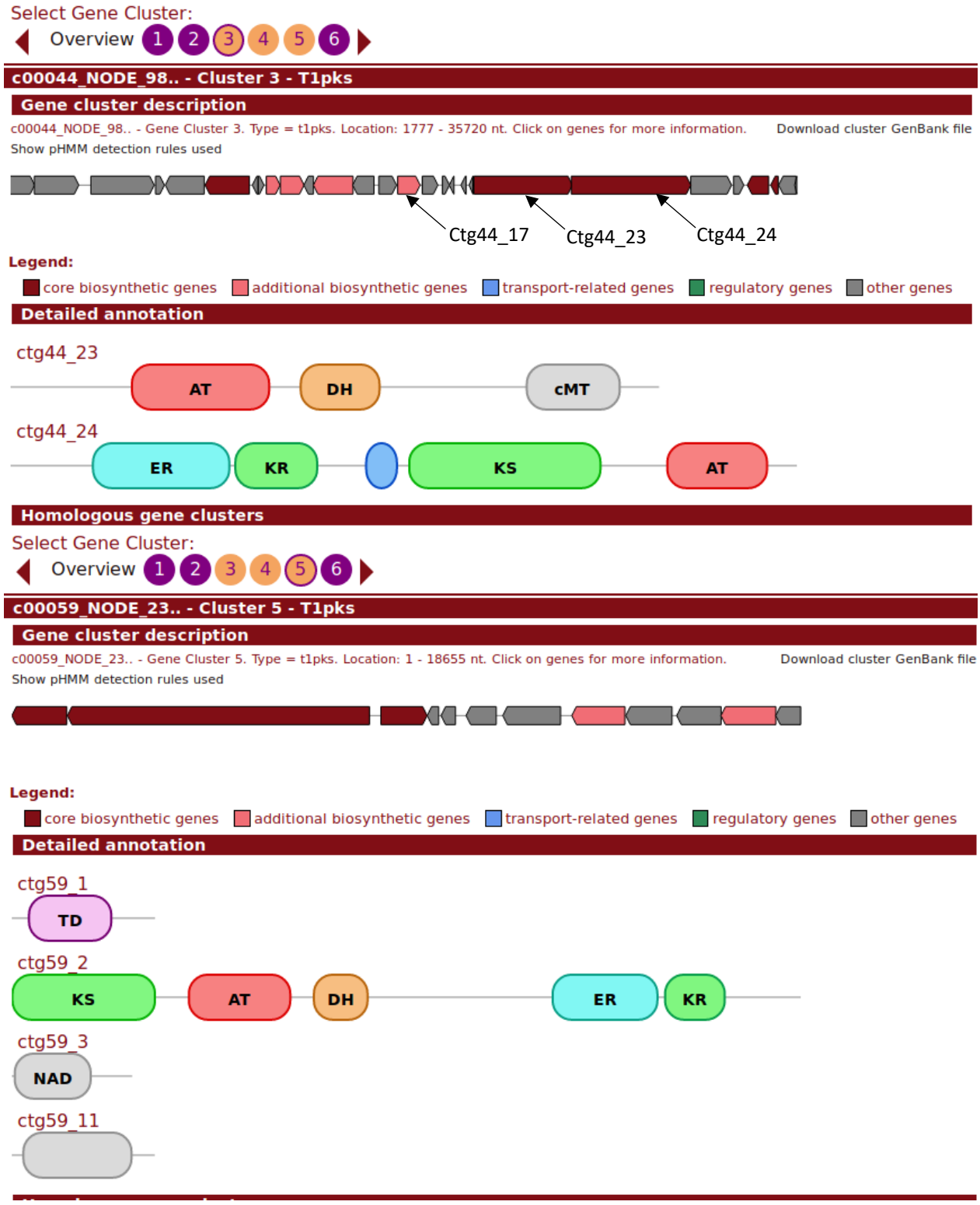


Figure 17: Candidate BGC of Iatrunculin A (Top) and other PKSI cluster identified by antiSMASH from the MAG bin.14.fa.maxbin2 recovered from the PE150_plus_Nano assembly.

4.6 Construction of large insert cosmid library from *C. mycofijiensis*

In order to capture and archive the bacterial diversity of the *C. mycofijiensis* microbiome for future functional studies, a large insert cosmid library of sufficient size to redundantly cover the microbiome was constructed. Metagenomic DNA was extracted from 40 g of a fresh *C. mycofijiensis* sample using the protocol described in Methods 2.9. Metagenomic DNA extraction yielded approximately 380 µg of crude material. To derive cloning-competent DNA, approximately 125 µg of this DNA was subjected to preparative agarose gel electrophoresis for contaminant removal and size-selection as described in Methods 2.3.1 to yield 19 µg of clean HMW DNA. Seven large scale packaging reactions were then carried out (Table 4, Methods 2.7) during which the size-selected DNA was end-repaired (NEBNext® End Repair Module) and ligated to digested and dephosphorylated pWEB-TNC vector using Quick Ligase Kit™ to the manufacturer's instructions.

Plate	Strain	Ligation ratio DNA:Vector	Ligation volume (µl)	# Ligations	Total insert DNA (ng)	Yield (#colonies)
CS_1	EC100	400ng:440ng	20	8	3,520	192,000
CS_2	Δ*entD*	125ng:250ng	10	16	2,000	25,000
CS_3	Δ*entD*	400ng:440ng	20	8	3,520	40,000
CS_T	EC100 & Δ*entD*	400ng:440ng	20	4	1,760	60,000
CS_4	Δ*entD*	200ng:220ng	10	8	1,600	41,000
CS_5	Δ*entD*	200ng:220ng	10	12	2,400	90,000
CS_6	Δ*entD*	200ng:220ng	10	8	1,600	35,000
Total					16.4µg	483,000

Table 4: *C. mycofijiensis* library construction Details for the large scale packaging reactions carried out and their respective yield, which resulted in the plates CS_1 - CS_6 & CS_T forming the *C. mycofijiensis* cosmid library.

In order to determine library size and confirm that clones contained inserts, dilutions from each packaging reaction were plated on agar plates and colonies counted to determine overall efficiency of the packaging reaction. Cosmid DNA was then isolated from 12 individual colonies for each of the packaging reactions and insert size assessed using agarose gel electrophoresis. All cosmids isolated (84/84) were visible considerably above the 10 kb marker of the Hyperladder 1 kb (Appenidx A9) indicating that they contain large inserts. The 483,000 clones resulting from 16.4 µg of HMW

DNA used for library construction indicate a packaging efficiency of 29,451 pfus/µg. This packaging efficiency is approximately 5 times higher than the 6464 pfus/µg observed during SWNZ library construction but still significantly lower than the packaging efficiencies observed during testing of the packaging extracts ($2.30*10^8$ pfus/µg). Lower efficiencies are common for complex metagenomic samples, though this is not typically reported in the literature (Jeremy Owen, personal communication).

In order to provide a final assessment of library diversity and composition, cosmid DNA from 32 of the clones isolated was sequenced using primers (2.14.11) targeting conserved sites flanking the insert site (Figure 4). Sequences were then analysed using BLASTx to give a rough characterization of the taxonomy and origin of the DNA (Appendix A14). With an average identity of 61.49%, 29 of the 32 inserts contained DNA inserts of bacterial origin that was not *E. coli* or any other common laboratory bacterium. None of the inserts were identified as those of a marine sponge.

Phylum level assignment of the genetic diversity captured includes notable taxonomic groups frequently found in sponges, such as Proteobacteria and Chloroflexi. Other phyla captured include Cyanobacteria, Actinobacteria, Bacteroidetes, Firmicutes, Acidobacteria, Candidate Phylum “Rukobacteria” and Candidate Phylum “Gemmatimonadetes”. Many of these phyla were also identified from the MAGs by checkM, most notably the abundant presence of Proteobacteria. Numerous of these phyla have been shown to yield natural products, in particular Actinobacteria and Proteobacteria (Yilmaz et al. 2015, De Mol et al. 2018, Yao et al. 2018). More specifically members of the *Nonomuraea* genus (Table 5, F11) have been shown to produce several biologically active natural products [233], as have members of the *Cystobacter* genus (Table 5, E05;[234]), and members of the *Cupriavidus* genus (Table 5, D09) are known to contain NRPS clusters [235]. Particularly noteworthy is the presence of an insert deriving from a putative *Sorangium* species (Table 5, D08), which is a prominent myxobacterial natural product producer [236]. Collectively these results indicated that the library constructed here was comprised overwhelmingly of large bacterial gDNA inserts obtained from the *C. mycofijiensis* microbiome and covered a breadth of taxonomic diversity.

Well	Description [Organism]	Accession	Pairwise Identity	Bit-Score	E-Value	Sequence Length
D08	hypothetical protein [Sorangium cellulosum]	WP_020733992	46.50%	153.30	9.57E-42	159
D09	hypothetical protein [Cupriavidus sp. amp6]	WP_029049114	71.90%	266.93	1.01E-82	185
E05	AAA family ATPase [Cystobacter ferrugineus]	WP_071900752	67.70%	202.99	7.35E-60	167
F11	radical SAM protein [Nonomuraea candida]	WP_043620474	66.70%	141.35	7.31E-36	99

Table 5: End-sequencing results for the *C. mycofijiensis* cosmid library (Appendix A14 for full table of results).

4.7 Phosphopantetheinyl-transferase enrichment functional screen of *C. mycofijiensis* library

As an initial screen for the biosynthetic diversity in the metagenome library, a phosphopantetheinyl-transferase (PPTase) complementation screen was employed. PPTases are enzymes that post-translationally modify NRPS and PKS enzymes by addition of a coenzyme-A-derived phosphopantetheine molecule to a conserved serine residue in the carrier protein. This is essential for the activity of the biosynthetic machinery as it allows tethering of the monomer substrates and the intermediate compound [237]. Using the host strain EC100 Δ*entD*, which has the native PPTase (entD) knocked out and is thus unable to complete the biosynthesis of the siderophore enterobactin, one can enrich the cosmid library by plating on iron-deficient media [120]. Since PPTases are frequently found in NRPS and PKS BGCs, complementation of PPTase activity using this survival screen is an efficient means for recovering BGCs from a metagenome libraries [120]. A significant advantage of this functional screen is that it relies on the expression of only one gene, thus increasing the probability of it being expressed, while some other functional screens require expression of whole BGCs [108,120,238].

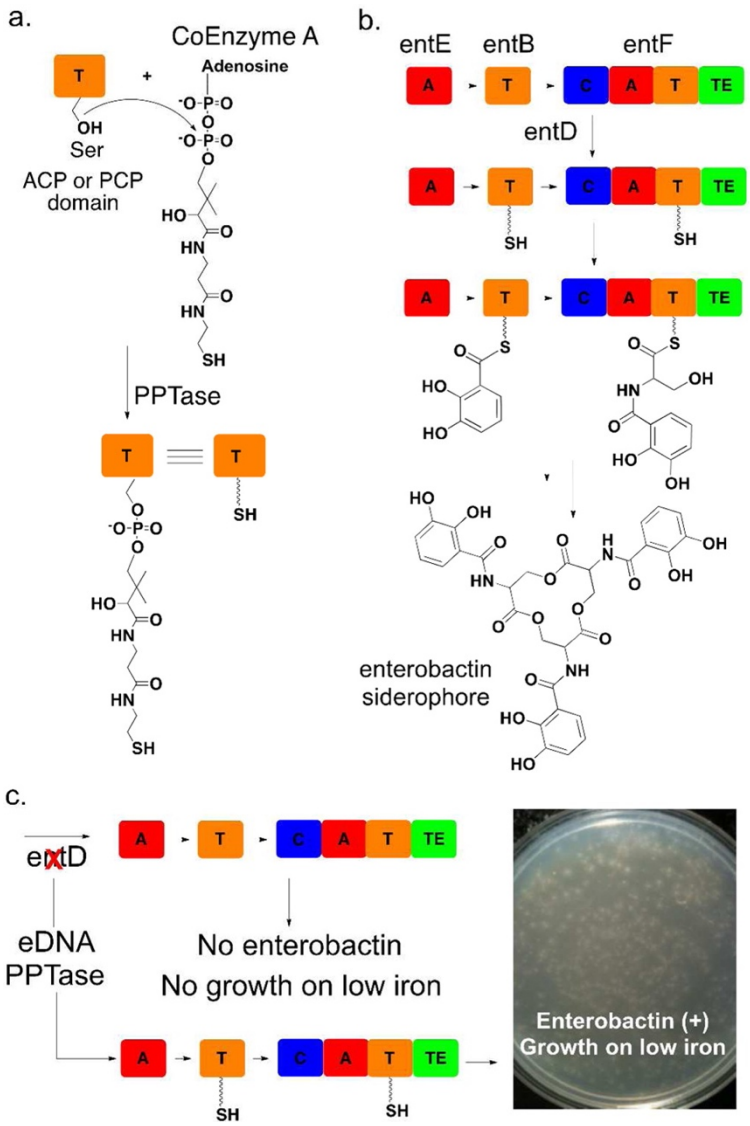


Figure 18: PPTase functional screening (Reprinted with permission from [120], Copyright 2013 American Chemical Society.) a. the CoEnzyme A-derived phosphopantetheinyl group is attached to the conserved serine residue on the carrier protein to form a functional thiolation (T) domain b. function of the *entD* encoded PPTase in the enterobactin BGC c. *entD* deletion causes mutants not to grow on low iron media as enterobactin is not being produced, which is rescued if a functional PPTase is expressed from the cosmid insert. Approximately 5 milllion eDNA clones from the metagenomic library constructed in *E. coli* in the paper were plated on iron-deficient media to produce the final image.

Complementation screening was carried out as described in 2.10 and 2.11 and several colonies were successfully recovered, indicating that siderophore production was complemented by a PPTase (Appendix A10). Four of the recovered clones were sequenced, of which one appeared to contain a fragment from an NRPS biosynthetic system (Table 6, E12) and another a possible alpha-sialidase, which is involved in the biosynthesis of sialic acid, a biosynthetic precursor [239]. Overall, the results from the preliminary PPTase enrichment indicated that PKS and NRPS BGCs were present in the library and could be captured.

Well	Description	Name	% Pairwise Identity	Bit-Score	E Value	Sequence Length
E10	HNH endonuclease [Nitrosospira multiformis]	WP_113068581	69.10%	419.08	8.12E-145	291
E12	Enterobactin synthase EntD component [Shigella dysenteriae 1617]	AHA63490	64.50%	88.58	3.26E-19	76
G01	thiamine pyrophosphate-binding protein [Sulfitobacter sp. AM1-D1]	WP_071970173	68.20%	325.09	1.22E-104	245
H01	exo-alpha-sialidase [Candidatus Poribacteria bacterium]	RKU35250	56.90%	125.56	2.03E-32	102

Table 6: End-sequencing results for cosmids recovered from the PPTase functional screen of the *C. mycofijiensis* cosmid library Sanger-sequenced cosmids were analyzed using Blastx with standard settings as part of Geneious Prime.

4.8 Discussion

Metagenome sequencing, genome binning and secondary metabolism analyses of the *C. mycofijiensis* marine sponge are, to the best of my knowledge, the first exploration of this chemically rich marine sponge species. As assembly quality was known to be a critical step in the process of recovering MAGs and identifying BGCs [166,171,240,241] several combinations of the different Illumina and Nanopore sequencing reads were trialed to obtain a high quality metagenomics assembly. The PE150_plus_Nano hybrid assembly was the best in terms of contiguity. This is consistent with the demonstrated ability of long reads to act as scaffolds for high-quality metagenomic assemblies [231,240,242]. The coverage achieved by Nanopore sequencing here was very low and in an attempt to generate further long read data Pacbio sequencing was attempted from 10 µg of HMW metagenomic DNA. Unfortunately this attempt failed during library construction, not permitting implementation of further long reads in the scope of this thesis. Nonetheless, the small number of long reads obtained was sufficient to slightly improve the short read assembly and consequently allow the recovery of 46 unique dRep dereplicated MAGS, of which 22 were near-complete genomes. Future work with this sponge will aim to further improve the quality and contiguity of this assembly by collecting further long and short read data.

161 BGCs were identified from the 5 kb length-filtered assembly, of which 123 could be attributed to one of the dRep dereplicated MAGs. This further demonstrates the utility of DNA sequencing to associate BGCs with MAGs, especially from marine

environment [4]. A diverse range of secondary metabolite clusters including Type I-III PKSs, RiPPs (namely lanthipeptides and a head-to-tail cyclized peptide) as well as other bacteriocin and terpene clusters were recovered. The vast majority of BGCs showed low homology to known clusters, indicating the presence of potentially novel chemistry.

Retrobiosynthetic analysis and search of antiSMASH outputs revealed putative BGCs with features that might indicate they encode one of the three target compounds (-)-zampanolide, laulimalide A and latrunculin A. While these results are preliminary, they provide a lead for future studies that aim to obtain more contiguous assemblies. It is possible, that due to the low identity to known references, part or all of the biosynthetic pathway may have been missed by the HMMs employed by the antiSMASH package. Future studies will include development of more sensitive profile HMMs that are based on a variety of symbiont specific polyketide sequences.

A large-insert cosmid library of 483,000 clones was also constructed from microbial DNA isolated from *C. mycofijiensis* and bacterial inserts confirmed by end-sequencing. Based on this end-sequencing data at least 9 phyla including Acidobacteria, Actinobacteria, Bacteriodetes, Chloroflexi, Cyanobacteria, Firmicutes, Proteobacteria, Candidate Phylum “Rukobacteria” and Candidate Phylum “Gemmatimonadetes”, were covered in the library. Most of these phyla were also identified by shotgun sequencing, indicating good phylogenetic and thus likely secondary metabolite coverage in the library. An initial attempt to enrich this library for NRPS and PKS cluster returned two cosmids related to BGCs from said enzyme families but sequencing of the whole cosmids as well as expression studies will have to be carried out to unequivocally confirm this. It can thus be concluded, that the functional screen worked but may not be optimal for library enrichment, as seen by the disparity between the number of clones recovered and the number of NRPS and PKS BGCs identified using antiSMASH. This discrepancy is likely due to low expression rates of the phylogenetically distant genes in *E. coli*.

This cosmid library will be an invaluable resource in future efforts to recover BGCs, discover potentially new natural products and design functional studies of this interesting sponge holobiont. Ultimately it is hoped that complete pathways for (-)-

zampanolide, laulimalide A and latrunculin A will be recovered from this metagenomic library and that these will serve as the basis for the sustainable production of these compounds by heterologous expression.

Chapter 5

Comparative analysis of genome resolved assemblies for six Tongan marine sponge metagenomes

Initial sequencing and assembly efforts with a single HMA marine sponge (*C. mycofijiensis*, Chapter 4) showed that high-quality, genome-resolved assemblies could be obtained from a relatively small amount (<20 Gbp) of paired-end 2x150 bp Illumina data. This work also showed that such assemblies provided fertile ground for the discovery of BGCs potentially encoding new medically or ecologically relevant natural products. This chapter builds on these results and describes metagenomic sequencing, assembly and binning of five further sponge microbiomes. A combined comparative analysis of the six microbiomes which were elucidated as part of this thesis along with seven other marine sponge microbiomes is also described. The collection of metagenomes examined contains two samples of *C. mycofijiensis*, which were collected at different locations, as well as five sponges from that same location, allowing the relative effects of geography and host species on microbiome composition to be examined. Key questions sought to be answered in this chapter were: 1) What are dominant bacterial taxa and how efficiently can their genomes be recovered from a metagenome? 2) How does the marine sponge microbiome vary with geography and among/within species? 3) What is the richness of BGCs encoding secondary metabolites in microbial sponge metagenomes? 4) How does the secondary metabolism of marine sponge microbiomes vary with geography and among/within species?

5.1 Isolation and direct shotgun sequencing of metagenomic DNA from five additional Tongan sponges

In order to maximise the chances of obtaining high quality metagenome assemblies, Tongan sponges with high microbial biomass needed to be identified. To this end, metagenomic DNA was isolated (2.9) from 20 different Tongan sponges collected from a different location (Pete's Cave, Appendix A 16) than *C. mycofijiensis* described in Chapter 4 (CS783). Visualizing these metagenomic DNA extracts using gel electrophoresis, clearly showed the differences of bacterial biomass between sponges (Figure 19). Five of the sponges that yielded large amounts of metagenomic DNA were chosen for shotgun sequencing. A sample (CS200) that was preliminarily identified as *C. mycofijiensis* was included among these to allow subsequent comparison of the microbiomes of the same species of sponge found at different locations as well as comparing the microbiome of different sponges species from the same location.

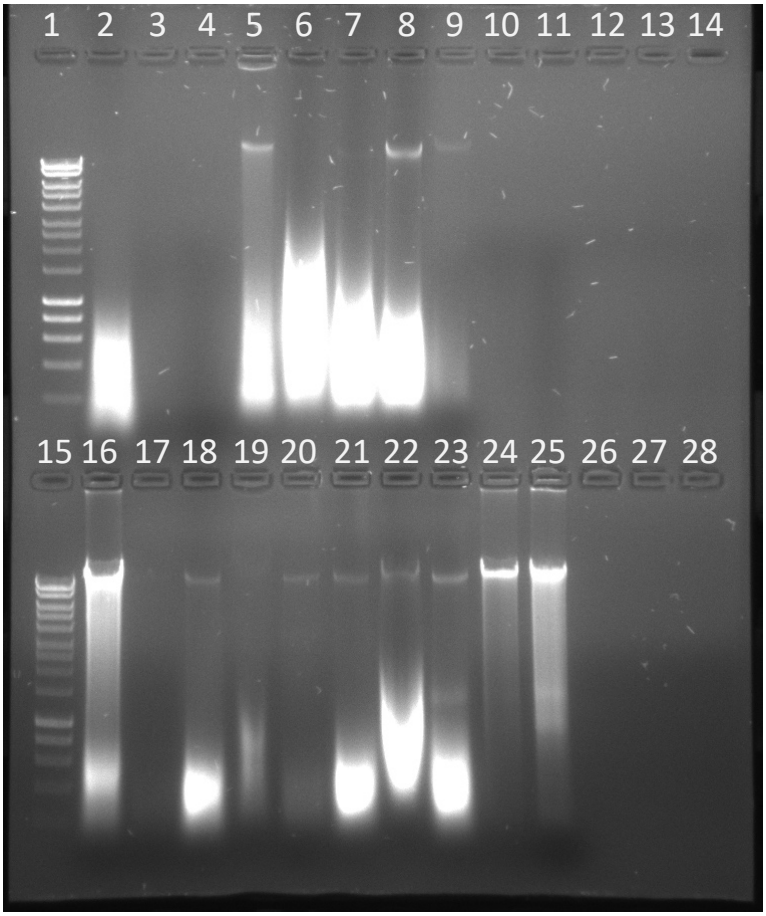


Figure 19: DNA extracts obtained from 20 different Tongan sponges using previously established extraction protocols (described in 2.9) scaled down to 2 g of sponge tissue (10 samples per row with HyperLadder^{TM} 1 kb in leftmost lane).

Figure 20: Photos of the 6 Tongan sponges (photos taken by Rob Keyzers) *C. mycofijiensis* = 783; CS200 = 834; CS202 = 837; CS203 = 839; CS204 = 841; CS211 = 854.

A larger scale metagenomic DNA was then carried out from approximately one gram of tissue for each of the five selected Tongan sponges as described in 2.9. This DNA was further cleaned and concentrated for Illumina sequencing as described in 2.12. Library preparation (TruSeq, PCR free, 500 bp insert size) and sequencing were carried out by Annoroad (China) with equal amounts of each of the five libraries split over an entire lane of paired-end (2x150) HiSeq4000 (Appendix A16). Datasets returned for each sponge ranged in size between 20-30 Gbp with the biggest dataset returned for the CS200 sponge (Appendix A16).

5.2 Assessment of sequence data quality

FastQC Version 0.11.5 (Babraham Bioinformatics) with default settings failed all five sponges on per sequence GC content. GC content, as defined by FastQC, should follow a bell-shaped normal distribution but all sponges had slightly shifted peaks, a narrower or wider bell-shape and some datasets showed two peaks. Although unusual, these uneven distributions were not considered problematic due to the fact that metagenomic DNA often contains a wide range of taxa and thus GC content. Some of the datasets were flagged with a warning for per tile sequence quality, which was found to be at the end of sequences, as well as irregular per base sequence content within the first 10 bps of the sequence. This was likely due to adapter content and both issues should be resolved during trimming.

5.3 Assembling the microbiome of five additional Tongan sponges

Trimming and assembly of raw reads was carried out using Trimmomatic v.0.36 as described in 2.14.1 resulting in 92.87% - 94.22% of paired reads surviving from the five datasets. All datasets were assembled independently using metaSPAdes v.3.12.0 (2.14.3) to yield the results summarized in Table 7 below. The metagenomic assembly of the CS200 sponge had a significantly higher N50, total length of larger contigs (≥5000 bp) and largest contig constructed (5.16 Mbp) than the metagenomic assemblies of the other four sponges (Table 7). Assemblies of the four remaining sponge metagenomes were still of good quality, with N50's above 2,000 bp despite the number of contigs approaching and in the case of CS211 exceeding 1 million. The average %GC content of the assemblies varied between 52% and 62% (Table

7), giving a preliminary indication that microbiome composition was broadly different among the samples examined.

	CS200	CS202	CS203	CS204	CS211
# contigs	238,671	399,539	426,214	397,797	440,822
# contigs (≥ 0 bp)	509,312	859,879	998,520	861,728	1,094,204
# contigs (≥ 1000 bp)	103,228	174,363	187,610	183,684	183,455
# contigs (≥ 5000 bp)	16,715	13,069	17,794	17,737	20,253
# contigs (≥ 10000 bp)	7,948	5,350	7,296	6,478	7,878
# contigs (≥ 25000 bp)	2,656	1,969	2,436	1,861	2,047
# contigs (≥ 50000 bp)	1,006	793	984	677	642
Largest contig	5,167,676	736,882	601,942	725,958	420,250
Total length	574,522,577	667,142,309	775,596,201	717,337,886	754,610,713
Total length (≥ 0 bp)	669,144,733	821,678,712	966,016,814	871,828,197	977,104,631
Total length (≥ 1000 bp)	481,113,822	509,765,864	610,356,293	568,125,031	577,200,279
Total length (≥ 5000 bp)	316,531,671	218,385,925	291,315259	253,933,041	274,487,893
Total length (≥ 10000 bp)	255,976,842	166,499,697	220,051,412	177,796,170	190,041,088
Total length (≥ 25000 bp)	175,281,647	115,315,894	146,745103	109,278,600	102,709,048
Total length (≥ 50000 bp)	118,494,071	74,308,892	96,300,168	68,983,871	54,682,750
N50	6,937	2,200	2,731	2,631	2,557
N75	1,507	1,039	1,126	1,141	1,045
L50	11,727	50,409	44,774	47,641	49,999
L75	61,735	165,130	160,575	155,463	172,454
GC (%)	58.98	52.64	58.71	57.76	61.85
Mismatches					
# N's	300	800	5,300	3,900	600
# N's per 100 kbp	0.05	0.12	0.68	0.54	0.08

Table 7: Results from the assemblies of the Tongan sponges CS200, CS202, CS203, CS204, CS211 quantified using quast v.5.0.0 with default settings.

5.4 Taxonomic identification of the six Tongan sponges

In order to obtain a preliminary taxonomic identification of the sponges from which metagenomic DNA was extracted, the barrnap v.0.9 algorithm (Torsten Seemann 2015) was used to recover eukaryotic ribosomal sequences from each of the five metagenomic assemblies. These were then analyzed using the SINA1.2.11 online tool as described in 2.14.7 (Table 8).

Sponge	# sequences returned	Length of sequence (bp)	Identity%	Taxonomic lowest common ancestor (Level of classification)
CS783	293	1,764	90.27	Dictyoceratida (Order)
CS200	401	1,764	90.27	Dictyoceratida (Order)
CS202	371	533	90.7	Verongiida (Order)
CS203	462	1,281	97.34	Heteroscleromorpha (Subclass)
CS204	454	1,798	89.86	Demospongiae (Class)
CS211	417	593	91.48	Dictyoceratida (Order)

Table 8: Sponge taxonomy based on 18S rRNA sequences extracted using barrnap v0.9 (Torsten Seemann 2015) with –kindom "euk" specified, returned reads analysed using the SILVA ACT web-tool (SINA 1.2.11, SILVA release 132) and top hit from the Porifera phylum listed here.

These extracted 18S rRNA sequences were then aligned to each other using BBMap v.38.31 with default parameters. The 18S rRNA sequence of CS200 and CS783 were 100% identical (Table 9), which, taking into account the morphological analysis, indicated that both these sponges were *C. mycofijiensis*. All other alignments were below the threshold for species assignment (99%), however it should be noted that the CS203 18S rRNA sequence was only partial (Table 9) and sequence similarity to the other three sponges may in fact be higher (Table 9).

	CS783	CS200	CS203	CS204
CS783				
CS200	100% 100%			
CS203	78.1% 92.4%	78.1% 92.4%		
CS204	100% 91.2%	100% 91.2%	100% 90.0%	

Table 9: 18S rRNA sequence alignment results Note that CS202 and CS211 were excluded due to the short 18S rRNA sequences returned (Table 8). Top number indicates the percentage of mapped bases and bottom number the percentage of matched bases (relative to mapped bases).

5.5 Binning and bin analysis of the six Tongan sponge microbiomes

In order to recover MAGs from the five additional Tongan sponge metagenome assemblies, contigs from each assembly were binned using Maxbin2, Metabat2 and Autometa (2.14.4) and the three bin sets dereplicated using dRep (2.14.5) as previously for *C. mycofijiensis*. The highest number of dereplicated genomes was

recovered from CS200 (76, Table 10). Approximately 50 dereplicated MAGs were recovered from each of the remaining metagenomic assemblies (Table 10).

Sponge	Metabat2	MaxBin2	Autometa	dRep	Near-complete MAGs	Most complete and least contaminated
CS783	81	96	101	46	22	97.44% and 1.71%
CS200	117	119	153	76	33	99.15% and 1.28%
CS202	87	121	103	55	26	97.80% and 0%
CS203	118	114	126	52	26	98.29% and 0.85%
CS204	114	122	98	50	19	98.29% and 0.85%
CS211	122	138	123	48	20	97.51% and 0.50%

Table 10: Number of bins recovered from the six Tongan sponges from the suite of binning algorithms used (Metabat2, Maxbin2 and Autometa) as well as number of dRep dereplicated genomes resulting from these. Note that dRep with default settings only considers genomes of ≥500 kb length, >75% completeness, <25% contamination and <25% strain heterogeneity. Also includes number of near-complete MAGs (≥90% complete and ≤5% contaminated) and the most complete and least contaminated MAG as identified by checkM.

5.5.1 Bin size and quality

Sizes of the dereplicated MAGs recovered from all six Tongan sponges varied between 1.46 Mbp and 7.47 Mbp (Appendix A15, A17-A21). While some of these MAGs may be incomplete, previous studies suggests that genomes of bacterial symbionts may be extremely reduced in size with genomes smaller than 300 Kbp confirmed [225,229]. Furthermore, Archaeal MAGs were recovered from each assembly and these are often <2 Mbp in size [243]. Cases of genome reduction aside, genomes can be considered near-complete when ≥90% complete and ≤5% contaminated [126]. The CS200 sponge yielded the most near-complete MAGs with 33, of which six had 0% contamination (Appendix A17). Five further near-complete MAGs from CS202 (Appendix A18) and a further six from CS211 had 0% contamination (Appendix A21). Fewer near-complete MAGs were recovered from the CS204 and CS211 than from CS202 and CS203 (Appendix A18-A21) despite assembly quality being comparable. CS200 also contained the most complete and least contaminated near-complete MAG (99.15% complete and 1.28% contaminated; Appendix A17).

5.5.2 Phylogenetic assignment of MAGs

Rough taxonomy was assigned to each MAG using checkM, which is run as part of dRep and taxonomically identifies bins based on a collection of 106 single copy marker

genes [199]. Referencing marker lineages to GTDB indicated the abundant presence of the bacteria from the Candidate Phylum “Patescibacteria” (k__Bacteria(UID1452)), Proteobacteria (k__Bacteria(UID2495)) and Acidobacteria (k__Bacteria(UID3187); Figure 21). The CS200 sponge was the only one to include members of the Veruccomicrobia phylum and members of the Planctomycete phylum in this sponge. Only the CS203 sponge’s metagenome included members of the Bacteriodetes phylum and members of the SAR86 order. The Chromatiales order was restricted to the CS204 metagenome, the Xanthomonadales order to the CS211 metagenome and Deltaproteobacteria to the CS202 metagenome (Figure 21). Based on the number of different marker lineages identified, the most diverse sponge microbiome was CS203 (Figure 21).

5.5.3 Analysis and comparison of microbiome composition

The collection of metagenomic assemblies from the six Tongan sponges presented the opportunity to examine the extent to which sponge species and location influence microbiome composition. As an initial means of visualization and qualitative comparison of microbiomes, percent GC content was plotted against coverage of contigs for the MAGs from each of the 6 sponges (Figure 21). Each of the points in the graph was colored by marker lineage (as identified by checkM) to further distinguish poorly resolved clusters in the graph. Distinct aggregations are visible in each sample and the CS200 metagenome is particularly well separated with several dense aggregations of contigs, indicating it contains a particularly diverse collection of well-resolved MAGs (Figure 21 A). GC content of aggregations and thus putative MAGS covers a wide range from approximately 35% to 70% across the six metagenomes (Figure 21). This analysis gave an initial indication that there were broad differences in the microbial community composition between sponge species. The blobplots for the two *C. mycofijiensis* samples (CS200 and CS783) appeared the most similar visually with four distinct vertically arranged aggregations of the same marker lineage at 50% GC and the vast majority of remaining aggregations at higher %GC (Figure 21, A and B).

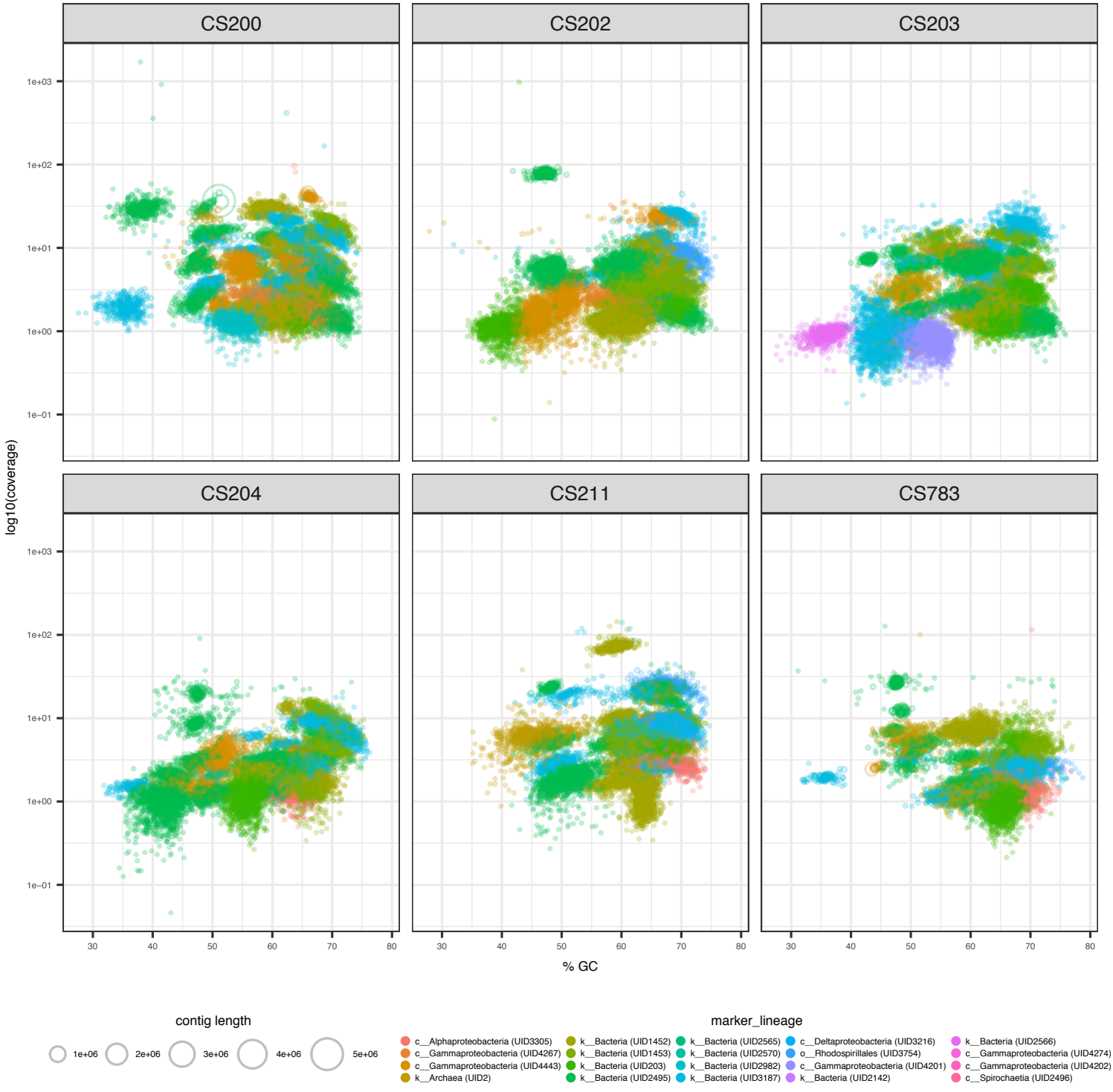


Figure 21: Blobplot of the contigs attributed to dRep dereplicated bins for each of the six Tongan sponges %GC content of contigs is plotted against log-transformed coverage of contigs (maximum k-mer coverage as calculated by metaSPAdes) and contigs then colored by marker lineage as identified by checkM with size of the circle indicating size of the contig (2.14.10).

A quantitative analysis of microbiome composition was then conducted using dRep compare, which uses marker genes and average nucleotide identity (ANI) to identify

unique genomes [167]. The dRep compare module initially clusters bins based on a 0.9 similarity cutoff using MASH [244] followed by separation of these initial, so-called primary clusters, using the more precise gANI [245] with a 0.99 similarity cutoff [167]. This analysis identified thirteen MAGs that were present in two sponges. Nine of these were shared between the two *C. mycofijiensis* samples (CS200 and CS783) and four between CS203 and CS204 (Appendix A22). The low number of shared MAGs indicates sponges may have distinct microbiome compositions. The fact that the two samples of *C. mycofijiensis* (CS200 and CS783) shared the most MAGs, in spite of being collected at different locations, suggests that sponge species may be a more important driver of microbiome composition than geographical location.

To investigate this apparent trend further, 16S rRNA sequences were extracted from each metagenomic assembly using barrnap v.0.9 (Torsten Seemann 2015) and unique sequences identified using CD-HIT [246] with a 99% identity cut-off (2.14.6, [134]). Abundance was assigned using coverage information extracted from the FASTA header for the contig on which the 16S rRNA sequence was found and taxonomy for each gene identified using the SILVA SSU database (2.14.6). By extracting these sequences from shotgun assemblies, rather than using amplicon sequencing, the true abundance of each microbiome member is more accurately reflected. To facilitate future analyses and ensure reproducibility, the processing pipeline for this analysis was compiled into an annotated jupyter notebook, which can be found here (https://github.com/MaxMeta/Vincent_Masters).

The establishment of a robust processing pipeline presented the opportunity to expand the scope of the comparative microbiome analysis. To this end, 16S sequences were extracted from seven further sponge metagenomes: Four New Zealand costal sponge metagenomes and three Mediterranean sponge metagenomes. The New Zealand sponge metagenomes (s0 - s3; Figure 22) are being examined as part of other projects in the Owen lab. Each of these had been identified as *Mycale hentscheli* and was collected from Doubtful Sound, NZ. The Mediterranean metagenomes were publicly available and were generated from three different species, *Petrosia ficiformis* (pf; Figure 22), *Sarcotragus foetidus* (sf; Figure 22) and *Aplysina aerophoba* (aa; Figure 22), collected in the Mediterranean sea [141]. The publicly available Mediterranean data

did not have coverage information, so sequences were in these metagenomes were assigned zeros or ones for binary presence/absence.

Distances between microbiomes were computed using cosine similarity, which is known to perform well on sparse data such as the collection analysed here, and is also applicable to binary data. Hierarchical clustermaps were then generated using average linkage to visualise groupings. Two separate analyses were carried out: In the first of these, the binary Mediterranean data were excluded, and relative abundance for each sequence was used for clustering. In the second analysis, all data were dichotomised (converted to binary presence absence), allowing inclusion of the Mediterranean data.

The strongest clustering in this analysis was between sponges from the same large scale geographical location (Tonga, Mediterranean and New Zealand). Each of these locations formed a distinct cluster in the binary clustering analysis (Figure 22 A). In the case of the New Zealand and Tongan sponges, this clustering was also observed when relative abundance was considered. Within each of the broad locations, there were shared species between different sponges. There were also some shared species between Mediterranean and Tongan sponges but no shared species between New Zealand sponges and any of the non-NZ samples (Figure 22 A, B). Large-scale geographic location thus appears to have significant impact on sponge microbiome composition.

The closest grouping in the presence/absence analysis of the Tongan sponges was between CS203 and CS204. The two *C. mycofijiensis* samples (CS200 and CS783) were also very similar. This result is congruent with the dRep comparison of MAGs recovered from the six Tongan sponges, which identified shared MAGs between sponges in these pairs (Appendix A22). The inclusion of abundance in the analysis altered the groupings (Figure 22 B) and in this case CS200 was more closely related to CS211 than to the other *C. mycofijiensis* specimen (CS783). This raises considerations about what defines two sponge microbiomes as being more similar, whether it is the mere number of shared taxa or whether the abundance of taxa should be accounted for [65]. It should be noted that clustering including abundance information could theoretically be influenced by one to a few dominant taxa that are shared

between two sponges. Based on the relative abundance estimates obtained here, this seems to be a trend present in Tongan and NZ sponges, where there is one to a few dominant species followed by a larger number significantly less abundant taxa (Figure 22 B). It can thus be concluded that the two *C. mycofijiensis* samples (CS200 and CS783) have a very similar microbiome composition based on the stringent comparison of dereplicated MAGs as well as binary 16S rRNA analysis. CS203 and CS204 also have a very similar microbiome composition, in particular based on 16S rRNA data. These sponges were not classified as the same species based on 18S rRNA sequences. Overall the results indicate that species may be an important driver of sponge microbiome composition but that other factors, e.g. large scale oceanographic location, are clearly important.

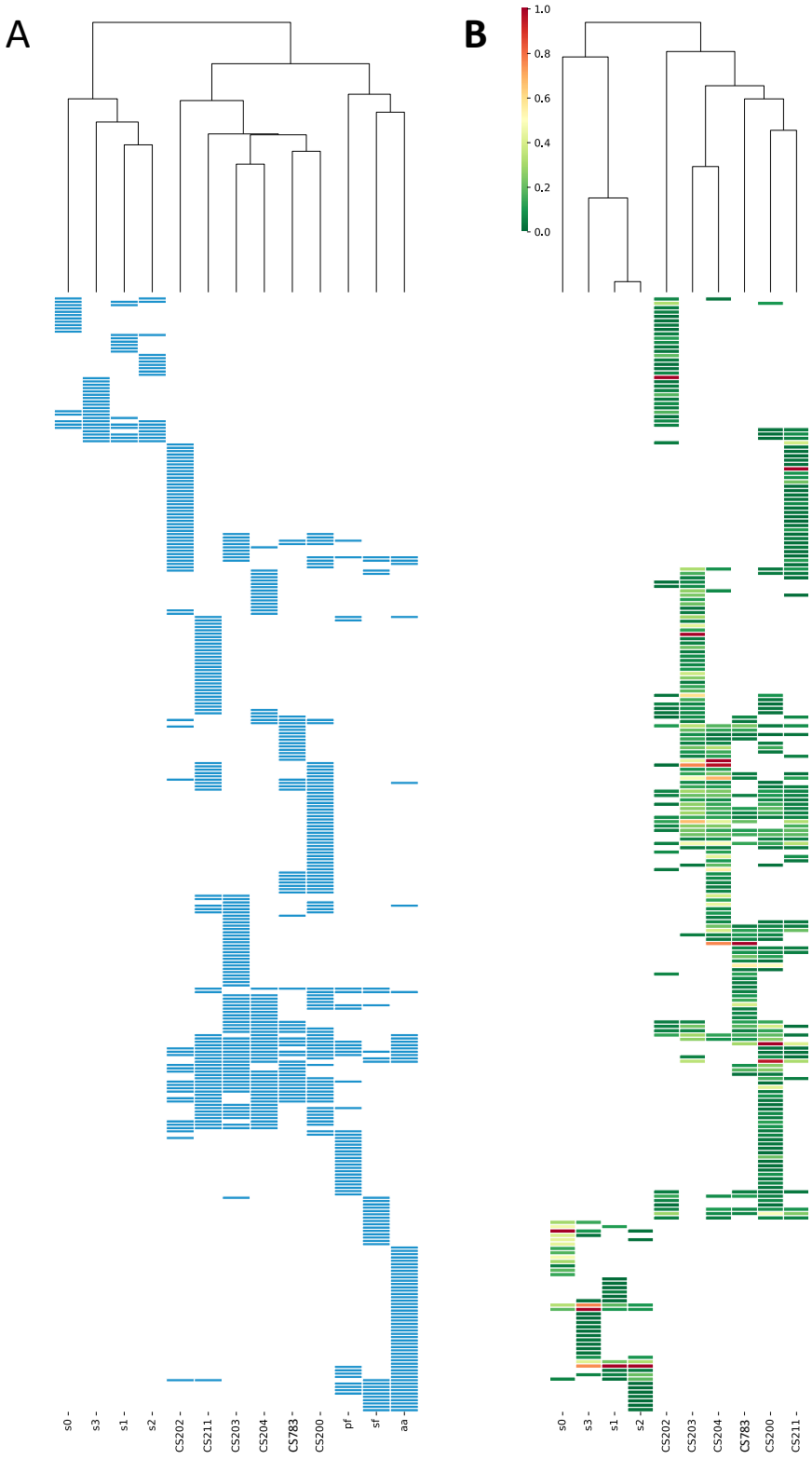


Figure 22: Cosine clustering of 16S rRNA sequences from the six Tongan sponges, three Mediterranean sponges (pf, sf, aa; Horn et al. 2016) and four New Zealand sponge samples (s0, s1, s2, s3). (CS783 = *C. mycofijiensis*) Plots produced using seaborn show cosine distance metric of standardized data with (A) clustering by presence (blue)/absence (white) and (B) showing clustering including abundance information. Note that abundance information could not be obtained for Mediterranean sponges (2.14.6).

5.6 Analysis of the secondary metabolite potential of six Tongan sponges

In order to identify and annotate secondary metabolite biosynthetic gene clusters within the MAGs recovered from the five Tongan sponges a standalone install of antiSMASH was used to analyse all the contigs (≥5 kb) from each of the metagenomic assemblies (2.14.8). A length-filter of ≥5 kb was chosen because secondary metabolite clusters are extremely unlikely to be shorter than 5 kb, and the runtime of antiSMASH was consequently dramatically reduced. A total of 1,343 BGCs were identified from the length-filtered assemblies. CS203 was the most metabolically diverse with 18 different classes of secondary metabolites identified, of which thiopeptides and oligosaccharides were only seen in this sponge (Figure 23).

Type I PKS clusters vastly outnumbered other PKS types in all of the six sponges, with Type III PKSs the next most abundant followed by Type II PKSs (Figure 23). Only two trans-AT PKS clusters were identified from the six sponges (Figure 23), an observation in stark contrast to the numerous trans-AT PKSs recovered from other sponge metagenomes [56,82]. One NRPS cluster was identified per sponge, with the exception of *C. mycofijiensis*, which contained none, and CS211, which contained two (Figure 23). Out of the 9 classes of RiPPs identified, lanthipeptides and lassopeptides were present universally in the six Tongan sponge metagenomes (Figure 23). The most common sub-classes of RiPPs were the lassopeptides and lanthipeptides, with 19 and 12 BGCs found respectively.

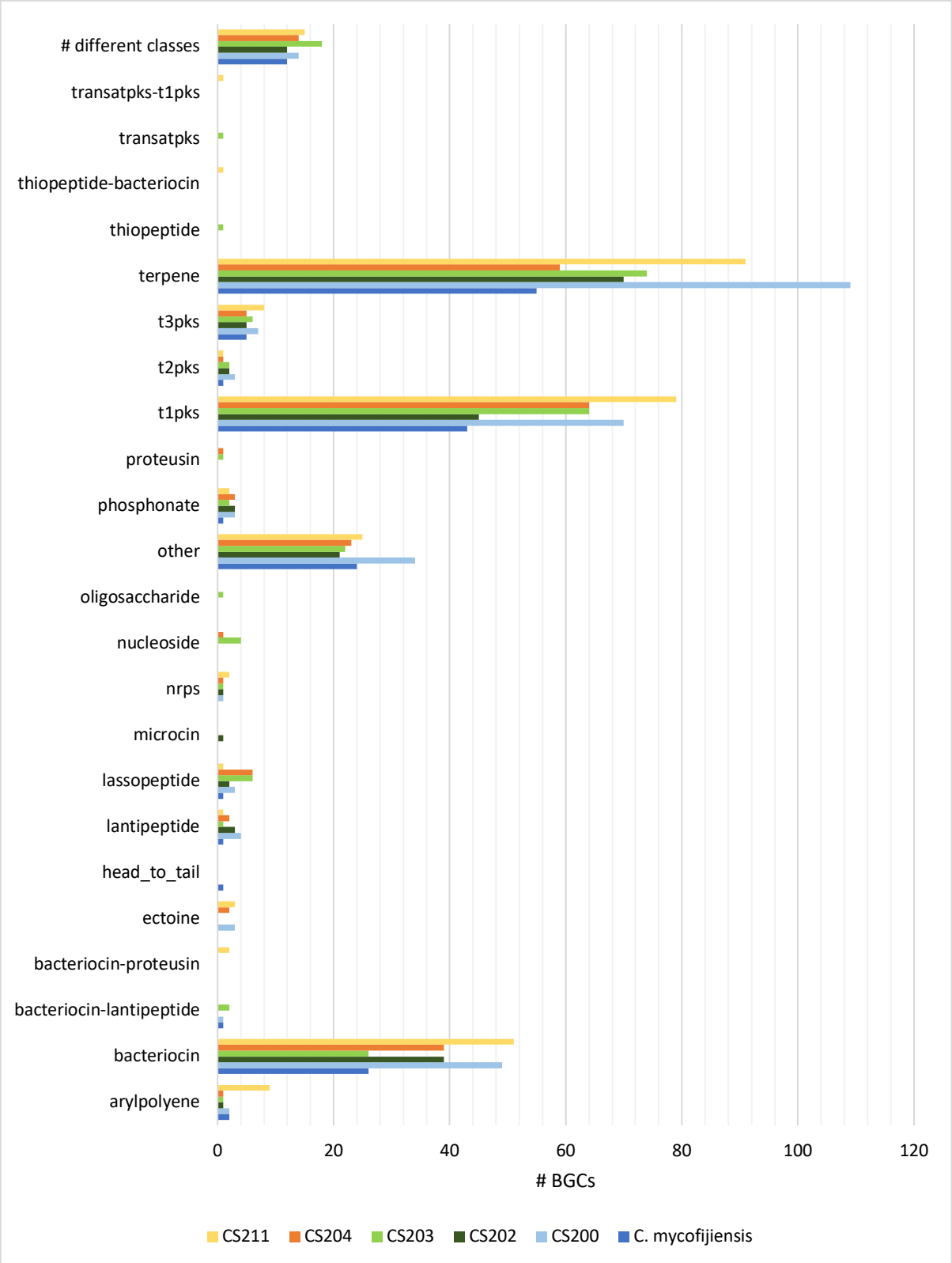
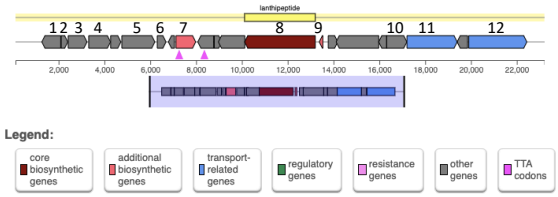


Figure 23: Number of BGCs identified per secondary metabolite class for the six Tongan sponges by standalone antiSMASH from the length-filtered (>5kb) assemblies of the six Tongan sponges. T1pks = Type 1 Polyketide synthase; t2pks = Type 2 Polyketide synthase; t3pks = Type 3 Polyketide synthase; nrps = Non-ribosomal peptide synthetase; transatpks = trans-AT polyketide synthase.

Due to the comparative completeness of RiPP BGCs identified in the Tongan sponge metagenomes, contigs containing these BGCs were re-analysed with the recently released antiSMASH5 online tool, which includes updated algorithms for detection of several RiPP classes including lanthipeptides and lassopeptides (https://docs.antismash.secondarymetabolites.org/glossary). Re-analysis resulted in almost identical BGC classifications, however precursor peptide predictions were notably different. AntiSMASH5 predicted 10 precursor peptides compared to 9 from antiSMASH4.1.0 (Appendix A24). All 6 lassopeptides with predicted precursor peptides obtained high RODEO scores between 19-26, indicating the precursor peptide prediction is highly likely to be accurate [92]. Manual analysis of the four predicted Class II and III lanthipeptide core AA-sequences, showed that these contained the Cys and Dha or Dhb residues required for formation of the characteristic lanthionine or methyl-lanthionine moieties. The corresponding BGCs each contained at least one lanthionine synthetase, providing strong evidence that the automated functional assignment of these BGCs was correct. Among the 10 RiPP BGCs for which precursor peptides were predicted, three were chosen for detailed manual analysis based on apparent completeness and novelty of predicted core peptides.

BGC143-CS200 was one of the largest and its predicted core peptide showed distant homology to the known lanthipeptides, Prochlorisin 3.3 (Figure 24), which was isolated from a marine cyanobacterium [223]. The BGC contained the peptidase required to cleave the leader peptide from the core peptide (Key 7, Figure 24) as well as putative transport (Key 10-12, Figure 24), regulatory (Key 2, Figure 24) and a self-resistance gene (Key 3, Figure 24). The presence of a self-resistance gene is particularly interesting as it indicates that the product might possess antibacterial activity.

antiSMASH5 prediction and annotation of BGC 143 – CS200



Key	Identification
1	nucleoside triphosphate pyrophosphohydrolase [bacterium] (BLASTp score 327; e-value: 1e-109)
2	DUF1844 domain-containing protein [Acidobacteria bacterium] (BLASTp score 101; e-value: 6e-26)
3	MBL fold metallo-hydrolase [bacterium] (BLASTp score 386; e-value: 6e-133)
4	ribF: riboflavin biosynthesis protein [uncultured bacterium] (BLASTp score 447; e-value: 6e-156)
5	cysteine--tRNA ligase [bacterium] (BLASTp score 734; e-value: 0.0)
6	YraN family protein [Candidatus Marinimicrobia bacterium] (BLASTp score 138; e-value: 1e-39)
7	biosynthetic-additional (rule-based-clusters) Peptidase_M42
8	SMCOG1070:lanthionine synthetase C family protein (Score: 692.8; E-value: 1.3e-209)
9	biosynthetic-additional (lanthipeptides) predicted lanthipeptide
10	NHLP bacteriocin system secretion protein [Acidobacteria bacterium] (BLASTp score: 286; e-value: 4e-91)
11	SMCOG1288:ABC transporter related protein (Score: 282.8; E-value: 9.4e-86)
12	SMCOG1288:ABC transporter related protein (Score: 236.1; E-value: 1.3e-71)

Peptide sequence predicted by antiSMASH5

VKFIEKDSDCDAMFVLPDPVATDELTPEQLEAVAGG - CFDIDIGDIDhbICWEDhbA

Leader peptide – Core peptide
Dha: Didehydroalanine
Dhb: Didehydrobutyrine

ctg6_15 - Class II
Cleavage pHMM score: 1.50
RODEO score: 0

Top hit of pairwise BLAST alignment using RiPPMiner

```
Aligment of input sequence UK\|P1\| with Prochlorosin 3.3:

 Prochlorosin@3.3
Length=87

 Score = 21.6 bits (44),  Expect = 0.068, Method: Compositional matrix adjust.
 Identities = 9/17 (53%), Positives = 12/17 (71%), Gaps = 0/17 (0%)

Sbjct  48  AAASELSDEELEAASGG  64
           A  EL+ E+LEA +GG
Query  20  VATDELTPEQLEAVAGG  36
```

Predicted structure based on antiSMASH5

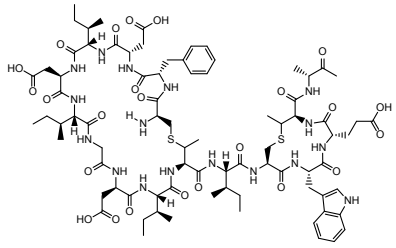
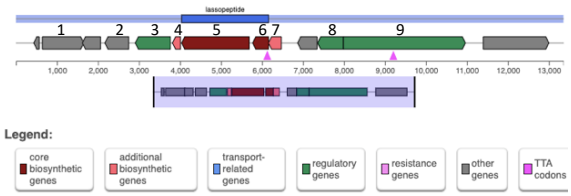


Figure 24: BGC 143 - CS200 antiSMASH5 BGC prediction and annotation (BLASTp was run subsequently on unidentified ORFs) as well as leader and core peptide prediction from RODEO run as part of antiSMASH5. Pairwise alignment of predicted peptide sequence was carried out using the RiPPMiner (Agrawal et al. 2017) online tool sequence similarity search. Chemical structure was created in ChemDraw using the peptide and macrolactam prediction from antiSMASH5.

BGC160-CS203 encodes a lassopeptide that is highly similar (BLASTp) to Chaxapeptin (Figure 25), a lassopeptide that possesses cell migration inhibitory activity (Elsayed et al. 2015). Chaxapeptin was isolated from *Streptomyces leeuwenhoekii* strain C58 [247], however BLASTp search of ORFs in BGC160-CS203 predominantly returned Acidobacteria as the taxonomic origin, indicating that in this case the producer is not an actinobacterial species. The BGC could not be assigned to any of the assembled MAGs in CS203, so the identity of the producing organism remains unknown but evidence suggests that new Chaxapeptin-like lassopeptides are produced by an as yet unidentified acidobacterium from the CS203 metagenome.

antiSMASH5 prediction and annotation of BGC 160 – CS203



Key	Identification
1	SMCOG1077:phage integrase family protein (Score: 65.5; E-value: 1.2e-19)
2	antitermination protein NusG [Acidobacteria bacterium] (BLASTp score 252; e-value: 1e-82)
3	SMCOG1008:response regulator (Score: 83.5; E-value: 2.8e-25)
4	Predicted Lassopeptide
5	SMCOG1177:asparagine synthase (glutamine-hydrolyzing) (Score: 168.2; E-value: 4.7e-51)
6	biosynthetic (rule-based-clusters) lassopeptide: PF13471
7	biosynthetic-additional (rule-based-clusters) PF05402
8	SMCOG1016:LuxR family DNA-binding response regulator (Score: 131; E-value: 6e-40)
9	SMCOG1048:sensor histidine kinase (Score: 151.5; E-value: 6.1e-46)

Peptide sequence predicted by antiSMASH5

MNHRDDPTMTRTNATAYEPPVLILIGDAENVVL - **GVPGGGDD**HFGFSPWQFE - **FEEDNDEGGAPRA**

Leader peptide – Core peptide
Red – Putative macrolactam
Purple – Putative cleaved off residues

ctg5_6 - Class II
Cleavage pHMM score: -5.70
RODEO score: 19

Top hit of pairwise BLAST alignment using RiPPMiner

```
Aligment of input sequence UK\|P1\| with Chaxapeptin:

 Chaxapeptin
Length=44

 Score = 28.9 bits (63),  Expect = 5e-05, Method: Compositional matrix adjust.
 Identities = 15/39 (38%), Positives = 18/39 (46%), Gaps = 0/39 (0%)

Sbjct  2   EPQMTELQPEAYEAPSLIEVGEFSEDTLGFGSKPLDSFG  40
           +P MT      AYE P LI +G+      LG      D FG
Query  6   DPTMTRTNATAYEPPVLILIGDAENVVLGVPGGGDDHFG  44
```

Predicted structure based on antiSMASH5

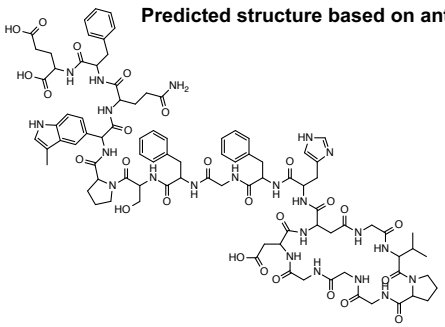
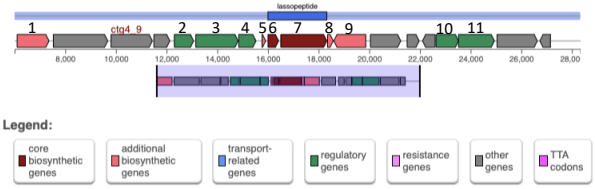


Figure 25: BGC 160 - CS203 antiSMASH5 BGC prediction and annotation (BLASTp was run subsequently on unidentified ORFs) as well as leader and core peptide prediction from RODEO run as part of antiSMASH5. Pairwise alignment of predicted peptide sequence was carried out using the RiPPMiner (Agrawal et al. 2017) online tool sequence similarity search. Chemical structure was created in ChemDraw using the peptide and macrolactam prediction from antiSMASH5.

BGC104-CS200 encodes a lassopeptide, which showed very low homology to known RiPPs (BLASTp) whilst obtaining a high RODEO score of 26, indicating it highly likely to encode a lassopeptide (Figure 26). While not all of the ORFs could be attributed a putative function, this BGC is well resolved and appears complete as it contains numerous regulatory genes and the two key biosynthetic genes (Figure 26, Key 6&7) required for lassopeptide synthesis.

antiSMASH5 prediction and annotation of BGC 104 – CS200



Legend:

core biosynthetic genes | additional biosynthetic genes | transport-related genes | regulatory genes | resistance genes | other genes | TTA codons

Key	Identification
1	SMCOG1139:aminotransferase class V (Score: 252.7; E-value: 9.7e-77)
2	SMCOG1133:Transcription regulator, crp (Score: 134.2; E-value: 6.5e-41)
3	SMCOG1003:sensor histidine kinase (Score: 221.1; E-value: 6.2e-67)
4	SMCOG1008:response regulator (Score: 222.5; E-value: 8.5e-68)
5	predicted lassopeptide
6	biosynthetic (rule-based-clusters) lassopeptide: PF13471
7	SMCOG1177:asparagine synthase (glutamine-hydrolyzing) (Score: 297.9; E-value: 2.6e-90)
8	biosynthetic-additional (rule-based-clusters) PF05402
9	SMCOG1053:beta-lactamase (Score: 201.3; E-value: 3.4e-61)
10	SMCOG1008:response regulator (Score: 203.2; E-value: 6.9e-62)
11	SMCOG1003:sensor histidine kinase (Score: 207.2; E-value: 9.8e-63)

Peptide sequence predicted by antiSMASH5

MTALRGRQRRKKPYVTPRVVDFGAIDAMTG -
DCFGLCLDGMNGGLFWGP

Leader peptide – Core peptide
Red – Putative macrolactam
Purple – Putative cleaved off residues

ctg4_14 - Class III
Cleavage pHMM score: -1.90
RODEO score: 26

Top hit of pairwise BLAST alignment using RiPPMiner

```
Aligment of input sequence UK\|P1\| with Thuricin CD beta:

 Thuricin@CD@beta
Length=49

 Score = 17.3 bits (33),  Expect = 1.2, Method: Compositional matrix adjust.
 Identities = 7/29 (24%), Positives = 12/29 (41%), Gaps = 0/29 (0%)

Sbjct  5   NKQNVNIIPESEEVGGWVACVGACGTVCL  33
           ++    + P     + G    A  G C  +CL
Query  9   RRKKPYVTPRVVDFGAIDAMTGDCFGLCL  37
```

Predicted structure based on antiSMASH5

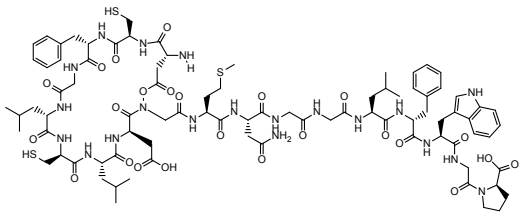


Figure 26: BGC 104 - CS200 antiSMASH5 BGC prediction and annotation (BLASTp was run subsequently on unidentified ORFs) as well as leader and core peptide prediction from RODEO run as part of antiSMASH5. Pairwise alignment of predicted peptide sequence was carried out using the RiPPMiner (Agrawal et al. 2017) online tool sequence similarity search. Chemical structure was created in ChemDraw using the peptide and macrolactam prediction from antiSMASH5.

The vast majority of BGCs from the length-filtered assemblies did not return any significant homology hits when analysed with the KnownClusterblast module, which compares BGCs to the MiBIG database, integrated in antiSMASH. Together with the number and diversity of BGCs identified, particularly those encoding RiPP molecules, this indicates that these marine sponges hold great potential to yield new bioactive small molecules via heterologous expression of BGCs.

5.7 Secondary metabolite potential of MAG taxa per sponge

In order to provide a picture of secondary metabolite distribution across bacterial phyla, each of the 327 recovered MAGs was individually analysed using a standalone install of antiSMASH (2.14.8). BGCs were then assigned to the marker lineage identified by checkM analysis of the corresponding MAG. This qualitative analysis is summarized in Figure 27 below and provides an interesting initial picture of variation in secondary metabolite richness across bacterial taxa and can be used for the

inferences such as secondary metabolite classes being specific to certain taxonomic groups or taxonomic groups which are generally rich in secondary metabolism.

One of the striking findings was that Type II PKS BGCs were restricted to the Candidate phylum “Patescibacteria”, which is found in all six Tongan sponges (Figure 27). Type 2 PKSs are iterative and usually synthesize aromatic polyketides and have not previously been identified in the Candidate phylum “Patescibacteria”. Proteusins were unique to CS203, where one BGC was found in a Deltaproteobacterium. Trans-AT PKSs were only found in CS211, where they were linked to a gammaproteobacterium within the Xanthomonadaceae family. Acidobacteria, which have recently been shown to be rich in secondary metabolite BGCs in soil [248], were linked to the highest number of different compound classes in each of the six Tongan sponges and thus present an excellent taxonomic group to target for natural product discovery. Secondary metabolite classes within this taxonomic group included Type I and III PKSs, bacteriocins, bacteriocin-lanthipeptides, lanthipeptides, lassopeptides and NRPSs (Figure 27).

Figure 27: Number of BGCs per secondary metabolite class per marker lineage per sponge Bars are colored by marker lineage as identified by checkM (taxonomic reference of marker lineage in Appendix A25). Secondary metabolite classes were identified by antiSMASH. (CS783 = *C. mycofijiensis*). T1pks = Type 1 Polyketide synthase; t2pks = Type 2 Polyketide synthase; t3pks = Type 3 Polyketide synthase; nrps = Non-ribosomal peptide synthetase; transatpks = trans-AT polyketide synthase.

5.8 Comparing the secondary metabolite profile of the six Tongan sponges

A global comparison of secondary metabolite profiles for each metagenome was conducted using the recently developed BiG-SCAPE algorithm [200]. BiG-SCAPE carries out similarity clustering and phylogenetic comparison of complete and incomplete BGCs [200]. Comparison is based on Pfam domain strings and the Jacard index (JI) as well as the adjacency index (AI) and the newly developed domain sequence similarity (DSS), which compares Pfam domain string differences and sequence identity [200]. Comparing the number of times BGCs from two sponges were identified in the same Gene Cluster Family (GCF) by BiG-SCAPE (2.14.9), showed that CS203 and CS204 were the most similar with 39 shared or related BGCs (Figure 28, Appendix A26). BGCs from CS200 and CS211 were found in the same GCF 37 times and BGCs from CS200 and *C. mycofijiensis* 31 times (Figure 28, Appendix A26). This provides further evidence that CS203 and CS204 have very similar microbiome composition. While the high similarity between the two *C. mycofijiensis* samples (CS200 and CS783) was expected, the high similarity of CS200 and CS211 BGCs was somewhat unexpected. It should be noted here that these two sponges also grouped closely in the 16s rRNA cosine similarity clustering including abundance.

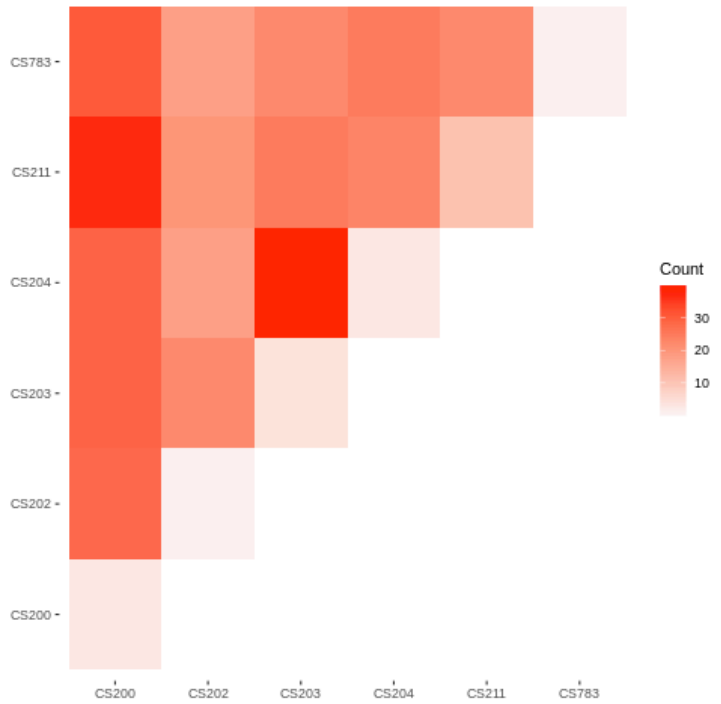


Figure 28: Heatmap of the number of times BGCs from two sponges were identified in the same GCF by BiG-SCAPE at the standard cutoff of 0.3 (2.14.9, Appendix A26, CS783 = *C. mycofijiensis*)

5.9 Discussion

5.9.1 Metagenomics as a tool for investigating sponge microbiomes

The bioinformatics workflow optimised in Chapter 4 incorporates new and powerful bioinformatic algorithms, such as metaSPAdes, Autometa and dRep. Applying this workflow to five further sponge metagenomes demonstrates the robustness and scalability of this workflow. A total of 327 dRep dereplicated MAGs were recovered from the six Tongan sponges sequenced. Close to half of these MAGs (146) were near-complete, demonstrating the ability of this workflow to recover high quality MAGs at a large scale. Based on taxonomy obtained from 16S rRNA sequences extracted from the metagenomic assemblies, Chloroflexi, Proteobacteria and Acidobacteria were the three most abundant phyla in the six Tongan sponges, closely followed by Poribacteria (Appendix A23). This is consistent with sponge microbiome compositions reported in the literature [60,61,65]. The detection of Entotheonellaeota in CS200 is significant as these bacteria have been identified as "natural product factories" in marine sponges [29,30] and this provides evidence that these bacteria appear to be present in Tongan sponges. It is possible that the low level of detection of these bacteria in sequencing data is due to loss during DNA isolation (2.9) as they are easily pelleted with sponge debris due to their large filamentous morphology (Jörn Piel, personal communication).

5.9.2 Factors influencing microbiome composition in marine sponges

This analysis included two samples of *C. mycofijiensis* (CS200 and CS783) collected from two different locations in 'Eue, Tonga. These samples shared the highest number of 0.99 ANI identical MAGs (9) out of any two sponges. It should be noted that the number of shared symbionts may in fact be higher as 76 unique MAGs were identified from CS200 and only 46 from CS783. This may reflect a true difference in bacterial diversity but may also be due to the difference in sequencing depth (>10.58 Gbp vs 29.24 Gbp respectively) between these two samples.

CS200 and CS783 grouped closely by hieratical clustering based on binary presence/absence of 16S rRNA sequences (Figure 22), however when abundance was included, CS200 clustered more closely with CS211. Abundance clustering may

have been influenced by the overwhelming abundance of one dominant species compared to all other species.

CS203 and CS204 sponges were the only other two sponges with 0.99 ANI identical genomes (4 shared MAGs; Appendix A22) as identified by dRep compare. These sponges also clustered together in the 16S rRNA cosine similarity comparison with and without abundance information, indicating that their microbiomes can also be considered very similar. CS203 and CS204 were not identified as the same sponge species by 18S rRNA comparison, with 1,281 bp and 1,789 bp 18S rRNA sequences respectively of which 100% were mapped and 90.02% matched. Based on the evidence collected here, it can thus be concluded that sponge species is an important driver of microbiome composition in some cases, however given that two closely related microbiomes came from putatively different sponge species, it is clearly not the only driver. Large scale oceanographic location also appears to be a driver of microbiome composition as sponges from the same geographical location clustered together in the presence/absence clustering, which was maintained in the clustering of NZ and Tongan sponges using abundance information (Figure B). The distinct separation of NZ sponges as well as the fact that Mediterranean and Tongan sponges shared some bacterial taxa (see Figure 22) suggest that tropical sponges may share more bacterial species than sponges from other regions, an idea previously reported in the literature [61].

5.9.3 Secondary metabolism in Tongan marine sponges: Ecological and biotechnological implications

A total of 1,328 BGCs were identified from the six sponge microbiomes, of which 909 could be attributed to one of the extracted MAGs (Appendix A25). This result demonstrates the power of shotgun sequencing metagenomes for identifying potential new sources of secondary metabolites. The Competibacterales order (Gammaproteobacteria), the Micavibrionales order (Alphaproteobacteria) and Acidobacteria were identified as the marker lineages with the most BGCs per MAG, making them relevant targets for natural product discovery (Appendix A25). The microbiomes were particularly rich in lassopeptides (19 identified) and lanthipeptides (12 identified). The high abundance of these secondary metabolite classes is interesting from a biotechnological and ecological standpoint. Both classes of

molecule frequently possess antibacterial activity [88,92,95], thus the sponge microbiomes investigated here are a potential source of new medically relevant compounds. From an ecological standpoint, these compounds may provide a selective advantage to their producers as they compete for an ecological niche or to the sponge holobiont in general by acting as a defense chemical [33,193,249].

Chapter 6

Concluding remarks

The work carried out as part of this thesis developed a scalable and reproducible workflow for the metagenomic investigation of marine sponge microbiomes using shotgun sequencing. It also resulted in the construction of a large insert metagenomic library that redundantly covers a sponge microbiome. A total of 327 dereplicated MAGs were recovered from six metagenomic assemblies and 146 of these were near-complete (Table 10). Many of the high-quality MAGs are phylogenetically distant from previously described species, providing a wealth of new material for future functional and bioinformatic studies. Metagenomic assemblies and consequent MAG recovery could likely be further improved by the incorporation of long read data [138,153] but this is not always guaranteed to help with the resolution of fragmented assemblies [29]. It is interesting to consider that in the not-so-distant future whole genomes may be sequenced in a single read and a post-assembly era of metagenomics may begin [170].

Microbial community composition of the six Tongan marine sponges was compared using marker-gene based and 16S rRNA-based approaches. The amplification bias usually observed in 16S rRNA amplicon studies [123,132,134,225,250] was resolved by extracting 16S rRNA sequences from metagenomic assemblies. This also addressed the susceptibility of 16S rRNA amplification to transient bacteria [131] as metagenomic assemblies require a certain amount of coverage and consequently bacterial DNA to assemble this ~1,550 bp gene.

The unique study design included two samples of *C. mycofijiensis* (CS783 and CS200) collected from two different locations in ‘Eua, Tonga. These were shown to have a very similar microbiome compositions based on the comparison of dereplicated MAGs (Appendix A 22) as well as presence/absence clustering of 16S rRNA sequences (Figure 22), which further supports the idea that host taxonomy is an important driver of microbiome composition in marine sponges [65,132]. However, large-scale oceanographic location was found to be the most significant driver of microbiome composition. Sponges from different oceanic regions shared few or no microbiome members based on 16S rRNA sequence analysis (Figure 22). One interesting observation was that some bacterial species were common to both the Mediterranean and Tongan microbiomes. By contrast, the New Zealand microbiomes did not share any species in common with the other two locations. This is might be due to the more similar ocean temperatures in the costal Mediterranean and Tongan waters, and

perhaps indicates that temperature is another important factor determining microbiome composition.

Metagenomic sequencing has proven useful for the discovery of natural product BGCs [54,187,248]. However, PKS and NRPS clusters can be hard to assemble from short-read data owing to repetitive elements [77,232]. This might explain the numerous seemingly partial PKS BGCs observed in the Tongan marine sponges. Nonetheless, the workflow developed here allowed the recovery of some relatively large Type I PKS clusters (>70 kb) as well as numerous seemingly complete RiPP clusters and a diverse collection of other secondary metabolite clusters. One drawback of the rule-based discovery used in antiSMASH, is that low homology biosynthetic genes and consequently BGCs might be missed [73,251,252]. A possible solution for this is 'EvoMining' [251,252], which is based on the observation that secondary metabolism genes often evolve from primary metabolism genes and thus looking at divergences of genes across several bacterial species can help identify new pathways [252].

A longer term goal of this work is the heterologous expression or native host production of the identified BGCs to form sustainable sources of new natural products. While no conclusive BGC was evidently responsible for the production of laulimalide A, latrunculin A or (-)-zampanolide, potential leads were found for both the laulimalide A and latrunuclin A BGC. The lack of conclusive identification may be due to: (1) potential loss of Entotheonellaeota during metagenomic DNA extraction as has been previously observed (Prof. Jörn Piel, personal communication), (2) the fragmentation of PKS and NRPS BGCs [231,232], (3) failure of current algorithms and underlying models to detect low homology BGCs [29,74].

There are numerous techniques for heterologous expression of BGCs [79]. Among these, synthetic construction of complete clusters from small synthetic DNA blocks is becoming progressively more attractive as the cost of accurate DNA synthesis decreases [79,253]. This also allows codon optimization of the BGC for the expression host and relatively facile introduction of regulatory elements [79,253]. Future work could involve using this approach to express small to medium-sized (<25 kb) BGCs accurately identified from bioinformatic analyses, such as the lanthipeptide and lassopeptide BGCs identified in this study. Alternatively, it might be possible to achieve

in vitro production of these RiPPs by modification of purified precursor proteins by purified tailoring enzymes. This “cell free” approach has previously been applied to several RiPP clusters [89,254,255] and provides a potential new avenue for compounds discovery where heterologous expression has failed. Yet another potential approach for compound production is to isolate and express complete BGCs from the metagenomic library that was constructed as part of this work using the CATCH method or other recently developed CRISPR/Cas-based methods [79,256].

It has been previously suggested that natural products act as quorum sensing, communication or defense compounds in prokaryote-eukaryote symbioses, but little functional evidence has been established [110,249,257,258]. The collection of new genomes elucidated here contains numerous natural product BGCs and presents the opportunity to isolate and experimentally verify the function of their small molecule products. Future studies might include examination of purified symbiont natural products for biological activity. These could focus on finding activities that are relevant to establishing or maintaining symbiosis. For example, direct interaction with eukaryotic signalling pathways, elimination of competing or pathogenic microbial species via antibiotic activity or deterring predation via toxicity to potential predators.

In conclusion, the workflow developed here is a powerful tool for elucidating new bacterial symbiont genomes without the need for cultivation. Application of this workflow successfully linked numerous BGCs to the complete or near-complete genomes. This work found that numerous factors shape microbiome composition in marine sponges, and highlighted the need for further studies aimed at understanding the complex biological system that is the sponge holobiont. The work also builds the foundation for functional studies aiming to elucidate the roles of natural products in sponge microbiomes. Finally, several of the identified BGCs are good candidates for encoding novel antibiotics, and might form the basis for future drug discovery efforts.

Appendix

Note that "$" at the start of a line denotes UNIX commands, ">" denotes R commands and ">>>" denotes python.

A1: Detailed instructions for λ-phage extract preparation.
Day 1: Streak out BHB2688 and NM759 glycerol stocks on NZY media plates and incubate at 30°C.
Day 2: Inoculate two NZY media plates with a single colony from each of the previous night's BHB2688 and NM759 plates and incubated one plate for each strain at 30°C and 42°C.
Day 3: Check 42°C plates for any growth. If clear, the streaked out single colony may be used to inoculate overnight cultures. Plates may be stored at 4°C and used for up to 1 week.
Day 4: Inoculate 80 ml of NZY media in a 250 mL culturing flask with a colony of BHB 2688 (30°C plate). Grow overnight in a shaking incubator (30°C, 265 rpm).
Day 5: Inoculate two 2 litre flasks (each containing 700 ml sterile NZY media) with 35 ml of the BHB 2688 overnight culture. Grow the culture in a shaking incubator (32°C, 170 rpm) until OD_{600} ~ 0.6. Transfer the flasks to a 65°C water bath and bring internal temperature to 45°C, gently swirling the flasks. Immediately transfer the flasks to a 45°C water bath for 15 minutes, swirling every 5 minutes. Return flasks to shaking incubator (38–39°C, 265 rpm, 2-3 h). After approximately 2 hours, remove 2 ml of the cell culture, split over two culturing tube, add 3–4 drops of chloroform and incubate at 37°C for 2–3 minutes. When chloroform/cell suspension clears, centrifuge cultures (5,000 rpm, 4°C, 15 min). Immediately decant supernatant, place centrifugation vessels on ice, dry the inside of the vessels with a lint-free towel and vortex for a few seconds. Resuspend all pelleted cell material in 3ml of sucrose solution (10% sucrose, 50 mM Tris-Cl [pH=8.0]) by adding it into one bottle, vortex to resuspend the pellet and transferring to the next bottle. Aliquot 500 µl of the solution into microfuge tubes and add 25 µl of lysozyme solution (2 mg/ml in 10 mM Tris-Cl [pH=8.0]) to each aliquot. Flash-freeze and store at -80°C.
Day 6: Thaw tubes from the day before on ice for at least 1 hour. Add 25 µl of packaging buffer (60 mM Tris-Cl [pH=8.0], 50 mM $MgCl_2$, 30 mM ATP [pH=7.0], 0.002% ß-mercaptoethanol [vol/vol], 3 µM putrescine dihydrochloride, 3 µM

spermidine trihydrochloride) to each tube. Scoop the material into a centrifugation tube and centrifuge (45,000 g, 4°C, >3 hours) Note that longer spin periods increased the yield. Transfer the supernatant was into a chilled 50 ml tube, aliquot 45 µl into microcentrifuge tubes and flash-freeze.
In the afternoon, inoculate 30 ml of NZY media with a colony of NM759 (30°C plate) and incubate overnight in a shaking incubator (30°C, 265 rpm).
Day 7: Inoculate a 2 litre flask containing 500 ml of NZY media with 25 ml of the NM759 overnight culture and grow in a shaking incubator (32°C, 170 rpm) until OD_{600} ~ 0.6. Transfer the flask to a 65°C water bath, swirling gently until the internal temperature reaches 45°C, then transfer it to a 45°C water bath for 15 min, swirling every 5 minutes. Return the flask to the shaking incubator (38–39°C, 2–3 h). After approximately 2 hours, remove 2 ml of the cell culture, split over two culturing tubes, add 3–4 drops of chloroform and incubate at 37°C for 2–3 minutes. When the chloroform/cell suspension clears, centrifuge cultures (5,000 rpm, 4°C,15 min). Immediately decant the supernatant, place the centrifugation vessels on ice, dry the inside of the vessels with a lint-free towel and vortex for a few seconds. Resuspend all pelleted cell material in 3.6 ml sonication buffer (20 mM Tris-Cl [pH=8.0], 1 mM EDTA, 0.0003% ß-mercaptoethanol [vol/vol]) and aliquot into microcentrifuge tubes. Sonicate these aliquots on ice for 1-2 minutes at 4 W power, 0.5 s pulse and 1 s rest until an energy of 4 kJ and centrifuge (17,000 g, 4°C, 20 min). Add supernatant to a chilled 15 ml tube and add 1/2 vol of sonication buffer (20 mM Tris-Cl [pH=8.0], 1 mM EDTA, 0.0003% ß-mercaptoethanol [vol/vol]) as well as 1/6 vol of packaging buffer (60 mM Tris-Cl [pH=8.0], 50 mM $MgCl_2$, 30 mM ATP [pH=7.0], 0.002% ß-mercaptoethanol [vol/vol], 3 µM putrescine dihydrochloride, 3 µM spermidine trihydrochloride). Mix by gentle inversion, aliquot 60 µl and flash-freeze for storage at -80°C.

A2: adap_ID.sh (written by Matt Storey).

```
#! /bin/bash
#This will find adapters in you sequence.fq.gz file. Once
found you can run trimmpmattic or bbduk to remove them
#the scrip will output a file that can be the input to either
of these progs!
#Usage: ./apap_ID.sh filename.fq.gz

#########################################################
```

```
#CLI arguements input .fq.gz file to be analysed for adapters
FILE="$1"

echo $FILE

#set up file for adap hits to be written into
ADAP_FA="adapter.$(basename ${1}.fa)"

# array of adapters (from trimmomatic etc)
declare -A
ADAP=([>Reverse_adapter]="AGATCGGAAGAGCACACGTCTGAACTCCAGTCACAT
CACGATCTCGTATGCCGTCTTCTGCTTG"
[>TruSeq_Universal_Adapter]="AATGATACGGCGACCACCGAGATCTACACTCTT
TCCCTACACGACGCTCTTCCGATCT"
[>pcr_dimer]="AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCT
CTTCCGATCTAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCTCGTATGCCGTC
TTCTGCTTG"
[>PCR_Primers]="AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACG
CTCTTCCGATCTCAAGCAGAAGACGGCATACGAGCTCTTCCGATCT"
[>TruSeq_Adapter_Index_1_6]="GATCGGAAGAGCACACGTCTGAACTCCAGTCAC
ATCACGATCTCGTATGCCGTCTTCTGCTTG"
[>TruSeq_Adapter_Index_2]="GATCGGAAGAGCACACGTCTGAACTCCAGTCACCG
ATGTATCTCGTATGCCGTCTTCTGCTTG"
[>TruSeq_Adapter_Index_3]="GATCGGAAGAGCACACGTCTGAACTCCAGTCACTT
AGGCATCTCGTATGCCGTCTTCTGCTTG"
[>TruSeq_Adapter_Index_4]="GATCGGAAGAGCACACGTCTGAACTCCAGTCACTG
ACCAATCTCGTATGCCGTCTTCTGCTTG"
[>TruSeq_Adapter_Index_5]="GATCGGAAGAGCACACGTCTGAACTCCAGTCACAC
AGTGATCTCGTATGCCGTCTTCTGCTTG"
[>TruSeq_Adapter_Index_6]="GATCGGAAGAGCACACGTCTGAACTCCAGTCACGC
CAATATCTCGTATGCCGTCTTCTGCTTG"
[>TruSeq_Adapter_Index_7]="GATCGGAAGAGCACACGTCTGAACTCCAGTCACCA
GATCATCTCGTATGCCGTCTTCTGCTTG"
[>TruSeq_Adapter_Index_8]="GATCGGAAGAGCACACGTCTGAACTCCAGTCACAC
TTGAATCTCGTATGCCGTCTTCTGCTTG"
[>TruSeq_Adapter_Index_9]="GATCGGAAGAGCACACGTCTGAACTCCAGTCACGA
TCAGATCTCGTATGCCGTCTTCTGCTTG"
[>TruSeq_Adapter_Index_10]="GATCGGAAGAGCACACGTCTGAACTCCAGTCACT
AGCTTATCTCGTATGCCGTCTTCTGCTTG"
[>TruSeq_Adapter_Index_11]="GATCGGAAGAGCACACGTCTGAACTCCAGTCACG
GCTACATCTCGTATGCCGTCTTCTGCTTG"
[>TruSeq_Adapter_Index_12]="GATCGGAAGAGCACACGTCTGAACTCCAGTCACC
TTGTAATCTCGTATGCCGTCTTCTGCTTG"
[>TruSeq_Adapter_Index_13]="GATCGGAAGAGCACACGTCTGAACTCCAGTCACA
GTCAACAATCTCGTATGCCGTCTTCTGCTTG"
[>TruSeq_Adapter_Index_14]="GATCGGAAGAGCACACGTCTGAACTCCAGTCACA
GTTCCGTATCTCGTATGCCGTCTTCTGCTTG"
[>TruSeq_Adapter_Index_15]="GATCGGAAGAGCACACGTCTGAACTCCAGTCACA
TGTCAGAATCTCGTATGCCGTCTTCTGCTTG"
```

[>TruSeq_Adapter_Index_16]="GATCGGAAGAGCACACGTCTGAACTCCAGTCACC
CGTCCCGATCTCGTATGCCGTCTTCTGCTTG"
[>TruSeq_Adapter_Index_18_7]="GATCGGAAGAGCACACGTCTGAACTCCAGTCA
CGTCCGCACATCTCGTATGCCGTCTTCTGCTTG"
[>TruSeq_Adapter_Index_19]="GATCGGAAGAGCACACGTCTGAACTCCAGTCACG
TGAAACGATCTCGTATGCCGTCTTCTGCTTG"
[>TruSeq_Adapter_Index_20]="GATCGGAAGAGCACACGTCTGAACTCCAGTCACG
TGGCCTTATCTCGTATGCCGTCTTCTGCTTG"
[>TruSeq_Adapter_Index_21]="GATCGGAAGAGCACACGTCTGAACTCCAGTCACG
TTTCGGAATCTCGTATGCCGTCTTCTGCTTG"
[>TruSeq_Adapter_Index_22]="GATCGGAAGAGCACACGTCTGAACTCCAGTCACC
GTACGTAATCTCGTATGCCGTCTTCTGCTTG"
[>TruSeq_Adapter_Index_23]="GATCGGAAGAGCACACGTCTGAACTCCAGTCACG
AGTGGATATCTCGTATGCCGTCTTCTGCTTG"
[>TruSeq_Adapter_Index_25]="GATCGGAAGAGCACACGTCTGAACTCCAGTCACA
CTGATATATCTCGTATGCCGTCTTCTGCTTG"
[>TruSeq_Adapter_Index_27]="GATCGGAAGAGCACACGTCTGAACTCCAGTCACA
TTCCTTTATCTCGTATGCCGTCTTCTGCTTG"
[>I5_Nextera_Transposase_1]="CTGTCTCTTATACACATCTGACGCTGCCGACGA
"
[>I7_Nextera_Transposase_1]="CTGTCTCTTATACACATCTCCGAGCCCACGAGA
C"
[>I5_Nextera_Transposase_2]="CTGTCTCTTATACACATCTCTGATGGCGCGAGG
GAGGC"
[>I7_Nextera_Transposase_2]="CTGTCTCTTATACACATCTCTGAGCGGGCTGGC
AAGGC"
[>I5_Primer_Nextera_XT_and_Nextera_Enrichment_[NSE]501]="GACGC
TGCCGACGAGCGATCTAGTGTAGATCTCGGTGGTCGCCGTATCATT"
[>I5_Primer_Nextera_XT_and_Nextera_Enrichment_[NSE]502]="GACGC
TGCCGACGAATAGAGAGGTGTAGATCTCGGTGGTCGCCGTATCATT"
[>I5_Primer_Nextera_XT_and_Nextera_Enrichment_[NSE]503]="GACGC
TGCCGACGAAGAGGATAGTGTAGATCTCGGTGGTCGCCGTATCATT"
[>I5_Primer_Nextera_XT_and_Nextera_Enrichment_[NSE]504]="GACGC
TGCCGACGATCTACTCTGTGTAGATCTCGGTGGTCGCCGTATCATT"
[>I5_Primer_Nextera_XT_and_Nextera_Enrichment_[NSE]505]="GACGC
TGCCGACGACTCCTTACGTGTAGATCTCGGTGGTCGCCGTATCATT"
[>I5_Primer_Nextera_XT_and_Nextera_Enrichment_[NSE]506]="GACGC
TGCCGACGATATGCAGTGTGTAGATCTCGGTGGTCGCCGTATCATT"
[>I5_Primer_Nextera_XT_and_Nextera_Enrichment_[NSE]507]="GACGC
TGCCGACGATACTCCTTGTGTAGATCTCGGTGGTCGCCGTATCATT"
[>I5_Primer_Nextera_XT_and_Nextera_Enrichment_[NSE]508]="GACGC
TGCCGACGAAGGCTTAGGTGTAGATCTCGGTGGTCGCCGTATCATT"
[>I5_Primer_Nextera_XT_and_Nextera_Enrichment_[NSE]517]="GACGC
TGCCGACGATCTTACGCGTGTAGATCTCGGTGGTCGCCGTATCATT"
[>I7_Primer_Nextera_XT_and_Nextera_Enrichment_N701]="CCGAGCCCA
CGAGACTAAGGCGAATCTCGTATGCCGTCTTCTGCTTG"
[>I7_Primer_Nextera_XT_and_Nextera_Enrichment_N702]="CCGAGCCCA
CGAGACCGTACTAGATCTCGTATGCCGTCTTCTGCTTG"
[>I7_Primer_Nextera_XT_and_Nextera_Enrichment_N703]="CCGAGCCCA
CGAGACAGGCAGAAATCTCGTATGCCGTCTTCTGCTTG"

[>I7_Primer_Nextera_XT_and_Nextera_Enrichment_N704]="CCGAGCCCA
CGAGACTCCTGAGCATCTCGTATGCCGTCTTCTGCTTG"
[>I7_Primer_Nextera_XT_and_Nextera_Enrichment_N705]="CCGAGCCCA
CGAGACGGACTCCTATCTCGTATGCCGTCTTCTGCTTG"
[>I7_Primer_Nextera_XT_and_Nextera_Enrichment_N706]="CCGAGCCCA
CGAGACTAGGCATGATCTCGTATGCCGTCTTCTGCTTG"
[>I7_Primer_Nextera_XT_and_Nextera_Enrichment_N707]="CCGAGCCCA
CGAGACCTCTCTACATCTCGTATGCCGTCTTCTGCTTG"
[>I7_Primer_Nextera_XT_and_Nextera_Enrichment_N708]="CCGAGCCCA
CGAGACCAGAGAGGATCTCGTATGCCGTCTTCTGCTTG"
[>I7_Primer_Nextera_XT_and_Nextera_Enrichment_N709]="CCGAGCCCA
CGAGACGCTACGCTATCTCGTATGCCGTCTTCTGCTTG"
[>I7_Primer_Nextera_XT_and_Nextera_Enrichment_N710]="CCGAGCCCA
CGAGACCGAGGCTGATCTCGTATGCCGTCTTCTGCTTG"
[>I7_Primer_Nextera_XT_and_Nextera_Enrichment_N711]="CCGAGCCCA
CGAGACAAGAGGCAATCTCGTATGCCGTCTTCTGCTTG"
[>I7_Primer_Nextera_XT_and_Nextera_Enrichment_N712]="CCGAGCCCA
CGAGACGTAGAGGAATCTCGTATGCCGTCTTCTGCTTG"
[>I5_Primer_Nextera_XT_Index_Kit_v2_S502]="GACGCTGCCGACGAATAGA
GAGGTGTAGATCTCGGTGGTCGCCGTATCATT"
[>I5_Primer_Nextera_XT_Index_Kit_v2_S503]="GACGCTGCCGACGAAGAGG
ATAGTGTAGATCTCGGTGGTCGCCGTATCATT"
[>I5_Primer_Nextera_XT_Index_Kit_v2_S505]="GACGCTGCCGACGACTCCT
TACGTGTAGATCTCGGTGGTCGCCGTATCATT"
[>I5_Primer_Nextera_XT_Index_Kit_v2_S506]="GACGCTGCCGACGATATGC
AGTGTGTAGATCTCGGTGGTCGCCGTATCATT"
[>I5_Primer_Nextera_XT_Index_Kit_v2_S507]="GACGCTGCCGACGATACTC
CTTGTGTAGATCTCGGTGGTCGCCGTATCATT"
[>I5_Primer_Nextera_XT_Index_Kit_v2_S508]="GACGCTGCCGACGAAGGCT
TAGGTGTAGATCTCGGTGGTCGCCGTATCATT"
[>I5_Primer_Nextera_XT_Index_Kit_v2_S510]="GACGCTGCCGACGAATTAG
ACGGTGTAGATCTCGGTGGTCGCCGTATCATT"
[>I5_Primer_Nextera_XT_Index_Kit_v2_S511]="GACGCTGCCGACGACGGAG
AGAGTGTAGATCTCGGTGGTCGCCGTATCATT"
[>I5_Primer_Nextera_XT_Index_Kit_v2_S513]="GACGCTGCCGACGACTAGT
CGAGTGTAGATCTCGGTGGTCGCCGTATCATT"
[>I5_Primer_Nextera_XT_Index_Kit_v2_S515]="GACGCTGCCGACGAAGCTA
GAAGTGTAGATCTCGGTGGTCGCCGTATCATT"
[>I5_Primer_Nextera_XT_Index_Kit_v2_S516]="GACGCTGCCGACGAACTCT
AGGGTGTAGATCTCGGTGGTCGCCGTATCATT"
[>I5_Primer_Nextera_XT_Index_Kit_v2_S517]="GACGCTGCCGACGATCTTA
CGCGTGTAGATCTCGGTGGTCGCCGTATCATT"
[>I5_Primer_Nextera_XT_Index_Kit_v2_S518]="GACGCTGCCGACGACTTAA
TAGGTGTAGATCTCGGTGGTCGCCGTATCATT"
[>I5_Primer_Nextera_XT_Index_Kit_v2_S520]="GACGCTGCCGACGAATAGC
CTTGTGTAGATCTCGGTGGTCGCCGTATCATT"
[>I5_Primer_Nextera_XT_Index_Kit_v2_S521]="GACGCTGCCGACGATAAGG
CTCGTGTAGATCTCGGTGGTCGCCGTATCATT"
[>I5_Primer_Nextera_XT_Index_Kit_v2_S522]="GACGCTGCCGACGATCGCA
TAAGTGTAGATCTCGGTGGTCGCCGTATCATT"

[>I7_Primer_Nextera_XT_Index_Kit_v2_N701]="CCGAGCCCACGAGACTAAG
GCGAATCTCGTATGCCGTCTTCTGCTTG"
[>I7_Primer_Nextera_XT_Index_Kit_v2_N702]="CCGAGCCCACGAGACCGTA
CTAGATCTCGTATGCCGTCTTCTGCTTG"
[>I7_Primer_Nextera_XT_Index_Kit_v2_N703]="CCGAGCCCACGAGACAGGC
AGAAATCTCGTATGCCGTCTTCTGCTTG"
[>I7_Primer_Nextera_XT_Index_Kit_v2_N704]="CCGAGCCCACGAGACTCCT
GAGCATCTCGTATGCCGTCTTCTGCTTG"
[>I7_Primer_Nextera_XT_Index_Kit_v2_N705]="CCGAGCCCACGAGACGGAC
TCCTATCTCGTATGCCGTCTTCTGCTTG"
[>I7_Primer_Nextera_XT_Index_Kit_v2_N706]="CCGAGCCCACGAGACTAGG
CATGATCTCGTATGCCGTCTTCTGCTTG"
[>I7_Primer_Nextera_XT_Index_Kit_v2_N707]="CCGAGCCCACGAGACCTCT
CTACATCTCGTATGCCGTCTTCTGCTTG"
[>I7_Primer_Nextera_XT_Index_Kit_v2_N710]="CCGAGCCCACGAGACCGAG
GCTGATCTCGTATGCCGTCTTCTGCTTG"
[>I7_Primer_Nextera_XT_Index_Kit_v2_N711]="CCGAGCCCACGAGACAAGA
GGCAATCTCGTATGCCGTCTTCTGCTTG"
[>I7_Primer_Nextera_XT_Index_Kit_v2_N712]="CCGAGCCCACGAGACGTAG
AGGAATCTCGTATGCCGTCTTCTGCTTG"
[>I7_Primer_Nextera_XT_Index_Kit_v2_N714]="CCGAGCCCACGAGACGCTC
ATGAATCTCGTATGCCGTCTTCTGCTTG"
[>I7_Primer_Nextera_XT_Index_Kit_v2_N715]="CCGAGCCCACGAGACATCT
CAGGATCTCGTATGCCGTCTTCTGCTTG"
[>I7_Primer_Nextera_XT_Index_Kit_v2_N716]="CCGAGCCCACGAGACACTC
GCTAATCTCGTATGCCGTCTTCTGCTTG"
[>I7_Primer_Nextera_XT_Index_Kit_v2_N718]="CCGAGCCCACGAGACGGAG
CTACATCTCGTATGCCGTCTTCTGCTTG"
[>I7_Primer_Nextera_XT_Index_Kit_v2_N719]="CCGAGCCCACGAGACGCGT
AGTAATCTCGTATGCCGTCTTCTGCTTG"
[>I7_Primer_Nextera_XT_Index_Kit_v2_N720]="CCGAGCCCACGAGACCGGA
GCCTATCTCGTATGCCGTCTTCTGCTTG"
[>I7_Primer_Nextera_XT_Index_Kit_v2_N721]="CCGAGCCCACGAGACTACG
CTGCATCTCGTATGCCGTCTTCTGCTTG"
[>I7_Primer_Nextera_XT_Index_Kit_v2_N722]="CCGAGCCCACGAGACATGC
GCAGATCTCGTATGCCGTCTTCTGCTTG"
[>I7_Primer_Nextera_XT_Index_Kit_v2_N723]="CCGAGCCCACGAGACTAGC
GCTCATCTCGTATGCCGTCTTCTGCTTG"
[>I7_Primer_Nextera_XT_Index_Kit_v2_N724]="CCGAGCCCACGAGACACTG
AGCGATCTCGTATGCCGTCTTCTGCTTG"
[>I7_Primer_Nextera_XT_Index_Kit_v2_N726]="CCGAGCCCACGAGACCCTA
AGACATCTCGTATGCCGTCTTCTGCTTG"
[>I7_Primer_Nextera_XT_Index_Kit_v2_N727]="CCGAGCCCACGAGACCGAT
CAGTATCTCGTATGCCGTCTTCTGCTTG"
[>I7_Primer_Nextera_XT_Index_Kit_v2_N728]="CCGAGCCCACGAGACTGCA
GCTAATCTCGTATGCCGTCTTCTGCTTG"
[>I7_Primer_Nextera_XT_Index_Kit_v2_N729]="CCGAGCCCACGAGACTCGA
CGTCATCTCGTATGCCGTCTTCTGCTTG"
[>I5_Adapter_Nextera]="CTGATGGCGCGAGGGAGGCGTGTAGATCTCGGTGGTCGC
CGTATCATT"

[>I7_Adapter_Nextera_No_Barcode]="CTGAGCGGGCTGGCAAGGCAGACCGATC
TCGTATGCCGTCTTCTGCTTG"
[>Nextera_LMP_Read1_External_Adapter]="GATCGGAAGAGCACACGTCTGAA
CTCCAGTCAC"
[>Nextera_LMP_Read2_External_Adapter]="GATCGGAAGAGCGTCGTGTAGGG
AAAGAGTGT"
[>RNA_Adapter_RA5_part_#_15013205]="GATCGTCGGACTGTAGAACTCTGAAC
"
[>RNA_Adapter_RA3_part_#_15013207]="CCTTGGCACCCGAGAATTCCA"
[>Stop_Oligo_STP_8]="CCACGGGAACGTGGTGGAATTC"
[>RNA_RT_Primer_RTP_part_#_15013981]="TGGAATTCTCGGGTGCCAAGGC"
[>RNA_PCR_Primer_RP1_part_#_15013198]="TCGGACTGTAGAACTCTGAACGT
GTAGATCTCGGTGGTCGCCGTATCATT"
[>RNA_PCR_Primer_Index_1_RPI1_2,9]="TGGAATTCTCGGGTGCCAAGGAACTC
CAGTCACATCACGATCTCGTATGCCGTCTTCTGCTTG"
[>RNA_PCR_Primer_Index_2_RPI2]="TGGAATTCTCGGGTGCCAAGGAACTCCAGT
CACCGATGTATCTCGTATGCCGTCTTCTGCTTG"
[>RNA_PCR_Primer_Index_3_RPI3]="TGGAATTCTCGGGTGCCAAGGAACTCCAGT
CACTTAGGCATCTCGTATGCCGTCTTCTGCTTG"
[>RNA_PCR_Primer_Index_4_RPI4]="TGGAATTCTCGGGTGCCAAGGAACTCCAGT
CACTGACCAATCTCGTATGCCGTCTTCTGCTTG"
[>RNA_PCR_Primer_Index_5_RPI5]="TGGAATTCTCGGGTGCCAAGGAACTCCAGT
CACACAGTGATCTCGTATGCCGTCTTCTGCTTG"
[>RNA_PCR_Primer_Index_6_RPI6]="TGGAATTCTCGGGTGCCAAGGAACTCCAGT
CACGCCAATATCTCGTATGCCGTCTTCTGCTTG"
[>RNA_PCR_Primer_Index_7_RPI7]="TGGAATTCTCGGGTGCCAAGGAACTCCAGT
CACCAGATCATCTCGTATGCCGTCTTCTGCTTG"
[>RNA_PCR_Primer_Index_8_RPI8]="TGGAATTCTCGGGTGCCAAGGAACTCCAGT
CACACTTGAATCTCGTATGCCGTCTTCTGCTTG"
[>RNA_PCR_Primer_Index_9_RPI9]="TGGAATTCTCGGGTGCCAAGGAACTCCAGT
CACGATCAGATCTCGTATGCCGTCTTCTGCTTG"
[>RNA_PCR_Primer_Index_10_RPI10]="TGGAATTCTCGGGTGCCAAGGAACTCCA
GTCACTAGCTTATCTCGTATGCCGTCTTCTGCTTG"
[>RNA_PCR_Primer_Index_11_RPI11]="TGGAATTCTCGGGTGCCAAGGAACTCCA
GTCACGGCTACATCTCGTATGCCGTCTTCTGCTTG"
[>RNA_PCR_Primer_Index_12_RPI12]="TGGAATTCTCGGGTGCCAAGGAACTCCA
GTCACCTTGTAATCTCGTATGCCGTCTTCTGCTTG"
[>RNA_PCR_Primer_Index_13_RPI13]="TGGAATTCTCGGGTGCCAAGGAACTCCA
GTCACAGTCAAATCTCGTATGCCGTCTTCTGCTTG"
[>RNA_PCR_Primer_Index_14_RPI14]="TGGAATTCTCGGGTGCCAAGGAACTCCA
GTCACAGTTCCATCTCGTATGCCGTCTTCTGCTTG"
[>RNA_PCR_Primer_Index_15_RPI15]="TGGAATTCTCGGGTGCCAAGGAACTCCA
GTCACATGTCAATCTCGTATGCCGTCTTCTGCTTG"
[>RNA_PCR_Primer_Index_16_RPI16]="TGGAATTCTCGGGTGCCAAGGAACTCCA
GTCACCCGTCCATCTCGTATGCCGTCTTCTGCTTG"
[>RNA_PCR_Primer_Index_17_RPI17]="TGGAATTCTCGGGTGCCAAGGAACTCCA
GTCACGTAGAGATCTCGTATGCCGTCTTCTGCTTG"
[>RNA_PCR_Primer_Index_18_RPI18]="TGGAATTCTCGGGTGCCAAGGAACTCCA
GTCACGTCCGCATCTCGTATGCCGTCTTCTGCTTG"
[>RNA_PCR_Primer_Index_19_RPI19]="TGGAATTCTCGGGTGCCAAGGAACTCCA
GTCACGTGAAAATCTCGTATGCCGTCTTCTGCTTG"

[>RNA_PCR_Primer_Index_20_RPI20]="TGGAATTCTCGGGTGCCAAGGAACTCCA
GTCACGTGGCCATCTCGTATGCCGTCTTCTGCTTG"
[>RNA_PCR_Primer_Index_22_RPI22]="TGGAATTCTCGGGTGCCAAGGAACTCCA
GTCACCGTACGATCTCGTATGCCGTCTTCTGCTTG"
[>RNA_PCR_Primer_Index_23_RPI23]="TGGAATTCTCGGGTGCCAAGGAACTCCA
GTCACGAGTGGATCTCGTATGCCGTCTTCTGCTTG"
[>RNA_PCR_Primer_Index_24_RPI24]="TGGAATTCTCGGGTGCCAAGGAACTCCA
GTCACGGTAGCATCTCGTATGCCGTCTTCTGCTTG"
[>RNA_PCR_Primer_Index_25_RPI25]="TGGAATTCTCGGGTGCCAAGGAACTCCA
GTCACACTGATATCTCGTATGCCGTCTTCTGCTTG"
[>RNA_PCR_Primer_Index_26_RPI26]="TGGAATTCTCGGGTGCCAAGGAACTCCA
GTCACATGAGCATCTCGTATGCCGTCTTCTGCTTG"
[>RNA_PCR_Primer_Index_27_RPI27]="TGGAATTCTCGGGTGCCAAGGAACTCCA
GTCACATTCCTATCTCGTATGCCGTCTTCTGCTTG"
[>RNA_PCR_Primer_Index_28_RPI28]="TGGAATTCTCGGGTGCCAAGGAACTCCA
GTCACCAAAAGATCTCGTATGCCGTCTTCTGCTTG"
[>RNA_PCR_Primer_Index_29_RPI29]="TGGAATTCTCGGGTGCCAAGGAACTCCA
GTCACCAACTAATCTCGTATGCCGTCTTCTGCTTG"
[>RNA_PCR_Primer_Index_30_RPI30]="TGGAATTCTCGGGTGCCAAGGAACTCCA
GTCACCACCGGATCTCGTATGCCGTCTTCTGCTTG"
[>RNA_PCR_Primer_Index_31_RPI31]="TGGAATTCTCGGGTGCCAAGGAACTCCA
GTCACCACGATATCTCGTATGCCGTCTTCTGCTTG"
[>RNA_PCR_Primer_Index_32_RPI32]="TGGAATTCTCGGGTGCCAAGGAACTCCA
GTCACCACTCAATCTCGTATGCCGTCTTCTGCTTG"
[>RNA_PCR_Primer_Index_33_RPI33]="TGGAATTCTCGGGTGCCAAGGAACTCCA
GTCACCAGGCGATCTCGTATGCCGTCTTCTGCTTG"
[>RNA_PCR_Primer_Index_34_RPI34]="TGGAATTCTCGGGTGCCAAGGAACTCCA
GTCACCATGGCATCTCGTATGCCGTCTTCTGCTTG"
[>RNA_PCR_Primer_Index_35_RPI35]="TGGAATTCTCGGGTGCCAAGGAACTCCA
GTCACCATTTTATCTCGTATGCCGTCTTCTGCTTG"
[>RNA_PCR_Primer_Index_36_RPI36]="TGGAATTCTCGGGTGCCAAGGAACTCCA
GTCACCCAACAATCTCGTATGCCGTCTTCTGCTTG"
[>RNA_PCR_Primer_Index_37_RPI37]="TGGAATTCTCGGGTGCCAAGGAACTCCA
GTCACCGGAATATCTCGTATGCCGTCTTCTGCTTG"
[>RNA_PCR_Primer_Index_38_RPI38]="TGGAATTCTCGGGTGCCAAGGAACTCCA
GTCACCTAGCTATCTCGTATGCCGTCTTCTGCTTG"
[>RNA_PCR_Primer_Index_39_RPI39]="TGGAATTCTCGGGTGCCAAGGAACTCCA
GTCACCTATACATCTCGTATGCCGTCTTCTGCTTG"
[>RNA_PCR_Primer_Index_40_RPI40]="TGGAATTCTCGGGTGCCAAGGAACTCCA
GTCACCTCAGAATCTCGTATGCCGTCTTCTGCTTG"
[>RNA_PCR_Primer_Index_41_RPI41]="TGGAATTCTCGGGTGCCAAGGAACTCCA
GTCACGACGACATCTCGTATGCCGTCTTCTGCTTG"
[>RNA_PCR_Primer_Index_42_RPI42]="TGGAATTCTCGGGTGCCAAGGAACTCCA
GTCACTAATCGATCTCGTATGCCGTCTTCTGCTTG"
[>RNA_PCR_Primer_Index_43_RPI43]="TGGAATTCTCGGGTGCCAAGGAACTCCA
GTCACTACAGCATCTCGTATGCCGTCTTCTGCTTG"
[>RNA_PCR_Primer_Index_44_RPI44]="TGGAATTCTCGGGTGCCAAGGAACTCCA
GTCACTATAATATCTCGTATGCCGTCTTCTGCTTG"
[>RNA_PCR_Primer_Index_45_RPI45]="TGGAATTCTCGGGTGCCAAGGAACTCCA
GTCACTCATTCATCTCGTATGCCGTCTTCTGCTTG"

```
[>RNA_PCR_Primer_Index_46_RPI46]="TGGAATTCTCGGGTGCCAAGGAACTCCA
GTCACTCCCGAATCTCGTATGCCGTCTTCTGCTTG"
[>RNA_PCR_Primer_Index_47_RPI47]="TGGAATTCTCGGGTGCCAAGGAACTCCA
GTCACTCGAAGATCTCGTATGCCGTCTTCTGCTTG"
[>RNA_PCR_Primer_Index_48_RPI48]="TGGAATTCTCGGGTGCCAAGGAACTCCA
GTCACTCGGCAATCTCGTATGCCGTCTTCTGCTTG"
[>PhiX_read1_adapter]="AGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGAT
CTCGTATGCCGTCTTCTGCTTGAAA"
[>PhiX_read2_adapter]="AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCT
CGGTGGTCGCCGTATCATTAAAAAA"
[>Bisulfite_R1]="AGATCGGAAGAGCACACGTCTGAAC"
[>Bisulfite_R2]="AGATCGGAAGAGCGTCGTGTAGGGA")

#loop opver the array of adapters and grep them against the
input file. zcat takes .gz as imput, could set up check for
file type and make allowences for uncompressed files?

for i in ${!ADAP[@]}; do
    VAL="$( zcat $FILE | head -400000 | grep "${ADAP[$i]}" |
wc -l )"
    if [ $VAL -gt 100 ]
    then
        echo "$VAL"
        echo "This adaptor was found:"
        echo "$i"
        echo "${ADAP[$i]}"
        echo "$i" >> $ADAP_FA
        echo "${ADAP[$i]}" >> $ADAP_FA
   fi
done

#Adapters from  bbmap resources
```

A3: Aligning and extracting Nanopore reads.

```
$ bbmap/bbmapskimmer.sh maxlen=1000
in=/path/to/basecalled_reads.fasta
ref=/path/to/PE_150_contigs.fasta out=output.sam
$ samtools view -S -b output.sam > output.bam
$ samtools view -F 4 output.bam > mapped_output.bam
$ samtools faidx /path/to/basecalled_reads.fasta
$ cat mapped_output.bam | awk '{print $1;}' >
mapped_seqeunce_identifiers.txt
$ for i in $(cat < mapped_seqeunce_identifiers.txt); do
samtools faidx basecalled_reads.fasta $i >>
mapped_Nanopore_reads.fasta; done
```

A4: fasta_len_filter.sh (written by Matt Storey).

```
#! /bin/bash
#unwraps INPUT then pipes to length filter with a cut off of >
$3

INPUT=$1
OUTPUT=$2

awk '/^>/ {printf("\n%s\n",$0);next; } { printf("%s",$0);}
END {printf("\n");}' < $INPUT | awk -v LEN="$3" '!/^>/ { next
} { getline seq } length(seq) ≥ LEN { print $0 "\n" seq }' >
$OUTPUT
```

A5: Workflow for creating coverage (maximum k-mer coverage as calculated by metaSPAdes) over %GC graphs. `Calc.gc.pl` script from [201].

```
$ perl ~/multi-metagenome/R.data.generation/calc.gc.pl -i
assembly.fa -o assembly.gc.tab
# these grep and paste commands are specific to the sequence
header created by the SPAdes assembler, which contains the
length and coverage of the contig
$ grep "^>" assembly.fa | awk -F'_' 'BEGIN {print "coverage"}
{print $6}' > coverage_column.tab
$ grep "^>" assembly.fa | awk -F'_' 'BEGIN {print "length"}
{print $4}' > assembly_length.tab
$ paste -d '\t' assembly.gc.tab coverage_column.tab >
assembly_copy.gc.coverage.tab
$ paste -d '\t' assembly_copy.gc.coverage.tab
assembly_length.tab > assembly.gc.coverage.length.tab
# checked head, tail and wc -l of the columns match
# create a table of all the contigs in each of the bin files
$ grep "^>" *.fasta | sed 's/:>/\t/' | awk -F'_' 'BEGIN {print
"bin""\t""contig"} {print $0}' > contigs_per_bin.tsv
# copied results.tsv created by checkM run as part of dRep and
located in dRep_output/data/checkM/checkM_outdir/results.tsv
> gc_coverage <-
read.delim("/path/to/assembly.gc.coverage.length.tab",
header=TRUE, sep="\t")
```

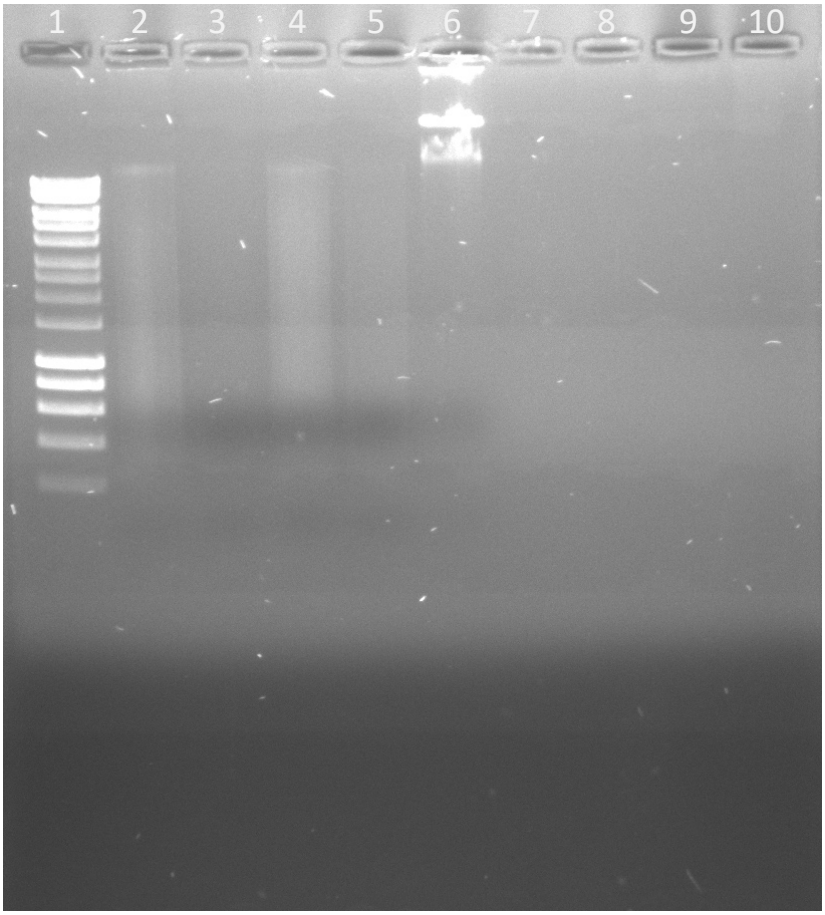
```
> contigs_per_bin <-
read.delim("/path/to/contigs_per_bin.tsv", header=TRUE,
sep="\t")
> bin_taxonomy <- read.delim("/path/to/results.tsv",
header=TRUE, sep="\t")
> gc_coverage$bin <-
contigs_per_bin$bin[match(gc_coverage$contig,
contigs_per_bin$contig)]
> gc_coverage$marker_lineage <-
bin_taxonomy$Marker.lineage[match(gc_coverage$bin,
bin_taxonomy$Bin.Id)]
# plotting the graph using the table created
> library(ggplot2)
> ggplot(gc_coverage[!is.na(gc_coverage$bin), ], aes(x = gc, y
= coverage, color = marker_lineage, size = length)) +
geom_point(alpha = 0.25, shape = 21, stroke = (0.8)) +
scale_y_log10(limits = c(0.1, 8000)) + theme_bw() +
theme(legend.position="right", legend.key.size = unit(0.25,
"cm")) + xlab("% GC") + ylab("log10(coverage)") +
scale_size_area(name = "Contig length", max_size = 20) +
guides(colour = guide_legend(override.aes = list(alpha = 1,
size = 3, shape = 19)))
# writing out plot file as png
> coverage_over_gc_plot <-
ggplot(gc_coverage[!is.na(gc_coverage$bin), ], aes(x = gc, y =
coverage, color = marker_lineage, size = length)) +
geom_point(alpha = 0.25, shape = 21, stroke = (0.8)) +
scale_y_log10(limits = c(0.1, 8000)) + theme_bw() +
theme(legend.position="right", legend.key.size = unit(0.25,
"cm")) + xlab("% GC") + ylab("log10(coverage)") +
scale_size_area(name = "Contig length", max_size = 20) +
guides(colour = guide_legend(override.aes = list(alpha = 1,
size = 3, shape = 19)))
> png(file = "/path/to/plot.png")
```

```
> plot(coverage_over_gc_plot)
> dev.off()
```

A6: Details regarding the samples and methodology used for DNA extraction from Seawater and Antarctic samples as well as visibility of HMW DNA on an agarose gel (Gel below for reference) and it's average concentration (measured using NanoPhotometer NP80 (Implen)).

Type of sample (quantity)	**Effective sample volume**	**Pre-treatment**	**Extraction Method**	**HMW DNA extracted**	**DNA conc. ng/µl**
NZ Seawater					
Seawater filter paper (2*1/2)	2*100 l	Biomass washed off filter paper using 1 ml of glycerol.	Method 1	4	n/a
Seawater filter paper (2*1/2)	2*100 l	Filter paper cut into small pieces using a sterile scalpel, submerged in 3 ml of 100 mM Tris-Cl [pH=8.0] and vortexed for 5 minutes.	Method 2	4	n/a
Seawater filter paper March 2017 (1/4)	50 l	Filter paper submerged in 2.7 ml of seawater lysis buffer and vortexed for 5 min	Method 4	2	~45
Seawater filter paper March 2017 (1/4)	50 l	Biomass washed off filter paper using 2.7 ml of seawater lysis buffer	Method 4	2	~40
Seawater filter paper March 2017 (1/4)	50 l	Filter paper cut into small pieces using a sterile scalpel and submerged in 5 ml of sponge lysis buffer.	Method 5	1	~95
Seawater filter paper March 2017 (1/4)	50 l	Filter paper submerged in 5 ml of sponge lysis buffer and vortexed for 5 min.	Method 5	1	~85
Seawater filter paper March 2017 (2*1/2)	2*100 l	Filter paper submerged in 5 ml of sponge lysis buffer and vortexed for 5 min.	Method 5	2	~60
Antarctic sea ice					
Biomass survey in glycerol from Granite Harbour 21/11/08 (2)	1 ml	Thawed on ice. Centrifuged at 4,000 g at 4°C for 15 minutes. Pellet resuspended in 1 ml of soil lysis buffer.	Method 1	4	n/a
Brine samples in glycerol from Granite Harbour 24/11/08 (45, 65, 85cm)	45 cm: 0.5 ml 65 cm: 0.5 ml 85 cm: 0.5 ml	Thawed on ice. Centrifuged at 4,000 g at 4°C for 15 minutes. Pellet resuspended in 1 ml of soil lysis buffer.	Method 1	4	n/a

Biomass survey in glycerol from Terra Nova Bay 25/11/07 (4)	2 ml	Centrifuged at 4,000 g at 4°C for 15 minutes. Pellet resuspended in enzymatic lysis buffer	Method 3	4	n/a
Filter paper from Cape Evans 03/12/10 (Bottom & Middle)	Middle: 880 ml Bottom: 850 ml	Filter paper cut into small pieces using a sterile scalpel. Submerged in 1 ml enzymatic lysis buffer.	Method 3	4	n/a
Filter paper from Cape Evans 06/12/10 (Top, middle and bottom)	Top: 680 ml Middle: 885 ml Bottom: 855 ml	Three filter papers cut into small pieces using a sterile scalpel, submerged in 2 ml Enzymatic lysis buffer and vortexed for 2 minutes.	Method 6	3	~105
Biomass survey in glycerol from Terra Nova Bay 30/11/07 (4)	2 ml	Samples combined, centrifuged at 4,000 g at 4°C for 15 minutes and resuspended in 1 ml of enzymatic lysis buffer.	Method 6	3	~50
Seawater samples in glycerol from Terra Nova Bay, 23/11/ and 05/12/07 (9 ea.)	5 m: 3 ml 25 m: 3 ml 50 m: 3 ml	Samples combined, centrifuged at 17,000 g at room temperature for 20 minutes and resuspended in 1 ml of enzymatic lysis buffer	Method 6	4	n/a
Filter papers from Cape Evans, unknown dates in 2010 (4)	n/a	Four filter papers cut into small pieces using a sterile scalpel, submerged in 2 ml Enzymatic lysis buffer and vortexed for 2 minutes.	Method 6	2	n/a
Biomass survey in glycerol from Terra Nova Bay, 16/11/ and 19/11/07 (4 ea.)	4 ml	Samples combined, centrifuged at 17,000 g at room temperature for 20 minutes and resuspended in 1 ml of enzymatic lysis buffer	Method 6	1	n/a
Filter papers from Cape Evans, 26/11/10 (Bottom), 30/11/10 (Middle) and unknown date (Bottom)	Bottom 26/11/10: 100 ml Middle 30/11/10: 845 ml	Three filter papers cut into small pieces using a sterile scalpel, submerged in 2 ml Enzymatic lysis buffer and vortexed for 5 minutes.	Method 6	n/a	n/a
Seawater samples in glycerol from Terra Nova Bay 16/11/07 (9) and 29/11/07 (8)	5 m: 3 ml 25 m: 2.5 ml 50 m: 3 ml	Samples combined, centrifuged at 17,000 g at room temperature for 20 minutes and resuspended in 1 ml of enzymatic lysis buffer	Method 6	n/a	n/a



Reference gel for column "HMW DNA extracted": Number scale (1-4) is used in Table above to denominate DNA quality. Lane 1 - Hyperladder (1 kb), Lane 2 - 1 (Distinct band of HMW DNA); Lane 3 - 4 (no DNA); Lane 4 - 2 (HMW DNA is present); Lane 5 - 3 (Mostly or all LMW DNA); Lane 6 = pWEB-TNC plus HMW insert; Lane 7-10 - empty.

A7: Describing the packaging reactions including size-selection method used for *Streptomyces albus* gDNA, vector stock used, colonies yielded and their dilutions, total colonies expected per µg of DNA and how many of the colonies appeared to have inserts on gel/colonies grown and tested.

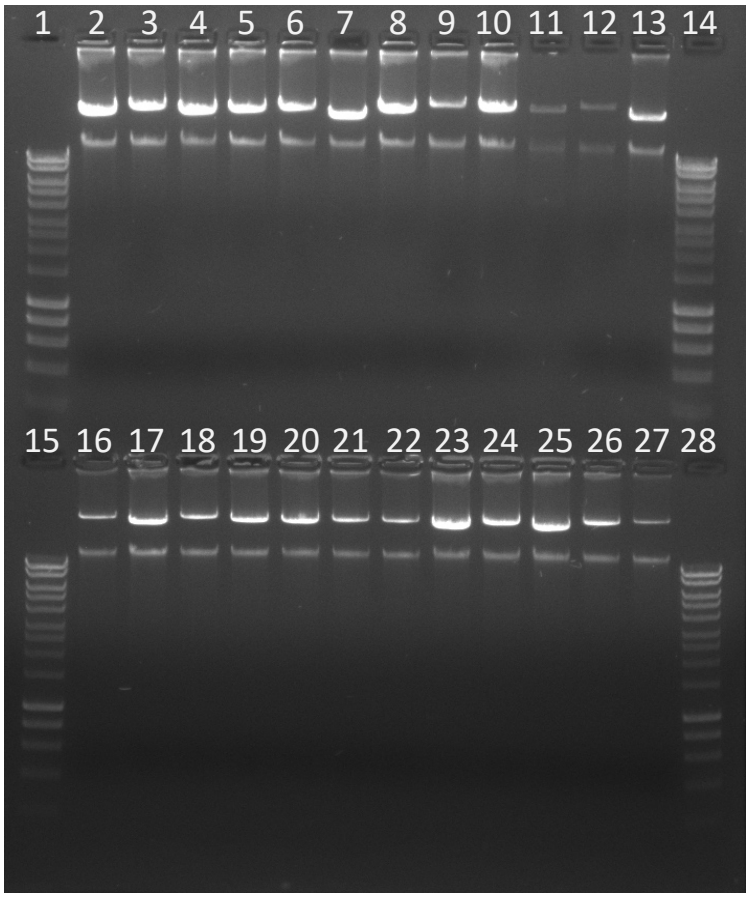
DNA sample	Vector	Colonies neat	Colonies 1:10	Colonies 1:100	Total colonies	Inserts confirmed by gel
Gel-El Classic	V7	220	36	6	61,989	6/6
Gel-El SYBR	V7	85	7	2	18,649	0/2
GELase Classic	V7	222	21	2	33,201	5/6
GELase SYBR	V7	95	10	0	10,244	2/2
GELase Classic	V4	119	4	0	8,353	1/4
GELase SYBR	V4	1200	85	14	181,240	4/32
GELase Classic	V5	208	41	3	48,226	2/6
GELase SYBR	V5	249	4	1	20,435	4/6
GELase Classic	V6	31	0	0	1,629	1/1
GELase Classic	V8	7	1	0	893	1/1
GELase Classic	V9	68	1	0	4,098	3/3
Gel-El. SYBR	V9	1000	57	2	92,984	1/27
GELase Classic	V4	188	15	3	33,516	1/3
GELase SYBR	V4	2400	149	45	44,0755	1/4
GELase Classic (s)	V4	1060	135	8	168,632	0/4
GELase SYBR (s)	V4	2	0	0	105	0/3
GELase Classic	V5	0	0	0	0	N/a
GELase SYBR	V5	2	0	0	105	N/a
GELase Classic (s)	V5	225	12	0	18,124	0/3
GELase SYBR (s)	V5	1200	141	21	247,432	0/4
GELase Classic	PCR	34	3	0	3,362	1/3
GELase SYBR	PCR	57	12	0	9,298	3/3

A8: Packaging reactions carried out for the construction of SWNZ cosmid library, including colonies yielded and number of inserts confirmed by gel electrophoresis.

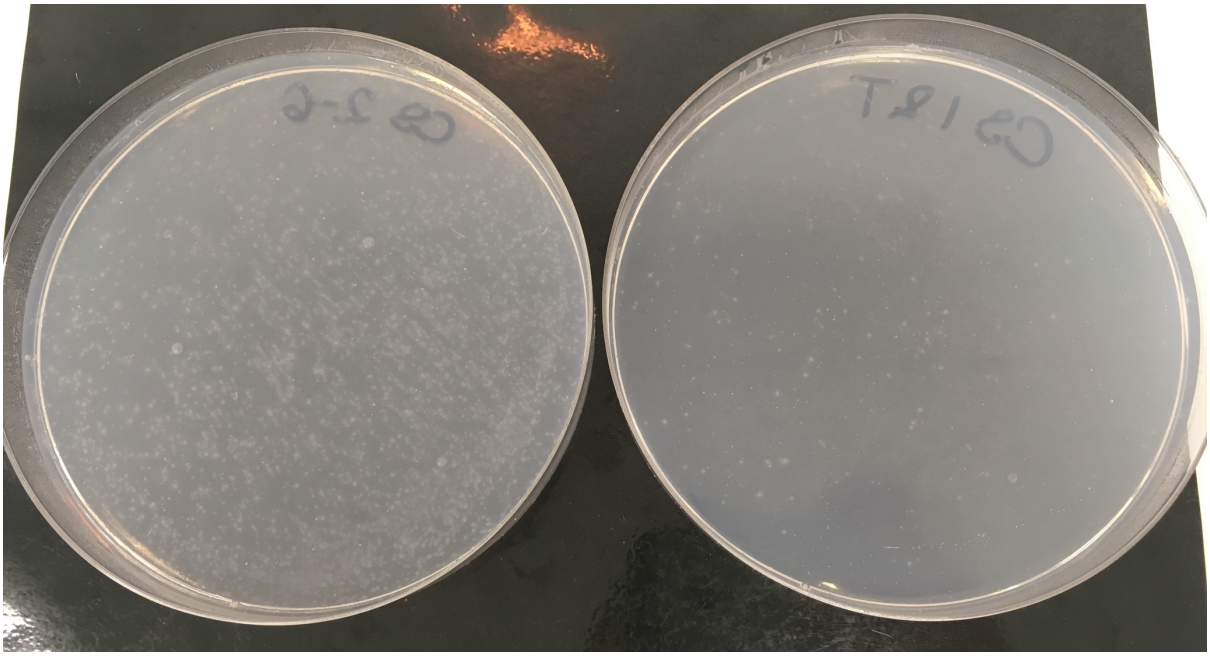
DNA sample	End-repaired	Ligation	Packaging reaction	PDB + chloroform	Colonies neat, 1:10, 1:100	Total colonies from ligation	Inserts confirmed by gel
2003 h1-3 (1)	No	50 ng DNA + 100 ng TNC (V2) in 10 µl	10 µl ligation + 45 µl PE	500 µl + 20 µl	70 9 0	1,960	16/24
2003 h1-3 (2)	No	50 ng DNA + 100 ng TNC (V2) in 10 µl	10 ul ligation + 45 µl PE	500 µl + 20 µl	92 3 1	2,576	8/16
2003 l1&2 (1)	No	50 ng DNA + 100 ng TNC (V2) in 10 µl	10 µl ligation + 45 µl PE	500 µl + 20 µl	8 0 1	224	2/2
2003 l1&2 (2)	No	50 ng DNA + 100 ng TNC (V2) in 10 µl	10 µl ligation + 45 µl PE	500 µl + 20 µl	3 1 0	84	0/2
2003 h1-3 (1)	No	50 ng DNA + 100 ng TNC (V2) in 10 µl	10 µl ligation + 45 µl PE	500 µl + 20 µl	2 2 0	100	1/12

2003 h1-3 (2)	No	50 ng DNA + 100 ng TNC (V2) in 10 µl	10 µl ligation + 45 µl PE	500 µl + 20 µl	65 1 1	3,250	1/12
2003 h1-3 (3)	No	50 ng DNA + 100 ng TNC (V2) in 10 µl	10 µl ligation + 45 µl PE	500 µl + 20 µl	177 66 13	8,850	1/12
2003 h1-3 (4)	No	50 ng DNA + 100 ng TNC (V2) in 10 µl	10 µl ligation + 45 µl PE	500 µl + 20 µl	1 0 0	50	1/12
2003 h1-3 (5)	No	50 ng DNA + 100 ng TNC (V2) in 10 µl	10 µl ligation + 70 µl PE	500 µl + 20 µl	3 0 0	150	1/12
2003 h1-3 (6)	No	50 ng DNA + 100 ng TNC (V2) in 10 µl	10 µl ligation + 70 µl PE	500 µl + 20 µl	3 0 0	150	1/12
2003 h1-3 (7)	No	50 ng DNA + 100 ng TNC (V8) in 10 µl	10 µl ligation + 70 µl PE	500 µl + 20 µl	0 1 0	1	1/12
2003 h1-3	Yes	180 ng DNA + 360 ng TNC (V7) in 25 µl	25 µl ligation + 130 µl PE	300 µl + 30 µl	19 2 0	741	1/2
0802 h1-3 MWCO conc	Yes	54 ng DNA + 180 ng TNC (V7) in 25 µl	25 µl ligation + 130 µl PE	300 µl + 30 µl	32 4 0	1,248	1/4
0802 h1-3 MWCO remn	Yes	79 ng DNA + 180 ng TNC (V7) in 25 µl	25 µl ligation + 130 µl PE	300 µl + 30 µl	210 14 3	8,190	1/18
0802 h1-3 MWCO conc	Yes	50 ng DNA + 100 ng TNC (V4) in 13 µl	13 µl ligation + 60 µl PE	350 µl + 30 µl	30 4 0	1,050	2/3
0802 h1-3 MWCO conc	Yes	50 ng DNA + 100 ng TNC (V4) in 13 µl	13 µl ligation + 60 µl PE	350 µl + 30 µl	1 0 0	35	1/1
0802 h1-3 MWCO conc	Yes	50 ng DNA + 100 ng TNC (V4) in 13 µl	13 µl ligation + 60 µl PE	350 µl + 30 µl	N/A 2 0		2/2
0802 h1-3 MWCO conc	Yes	50 ng DNA + 100 ng TNC (V4) in 14 µl	14 µl ligation + 60 µl PE	350 µl + 30 µl	2 0 0	70	0/1
0802 h1-3 MWCO conc	Yes	50 ng DNA + 100 ng TNC (V4) in 14 µl	14 µl ligation + 60 µl PE	350 µl + 30 µl	34 5 0	1,190	1/3
0802 h1-3 MWCO conc	Yes	50 ng DNA + 100 ng TNC (V4) in 14 µl	14 µl ligation + 60 µl PE	350 µl + 30 µl	328 31 3	11,480	2/11
0802 h1-3 MWCO conc	Yes	50 ng DNA + 100 ng TNC (V4) in 14 µl	14 µl ligation + 60 µl PE	350 µl + 30 µl	21 5 0	735	1/3

A9: Preparative gel electrophoresis of 24 cosmid clones isolated from the *C. mycofijiensis* library constructed.



A10: Colonies resulting from the first round of selection carried out as part of the PPTase enrichment screen described in 4.7 (Methods 2.11 for details).



A11: Number of bins produced by Metabat2, Maxbin2 and Autometa binning tools as well as number of bins resulting from dRep dereplication of these bins for two different *C.mycofijiensis* assemblies. Note that dRep with default settings only considers genomes of ≥500kb length, >75% completeness and <25% contamination.

Sponge	**Metabat2**	**MaxBin2**	**Autometa**	**dRep dereplicated**
PE150_plus_Nano	81	96	101	46
PE250_on_PE150	87	101	90	43

A12: Description of *C. mycofijiensis* metadata, date and type of sequencing carried out as well as number of reads (identified by FastQC Version 0.11.5 with default settings) and number of bases returned (in Gbp).

Sample (Reference)	**Metadata**	**Date**	**Sequencing technology (provider)**	**# sequences (FastQC)**	**Gbp**
C. mycofijiensis (PE150)	Collected from Cathedral Cave, 'Eua, Tonga on 07/06/2016	11/09/ 2017	Illumina HiSeq 4000, 2x150bp paired-end (Anorad)	35,262,051	10.58
C. mycofijiensis (PE250)	Collected from Cathedral Cave, 'Eua, Tonga on 07/06/2016	05/12/ 2017	Illumina HiSeq 4000, 2x250bp paired-end (Anorad)	25,162,970	12.58
C. mycofijiensis and *Mycale hentscheli* (Nano_reads)	Collected from Cathedral Cave, 'Eua, Tonga on 07/06/2016 and collected from Doubtful Sound, New Zealand	01/09/ 2017	Oxford Nanopore MinION (R9.4.1 Flow Cell); Ligation Sequencing Kit 1D R9 Version (SQK_LSK108)	318,814	

A13: Assembly results quantified using quast v.5.0.0 with default settings for individual assemblies of PE150, PE250, PE_both (PE150 and PE250 concatenated into one dataset), PE150_plus_Nano, PE250_on_PE150 (PE150 assembly supplied as trusted contigs for PE250 assembly) and PE250_on_PE150_plus_Nano (PE150_plus_Nano supplied as trusted contigs for PE250_assembly).

	PE150	**PE250**	**PE_both**	**PE150_ plus_Nano**	**PE250_on _PE150**	**PE250_on _PE150_ plus_Nano**
# contigs	275,848	610,748	857,815	274,771	837,749	837,012
# contigs (≥ 0 bp)	547,761	1,292,838	1,612,819	546,669	1,538,731	1,537,336
# contigs (≥ 1000 bp)	125,531	63,096	182,821	124,531	190,032	190,020
# contigs (≥ 5000 bp)	12,487	249	13,252	12,256	17,851	17,761
# contigs (≥ 10000 bp)	4,944	52	5,160	4,967	6,659	6,565
# contigs (≥ 25000 bp)	1,574	2	1,635	1,631	2,055	2,049
# contigs (≥ 50000 bp)	593	0	618	619	792	807
Largest contig	1,452,608	26,536	1,644,273	2,765,707	1,121,381	1,301,649
Total length	525,812,756	441,609,279	953,513,948	525,805,504	1,004,619,775	1,004,988,585
Total length (≥ 0 bp)	621,792,397	731,092,925	1,273,532,229	621,776,526	1,304,113,051	1,304,189,864
Total length (≥ 1000 bp)	421,185,594	85,057,041	508,123,310	421,228,225	575,252,410	576,142,958
Total length (≥ 5000 bp)	207,685,886	1,989,611	218,261,023	210,311,935	278,981,443	279,904,488
Total length (≥ 10000 bp)	156,570,789	733,405	163,529,057	160,876,357	202,721,479	203,583,034
Total length (≥ 25000 bp)	105,713,102	52,361	110,322,500	110,472,675	134,300,197	136,751,924
Total length (≥ 50000 bp)	72,216,924	0	75,422,370	75,882,121	90,961,553	94,024,851
N50	2,987	696	1,093	3,008	1,246	1,249
N75	1,185	580	672	1,186	700	700
L50	27,104	230,462	152,770	26,140	124,116	123,545
L75	100,837	405,094	441,976	99,798	407,150	406,325
GC (%)	56.82	43.35	50.58	56.82	50.05	50.05
Mismatches						
# N's	500	0	500	500	0	0
# N's per 100 kbp	0.1	0	0.05	0.1	0	0
Spades warnings	no warnings	no warnings	2 warnings	no warnings	6 warnings,	4 warnings

A14: End-sequencing results from 32 Sanger-sequenced cosmids that were consequently analyzes using Blastx with standard settings as part of Geneious Prime. 3 cosmids did not return any results for both primers and 11 reactions failed on either the forward or reverse primer. Two E. coli sequences and one fungal sequence was omitted from the table.

Well	Description [Organism]	Accession	Pairwise Identity	Bit-Score	E-Value	Sequence Length
A06	ferrous iron transporter B [Fuerstia marisgermanicae]	WP_077025374	71.40%	171.79	1.13E-46	133
A07	hypothetical protein [Afipia sp. 1NLS2]	WP_009338085	37.60%	113.24	5.90E-27	173
A08	hypothetical protein AMS19_06970 [Gemmatimonas sp. SG8_23]	KPJ82711	64.00%	224.56	4.34E-71	164
A09	hypothetical protein AMS19_09615 [Gemmatimonas sp. SG8_23]	KPJ80334	58.80%	177.95	5.74E-48	187
A10	ribulose-phosphate 3-epimerase [Acidobacteria bacterium]	PYR75947	68.70%	191.05	8.33E-58	147
A12	hypothetical protein [uncultured Gemmatimonadales bacterium HF0770_41L09]	ADI23526	44.40%	84.34	1.99E-15	117
B05	hypothetical protein DMD75_18550 [Candidatus Rokubacteria bacterium]	PYO08398	60.40%	63.16	5.57E-08	96
B07	DNA polymerase [Candidatus Synechococcus spongiarum 15L]	KKZ12738	65.80%	209.53	6.10E-63	161
B08	FAD-binding oxidoreductase, partial [Actinobacteria bacterium]	REK22985	55.00%	93.20	1.87E-19	80
B10	beta-lactamase TEM-1, partial [Acinetobacter baumannii]	APX42443	94.40%	379.02	1.16E-129	196
B12	RNA polymerase sigma factor RpoD [uncultured bacterium W5-15b]	AFD03342	67.30%	284.65	1.06E-88	205
C06	Gfo/Idh/MocA family oxidoreductase [Emticicia sp. MM]	WP_015029295	78.60%	51.99	2.68E-04	28
C08	DNA-binding protein [Reinekea blandensis]	WP_008048371	71.00%	45.05	8.34E-03	31
C09	ATP-binding protein [Candidatus Thiodictyon syntrophicum]	WP_100918879	65.30%	110.92	4.01E-27	75
C10	Uncharacterized conserved protein, LabA/DUF88 family [Thermoleophilum album]	SEH12861	42.90%	115.16	2.69E-27	203
C11	site-2 protease family protein [Halioglobus lutimaris]	WP_101518222	35.40%	56.610	3.97E-06	113
C12	beta-lactamase TEM-1, partial [Acinetobacter baumannii]	APX42443	93.30%	315.85	1.19E-105	165
D05	carbon monoxide dehydrogenase [Blastocatellia bacterium AA13]	PYP86095	67.60%	291.20	5.84E-96	213
D06	hypothetical protein [Afipia sp. 1NLS2]	WP_009338085	37.30%	108.23	2.99E-25	169
D07	response regulator [Paenibacillus tuaregi]	WP_068619831	50.00%	85.11	1.17E-16	80
D08	hypothetical protein [Sorangium cellulosum]	WP_020733992	46.50%	153.30	9.57E-42	159
D09	hypothetical protein [Cupriavidus sp. amp6]	WP_029049114	71.90%	266.93	1.01E-82	185
E05	AAA family ATPase [Cystobacter ferrugineus]	WP_071900752	67.70%	202.99	7.35E-60	167
E06	hypothetical protein [Phyllobacterium sp. YR531]	WP_008123265	52.60%	75.10	1.10E-14	76
E07	DEAD/DEAH box helicase [Gemmatimonadetes bacterium]	RMH20306	38.20%	89.74	3.69E-17	217
E08	hypothetical protein AMS19_02000 [Gemmatimonas sp. SG8_23]	KPJ84095	72.60%	127.49	3.61E-31	117
E09	methionine--tRNA ligase [Rhodothermus marinus]	WP_012844066	81.00%	333.18	9.89E-108	189

E10	methionine--tRNA ligase [Rhodothermus marinus]	WP_012844066	80.70%	318.93	2.05E-102	181
E11	hypothetical protein DCC55_22245 [Chloroflexi bacterium]	RIK38088	35.50%	55.07	8.84E-06	93
E12	hypothetical protein AMS25_04205 [Gemmatimonas sp. SM23_52]	KPK81992	66.30%	211.85	3.37E-64	175
F06	Type I restriction-modification system, DNA-methyltransferase subunit M [Anaerolineae bacterium]	RCK72530	81.10%	320.09	1.33E-104	206
F07	amino acid ABC transporter ATP-binding protein, partial [Candidatus Thermofonsia Clade 2 bacterium]	PJF24576	78.40%	171.79	6.94E-51	102
F08	hypothetical protein DCC55_40445 [Chloroflexi bacterium]	RIK25727	60.50%	61.62	4.31E-08	43
F11	radical SAM protein [Nonomuraea candida]	WP_043620474	66.70%	141.35	7.31E-36	99
F12	hypothetical protein ABS36_06335 [Acidobacteria bacterium SCN 69-37]	ODS56047	35.60%	63.54	1.05E-08	90
G05	methylmalonyl-CoA mutase [Gemmatimonas sp. SG8_23]	KPJ77104	81.00%	337.42	1.73E-111	211
G06	hypothetical protein COB20_15440 [SAR86 cluster bacterium]	PCI74174	66.50%	274.63	1.36E-87	185
G07	S24 family peptidase [Anderseniella sp. Alg231-50]	WP_108882341	50.00%	70.86	4.31E-11	64
G08	hypothetical protein [Cellulomonas flavigena]	WP_048771985	54.10%	131.72	1.32E-31	109
G09	oxidoreductase [SAR202 cluster bacterium Ae2-Chloro-G2]	PKB60861	52.70%	203.76	1.14E-57	203
G10	cupin [Leisingera methylohalidivorans DSM 14336]	AHD02121	62.80%	121.32	4.24E-33	86
H05	hypothetical protein [Phyllobacterium sp. YR531]	WP_008123265	50.80%	58.15	2.30E-08	61

A15: Part of the Widb.csv table produced by dRep detailing the dereplicated genomes from the PE150_plus_Nano assembly.

genome	score	completeness	contamination	strain_heterogeneity	size
bin.47.fa.metabat2.fasta	96.4943254	94.07	0	0	1597125
bin.76.fa.metabat2.fasta	81.3976366	95.44	3.3	0	5787251
bin.1.fa.maxbin2.fasta	61.4803018	96.54	7.69	10	6237295
bin.18.fa.metabat2.fasta	76.1064968	95.44	4.4	0	5117247
bin.44.fa.maxbin2.fasta	67.0526264	89.56	4.95	0	6165426
bin.30.fa.maxbin2.fasta	85.1474663	94.99	2.5	0	4860969
bin.45.fa.metabat2.fasta	53.5615816	81.8	6.49	36.36	3591616
bin.72.fa.metabat2.fasta	85.2519079	95.44	2.55	0	5226415
cluster_DBSCAN_round3_2.fasta	91.8653161	91.88	0.43	0	3472899
bin.12.fa.maxbin2.fasta	49.3283351	98.18	11.36	47.06	7067083
bin.16.fa.metabat2.fasta	62.1695268	87.56	5.65	17.65	3233210
cluster_DBSCAN_round6_7.fasta	77.337828	86.98	2.41	16.67	3356316
bin.14.fa.metabat2.fasta	87.4414816	95.23	1.99	0	3166245
bin.13.fa.maxbin2.fasta	54.1664903	75.07	4.77	21.43	2885972
cluster_DBSCAN_round611_9.fasta	91.8467035	90	0.08	100	4330720
cluster_DBSCAN_round1_1.fasta	86.4335741	85.45	0.23	0	2960969

cluster_DBSCAN_round7_9.fasta	91.1121106	93.38	0.9	0	5856010
cluster_DBSCAN_round79_5.fasta	75.8933625	79.68	1.19	0	2440119
bin.37.fa.metabat2.fasta	-2.9208835	85.2	18.19	3.66	5179247
bin.50.fa.metabat2.fasta	92.5113581	97.15	1.47	25	2526738
bin.4.fa.metabat2.fasta	1.38928687	83.07	16.69	0	2793746
bin.78.fa.metabat2.fasta	74.1921396	93.73	4.53	22.22	4591044
bin.66.fa.metabat2.fasta	74.5264854	95.6	4.72	0	4015921
bin.14.fa.maxbin2.fasta	62.7465626	100	7.98	3.45	4760373
bin.73.fa.metabat2.fasta	92.4854647	95.54	1.1	0	5389559
cluster_DBSCAN_round14_0.fasta	84.4709031	81.25	0	0	3309883
bin.6.fa.metabat2.fasta	83.421599	90.16	1.76	0	4726161
bin.71.fa.metabat2.fasta	27.8556292	83.01	11.4	0	3328995
bin.27.fa.metabat2.fasta	91.4875899	97.44	1.71	0	4375664
bin.70.fa.metabat2.fasta	88.6769999	93.77	1.6	42.86	5479390
cluster_DBSCAN_round1_2.fasta	93.9086454	94.09	0.45	0	3636941
bin.17.fa.metabat2.fasta	87.1234957	94.61	2.09	33.33	4006340
bin.10.fa.metabat2.fasta	94.3413756	96.7	0.99	0	3266272
bin.2.fa.metabat2.fasta	72.6972585	85.86	3.01	0	1732882
bin.9.fa.metabat2.fasta	75.0957971	82.15	1.88	0	3736075
bin.80.fa.maxbin2.fasta	-9.7882341	77.13	18.58	22.73	4372242
bin.57.fa.metabat2.fasta	74.9663978	78.32	1.16	50	2502743
bin.33.fa.metabat2.fasta	84.8183829	87.64	0.99	0	2846939
bin.21.fa.metabat2.fasta	64.076347	91.75	5.94	0	4134361
cluster_DBSCAN_round2_1.fasta	88.4769636	90.56	0.99	0	3000250
bin.74.fa.metabat2.fasta	75.4856979	91.51	3.7	0	3506183
bin.44.fa.metabat2.fasta	64.4781079	87.15	4.95	0	3413008
bin.43.fa.metabat2.fasta	77.9325692	76.23	0.19	0	2256917
bin.21.fa.maxbin2.fasta	79.192398	83.95	1.48	0	3389051
bin.68.fa.metabat2.fasta	84.1782082	86.37	0.89	0	1965016
bin.10.fa.maxbin2.fasta	97.0486136	95.31	0.07	100	1463548

A16: Metadata of the five sponges investigated, type and provider of sequencing as well as number of reads (identified by FastQC Version 0.11.5 with default settings) and number of bases returned (in Gbp).

Sample	**Metadata**	**Date sequenced**	**Sequencing technology (provider)**	**Type of sequencing (library)**	**Total sequences (FastQC)**	**Gbp returned**
C. mycofijiensis (783 in Figure 20)	Collected from Cathedral Cave, 'Eua, Tonga on 07/06/2016	11/09/2017	Illumina HiSeq 4000 (Annorad) + Oxford Nanopore MinION (in house, 2.14)	2x150 bp paired-end (TruSeq, PCR free) + Nanopore long read	35,262,051 + 19,486	>10.58
CS200 (834 in Figure 20)	Collected from Pete's cave, 'Eua, Tonga on 08/06/2016	01/05/2018	Illumina HiSeq 4000 (Annorad)	2x150 bp paired-end (TruSeq, PCR free)	97,472,750	29.24
CS202 (837 in Figure 20)	Collected from Pete's cave, 'Eua, Tonga on 08/06/2016	01/05/2018	Illumina HiSeq 4000 (Annorad)	2x150 bp paired-end (TruSeq, PCR free)	74,789,106	22.44
CS203 (839 in Figure 20)	Collected from Pete's cave, 'Eua, Tonga on 08/06/2016	01/05/2018	Illumina HiSeq 4000 (Annorad)	2x150 bp paired-end (TruSeq, PCR free)	81,926,823	24.58
CS204 (841 in Figure 20)	Collected from Pete's cave, 'Eua, Tonga on 08/06/2016	01/05/2018	Illumina HiSeq 4000 (Annorad)	2x150 bp paired-end (TruSeq, PCR free)	68,468,774	20.54
CS211 (854 in Figure 20)	Collected from Pete's cave, 'Eua, Tonga on 09/06/2016	01/05/2018	Illumina HiSeq 4000 (Annorad)	2x150 bp paired-end (TruSeq, PCR free)	93,447,634	28.04

A17: Part of the Widb.csv table produced by dRep detailing the dereplicated genomes from the CS200 assembly.

genome	score	completeness	contamination	strain_heterogeneity	size
bin.25.fa.maxbin2.fa.fasta	81.0243082	81.79	0.79	0	3497278
bin.3.fa.maxbin2.fa.fasta	88.5215754	97.58	2.25	0	5597713
bin.69.fa.metabat2.fasta	83.8390635	93.89	2.5	0	4753388
bin.45.fa.metabat2.fasta	86.5365947	96.64	2.5	0	5889702

bin.8.fa.metabat2.fasta	62.952496	93.24	6.59	0	5205666
bin.44.fa.metabat2.fasta	82.7541006	96.54	3.3	0	5743470
bin.15.fa.metabat2.fasta	62.8391141	88.89	5.62	0	5222334
bin.59.fa.metabat2.fasta	55.5653646	95.44	8.79	20	5984513
bin.6.fa.metabat2.fasta	74.5124061	89.23	3.5	14.29	5939582
bin.106.fa.metabat2.fasta	97.2124654	94.44	0	0	5486947
bin.74.fa.metabat2.fasta	73.1271855	79.23	1.84	66.67	1802288
bin.11.fa.maxbin2.fa.fasta	100.881567	97.8	0	0	3792225
cluster_DBSCAN_round3_12.fasta	97.6979733	95.45	0	0	4275371
cluster_DBSCAN_round3_0.fasta	98.5566476	96.7	0.3	0	5887900
bin.33.fa.metabat2.fasta	25.8998046	97.19	17.03	65	6788020
bin.111.fa.metabat2.fasta	31.8930458	75.81	9.5	18.18	4945995
cluster_DBSCAN_round4_12.fasta	75.9780169	83.77	1.98	0	2417334
bin.73.fa.metabat2.fasta	65.7757657	82.97	3.85	0	3195058
bin.52.fa.maxbin2.fa.fasta	58.5954089	85.65	6.04	16.67	3336958
bin.23.fa.metabat2.fasta	92.5504815	95.42	1.13	0	2836076
bin.84.fa.metabat2.fasta	75.6776416	86.91	2.82	9.09	4915411
bin.41.fa.maxbin2.fa.fasta	73.2228437	86.65	3.15	0	4679005
bin.50.fa.metabat2.fasta	66.9075997	87.3	4.69	20	4023997
cluster_DBSCAN_round652_9.fasta	72.687291	82.69	2.53	16.67	3200754
cluster_DBSCAN_round7_66.fasta	42.8123515	83.03	9.21	44.44	2691063
bin.20.fa.maxbin2.fa.fasta	83.6565556	94.02	2.56	0	3924071
bin.25.fa.metabat2.fasta	39.4797016	83.32	9.94	39.02	3055828
bin.7.fa.metabat2.fasta	95.2426054	99.15	1.28	0	2523133
cluster_DBSCAN_round651_23.fasta	72.896757	77.56	1.47	28.57	3198834
bin.109.fa.metabat2.fasta	97.20053	95.75	0.17	0	4136733
bin.23.fa.maxbin2.fa.fasta	93.9958245	95.75	0.84	0	1665192
bin.63.fa.metabat2.fasta	81.7241135	96.05	3.41	0	4777659
bin.21.fa.metabat2.fasta	74.1889174	94.49	4.77	25	2585210
bin.28.fa.metabat2.fasta	84.5251139	92.31	2.14	33.33	3523362
bin.64.fa.metabat2.fasta	75.3714144	84.9	2.42	25	2560203
bin.79.fa.metabat2.fasta	83.742503	88.03	1.36	33.33	3150644
bin.19.fa.maxbin2.fa.fasta	89.1088308	97.44	2.14	0	3119526
cluster_DBSCAN_round3_9.fasta	85.077933	93.47	2.14	0	3649547
bin.65.fa.metabat2.fasta	92.3213117	94.23	0.85	0	6171084
cluster_DBSCAN_round650_12.fasta	66.6120472	75.86	2.59	50	2657375
cluster_DBSCAN_round7_19.fasta	67.2852796	79.02	2.75	0	4558018
bin.107.fa.metabat2.fasta	59.1066065	76.86	4.15	28.57	4066304
cluster_DBSCAN_round6_10.fasta	87.2533947	90.32	1.08	0	2774845
cluster_DBSCAN_round7_20.fasta	89.8353785	88.03	0.07	0	4402453
bin.61.fa.metabat2.fasta	63.4724016	92.88	6.62	23.08	5221113
cluster_DBSCAN_round45_17.fasta	37.4386052	88.48	10.69	2.78	4595405

bin.99.fa.metabat2.fasta	69.5204672	91.94	4.94	0	3294144
bin.15.fa.maxbin2.fa.fasta	85.7840917	90.12	1.31	0	3551145
bin.16.fa.metabat2.fasta	81.1073329	90.81	2.56	33.33	6081149
bin.110.fa.metabat2.fasta	81.9394189	84.66	0.93	0	1976679
cluster_DBSCAN_round646_3.fasta	81.6343254	88.12	1.98	50	2870490
bin.88.fa.metabat2.fasta	75.4855801	83.23	1.98	0	2695444
bin.87.fa.metabat2.fasta	93.9676626	91.82	0	0	3715394
bin.9.fa.metabat2.fasta	87.0170162	98.18	2.73	0	5418736
cluster_DBSCAN_round648_11.fasta	74.4444933	81.7	1.88	0	4108008
bin.14.fa.metabat2.fasta	61.2624912	85.47	5.35	10	5772692
bin.91.fa.metabat2.fasta	90.9255587	88.94	0.09	0	3490796
bin.80.fa.metabat2.fasta	75.146415	89.44	3.3	0	3238367
bin.89.fa.metabat2.fasta	86.7923645	89.09	0.91	0	3030342
bin.58.fa.metabat2.fasta	95.1444427	92.73	0	0	4516423
bin.1.fa.metabat2.fasta	75.0456484	77.58	0.99	50	3612358
cluster_DBSCAN_round5_13.fasta	79.7475621	91.06	2.92	40	2365421
bin.39.fa.maxbin2.fa.fasta	57.5685257	76.36	4.89	77.27	2518982
cluster_DBSCAN_round3_14.fasta	89.2743222	87.79	0.09	0	1531290
bin.78.fa.metabat2.fasta	78.7302353	94.61	3.96	50	2416037
bin.31.fa.metabat2.fasta	82.5609211	84.82	0.99	0	3233497
bin.55.fa.metabat2.fasta	67.4414549	81.49	3.5	42.86	3509502
bin.77.fa.metabat2.fasta	58.7879086	76.17	4.13	33.33	1462641
bin.12.fa.metabat2.fasta	86.1270658	89.57	1.08	0	3166300
cluster_DBSCAN_round5_6.fasta	97.0702573	94.61	0	0	3826423
bin.52.fa.metabat2.fasta	84.1993849	90.15	1.98	100	2963239
bin.10.fa.metabat2.fasta	74.8897594	81.96	1.98	50	2077380
bin.45.fa.maxbin2.fa.fasta	92.234203	94.72	0.99	0	2927243
bin.32.fa.metabat2.fasta	87.8547261	90.56	0.99	0	3809133
bin.115.fa.metabat2.fasta	21.370188	79.19	12.04	4.55	1774875
bin.67.fa.metabat2.fasta	95.0596523	92.56	0	0	1466004

A18: Part of the Widb.csv table produced by dRep detailing the dereplicated genomes from the CS202 assembly.

genome	score	completeness	contamination	strain_heterogeneity	size
bin.82.fa.maxbin22.fasta	78.7311624	92.73	3.26	14.29	4971293
bin.52.fa.metabat2.fasta	71.8476118	97.58	5.62	0	5851677
bin.59.fa.metabat2.fasta	70.2319241	95.34	5.49	0	5132469
bin.35.fa.metabat2.fasta	87.1454402	97.19	2.5	0	5127097
bin.13.fa.maxbin22.fasta	64.0345518	84	4.95	60	2945693
cluster_DBSCAN_round36_5.fasta	27.798952	80.22	12.91	78.12	3075533
bin.54.fa.metabat2.fasta	74.224268	75.22	0.73	50	2666495

cluster_DBSCAN_round3_28.fasta	28.665704	76.65	11.5	67.27	2941405
cluster_DBSCAN_round3_8.fasta	83.5950423	86.14	0.99	0	2724608
bin.10.fa.maxbin22.fasta	21.5111117	85.32	15.4	73.91	2530345
bin.74.fa.metabat2.fasta	77.6262779	97.8	4.5	0	4206105
bin.80.fa.metabat2.fasta	57.6484352	80.38	5.56	57.14	3038571
bin.67.fa.metabat2.fasta	49.459674	75.26	5.56	0	2794795
bin.87.fa.metabat2.fasta	83.8464113	90.6	2.09	75	2752052
bin.3.fa.metabat2.fasta	70.2506178	88.89	4.78	66.67	3341797
bin.84.fa.metabat2.fasta	86.5522291	91.03	1.36	0	3008757
bin.35.fa.maxbin22.fasta	44.2258026	79.79	8.81	76.92	2135046
bin.3.fa.maxbin22.fasta	78.4120011	86.48	2.08	0	5283936
cluster_DBSCAN_round4_9.fasta	81.6195222	84.04	0.94	0	4613590
bin.64.fa.metabat2.fasta	60.8361507	84.48	5.17	4.35	3931765
bin.41.fa.metabat2.fasta	85.5033228	91.63	1.74	25	3914377
bin.61.fa.metabat2.fasta	65.8015027	79.59	3.23	12.5	2973090
cluster_DBSCAN_round2_8.fasta	84.3207981	86.82	1	33.33	3732548
bin.71.fa.metabat2.fasta	90.1354788	92.74	0.99	0	2869548
bin.70.fa.metabat2.fasta	93.4306942	90.9	0	0	3294646
bin.79.fa.metabat2.fasta	81.566323	87.91	1.65	0	4080330
bin.6.fa.maxbin22.fasta	86.5632891	92.63	1.71	0	5152344
bin.78.fa.metabat2.fasta	74.1192468	91.24	3.85	0	5370506
bin.76.fa.metabat2.fasta	83.4210264	95.11	2.79	0	5832935
bin.43.fa.metabat2.fasta	-8.7537564	86.21	20.09	18.31	4021477
bin.85.fa.metabat2.fasta	95.7759904	97.44	0.85	0	4651935
cluster_DBSCAN_round3_2.fasta	95.7342767	93.16	0	0	4614246
bin.7.fa.metabat2.fasta	76.1723444	94.84	4.25	0	4823908
bin.19.fa.metabat2.fasta	90.479417	96.59	1.68	0	1859783
bin.33.fa.metabat2.fasta	79.9405044	96.82	3.86	0	2328647
bin.40.fa.metabat2.fasta	57.6661382	92.98	7.79	20	4676036
bin.103.fa.maxbin22.fasta	70.3030946	84.95	3.88	75	1608114
bin.77.fa.metabat2.fasta	65.0556493	94.01	6.27	4.35	4028374
bin.90.fa.maxbin22.fasta	60.1479175	75.4	4.02	77.78	2103784
bin.56.fa.metabat2.fasta	91.8141058	94.88	1.13	0	2741796
bin.8.fa.maxbin22.fasta	100.48063	97.8	0	0	3677345
bin.15.fa.maxbin22.fasta	65.3002191	75.56	2.5	20	3647960
cluster_DBSCAN_round1_1.fasta	96.4895203	95.6	0.3	0	5074705
cluster_DBSCAN_round3_15.fasta	97.8017302	95.54	0	0	4052258
bin.17.fa.metabat2.fasta	92.7152283	90.51	0	0	4077588
cluster_DBSCAN_round29_0.fasta	91.4888111	88.99	0	0	3465248
cluster_DBSCAN_round3_13.fasta	78.8077095	91.31	2.96	0	3807563
bin.30.fa.metabat2.fasta	78.1154791	82.37	1.42	50	1520947
cluster_DBSCAN_round2_2.fasta	96.0115784	93.64	0	0	4210483

bin.16.fa.metabat2.fasta	79.2543266	85.47	1.9	62.5	1745419
cluster_DBSCAN_round28_4.fasta	78.6028673	92.95	3.45	25	3790304
cluster_DBSCAN_round29_11.fasta	67.584156	75.68	1.98	0	1905597
bin.4.fa.metabat2.fasta	94.0769197	96.7	0.99	0	3338076
cluster_DBSCAN_round4_8.fasta	97.3690399	95.6	0.11	0	2773482
bin.26.fa.metabat2.fasta	77.1646493	89.49	3.16	50	3521600

A19: Part of the Widb.csv table produced by dRep detailing the dereplicated genomes from the CS203 assembly.

genome	score	completeness	contamination	strain_heterogeneity	size
bin.63.fa.metabat2.fasta	88.0589303	96.54	2.2	0	5250144
bin.94.fa.metabat2.fasta	82.5079303	96.44	3.3	0	5570320
cluster_DBSCAN_round1_20.fasta	81.6714563	91.49	2.52	40	1757438
bin.93.fa.metabat2.fasta	88.0578072	92.24	1.3	0	5678024
bin.36.fa.metabat2.fasta	81.625935	85.1	1.26	66.67	4477596
bin.91.fa.maxbin2.fasta	-34.253872	76.22	22.54	2.22	4516069
bin.104.fa.metabat2.fasta	84.8599878	94.99	2.5	0	4711288
bin.105.fa.metabat2.fasta	96.3071956	94.07	0	0	3645710
bin.52.fa.maxbin2.fasta	71.5339988	93.85	4.85	0	2390003
bin.23.fa.maxbin2.fasta	93.9618675	95.75	0.84	0	1768037
bin.20.fa.metabat2.fasta	80.9032607	96.82	3.86	20	2599401
bin.99.fa.metabat2.fasta	95.3643323	96.7	0.99	100	3379867
bin.48.fa.metabat2.fasta	42.7687618	83.23	8.99	27.27	5651773
bin.20.fa.maxbin2.fasta	34.1656413	78.21	9.97	40	4791268
bin.23.fa.metabat2.fasta	77.8207883	80.32	0.97	0	3632427
bin.72.fa.metabat2.fasta	68.9274434	88.46	4.48	10	3524367
bin.7.fa.maxbin2.fasta	83.1913962	85.04	0.85	50	2057102
bin.17.fa.metabat2.fasta	81.412537	88.54	1.98	28.57	2901115
bin.21.fa.metabat2.fasta	66.1387432	81.67	3.85	50	1801320
bin.50.fa.maxbin2.fasta	86.0973869	94.44	2.14	0	3585368
bin.110.fa.metabat2.fasta	87.742888	90.91	1.14	50	4305062
bin.60.fa.metabat2.fasta	88.9528645	95.83	1.98	50	3492868
bin.98.fa.metabat2.fasta	85.2370614	82.64	0	0	3417684
cluster_DBSCAN_round715_7.fasta	81.0000835	81.53	0.51	0	2359256
cluster_DBSCAN_round1_12.fasta	85.3372959	93.04	2.2	50	4086343
bin.38.fa.metabat2.fasta	83.3991567	97.27	3.64	50	5331455
bin.101.fa.metabat2.fasta	93.7945385	95.21	0.91	100	5311215
cluster_DBSCAN_round1_8.fasta	77.4454206	85.89	2.35	42.86	4949258
bin.10.fa.maxbin2.fasta	82.2008606	92.47	2.58	14.29	3688913
bin.106.fa.metabat2.fasta	87.5416456	89.68	1.1	100	3681039
bin.50.fa.metabat2.fasta	88.5703852	90.76	0.99	0	3175706

bin.31.fa.metabat2.fasta	82.2698081	84.82	0.99	0	3403073
cluster_DBSCAN_round1_7.fasta	85.9261729	88.45	0.99	0	2776425
bin.109.fa.maxbin2.fasta	-25.170427	87.72	23.08	3.53	4766829
bin.91.fa.metabat2.fasta	71.3717803	79.38	2.07	22.22	4312780
bin.102.fa.metabat2.fasta	79.5647244	94.23	3.42	0	6113300
bin.80.fa.metabat2.fasta	96.5718055	98.29	0.85	0	4564175
bin.3.fa.metabat2.fasta	93.1049872	90.76	0	0	3316607
cluster_DBSCAN_round1_3.fasta	91.7910953	95.73	1.28	0	3864809
bin.4.fa.metabat2.fasta	90.4073234	91.82	0.91	100	3120456
bin.41.fa.metabat2.fasta	84.9150136	88.89	1.28	0	3901073
bin.15.fa.metabat2.fasta	78.2630175	91.14	2.99	0	3375576
bin.68.fa.metabat2.fasta	73.4420691	86.32	2.99	0	3069504
bin.35.fa.metabat2.fasta	8.9558501	89.01	16.48	0	4425598
bin.14.fa.metabat2.fasta	55.702944	97.25	8.79	0	4381579
cluster_DBSCAN_round690_111.fasta	77.268285	80.28	1.1	0	2573309
bin.2.fa.maxbin2.fasta	80.9013754	96.58	3.59	0	5904432
bin.30.fa.metabat2.fasta	82.7539785	89.74	1.88	0	4502986
bin.10.fa.metabat2.fasta	31.2259564	75.86	9.56	14.29	4388929
bin.52.fa.metabat2.fasta	77.103371	96.37	4.34	0	6189831
bin.7.fa.metabat2.fasta	-6.9945105	83.15	18.47	2.17	5990961
bin.70.fa.metabat2.fasta	68.9689717	83.12	3.36	25	3848662

A20: Part of the Widb.csv table produced by dRep detailing the dereplicated genomes from the CS204 assembly.

Genome	score	completeness	contamination	strain_heterogeneity	size
bin.103.fa.metabat2.fasta	96.6139692	94.72	0.11	100	3411506
bin.36.fa.metabat2.fasta	95.0339478	94.82	0.49	100	2094037
bin.42.fa.metabat2.fasta	92.7157573	90.78	0	0	1501255
bin.25.fa.metabat2.fasta	70.5079459	94.59	5.72	42.86	5356430
bin.98.fa.metabat2.fasta	74.5244422	93.49	4.4	20	5284590
bin.10.fa.maxbin2.fasta	60.5554337	96.44	7.69	0	6020746
bin.3.fa.maxbin2.fasta	66.9138238	96.54	6.59	11.11	6399624
bin.96.fa.maxbin2.fasta	28.8876355	91.24	12.94	4.35	5638569
bin.81.fa.metabat2.fasta	96.5231342	98.29	0.85	0	4542703
bin.90.fa.metabat2.fasta	77.7796802	76.9	0.31	0	4313417
cluster_DBSCAN_round6_6.fasta	88.3617532	86.42	0.08	0	2980254
bin.33.fa.metabat2.fasta	67.9294189	78.25	2.45	0	3780727
bin.30.fa.metabat2.fasta	93.4815164	94.07	0.56	0	3519701
bin.36.fa.maxbin2.fasta	65.2934207	84.98	4.44	3.33	4623260
bin.58.fa.metabat2.fasta	62.8690089	75.73	3.38	58.33	3327082
bin.49.fa.metabat2.fasta	70.6590262	79.26	2.31	36.36	3515777

cluster_DBSCAN_round44_11.fasta	74.2771856	80	1.62	0	3005023
cluster_DBSCAN_round840_2.fasta	46.2575251	82.07	7.61	4.35	3356006
bin.101.fa.metabat2.fasta	90.2250642	94.02	1.28	0	3520880
cluster_DBSCAN_round6_12.fasta	72.0375605	82.91	2.71	33.33	1763111
cluster_DBSCAN_round4_1.fasta	92.5547909	96.58	1.28	0	3750153
bin.76.fa.metabat2.fasta	47.856894	94.87	9.83	0	4218307
bin.95.fa.metabat2.fasta	39.5856739	82.48	9.05	4.76	3620918
bin.72.fa.metabat2.fasta	99.2169051	97.27	0.08	100	4290307
bin.26.fa.metabat2.fasta	88.2339124	98	2.73	50	5155438
bin.85.fa.metabat2.fasta	5.53305881	93.16	18.4	12.86	4192481
cluster_DBSCAN_round754_100.fasta	70.1271011	81.99	3.26	80	3450082
cluster_DBSCAN_round5_11.fasta	75.5568348	83.18	2.26	80	2597440
bin.30.fa.maxbin2.fasta	94.5921277	91.82	0	0	4304963
cluster_DBSCAN_round753_58.fasta	55.9827401	78.17	5.31	44.44	5223360
bin.105.fa.metabat2.fasta	20.2332668	92.6	15.01	3.03	7472410
bin.39.fa.metabat2.fasta	91.4051807	93.73	0.99	0	2969028
cluster_DBSCAN_round7_7.fasta	85.5901207	88.46	1.03	0	3450360
bin.97.fa.metabat2.fasta	76.1427609	87.61	2.78	16.67	4605413
bin.19.fa.maxbin2.fasta	79.7103857	93.96	3.3	0	3416065
bin.71.fa.metabat2.fasta	58.3682738	77.04	4.12	0	6113256
cluster_DBSCAN_round5_5.fasta	85.813688	88.45	0.99	0	2782311
cluster_DBSCAN_round2_1.fasta	76.0629394	87.8	2.99	33.33	4900063
bin.16.fa.metabat2.fasta	30.120313	91.9	13.09	13.04	5005461
bin.8.fa.metabat2.fasta	75.8907359	95.6	4.4	0	4153772
bin.38.fa.metabat2.fasta	73.6467701	97.8	5.49	20	4229717
bin.57.fa.maxbin2.fasta	83.7540616	97.8	3.38	7.14	4107734
cluster_DBSCAN_round6_5.fasta	79.123189	95.91	3.86	0	2388948
bin.17.fa.metabat2.fasta	99.1511997	96.59	0	0	3269809
cluster_DBSCAN_round754_308.fasta	87.126897	95.6	2.48	66.67	3696902
cluster_DBSCAN_round4_2.fasta	84.9796395	87.79	0.99	0	2144364
bin.84.fa.metabat2.fasta	83.0907262	90.96	2.2	50	4850361
bin.85.fa.maxbin2.fasta	-29.483108	82.18	24.26	32.56	4203088
bin.66.fa.metabat2.fasta	77.3062137	80.49	1.15	14.29	1683023
bin.89.fa.metabat2.fasta	79.504024	86.31	1.79	0	1637260

A21: Part of the Widb.csv table produced by dRep detailing the dereplicated genomes from the CS211 assembly.

genome	score	completeness	contamination	strain_heterogeneity	size
bin.20.fa.metabat2.fasta	46.3709592	82.25	8.17	35.71	5925520
bin.49.fa.metabat2.fasta	42.1356474	77.46	7.56	6.25	4035903
bin.3.fa.metabat2.fasta	63.2329103	94.99	6.79	0	4532775

cluster_DBSCAN_round4_13.fasta	75.4531687	80.67	1.5	33.33	3515739
bin.23.fa.maxbin2.fasta	88.6544255	86.57	0	0	1679240
bin.27.fa.maxbin2.fasta	55.6878885	90.45	7.77	27.27	2311693
bin.109.fa.maxbin2.fasta	-19.17718	81.58	22.35	41.18	3088071
bin.45.fa.metabat2.fasta	88.3866086	90.76	0.99	0	2970852
bin.100.fa.metabat2.fasta	91.1299456	89.18	0.08	100	4637810
bin.90.fa.metabat2.fasta	55.7842904	77.09	5.1	44.44	4295205
bin.62.fa.metabat2.fasta	57.6797718	91.45	7.26	5.13	4037374
cluster_DBSCAN_round1_2.fasta	80.5918518	90	2.88	100	3806772
cluster_DBSCAN_round3_1.fasta	93.4428757	90.91	0	0	3042972
bin.31.fa.metabat2.fasta	65.8808614	96.37	6.59	0	5139188
bin.34.fa.metabat2.fasta	77.8884302	92.14	3.3	0	5567835
cluster_DBSCAN_round798_24.fasta	83.7209789	88.03	1.28	0	2852525
bin.12.fa.metabat2.fasta	87.0903479	89.09	0.91	50	3450881
bin.102.fa.metabat2.fasta	93.0693342	90.91	0	0	4084340
bin.6.fa.maxbin2.fasta	89.2920065	96.17	2.1	66.67	1721180
bin.46.fa.metabat2.fasta	76.5646584	90.42	3.3	14.29	4197716
bin.5.fa.metabat2.fasta	98.2782129	96.04	0	0	2461597
cluster_DBSCAN_round800_189.fasta	95.1175619	92.63	0	0	3733453
bin.9.fa.metabat2.fasta	96.6162827	94.32	0	0	2123357
bin.69.fa.metabat2.fasta	87.629564	97.69	2.48	0	2814527
cluster_DBSCAN_round1_1.fasta	86.4381432	84.13	0	0	3229949
bin.108.fa.metabat2.fasta	29.039157	86.51	12.18	12.5	5508138
cluster_DBSCAN_round796_3.fasta	72.8665069	81.98	2.39	18.18	4923445
cluster_DBSCAN_round799_88.fasta	77.0084585	80.22	1.1	0	2681686
bin.4.fa.maxbin2.fasta	79.2531572	93.41	3.3	0	4610359
bin.41.fa.metabat2.fasta	83.5118686	97.8	3.3	0	4149530
bin.13.fa.metabat2.fasta	98.0658446	95.71	0	0	3902868
bin.6.fa.metabat2.fasta	83.3415487	89.2	1.85	50	3187767
bin.25.fa.metabat2.fasta	75.8893702	85.74	2.45	0	5381482
bin.80.fa.metabat2.fasta	35.1888745	89.47	11.4	7.14	3903901
bin.19.fa.maxbin2.fasta	82.2392418	93.73	2.74	0	4302236
bin.38.fa.metabat2.fasta	69.3319223	88.49	4.29	0	5553122
bin.1.fa.metabat2.fasta	59.9060594	81.75	4.99	23.81	3312489
bin.47.fa.metabat2.fasta	97.2122971	97.51	0.5	0	4601540
bin.39.fa.metabat2.fasta	71.0454322	84.49	3.4	47.06	3674734
bin.19.fa.metabat2.fasta	72.6565183	78.14	1.59	0	2627513
bin.44.fa.metabat2.fasta	85.280006	91.1	1.69	16.67	3100103
bin.50.fa.metabat2.fasta	85.3246314	88.19	0.99	0	1853071
bin.60.fa.metabat2.fasta	59.2975833	84.91	5.79	22.22	5417576
bin.7.fa.metabat2.fasta	65.7933833	91.81	5.98	25	5672129
bin.21.fa.maxbin2.fasta	79.2163811	85.47	1.88	66.67	2293323

cluster_DBSCAN_round4_0.fasta	90.5141982	92.25	0.85	0	5197823
bin.18.fa.maxbin2.fasta	92.2443501	94.44	0.85	0	4735795
bin.103.fa.metabat2.fasta	99.643748	97.49	0	0	3445428

A22: Figure produce by dRep compare as the primary dendrogram with color of writing separating MASH clusters. First number in brackets is the unique primary cluster and the second number the unique bins within the secondary cluster; i.e. if both numbers are the same, the bins are identified as identical at the default 99% ANI level.


MASH clustering
100.0
99.5
99.0
98.5
98.0
MASH Average Nucleotide Identity (ANI)

A23: 16S rRNA raw coverage information of bacterial taxa, summed up per Class (or Phylum if Class could not be identified) for the six Tongan sponges (CS783 = *C. mycofijiensis*), the four New Zealand sponges (s0, s1, s2, s3) and the three Mediterranean (sf, pf, aa). Class was chosen as most 16S rRNA sequences were classified at least at the class level but tables for family and genus level abundances as well as details of the code used to create these tables can be found on the GitHub.

	CS 783	CS 200	CS 202	CS 203	CS 204	CS 211	s0	s1	s2	s3	sf	pf	aa
Planctomycetes-Planctomycetacia	0	0	0	0	0	0	0	34.77	1.60	1.96	0	0	0
Gemmatimonadetes-PAUC43f marine benthic group	0	5.32	0	0	0	0	0	0	0	0	0	1	0
Chloroflexi-JG30-KF-CM66	0	0	0	1.96	0	0	0	0	0	0	0	0	2
PAUC34f-	0	72.00	6.20	0	0	0	0	0	0	0	0	0	1
Acidobacteria-Subgroup 9	0	17.66	6.43	0	0	0	0	0	0	0	0	0	0
Acidobacteria-Subgroup 6	2.01	42.83	3.79	46.99	0.93	12.00	0	0	0	0	0	0	2
Proteobacteria-Alphaproteobacteria	1.82	8.22	7.64	0.77	2.04	12.43	208.49	0	4.36	9.28	2	0	7
Kiritimatiellaeota-Kiritimatiellae	0	0	0	0	0	0	134.49	0.91	0	0	0	0	0
Verrucomicrobia-Verrucomicrobiae	0	2.12	0	0	0	0	0	0	0	0	0	0	0
Nitrospirae-Nitrospira	1.23	0	0	8.70	0	0	40.26	0	0	0	0	0	0
Deinococcus-Thermus-Deinococci	0	0	0	2.32	0	0	0	0	0	0	0	0	0
Patescibacteria-Parcubacteria	0	0	0	0	0	0	0	0	0	0	0	1	1
Bacteroidetes-Bacteroidia	0	0	0	1.12	0	0	54.30	1.80	1.37	2.34	0	1	1
Acidobacteria-Subgroup 11	0	3.32	0	0	0	0	0	0	0	0	0	0	0
Acidobacteria-Acidobacteriia	0	27.38	25.26	1.58	0	0	0	0	0	0	0	1	1
Actinobacteria-Actinobacteria	0	0	0	0	0	0	0	149.58	0	4.87	0	0	0
Spirochaetes-Spirochaetia	2.02	4.91	0	0	0	0	0	0	0	0	0	2	1
Acidobacteria-Subgroup 21	0	0	3.90	0	0	0	0	0	0	0	0	0	0
Bacteroidetes-Ignavibacteria	0	0	0	0	0	0	0	0	0	0	0	1	0
Dadabacteria-Dadabacteriia	3.07	0	0	0	0	0	39.42	0	0	0	0	0	1
Patescibacteria-Saccharimonadia	0	1.57	0	0	0	0	0	0	0	0	0	0	0
Bacteroidetes-Rhodothermia	0	6.51	0	7.92	0	1.15	0	0	0	0	1	1	1

Nitrospinae-P9X2b3D02	0	0	1.54	1.38	0	0	0	0	0	0	0	0	1
Chloroflexi-	0	7.21	0	0	2.06	1.09	0	0	0	0	1	0	1
Acidobacteria-Thermoanaerobaculia	0	0	2.63	11.00	12.25	0	0	0	0	0	0	0	1
Gemmatimonadetes-BD2-11 terrestrial group	3.55	23.19	8.80	21.38	0	31.02	0	0	0	0	1	2	3
Chloroflexi-Dehalococcoidia	2.65	29.87	4.40	9.32	4.07	19.70	0	0	0	0	0	4	4
Acidobacteria-Subgroup 26	0	0	0	0	0	0	24.76	0	0	0	0	0	0
Proteobacteria-Gammaproteobacteria	6.91	94.59	16.72	26.99	12.71	23.88	178.34	13.09	15.70	12.88	6	2	7
Entotheonellaeota-Entotheonellia	0	13.16	0	0	0	0	0	0	0	0	1	0	0
Cyanobacteria-Oxyphotobacteria	0	0	0.89	0	0	0	0	8.46	0	2.15	0	0	1
Proteobacteria-Deltaproteobacteria	0.79	16.83	7.34	1.55	1.64	6.14	0	0	1.00	0	0	2	2
Actinobacteria-Acidimicrobiia	0	33.05	38.93	42.19	11.16	6.94	0	0	3.98	0	1	0	3
Chloroflexi-Anaerolineae	10.43	67.03	13.42	11.01	15.36	111.31	0	0	0	0	2	0	3
Chloroflexi-TK17	0	0	3.46	9.35	11.03	0	0	0	0	0	1	1	1
Poribacteria-	11.36	57.05	94.12	23.24	21.50	7.55	0	0	0	0	1	1	1
Chloroflexi-TK10	1.23	51.70	19.08	0	0	0	0	0	0	0	0	0	1
Chloroflexi-SHA-26	1.36	0	0	0	0	0	0	0	0	0	0	0	0
AncK6-	0	17.54	0	0	0	0	0	0	0	0	0	0	0

A24: RiPP clusters identified from the six Tongan sponges analysed with antiSMASH4.1.0 standalone and reanalysed with antiSMASH5.

Cluster number (Sponge)	**Identified by antiSMASH4.1.0**	**Precursor peptide predicted by antiSMASH4.1.0**	**Identified by antiSMASH5**	**Precursor peptide predicted by antiSMASH5**
31 (*C. mycofijiensis*)	Bacteriocin - Lanthipeptide	No	Bacteriocin - Lanthipeptide	Yes
73 (*C. mycofijiensis*)	Head-to-tail	No	Head-to-tail	No
82 (*C. mycofijiensis*)	Lanthipeptide	Yes	Lanthipeptide	No
138 (*C. mycofijiensis*)	Lassopeptide	No	Lassopeptide	No
40 (CS200)	Lanthipeptide	No	N/a	N/a
59 (CS200)	Lassopeptide	No	Lassopeptide	No
63 (CS200)	Lanthipeptide	Yes	Lanthipeptide	No
104 (CS200)	Lassopeptide	Yes	Lassopeptide	Yes
124 (CS200)	Lanthipeptide	No	Lanthipeptide	No

143 (CS200)	Lanthipeptide	No	Lanthipeptide	Yes
145 (CS200)	Bacteriocin - Lanthipeptide	No	Bacteriocin - Lanthipeptide	Yes
163 (CS200)	Lassopeptide	Yes	Lassopeptide	Yes
87 (CS202)	Lanthipeptide	No	Lanthipeptide	No
93 (CS202)	Lanthipeptide	No	Lanthipeptide	No
108 (CS202)	Microcin	No	Lassopeptide	No
119 (CS202)	Lanthipeptide	Yes	Lanthipeptide	No
153 (CS202)	Lassopeptide	No	Lassopeptide	No
171 (CS202)	Lassopeptide	No	N/a	N/a
36 (CS203)	Bacteriocin - Lanthipeptide	No	Bacteriocin - Lanthipeptide	Yes
136 (CS203)	Lanthipeptide	No	Lanthipeptide	No
137 (CS203)	Thiopeptide	No	Thiopeptide & LAP	No
141 (CS203)	Lassopeptide	No	Lassopeptide	No
160 (CS203)	Lassopeptide	Yes	Lassopeptide	Yes
163 (CS203)	Bacteriocin - Lanthipeptide	No	Bacteriocin - Lanthipeptide	Yes
183 (CS203)	Lassopeptide	No	Lassopeptide	No
186 (CS203)	Lassopeptide	No	Lassopeptide	No
188 (CS203)	Lassopeptide	Yes	Lassopeptide	Yes
197 (CS203)	Lassopeptide	No	Lassopeptide	No
215 (CS203)	Proteusin	No	Proteusin	No
48 (CS204)	Lanthipeptide	No	Lanthipeptide	No
62 (CS204)	Lanthipeptide	No	Lanthipeptide	No
79 (CS204)	Lassopeptide	No	Lassopeptide	No
97 (CS204)	Lassopeptide	Yes	Lassopeptide	Yes
118 (CS204)	Lassopeptide	No	Lassopeptide	No
146 (CS204)	Proteusin	No	LAP & Proteusin	No
159 (CS204)	Lassopeptide	No	Lassopeptide	Yes
182 (CS204)	Lassopeptide	No	Lassopeptide	No
206 (CS204)	Lassopeptide	No	Lassopeptide	No
25 (CS211)	Lanthipeptide	Yes	Lanthipeptide	No
60 (CS211)	Lassopeptide	No	Lassopeptide	No
127 (CS211)	Thiopeptide -Bactericion	No	Thiopeptide - Bacteriocin	No
153 (CS211)	Bacteriocin - Proteusin	No	Bacteriocin - Proteusin	No

A25: Summary of BGCs and MAGs identified per marker lineage across all six Tongan sponges. Note that GTDB taxonomy is given as the highest level that differentiates the marker lineages identified here.

Marker lineage	GTDB taxonomy	#BGCs	#MAGs	BGCs/MAG
o__Rhodospirillales (UID3754)	Acetobacteraceae	30	8	3.75
k__Bacteria (UID3187)	Acidobacteria	310	56	5.54
k__Bacteria (UID2982)	Verrucomicrobia	0	1	0
k__Bacteria (UID2570)	Bacteriodetes; Bacteroidia	14	4	3.5
k__Bacteria (UID2566)	Bacteriodetes; Flavobacteria	2	1	2
k__Bacteria (UID2565)	Planctomycetota	6	1	6
k__Bacteria (UID2495)	Proteobacteria; Gammproteobacteria; Enterobacterales	202	70	2.89
k__Bacteria (UID2142)	Deinococcota	3	2	1.5
k__Bacteria (UID203)	Candidate Phylum “Patescibacteria”; Microgenomatia	57	15	3.80
k__Bacteria (UID1453)	Candidate Phylum “Patescibacteria”; Paceibacteria	71	29	2.45
k__Bacteria (UID1452)	Candidate Phylum “Patescibacteria”; ABY1	175	90	1.94
c__Spirochaetia (UID2496)	Spirochaetales	5	1	5
c__Gammaproteobacteria (UID4443)	SAR86; D2472; SCGC-AAA076-P13	83	24	3.46
c__Gammaproteobacteria (UID4274)	Chromatiales	4	1	4
c__Gammaproteobacteria (UID4267)	Competibacterales	44	7	6.29
c__Gammaproteobacteria (UID4202)	Xanthomonadales	5	1	5
c__Gammaproteobacteria (UID4201)	SAR86 - D2472 - D2472	3	1	3
c__Deltaproteobacteria (UID3216)	Desulfobacterales	4	1	4
c__Alphaproteobacteria (UID3305)	Micavibrionales	35	6	5.83

A26: Raw counts for Figure 28 showing the number of times a BGC from two sponges was identified in the same BiG-SCAPE family at the default cutoff of 0.3.

Sponge_1	Sponge_2	Count
CS200	CS200	3
CS200	CS202	28
CS200	CS203	29
CS200	CS204	29
CS200	CS211	37
CS200	CS783	31
CS202	CS202	1
CS202	CS203	22
CS202	CS204	18
CS202	CS211	20
CS202	CS783	18
CS203	CS203	4
CS203	CS204	39
CS203	CS211	25
CS203	CS783	22
CS204	CS204	3
CS204	CS211	23
CS204	CS783	25
CS211	CS211	11
CS211	CS783	22
CS783	CS783	1

References

1 de Kraker, M. E., Stewardson, A. J. & Harbarth, S. Will 10 million people die a year due to antimicrobial resistance by 2050? *PLOS Med* **13.11** (2016).

2 Chevrette, M. G. & Currie, C. R. Emerging evolutionary paradigms in antibiotic discovery. *J Ind Microbiol Biotechnol* **46**, 257-271 (2019).

3 Newman, D. J. & Cragg, G. M. Natural Products as Sources of New Drugs from 1981 to 2014. *J Nat Prod* **79**, 629-661 (2016).

4 De Mol, M. L., Snoeck, N., De Maeseneire, S. L. & Soetaert, W. K. Hidden antibiotics: Where to uncover? *Biotechnol Adv* **36**, 2201-2218 (2018).

5 Loureiro, C., Medema, M. H., van der Oost, J. & Sipkema, D. Exploration and exploitation of the environment for novel specialized metabolites. *Curr Opin Biotechnol* **50**, 206-213 (2018).

6 Blunt, J. W. *et al.* Marine natural products. *Nat Prod Rep* **35**, 8-53 (2018).

7 Ercolano, G., De Cicco, P. & Ianaro, A. New Drugs from the Sea: Pro-Apoptotic Activity of Sponges and Algae Derived Compounds. *Mar Drugs* **17** (2019).

8 Soldatou, S. & Baker, B. J. Cold-water marine natural products, 2006 to 2016. *Nat Prod Rep* **34**, 585-626 (2017).

9 Hartmann, A. C. *et al.* Meta-mass shift chemical profiling of metabolomes from coral reefs. *Proc Natl Acad Sci U S A* **114**, 11685-11690 (2017).

10 Miller, J. H. *et al.* Marine Invertebrate Natural Products that Target Microtubules. *J Nat Prod* **81**, 691-702 (2018).

11 Liu, J., Jung, J. H. & Liu, Y. Antimicrobial Compounds from Marine Invertebrates-Derived Microorganisms. *Curr Med Chem* **23**, 2892-2905 (2016).

12 Chen, L., Hu, J. S., Xu, J. L., Shao, C. L. & Wang, G. Y. Biological and Chemical Diversity of Ascidian-Associated Microorganisms. *Mar Drugs* **16** (2018).

13 Butler, M. S., Blaskovich, M. A. & Cooper, M. A. Antibiotics in the clinical pipeline at the end of 2015. *J Antibiot* **70**, 3 (2017).

14 Gogineni, V. & Hamann, M. T. Marine natural product peptides with therapeutic potential: Chemistry, biosynthesis, and pharmacology. *Biochim Biophys Acta Gen Subj* **1862**, 81-196 (2018).

15 Mayer, A. M. S., Rodriguez, A. D., Taglialatela-Scafati, O. & Fusetani, N. Marine Pharmacology in 2012-2013: Marine Compounds with Antibacterial, Antidiabetic, Antifungal, Anti-Inflammatory, Antiprotozoal, Antituberculosis, and Antiviral Activities; Affecting the Immune and Nervous Systems, and Other Miscellaneous Mechanisms of Action. *Mar Drugs* **15** (2017).

16 Calcott, M. J., Ackerley, D. F., Knight, A., Keyzers, R. A. & Owen, J. G. Secondary metabolism in the lichen symbiosis. *Chem Soc Rev* **47**, 1730-1760 (2018).

17 Chevrette, M. G. *et al.* The antimicrobial potential of Streptomyces from insect microbiomes. *Nat Commun* **10**, 516 (2019).

18 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).

19 Wright, G. D. Opportunities for natural products in 21(st) century antibiotic discovery. *Nat Prod Rep* **34**, 694-701 (2017).

20 Dhakal, D., Pokhrel, A. R., Shrestha, B. & Sohng, J. K. Marine Rare Actinobacteria: Isolation, Characterization, and Strategies for Harnessing Bioactive Compounds. *Front Microbiol* **8**, 1106 (2017).

21 Rodriguez Estevez, M., Myronovskyi, M., Gummerlich, N., Nadmid, S. & Luzhetskyy, A. Heterologous Expression of the Nybomycin Gene Cluster from the Marine Strain Streptomyces albus subsp. chlorinus NRRL B-24108. *Mar Drugs* **16** (2018).
22 Rodriguez, V. *et al.* Anthracimycin B, a Potent Antibiotic against Gram-Positive Bacteria Isolated from Cultures of the Deep-Sea Actinomycete Streptomyces cyaneofuscatus M-169. *Mar Drugs* **16** (2018).
23 Contador, C. A., Rodriguez, V., Andrews, B. A. & Asenjo, J. A. Use of genome-scale models to get new insights into the marine actinomycete genus Salinispora. *BMC Syst Biol* **13**, 11 (2019).
24 Jensen, P. R., Moore, B. S. & Fenical, W. The marine actinomycete genus Salinispora: a model organism for secondary metabolite discovery. *Nat Prod Rep* **32**, 738-751 (2015).
25 Ding, C. Y. G. *et al.* MS/MS-Based Molecular Networking Approach for the Detection of Aplysiatoxin-Related Compounds in Environmental Marine Cyanobacteria. *Mar Drugs* **16** (2018).
26 Dittmann, E., Gugger, M., Sivonen, K. & Fewer, D. P. Natural Product Biosynthetic Diversity and Comparative Genomics of the Cyanobacteria. *Trends Microbiol* **23**, 642-652 (2015).
27 Garcia, R., La Clair, J. J. & Muller, R. Future Directions of Marine Myxobacterial Natural Product Discovery Inferred from Metagenomics. *Mar Drugs* **16** (2018).
28 Hoffmann, T. *et al.* Correlating chemical diversity with taxonomic distance for discovery of natural products in myxobacteria. *Nat Commun* **9**, 803 (2018).
29 Lackner, G., Peters, E. E., Helfrich, E. J. & 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).
30 Wilson, M. C. *et al.* An environmental bacterial taxon with a large and distinct metabolic repertoire. *Nature* **506**, 58-62 (2014).
31 Huang, C. *et al.* Albumycin, a new isoindolequinone from Streptomyces albus J1074 harboring the fluostatin biosynthetic gene cluster. *J Antibiot (Tokyo)* **72**, 311-315 (2019).
32 Vuillemin, A. *et al.* Metabolic potential of microbial communities from ferruginous sediments. *Environ Microbiol* **20**, 4297-4313 (2018).
33 Wakimoto, T. *et al.* Calyculin biogenesis from a pyrophosphate protoxin produced by a sponge symbiont. *Nat Chem Biol* **10**, 648-655 (2014).
34 Zhang, W. *et al.* Marine biofilms constitute a bank of hidden microbial diversity and functional potential. *Nat Commun* **10**, 517 (2019).
35 Zhao, Y. *et al.* Genome-Centered Metagenomics Analysis Reveals the Symbiotic Organisms Possessing Ability to Cross-Feed with Anammox Bacteria in Anammox Consortia. *Environ Sci Technol* **52**, 11285-11296 (2018).
36 Laport, M. S., Santos, O. C. & Muricy, G. Marine sponges: potential sources of new antimicrobial drugs. *Curr Pharm Biotechnol* **10**, 86-105 (2009).
37 Brinkmann, C., Marker, A. & Kurtböke, D. An overview on marine sponge-symbiotic bacteria as unexhausted sources for natural product discovery. *Mar Drugs* **9**, 40 (2017).
38 Kanase, H. R. & Singh, K. N. M. Marine pharmacology: Potential, challenges, and future in India. *Am J Med Sci* **38**, 49 (2018).
39 Ruiz-Torres, V. *et al.* An Updated Review on Marine Anticancer Compounds: The Use of Virtual Screening for the Discovery of Small-Molecule Cancer Drugs. *Molecules* **22** (2017).

40 Bewley, C. A. & Faulkner, D. J. Lithistid Sponges: Star Performers or Hosts to the Stars. *Angew Chem Int Ed Engl* **37**, 2162-2178 (1998).
41 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).
42 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).
43 Hardoim, C. C. & Costa, R. Microbial communities and bioactive compounds in marine sponges of the family irciniidae-a review. *Mar Drugs* **12**, 5089-5122 (2014).
44 El-Demerdash, A. *et al.* Chemistry and Biological Activities of the Marine Sponges of the Genera Mycale (Arenochalina), Biemna and Clathria. *Mar Drugs* **16** (2018).
45 Shady, N. H., Fouad, M. A., Salah Kamel, M., Schirmeister, T. & Abdelmohsen, U. R. Natural Product Repertoire of the Genus Amphimedon. *Mar Drugs* **17** (2018).
46 Bayona, L. M., Videnova, M. & Choi, Y. H. Increasing Metabolic Diversity in Marine Sponges Extracts by Controlling Extraction Parameters. *Mar Drugs* **16** (2018).
47 Wakimoto, T. Toward the Dark Matter of Natural Products. *Chem Rec* **17**, 1124-1134 (2017).
48 Fisch, K. M. *et al.* Polyketide assembly lines of uncultivated sponge symbionts from structure-based gene targeting. *Nat Chem Biol* **5**, 494-501 (2009).
49 Hentschel, U., Piel, J., Degnan, S. M. & Taylor, M. W. Genomic insights into the marine sponge microbiome. *Nat Rev Microbiol* **10**, 641-654 (2012).
50 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).
51 Woodhouse, J. N., Fan, L., Brown, M. V., Thomas, T. & Neilan, B. A. Deep sequencing of non-ribosomal peptide synthetases and polyketide synthases from the microbiomes of Australian marine sponges. *ISME J* **7**, 1842-1851 (2013).
52 Abdelmohsen, U. R., Bayer, K. & Hentschel, U. Diversity, abundance and natural products of marine sponge-associated actinomycetes. *Nat Prod Rep* **31**, 381-399 (2014).
53 Indraningrat, A. A., Smidt, H. & Sipkema, D. Bioprospecting Sponge-Associated Microbes for Antimicrobial Compounds. *Mar Drugs* **14** (2016).
54 Trindade, M., van Zyl, L. J., Navarro-Fernandez, J. & Abd Elrazak, A. Targeted metagenomics as a tool to tap into marine natural product diversity for the discovery and production of drug candidates. *Front Microbiol* **6**, 890 (2015).
55 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**, 14002-14007 (2002).
56 Helfrich, E. J. & Piel, J. Biosynthesis of polyketides by trans-AT polyketide synthases. *Nat Prod Rep* **33**, 231-316 (2016).
57 Dyshlovoy, S. A. *et al.* Mycalamide A shows cytotoxic properties and prevents EGF-induced neoplastic transformation through inhibition of nuclear factors. *Mar Drugs* **10**, 1212-1224 (2012).
58 Gurel, G., Blaha, G., Steitz, T. A. & Moore, P. B. Structures of triacetyloleandomycin and mycalamide A bind to the large ribosomal subunit of Haloarcula marismortui. *Antimicrob Agents Chemother* **53**, 5010-5014 (2009).

59 Wu, C. Y. *et al.* Studies toward the unique pederin family member psymberin: structure-activity relationships, biochemical studies, and genetics identify the mode-of-action of psymberin. *J Am Chem Soc* **134**, 18998-19003 (2012).
60 Bayer, K., Jahn, M. T., Slaby, B. M., Moitinho-Silva, L. & Hentschel, U. Marine Sponges as Chloroflexi Hot Spots: Genomic Insights and High-Resolution Visualization of an Abundant and Diverse Symbiotic Clade. *mSystems* **3** (2018).
61 Schmitt, S. *et al.* Assessing the complex sponge microbiota: core, variable and species-specific bacterial communities in marine sponges. *ISME J* **6**, 564-576 (2012).
62 Freeman, M. F. *et al.* Metagenome mining reveals polytheonamides as posttranslationally modified ribosomal peptides. *Science* **338**, 387-390 (2012).
63 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).
64 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).
65 Webster, N. S. & Thomas, T. The Sponge Hologenome. *MBio* **7**, e00135-00116 (2016).
66 El-Gendy, M. M. & El-Bondkly, A. M. Production and genetic improvement of a novel antimycotic agent, saadamycin, against dermatophytes and other clinical fungi from endophytic Streptomyces sp. Hedaya48. *J Ind Microbiol Biotechnol* **37**, 831-841 (2010).
67 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).
68 Nagai, K. *et al.* YM-266183 and YM-266184, novel thiopeptide antibiotics produced by Bacillus cereus isolated from a marine sponge. I. Taxonomy, fermentation, isolation, physico-chemical properties and biological properties. *J Antibiot (Tokyo)* **56**, 123-128 (2003).
69 Suzumura, K. *et al.* YM-266183 and YM-266184, novel thiopeptide antibiotics produced by Bacillus cereus isolated from a marine sponge II. Structure elucidation. *J Antibiot (Tokyo)* **56**, 129-134 (2003).
70 Waters, A. L. *et al.* An analysis of the sponge Acanthostrongylophora igens' microbiome yields an actinomycete that produces the natural product manzamine A. *Front Mar Sci* **1** (2014).
71 Leal, M. C. *et al.* Marine microorganism-invertebrate assemblages: perspectives to solve the "supply problem" in the initial steps of drug discovery. *Mar Drugs* **12**, 3929-3952 (2014).
72 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).
73 Blin, K., Kim, H. U., Medema, M. H. & Weber, T. Recent development of antiSMASH and other computational approaches to mine secondary metabolite biosynthetic gene clusters. *Brief Bioinform* (2017).
74 Del Carratore, F. *et al.* Computational identification of co-evolving multi-gene modules in microbial biosynthetic gene clusters. *Commun Biol* **2**, 83 (2019).
75 Medema, M. H. & Osbourn, A. Computational genomic identification and functional reconstitution of plant natural product biosynthetic pathways. *Nat Prod Rep* **33**, 951-962 (2016).

76 Laureti, L. *et al.* Identification of a bioactive 51-membered macrolide complex by activation of a silent polyketide synthase in Streptomyces ambofaciens. *Proc Natl Acad Sci U S A* **108**, 6258-6263 (2011).
77 Dodge, G. J., Maloney, F. P. & Smith, J. L. Protein-protein interactions in "cis-AT" polyketide synthases. *Nat Prod Rep* **35**, 1082-1096 (2018).
78 Finking, R. & Marahiel, M. A. Biosynthesis of nonribosomal peptides1. *Annu Rev Microbiol* **58**, 453-488 (2004).
79 Huo, L. *et al.* Heterologous expression of bacterial natural product biosynthetic pathways. *Nat Prod Rep* (2019).
80 Brown, A. S., Calcott, M. J., Owen, J. G. & Ackerley, D. F. Structural, functional and evolutionary perspectives on effective re-engineering of non-ribosomal peptide synthetase assembly lines. *Nat Prod Rep* **35**, 1210-1228 (2018).
81 Kosol, S., Jenner, M., Lewandowski, J. R. & Challis, G. L. Protein-protein interactions in trans-AT polyketide synthases. *Nat Prod Rep* **35**, 1097-1109 (2018).
82 Piel, J. Biosynthesis of polyketides by trans-AT polyketide synthases. *Nat Prod Rep* **27**, 996-1047 (2010).
83 Chen, A., Re, R. N. & Burkart, M. D. Type II fatty acid and polyketide synthases: deciphering protein-protein and protein-substrate interactions. *Nat Prod Rep* **35**, 1029-1045 (2018).
84 Palmer, C. M. & Alper, H. S. Expanding the Chemical Palette of Industrial Microbes: Metabolic Engineering for Type III PKS-Derived Polyketides. *Biotechnol J* **14**, e1700463 (2019).
85 Villebro, R., Shaw, S., Blin, K. & Weber, T. Sequence-based classification of type II polyketide synthase biosynthetic gene clusters for antiSMASH. *J Ind Microbiol Biotechnol* **46**, 469-475 (2019).
86 Parvez, A. *et al.* Novel Type III Polyketide Synthases Biosynthesize Methylated Polyketides in Mycobacterium marinum. *Sci Rep* **8**, 6529 (2018).
87 Shimizu, Y., Ogata, H. & Goto, S. Type III Polyketide Synthases: Functional Classification and Phylogenomics. *Chembiochem* **18**, 50-65 (2017).
88 Arnison, P. G. *et al.* Ribosomally synthesized and post-translationally modified peptide natural products: overview and recommendations for a universal nomenclature. *Nat Prod Rep* **30**, 108-160 (2013).
89 Ortega, M. A. & van der Donk, W. A. New Insights into the Biosynthetic Logic of Ribosomally Synthesized and Post-translationally Modified Peptide Natural Products. *Cell Chem Biol* **23**, 31-44 (2016).
90 Scheidler, C. M., Kick, L. M. & Schneider, S. Ribosomal Peptides and Small Proteins on the Rise. *Chembiochem* (2019).
91 Schwalen, C. J., Hudson, G. A., Kille, B. & Mitchell, D. A. Bioinformatic Expansion and Discovery of Thiopeptide Antibiotics. *J Am Chem Soc* **140**, 9494-9501 (2018).
92 Tietz, J. I. *et al.* A new genome-mining tool redefines the lasso peptide biosynthetic landscape. *Nat Chem Biol* **13**, 470-478 (2017).
93 Son, S. *et al.* Ulleungdin, a Lasso Peptide with Cancer Cell Migration Inhibitory Activity Discovered by the Genome Mining Approach. *J Nat Prod* **81**, 2205-2211 (2018).
94 Hegemann, J. D., Zimmermann, M., Xie, X. & Marahiel, M. A. Lasso peptides: an intriguing class of bacterial natural products. *Acc Chem Res* **48**, 1909-1919 (2015).
95 Repka, L. M., Chekan, J. R., Nair, S. K. & van der Donk, W. A. Mechanistic Understanding of Lanthipeptide Biosynthetic Enzymes. *Chem Rev* **117**, 5457-5520 (2017).

96 Crone, W. J. *et al.* Dissecting Bottromycin Biosynthesis Using Comparative Untargeted Metabolomics. *Angew Chem Int Ed Engl* **55**, 9639-9643 (2016).
97 Czekster, C. M., Ge, Y. & Naismith, J. H. Mechanisms of cyanobactin biosynthesis. *Curr Opin Chem Biol* **35**, 80-88 (2016).
98 Mo, T. *et al.* Biosynthetic Insights into Linaridin Natural Products from Genome Mining and Precursor Peptide Mutagenesis. *ACS Chem Biol* **12**, 1484-1488 (2017).
99 Ghilarov, D. *et al.* Architecture of Microcin B17 Synthetase: An Octameric Protein Complex Converting a Ribosomally Synthesized Peptide into a DNA Gyrase Poison. *Mol Cell* **73**, 749-762 e745 (2019).
100 Helf, M. J., Freeman, M. F. & Piel, J. Investigations into PoyH, a promiscuous protease from polytheonamide biosynthesis. *J Ind Microbiol Biotechnol* **46**, 551-563 (2019).
101 Rocha-Martin, J., Harrington, C., Dobson, A. D. & O'Gara, F. Emerging strategies and integrated systems microbiology technologies for biodiscovery of marine bioactive compounds. *Mar Drugs* **12**, 3516-3559 (2014).
102 Timmermans, M. L., Paudel, Y. P. & Ross, A. C. Investigating the Biosynthesis of Natural Products from Marine Proteobacteria: A Survey of Molecules and Strategies. *Mar Drugs* **15** (2017).
103 Ward, A. C. & Allenby, N. E. Genome mining for the search and discovery of bioactive compounds: the Streptomyces paradigm. *FEMS Microbiol Lett* **365** (2018).
104 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* **35**, 725-731 (2017).
105 Brady, S. F. Construction of soil environmental DNA cosmid libraries and screening for clones that produce biologically active small molecules. *Nat Protoc* **2**, 1297-1305 (2007).
106 Owen, J. G. *et al.* Mapping gene clusters within arrayed metagenomic libraries to expand the structural diversity of biomedically relevant natural products. *Proc Natl Acad Sci U S A* **110**, 11797-11802 (2013).
107 Parages, M. L., Gutierrez-Barranquero, J. A., Reen, F. J., Dobson, A. D. & O'Gara, F. Integrated (Meta) Genomic and Synthetic Biology Approaches to Develop New Biocatalysts. *Mar Drugs* **14** (2016).
108 Cohen, L. J., Han, S., Huang, Y. H. & Brady, S. F. Identification of the Colicin V Bacteriocin Gene Cluster by Functional Screening of a Human Microbiome Metagenomic Library. *ACS Infect Dis* **4**, 27-32 (2018).
109 Gilbert, J. A., Jansson, J. K. & Knight, R. The Earth Microbiome project: successes and aspirations. *BMC Biol* **12**, 69 (2014).
110 Helber, S. B. *et al.* Sponges from Zanzibar host diverse prokaryotic communities with potential for natural product synthesis. *FEMS Microbiol Ecol* **95** (2019).
111 Moitinho-Silva, L. *et al.* The sponge microbiome project. *Gigascience* **6**, 1-7 (2017).
112 Rubin-Blum, M. *et al.* Fueled by methane: deep-sea sponges from asphalt seeps gain their nutrition from methane-oxidizing symbionts. *ISME J* (2019).
113 Turnbaugh, P. J. *et al.* The human microbiome project. *Nature* **449**, 804-810 (2007).
114 Iqbal, H. A., Feng, Z. & Brady, S. F. Biocatalysts and small molecule products from metagenomic studies. *Curr Opin Chem Biol* **16**, 109-116 (2012).
115 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).

116 Collins, J. & Hohn, B. Cosmids: a type of plasmid gene-cloning vector that is packageable in vitro in bacteriophage lambda heads. *Proc Natl Acad Sci U S A* **75**, 4242-4246 (1978).
117 Winn, R. N. & Norris, M. B. Analysis of mutations in l transgenic medaka using the cII mutation assay. *Techniques in Aquatic Toxicology* **2**, 725-754 (2005).
118 Peek, J. *et al.* Rifamycin congeners kanglemycins are active against rifampicin-resistant bacteria via a distinct mechanism. *Nat Commun* **9**, 4147 (2018).
119 Fieseler, L., Quaiser, A., Schleper, C. & Hentschel, U. Analysis of the first genome fragment from the marine sponge-associated, novel candidate phylum Poribacteria by environmental genomics. *Environ Microbiol* **8**, 612-624 (2006).
120 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).
121 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).
122 Tyson, G. W. *et al.* Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **428**, 37-43 (2004).
123 Alneberg, J. *et al.* Genomes from uncultivated prokaryotes: a comparison of metagenome-assembled and single-amplified genomes. *Microbiome* **6**, 173 (2018).
124 Goordial, J. & Ronholm, J. Metagenomics meets read clouds. *Nat Biotechnol* **36**, 1049-1051 (2018).
125 Mukherjee, S. *et al.* 1,003 reference genomes of bacterial and archaeal isolates expand coverage of the tree of life. *Nat Biotechnol* **35**, 676-683 (2017).
126 Parks, D. H. *et al.* Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nature Microbiology* **2**, 1533-1542 (2017).
127 Robinson, S. L., Christenson, J. K. & Wackett, L. P. Biosynthesis and chemical diversity of beta-lactone natural products. *Nat Prod Rep* **36**, 458-475 (2019).
128 Solden, L. M. & Wrighton, K. C. Finding life's missing pieces. *Nature Microbiology* **2**, 1458-1459 (2017).
129 Thompson, L. R. *et al.* A communal catalogue reveals Earth's multiscale microbial diversity. *Nature* **551**, 457-463 (2017).
130 Simister, R. L., Deines, P., Botte, E. S., Webster, N. S. & Taylor, M. W. Sponge-specific clusters revisited: a comprehensive phylogeny of sponge-associated microorganisms. *Environ Microbiol* **14**, 517-524 (2012).
131 Blanquer, A., Uriz, M. J. & Galand, P. E. Removing environmental sources of variation to gain insight on symbionts vs. transient microbes in high and low microbial abundance sponges. *Environ Microbiol* **15**, 3008-3019 (2013).
132 Podell, S. *et al.* Pangenomic comparison of globally distributed Poribacteria associated with sponge hosts and marine particles. *ISME J* **13**, 468-481 (2019).
133 Taylor, M. W. *et al.* 'Sponge-specific' bacteria are widespread (but rare) in diverse marine environments. *ISME J* **7**, 438-443 (2013).
134 Edgar, R. C. Updating the 97% identity threshold for 16S ribosomal RNA OTUs. *Bioinformatics* **34**, 2371-2375 (2018).
135 Fan, L. *et al.* Functional equivalence and evolutionary convergence in complex communities of microbial sponge symbionts. *Proc Natl Acad Sci U S A* **109**, E1878-1887 (2012).

136 Reveillaud, J. *et al.* Host-specificity among abundant and rare taxa in the sponge microbiome. *ISME J* **8**, 1198-1209 (2014).
137 Gao, Z. M. *et al.* Symbiotic adaptation drives genome streamlining of the cyanobacterial sponge symbiont "Candidatus Synechococcus spongiarum". *MBio* **5**, e00079-00014 (2014).
138 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. *ISME J* **11**, 2465-2478 (2017).
139 Burgsdorf, I. *et al.* Lifestyle evolution in cyanobacterial symbionts of sponges. *MBio* **6**, e00391-00315 (2015).
140 Gauthier, M. E. A., Watson, J. R. & Degnan, S. M. Draft genomes shed light on the dual bacterial symbiosis that dominates the microbiome of the coral reef sponge Amphimedon queenslandica. *Front Mar Sci* **3**, 196 (2016).
141 Horn, H. *et al.* An Enrichment of CRISPR and Other Defense-Related Features in Marine Sponge-Associated Microbial Metagenomes. *Front Microbiol* **7**, 1751 (2016).
142 Ryu, T. *et al.* Hologenome analysis of two marine sponges with different microbiomes. *BMC Genomics* **17**, 158 (2016).
143 Sunagawa, S. *et al.* Ocean plankton. Structure and function of the global ocean microbiome. *Science* **348**, 1261359 (2015).
144 Venter, J. C. *et al.* Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304**, 66-74 (2004).
145 Yergeau, E. *et al.* Metagenomic survey of the taxonomic and functional microbial communities of seawater and sea ice from the Canadian Arctic. *Sci Rep* **7**, 42242 (2017).
146 Zhang, H. *et al.* Microbial Community Dynamics and Assembly Follow Trajectories of an Early-Spring Diatom Bloom in a Semienclosed Bay. *Appl Environ Microbiol* **84** (2018).
147 Lebar, M. D., Heimbegner, J. L. & Baker, B. J. Cold-water marine natural products. *Nat Prod Rep* **24**, 774-797 (2007).
148 Lozupone, C. & Knight, R. UniFrac: a new phylogenetic method for comparing microbial communities. *Appl Environ Microbiol* **71**, 8228-8235 (2005).
149 Nunez-Montero, K. & Barrientos, L. Advances in Antarctic Research for Antimicrobial Discovery: A Comprehensive Narrative Review of Bacteria from Antarctic Environments as Potential Sources of Novel Antibiotic Compounds Against Human Pathogens and Microorganisms of Industrial Importance. *Antibiotics (Basel)* **7** (2018).
150 Poli, A. *et al.* Microbial Diversity in Extreme Marine Habitats and Their Biomolecules. *Microorganisms* **5** (2017).
151 Amos, G. C. *et al.* Designing and Implementing an Assay for the Detection of Rare and Divergent NRPS and PKS Clones in European, Antarctic and Cuban Soils. *PLoS One* **10**, e0138327 (2015).
152 Asencio, G. *et al.* Antibacterial activity of the Antarctic bacterium Janthinobacterium sp: SMN 33.6 against multi-resistant Gram-negative bacteria. *Electronic Journal of Biotechnology* **17**, 1 (2014).
153 Benaud, N. *et al.* Harnessing long-read amplicon sequencing to uncover NRPS and Type I PKS gene sequence diversity in polar desert soils. *FEMS Microbiol Ecol* **95** (2019).
154 Danilovich, M. E., Sánchez, L. A., Acosta, F. & Delgado, O. D. Antarctic bioprospecting: in pursuit of microorganisms producing new antimicrobials and enzymes. *Polar Biol*, 1-17 (2018).

155 Fukuda, W. *et al.* Lysobacter oligotrophicus sp. nov., isolated from an Antarctic freshwater lake in Antarctica. *Int J Syst Evol Microbiol* **63**, 3313-3318 (2013).
156 Leiva, S., Alvarado, P., Huang, Y., Wang, J. & Garrido, I. Diversity of pigmented Gram-positive bacteria associated with marine macroalgae from Antarctica. *FEMS Microbiol Lett* **362**, fnv206 (2015).
157 Lo Giudice, A. *et al.* Bacterium-bacterium inhibitory interactions among psychrotrophic bacteria isolated from Antarctic seawater (Terra Nova Bay, Ross Sea). *FEMS Microbiol Ecol* **60**, 383-396 (2007).
158 Zhang, Y., Zhao, J. & Zeng, R. Expression and characterization of a novel mesophilic protease from metagenomic library derived from Antarctic coastal sediment. *Extremophiles* **15**, 23-29 (2011).
159 von Salm, J. L. *et al.* Darwinolide, a New Diterpene Scaffold That Inhibits Methicillin-Resistant Staphylococcus aureus Biofilm from the Antarctic Sponge Dendrilla membranosa. *Org Lett* **18**, 2596-2599 (2016).
160 Arrigo, K. R. & Thomas, D. N. Large scale importance of sea ice biology in the Southern Ocean. *Antarctic Sci* **16**, 471-486 (2004).
161 Garrison, D. L. *et al.* Sea-ice microbial communities in the Ross Sea: autumn and summer biota. *Mar Ecol* **300**, 39-52 (2005).
162 Maas, E. W. *et al.* Phylogenetic analyses of bacteria in sea ice at Cape Hallett, Antarctica. *New Zeal J Mar Fresh* **46**, 3-12 (2012).
163 Mock, T. & Thomas, D. N. Recent advances in sea-ice microbiology. *Environ Microbiol* **7**, 605-619 (2005).
164 Ryan, K. G. *et al.* Comparison of the microalgal community within fast ice at two sites along the Ross Sea coast, Antarctica. *Antarctic Sci* **18**, 583-594 (2006).
165 Torstensson, A. *et al.* Physicochemical control of bacterial and protist community composition and diversity in Antarctic sea ice. *Environ Microbiol* **17**, 3869-3881 (2015).
166 Nurk, S., Meleshko, D., Korobeynikov, A. & Pevzner, P. A. metaSPAdes: a new versatile metagenomic assembler. *Genome Res* **27**, 824-834 (2017).
167 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. *ISME J* **11**, 2864-2868 (2017).
168 Sangwan, N., Xia, F. & Gilbert, J. A. Recovering complete and draft population genomes from metagenome datasets. *Microbiome* **4**, 8 (2016).
169 Sieber, C. M. K. *et al.* Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. *Nat Microbiol* **3**, 836-843 (2018).
170 Uritskiy, G. & DiRuggiero, J. Applying Genome-Resolved Metagenomics to Deconvolute the Halophilic Microbiome. *Genes (Basel)* **10** (2019).
171 Uritskiy, G. V., DiRuggiero, J. & Taylor, J. MetaWRAP-a flexible pipeline for genome-resolved metagenomic data analysis. *Microbiome* **6**, 158 (2018).
172 Peng, Y., Leung, H. C., Yiu, S. M. & Chin, F. Y. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* **28**, 1420-1428 (2012).
173 Boisvert, S., Raymond, F., Godzaridis, E., Laviolette, F. & Corbeil, J. Ray Meta: scalable de novo metagenome assembly and profiling. *Genome Biol* **13**, R122 (2012).
174 Namiki, T., Hachiya, T., Tanaka, H. & Sakakibara, Y. MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads. *Nucleic Acids Res* **40**, e155 (2012).

175 Li, D., Liu, C. M., Luo, R., Sadakane, K. & Lam, T. W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **31**, 1674-1676 (2015).
176 Peng, Y., Leung, H. C., Yiu, S. M. & Chin, F. Y. IDBA–a practical iterative de Bruijn graph de novo assembler. *In Annual international conference on research in computational molecular biology, Springer, Berlin, Heidelberg.*, 426-440 (2010).
177 Bankevich, A. *et al.* SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* **19**, 455-477 (2012).
178 Kang, D. *et al.* MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ Preprints* **7**, e27522v27521 (2019).
179 Blin, K. *et al.* The antiSMASH database version 2: a comprehensive resource on secondary metabolite biosynthetic gene clusters. *Nucleic Acids Res* **47**, D625-D630 (2019).
180 Blin, K. *et al.* antiSMASH 4.0-improvements in chemistry prediction and gene cluster boundary identification. *Nucleic Acids Res* **45**, W36-W41 (2017).
181 Eddy, S. R. Accelerated Profile HMM Searches. *PLoS Comput Biol* **7**, e1002195 (2011).
182 Skinnider, M. A. *et al.* Genomic charting of ribosomally synthesized natural product chemical space facilitates targeted mining. *Proc Natl Acad Sci U S A* **113**, E6343-E6351 (2016).
183 Hardy, C. D. & Butler, A. Ambiguity of NRPS Structure Predictions: Four Bidentate Chelating Groups in the Siderophore Pacifibactin. *J Nat Prod* (2019).
184 Shirley, W. A. *et al.* Unzipping Natural Products: Improved Natural Product Structure Predictions by Ensemble Modeling and Fingerprint Matching. *chemRxiV* (2018).
185 Hadjithomas, M. *et al.* IMG-ABC: new features for bacterial secondary metabolism analysis and targeted biosynthetic gene cluster discovery in thousands of microbial genomes. *Nucleic Acids Res* **45**, D560-D565 (2017).
186 Boddy, C. N. Bioinformatics tools for genome mining of polyketide and non-ribosomal peptides. *J Ind Microbiol Biotechnol* **41**, 443-450 (2014).
187 Doroghazi, J. R. *et al.* A roadmap for natural product discovery based on large-scale genomics and metabolomics. *Nat Chem Biol* **10**, 963-968 (2014).
188 Moghaddam, J. A. *et al.* Analysis of the Genome and Metabolome of Marine Myxobacteria Reveals High Potential for Biosynthesis of Novel Specialized Metabolites. *Sci Rep* **8**, 16600 (2018).
189 D'Agostino, P. M. & Gulder, T. A. M. Direct Pathway Cloning Combined with Sequence- and Ligation-Independent Cloning for Fast Biosynthetic Gene Cluster Refactoring and Heterologous Expression. *ACS Synth Biol* **7**, 1702-1708 (2018).
190 Dejong, C. A. *et al.* Polyketide and nonribosomal peptide retro-biosynthesis and global gene cluster matching. *Nat Chem Biol* **12**, 1007-1014 (2016).
191 Johnston, C. W. *et al.* Assembly and clustering of natural antibiotics guides target identification. *Nat Chem Biol* **12**, 233-239 (2016).
192 Maansson, M. *et al.* An Integrated Metabolomic and Genomic Mining Workflow To Uncover the Biosynthetic Potential of Bacteria. *mSystems* **1** (2016).
193 Tianero, M. D., Balaich, J. N. & Donia, M. S. Localized production of defence chemicals by intracellular symbionts of Haliclona sponges. *Nat Microbiol* (2019).
194 Koh, Y. E. Phototrophic Bacteria in Antarctic Sea Ice (2011).
195 Cowie, R. O. M. Bacterial community structure, function and diversity in Antarctic sea ice. (2011).
196 Streit, W. R., & Daniel, R. Metagenomics. (2017).

197 Gurgui, C. & Piel, J. Metagenomic approaches to identify and isolate bioactive natural products from microbiota of marine sponges. *Methods Mol Biol* **668**, 247-264 (2010).
198 Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114-2120 (2014).
199 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**, 1043-1055 (2015).
200 Navarro-Muñoz, J. *et al.* A computational framework for systematic exploration of biosynthetic diversity from large-scale genomic data. *bioRxiv* **445270** (2018).
201 Albertsen, M. *et al.* Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nat Biotechnol* **31**, 533-538 (2013).
202 Gionfriddo, C. M. *et al.* Microbial mercury methylation in Antarctic sea ice. *Nat Microbiol* **1**, 16127 (2016).
203 Sainis, I. *et al.* Cyanobacterial cyclopeptides as lead compounds to novel targeted cancer drugs. *Mar Drugs* **8**, 629-657 (2010).
204 Subramani, R. & Aalbersberg, W. Culturable rare Actinomycetes: diversity, isolation and marine natural product discovery. *Appl Microbiol Biotechnol* **97**, 9291-9321 (2013).
205 Field, J. J. *et al.* Microtubule-stabilizing activity of zampanolide, a potent macrolide isolated from the Tongan marine sponge Cacospongia mycofijiensis. *J Med Chem* **52**, 7328-7332 (2009).
206 Przeslak, A. D., Inman, M., Lewis, W. & Moody, C. J. Origin of the Thiopyrone CTP-431 "Unexpectedly" Isolated from the Marine Sponge Cacospongia mycofijiensis. *J Org Chem* **83**, 10595-10601 (2018).
207 Rueda, A., Zubía, E., Ortega, M. J., Carballo, J. L. & Salvá, J. New cytotoxic metabolites from the sponge Cacospongia scalaris. *J Org Chem* **62**, 1481-1485 (1997).
208 Taufa, T. *et al.* Zampanolides B-E from the Marine Sponge Cacospongia mycofijiensis: Potent Cytotoxic Macrolides with Microtubule-Stabilizing Activity. *J Nat Prod* **81**, 2539-2544 (2018).
209 Field, J. J. *et al.* Zampanolide, a Microtubule-Stabilizing Agent, Is Active in Resistant Cancer Cells and Inhibits Cell Migration. *Int J Mol Sci* **18** (2017).
210 Gava, F. *et al.* Gap junctions contribute to anchorage-independent clustering of breast cancer cells. *BMC Cancer* **18**, 221 (2018).
211 Prota, A. E. *et al.* Structural basis of microtubule stabilization by laulimalide and peloruside A. *Angew Chem Int Ed Engl* **53**, 1621-1625 (2014).
212 Ueoka, R. *et al.* Metabolic and evolutionary origin of actin-binding polyketides from diverse organisms. *Nature Chemical Biology* **11**, 705 (2015).
213 Poplau, P., Frank, S., Morinaka, B. I. & Piel, J. An enzymatic domain for the formation of cyclic ethers in complex polyketides. *Angew Chem Int Ed Engl* **52**, 13215-13218 (2013).
214 Ogura, H. *et al.* EpoK, a cytochrome P450 involved in biosynthesis of the anticancer agents epothilones A and B. Substrate-mediated rescue of a P450 enzyme. *Biochemistry* **43**, 14712-14721 (2004).
215 Podust, L. M. & Sherman, D. H. Diversity of P450 enzymes in the biosynthesis of natural products. *Nat Prod Rep* **29**, 1251-1266 (2012).
216 Maier, S. *et al.* Insights into the bioactivity of mensacarcin and epoxide formation by MsnO8. *Chembiochem* **15**, 749-756 (2014).

217 Thibodeaux, C. J., Chang, W. C. & Liu, H. W. Enzymatic chemistry of cyclopropane, epoxide, and aziridine biosynthesis. *Chem Rev* **112**, 1681-1709 (2012).
218 Walsh, C. T. & Wencewicz, T. A. Flavoenzymes: versatile catalysts in biosynthetic pathways. *Nat Prod Rep* **30**, 175-200 (2013).
219 Julien, B. *et al.* Isolation and characterization of the epothilone biosynthetic gene cluster from Sorangium cellulosum. *Gene* **249**, 153-160 (2000).
220 Reeves, C. D., Hu, Z., Reid, R. & Kealey, J. T. Genes for the biosynthesis of the fungal polyketides hypothemycin from Hypomyces subiculosus and radicicol from Pochonia chlamydosporia. *Appl Environ Microbiol* **74**, 5121-5129 (2008).
221 Ames, B. D., Liu, X. & Walsh, C. T. Enzymatic processing of fumiquinazoline F: a tandem oxidative-acylation strategy for the generation of multicyclic scaffolds in fungal indole alkaloid biosynthesis. *Biochemistry* **49**, 8564-8576 (2010).
222 Minami, A. *et al.* Sequential enzymatic epoxidation involved in polyether lasalocid biosynthesis. *J Am Chem Soc* **134**, 7246-7249 (2012).
223 Tang, G. L., Cheng, Y. Q. & Shen, B. Leinamycin biosynthesis revealing unprecedented architectural complexity for a hybrid polyketide synthase and nonribosomal peptide synthetase. *Chem Biol* **11**, 33-45 (2004).
224 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).
225 Miller, I. J. *et al.* Autometa: automated extraction of microbial genomes from individual shotgun metagenomes. *Nucleic Acids Res* (2019).
226 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).
227 Lin, H. H. & Liao, Y. C. Accurate binning of metagenomic contigs via automated clustering sequences using information of genomic signatures and marker genes. *Sci Rep* **6**, 24175 (2016).
228 Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119 (2010).
229 McCutcheon, J. P. & Moran, N. A. Extreme genome reduction in symbiotic bacteria. *Nat Rev Microbiol* **10**, 13-26 (2011).
230 Tully, B. J., Graham, E. D. & Heidelberg, J. F. The reconstruction of 2,631 draft metagenome-assembled genomes from the global oceans. *Sci Data* **5**, 170203 (2018).
231 Del Angel, V. D. *et al.* Ten steps to get started in Genome Assembly and Annotation. *F1000Research* **7** (2018).
232 Klassen, J. L. & Currie, C. R. Gene fragmentation in bacterial draft genomes: extent, consequences and mitigation. *BMC Genomics* **13**, 14 (2012).
233 Nazari, B. *et al.* Nonomuraea sp. ATCC 55076 harbours the largest actinomycete chromosome to date and the kistamicin biosynthetic gene cluster. *Medchemcomm* **8**, 780-788 (2017).
234 Treuner-Lange, A., Bruckskotten, M., Rupp, O., Goesmann, A. & Sogaard-Andersen, L. Whole-Genome Sequence of the Fruiting Myxobacterium Cystobacter fuscus DSM 52655. *Genome Announc* **5** (2017).
235 Kreutzer, M. F., Kage, H. & Nett, M. Structure and biosynthetic assembly of cupriachelin, a photoreactive siderophore from the bioplastic producer Cupriavidus necator H16. *J Am Chem Soc* **134**, 5415-5422 (2012).
236 Rachid, S., Gerth, K., Kochems, I. & Muller, R. Deciphering regulatory mechanisms for secondary metabolite production in the myxobacterium Sorangium cellulosum So ce56. *Mol Microbiol* **63**, 1783-1796 (2007).

237 Beld, J., Sonnenschein, E. C., Vickery, C. R., Noel, J. P. & Burkart, M. D. The phosphopantetheinyl transferases: catalysis of a post-translational modification crucial for life. *Nat Prod Rep* **31**, 61-108 (2014).
238 Nasrin, S. *et al.* Chloramphenicol Derivatives with Antibacterial Activity Identified by Functional Metagenomics. *J Nat Prod* **81**, 1321-1332 (2018).
239 Lundgren, B. R. & Boddy, C. N. Sialic acid and N-acyl sialic acid analog production by fermentation of metabolically and genetically engineered Escherichia coli. *Org Biomol Chem* **5**, 1903-1909 (2007).
240 Bishara, A. *et al.* High-quality genome sequences of uncultured microbes by assembly of read clouds. *Nat Biotechnol* (2018).
241 Chiara, M. *et al.* A-GAME: improving the assembly of pooled functional metagenomics sequence data. *BMC Genomics* **19**, 44 (2018).
242 Koren, S. *et al.* Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res* **27**, 722-736 (2017).
243 Koonin, E. V. & Wolf, Y. I. Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. *Nucleic Acids Res* **36**, 6688-6719 (2008).
244 Ondov, B. D. *et al.* Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol* **17**, 132 (2016).
245 Varghese, N. J. *et al.* Microbial species delineation using whole genome sequences. *Nucleic Acids Res* **43**, 6761-6771 (2015).
246 Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150-3152 (2012).
247 Elsayed, S. S. *et al.* Chaxapeptin, a Lasso Peptide from Extremotolerant Streptomyces leeuwenhoekii Strain C58 from the Hyperarid Atacama Desert. *J Org Chem* **80**, 10252-10260 (2015).
248 Crits-Christoph, A., Diamond, S., Butterfield, C. N., Thomas, B. C. & Banfield, J. F. Novel soil bacteria possess diverse genes for secondary metabolite biosynthesis. *Nature* **558**, 440-444 (2018).
249 Florez, L. V., Biedermann, P. H., Engl, T. & Kaltenpoth, M. Defensive symbioses of animals with prokaryotic and eukaryotic microorganisms. *Nat Prod Rep* **32**, 904-936 (2015).
250 Staley, C. & Sadowsky, M. J. Practical considerations for sampling and data analysis in contemporary metagenomics-based environmental studies. *J Microbiol Methods* **154**, 14-18 (2018).
251 Alanjary, M. *et al.* The Antibiotic Resistant Target Seeker (ARTS), an exploration engine for antibiotic cluster prioritization and novel drug target discovery. *Nucleic Acids Res* **45**, W42-W48 (2017).
252 Cruz-Morales, P. *et al.* Phylogenomic Analysis of Natural Products Biosynthetic Gene Clusters Allows Discovery of Arseno-Organic Metabolites in Model Streptomycetes. *Genome Biol Evol* **8**, 1906-1916 (2016).
253 Smanski, M. J. *et al.* Synthetic biology to access and expand nature's chemical diversity. *Nat Rev Microbiol* **14**, 135-149 (2016).
254 Funk, M. A. & van der Donk, W. A. Ribosomal Natural Products, Tailored To Fit. *Acc Chem Res* **50**, 1577-1586 (2017).
255 Wang, W. *et al.* Identification of anti-Gram-negative bacteria agents targeting the interaction between ribosomal proteins L12 and L10. *Acta Pharm Sin B* **8**, 772-783 (2018).
256 Jiang, W. *et al.* Cas9-Assisted Targeting of CHromosome segments CATCH enables one-step targeted cloning of large gene clusters. *Nat Commun* **6**, 8101 (2015).

257 Karimi, E. *et al.* Comparative Metagenomics Reveals the Distinctive Adaptive Features of the Spongia officinalis Endosymbiotic Consortium. *Front Microbiol* **8**, 2499 (2017).
258 Keyzers, R. A., Northcote, P. T. & Davies-Coleman, M. T. Spongian diterpenoids from marine sponges. *Nat Prod Rep* **23**, 321-334 (2006).