Investigating the biosynthetic potential of an Antarctic soil through metagenomics, cultivation, and heterologous expression

Valentin Waschulin

A thesis submitted to the University of Warwick for the degree of
Doctor of Philosophy

School of Life Sciences,
University of Warwick,
Coventry, CV4 7AL

April 2022

Table of Contents

List of Figures

List of Tables

Acknowledgements

I would like to thank my supervisors Prof. Elizabeth Wellington and Prof. Christophe Corre at the University of Warwick for their support and guidance. I would also like to thank the Wellington, Corre and Alberti groups for their support, especially Dr. Chiara Borsetto, without whose theoretical guidance, practical help and cloning vectors this thesis would look much worse, as well as Dr. Rob James whose knowledge was indispensable for my sequencing work. I furthermore want to thank my advisory bord, Prof. Yin Chen and Dr. Munehiro Ally for their helpful and critical discussion of my reports. I also want to thank my funder the European Union (Horizon 2020 grant agreement number 765147) and the entire CARTNET cohort for great meetings, discussions and collaborations – in particular, I'd like to thank Dr. Yong Kai (Duncan) Ng and Kristiina Vind. My thanks also go to Dr. Stefano Donadio and Dr. Margherita Sosio for welcoming me for my secondment at NAICONS in Milano. Lastly, I would like to thank my family and friends for their support in all my endeavours; and my wonderful partner Leigh for cheering me on when I was down, helping me when I was in need, being kind when I was grumpy, and putting me on the sofa with a blanket and a cup of tea when nothing else would do.

Declaration

This thesis is submitted to the University of Warwick in support of my application for the degree of Doctor of Philosophy. It has been composed by myself and has not been submitted in any previous application for any degree.

The work presented (including data generated and data analysis) was carried out by the author except in the cases outlined below:

Chapter 2: The 16S rRNA gene sequencing library preparation was carried out by Dr. Chiara Borsetto.

Chapter 3: Some BGC amplifications were carried out by undergraduate student Oliver Pearson under my supervision. LC-MS samples were run by Prof. Lijiang Song.

Icons for diagrams were obtained from BioRender.com

Parts of this thesis have been published by the author:

Chapter 1 and some parts of the abstract and discussion have been published under

Waschulin, V., Borsetto, C., James, R. *et al.* Biosynthetic potential of uncultured Antarctic soil bacteria revealed through long-read metagenomic sequencing. *ISME J* **16,** 101–111 (2022). doi.org/10.1038/s41396-021-01052-3

as well as the associated preprint on biorXiv (doi.org/10.1101/2020.12.09.416412)

Abstract

The growing problem of antibiotic resistance has led to the exploration of uncultured bacteria as sources of new antimicrobials. Metagenomic sequencing studies of samples from different environments have reported evidence of high biosynthetic gene cluster (BGC) diversity in metagenomes, and metagenomic library studies have yielded several novel natural products. However, accessing these compounds remains challenging. The constraints of short-read sequencing mean that the assembly of full-length BGC sequences from uncultured bacteria is nigh impossible, thus making assessment of BGC diversity difficult and downstream cloning infeasible. Conversely, metagenomic library approaches suffer from a bias towards known compounds as well as difficulties with expressing recovered BGCs. In the present work, a three-pronged approach was taken to access the biosynthetic diversity of bacteria from an Antarctic soil: A hybrid shotgun metagenome was sequenced and BGCs cloned and expressed, a novel regulatory gene-based screen for libraries was developed, and a number of isolates were obtained by culturing. Through metagenomic sequencing, many highly divergent BGCs were found in phyla such as Acidobacteriota and Verrucomicrobiota, but also the methanotrophic gammaproteobacterial order UBA7966. Sequencing of isolates obtained from the same soil indicated little overlap between the biosynthetic potential of readily cultured and uncultured bacteria. Several metagenomic BGCs were PCR-amplified, cloned and expressed in *Pseudomonas* and *Streptomyces*. While the sequencing of *Streptomyces* exconjugants showed that many inserts were truncated, a phenotype was observed in *Pseudomonas.* The library screening approach was validated in isolates, but the targets were absent in the metagenomic library used. In conclusion, the results uncover the rich diversity of BGCs from uncultured lineages present in the soil, show the potential of long-read sequencing to recover full-length BGCs from uncultured soil bacteria and demonstrate the feasibility of cloning them. However, they also indicate the necessity of refined molecular tools for successful heterologous expression of metagenomic BGCs.

List of Abbreviations

A	Adenylation
ANI	Average nucleotide identity
AT	Acyltransferase
BAC	Bacterial artificial chromosome
BGC	Biosynthetic gene cluster
BiG-FAM	Biosynthetic gene cluster families database
BiG-SCAPE	Biosynthetic Gene Similarity Clustering and Prospecting Engine
BiG-SLiCE	Biosynthetic Genes Super-Linear Clustering Engine
BLAST	Basic Local Alignment Search Tool
BLASTn	Nucleotide BLAST
BLASTp	Protein BLAST
BLASTx	Translated nucleotide to protein BLAST
bp	Basepair
BSA	Bovine serum albumin
CAT	Contig Assignment Tool
CFU	Colomy forming unit
CoA	Coenzyme A
CRISPR	Clustered regularly interspaced short palindromic repeats
CTAB	Cetyltrimethylammonium bromide = hexadecyltrimethylammonium bromide
DiPaC	Direct Pathway Cloning
DNA	Deoxyribonucleic acid
DUF	Domain of unknown function
E-value	Expect value
eDNA	Environmental deoxyribonucleic acid
EDTA	Ethylenediaminetetraacetic acid
Gb	Gigabase
GC	Gas chromatography
GCF	Gene cluster family
GNPS	Global Natural Product Social Molecular Networking
GTDB	Genome Taxonomy Database
HGT	Horizontal gene transfer
HMM	Hidden Markov Model
HMW	High molecular weight
IMG	Integrated Microbial Genomes
JGI	Joint Genome Institute
Kb	Kilobase
KS	Ketosynthase
LAL	Large ATP-binding regulator of the LuxR family
LC	Liquid chromatography
MAG	Metagenome-assembled genome
MARE	Methylenomycin autoregulatory element
Mb	Megabase

MCS	Multiple cloning site
MIBiG	Minimal Information about a Biosynthetic Gene Cluster
MMF	Methylenomycin furan
mRNA	Messenger RNA
MS	Mass spectrometry
NCBI	National Center for Biotechnology Information
NEB	New England Biolabs
NGS	Next-generation sequencing
NRPS	Nonribosomal peptide synthetase
ONT	Oxford Nanopore Technologies
ORF	Open reading frame
OTU	Operational taxonomic unit
PCP	Peptidyl carrier protein
PCR	Polymerase chain reaction
PFGE	Pulse field gel electrophoresis
PKS	Polyketide synthase
pMMO	Particulate methane monooxygenase
Q-TOF	Quadrupole time-of-flight
QIIME	Quantitative insights into microbial ecology
qPCR	Quantitative polymerase chain reaction
RBS	Ribosome binding site
RefSeq	NCBI Reference Sequence Database
RF	Radio frequency
RiPP	Ribosomally synthesised, posttranslationally modified peptide
RNA	Ribonucleic acid
rRNA	Ribosomal ribonucleic acid
RT	Room temperature
SARP	Streptomyces antibiotic regulatory protein
SCIFF	Six cysteines in forty-five amino acids
SLIC	Sequence and ligation-independent cloning
SMRT	Single molecule rapid real-time sequencing
SOC	Super optimal broth with catabolite repression
SPRI	Solid phase reversible immobilization
TAE	Tris-acetate-EDTA
TAR	Transformation-associated recombination
TBE	Tris-Borate-EDTA
TE	Tris-EDTA
tRNA	Transfer-RNA
TSS	Transformation and storage solution
UBA	Uncultured bacteria and archaea
USCγ	Upland soil cluster gamma
UV-Vis	Ultraviolet-visible
VEPE	Vinyl ether lipid phosphatidylethanolamine

1 Introduction

1.1 Antibiotics and humans

It is hard to overstate the benefits that microbial natural products have brought to humanity. They are used as antibiotics, antifungals, antihelmintics, antivirals and immunosuppressants; but also as herbicides, insecticides and food conservation agents. Out of all these diverse applications, their use as antibiotics is probably the most important. Antibiotics have greatly reduced infant mortality rates, vastly increased the survivability of surgery and chemotherapy, and made death from bacterial infections such as pneumonia a rare phenomenon in the developed world (Zaffiri, Gardner, and Toledo-Pereyra 2012). Around the middle of the 20th century, many new classes of natural product antibiotics were discovered and developed as drugs. Usually, resistance would emerge and spread among pathogens soon after introduction of a novel antibiotic (Finland, Frank, and Wilcox 1950). While this was not an important problem as long as the development pipeline was "filled" and novel antibiotics were brought to market every few years, it became an issue when antibiotic discovery stalled in the 1970s. As of today, the ever-increasing trend of multidrug resistance in bacteria threatens to lead to a post-antibiotic era and it has been estimated that by 2050, 10 million people could die every year due to infections with drug resistant bacteria (Neil 2016). The drying up of the pipeline has been explained by the increased rate of rediscovery, meaning that the same compounds were discovered again and again in screening programs (Watve et al. 2001; Baltz 2006). This could be explained by the unequal distribution of antibiotic biosynthetic gene clusters (BGCs). While the vast majority of antibiotics would occur very infrequently ($<10^{-6}$), a smaller number of extremely widespread antibiotics such as streptothricin (c. 10^{-1}), streptomycin or tetracycline (both c. 10^{-2}) would make discovery difficult (Baltz 2006). Coupled with a lack of profitability of new antibiotic drugs for pharmaceutical companies compared to drugs for chronic illnesses, this therefore led to a divestment of natural products research by major pharmaceutical players

(Årdal et al. 2020). However, growing awareness of the antibiotic resistance crisis and the genomic insights brought by the sequencing revolution have led to a renewed interest in natural products (Bachmann, Van Lanen, and Baltz 2014; J. W.-H. Li and Vederas 2009).

1.2 Specialised metabolite discovery and analysis

Specialised metabolite discovery methods have evolved greatly over time. While the discovery of penicillin was purely serendipitous, further discoveries were made using screening processes such as the Waksman screening platform (Fleming 1929; Schatz, Bugie, and Waksman 2005). In this approach, antibiotic producing strains (often *Streptomyces*) were identified by zones of inhibition on agar overlay plates. Important broad-spectrum antibiotics discovered this way in the 1940s to 1960s include streptomycin, chloramphenicol, vancomycin and rifampicin (Lewis 2013). However, the problem of re-discovery led to the widespread abandonment of the approach in favour of synthetic chemistry. Unfortunately, the synthetic approach did not yield as many successful compounds as hoped, with only fluoroquinolones emerging as broad-spectrum antibiotics of major clinical importance (Lewis 2013). This failure of purely synthetic chemistry together with advances in molecular biology led to renewed scientific interest in natural products as sources of pharmaceuticals (Demain 2002; Lewis 2013). Even before genome sequencing became widespread, it was speculated that *Streptomyces* harboured many more antibiotics than the ones already discovered, with a study estimating up to 100,000 mostly undiscovered antibiotics in the genus (Watve et al. 2001). When the *Streptomyces coelicolor* genome was sequenced in 2002 and other antibiotic producers followed, it became clear that even in characterised producer strains, many compounds encoded in the genome were not actually known (Weber et al. 2003; Bentley et al. 2002). This indicated that not only were there many more specialised metabolites to discover, but that they were encoded in the genomes of bacteria already in culture. Further sequencing has confirmed that silence of many BGCs is the

norm rather than an exception (Rutledge and Challis 2015). Exploiting these inactive and cryptic BGCs promises to be a much more focused route for antibiotic discovery than random screening of an exponentially increasing number of isolates (Aigle and Corre 2012; Rutledge and Challis 2015). At the same time, advances in analytical methods such as high-resolution mass spectrometry, structure prediction, automated fractionation and screening platforms, as well as computational advances such as GNPS (Global Natural Product Social Molecular Networking) have enabled quicker dereplication, i.e. elimination of known compounds (Atanasov et al. 2021).

1.3 Bioinformatic analysis of specialised metabolism

In the last 15 years, the bioinformatic analysis of specialised metabolite gene clusters matured from simple gene detection to sophisticated predictions of structure and function as well as global analyses of diversity (Figure 1.1). The increasing amount of sequence data led to the development of bioinformatic tools related to natural product biosynthesis, most importantly genome mining tools such as BAGEL and antiSMASH (de Jong et al. 2006; Medema et al. 2011). These tools scan genomes for the presence of homologues of genes known to be involved in specialised metabolite biosynthesis using hidden markov models (HMMs). Further advances in genome mining include substrate prediction of NRPS and PKS enzymes (Röttig et al. 2011; Bachmann and Ravel 2009), cluster border prediction (Cimermancic et al. 2014) precursor prediction for lassopeptides (Tietz et al. 2017), antibiotic target prediction (Alanjary et al. 2017), improved NRPS/PKS product structure prediction (Skinnider et al. 2017) and neural network based genome mining (Hannigan et al. 2019). However, the falling cost of sequencing and thereby exponentially increasing number of publicly available sequences also necessitated the development of tools that allowed a systematic comparison of BGCs exceeding simple BLAST homology scores. The MiBiG database was introduced as a community-

curated, central repository of BGCs with known products (Kautsar, Blin, et al. 2020). BiG-SCAPE is a networking tool that functions by pairwise comparison of antiSMASH-derived BGCs based on PFAM protein domains and thereby enables the construction of similarity networks of BGCs, allowing e.g. the determination of common and rare BGCs in culture collections (Kautsar et al. 2021). Systematic investigation of BGCs from the many thousand publicly available genomes also promises to lead to more generalisable insights about the distribution of BGCs. However, pairwise comparison is computationally expensive for large datasets, since it scales quadratically with sequence number. The recently published tool BiG-SLiCE (and its sister database BiG-FAM) attempts to solve this problem by comparing antiSMASH-derived BGCs based on the presence or absence of a large, pre-computed set of HMMs (Kautsar, Hooft, et al. 2020). These comparisons get converted into distance scores between BGCs, enabling the grouping of BGCs into gene cluster families (GCFs) based on distance cutoffs. Query BGCs can then be scored according to their distance to these GCFs, enabling both a classification of a query BGC as well as a quantification of sequence novelty. A study anchoring the BiG-SLiCE-derived GCFs in BGC-linked compounds from NPAtlas led to the extrapolation that only 3% of the biosynthetic diversity in the world was experimentally characterised and found evidence for biome-specific distribution of BGCs (Gavriilidou et al. 2021).

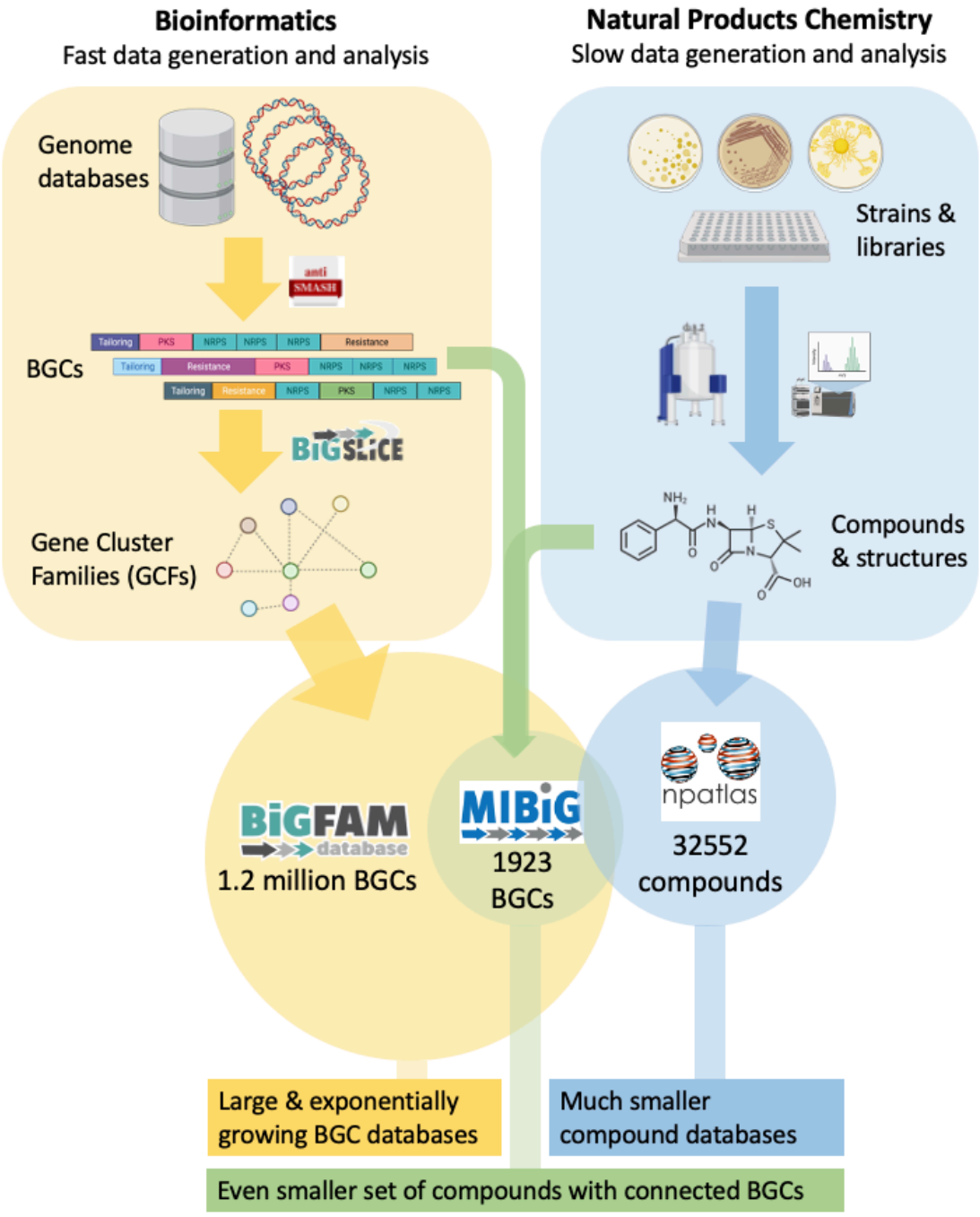


Figure 1.1: An overview of selected bioinformatic tools and databases used to mine, store and connect information about BGCs, compound structure, and producing organisms.

1.4 Culture-based approaches for natural product discovery

The vast majority of antibiotics and other natural products discovered to date come from isolated organisms. It was only in the 1990s the first metagenomic 16S rRNA gene amplicon

sequencing studies of different environments enabled us to systematically discover and characterise the diversity of uncultured bacteria (Schmidt, DeLong, and Pace 1991). As a consequence of these early metagenomic studies, uncultured bacteria were speculated to harbour novel natural products and thereby novel antibiotics (Seow et al. 1997; Rondon et al. 2000). While this "microbial dark matter" was often called unculturable, efforts to bring these as-of-yet uncultivated organisms into culture showed that it is indeed possible to isolate many of them. The two main obstacles to overcome were firstly the inability of some bacteria to grow on the provided substrate, and secondly the competition of fast-growing bacteria that would crowd out the slow growers (Vartoukian, Palmer, and Wade 2010). Methods developed to overcome these hurdles include low-nutrient media and long incubation times to select for oligotrophic, slow-growing bacteria that would normally be outcompeted; adapted media preparation to avoid the formation of toxic compounds such as H_2O_2; use of different gelling agents such as gellan gum or carrageenan; as well as use of microscopes to detect small colonies (George et al. 2011; Kato et al. 2018; Vartoukian, Palmer, and Wade 2010; Pulschen et al. 2017; Janssen et al. 2002). Other novel microorganisms previously only known from metagenomic analysis were isolated by stepwise enrichment and purification processes featuring specific selection steps such as centrifugation, antibiotic treatment or growth on floating filter discs (H. Zhang et al., 2003.; Könneke et al. 2005; Tveit et al. 2019). Further innovations include high-throughput dilution to extinction approaches using liquid medium or polymer matrixes such as alginate. More recently, microfluidic approaches using water-in-oil emulsions have enabled extremely high throughput cultivation that can even be coupled with on-chip screening for antibiotic activity and subsequent cultivation on plates (Mahler et al. 2021). Furthermore, in-situ cultivation of encapsulated microorganisms has emerged as a promising approach. The principle behind it is to culture bacteria in encapsulated compartments that keep bacterial cells separate from the environment, but at the same time

allow diffusion of nutrients and waste products through sufficiently small pores (Aoi et al. 2009; Ben-Dov, Kramarsky-Winter, and Kushmaro 2009; Gavrish et al. 2008; Jung, Aoi, and Epstein 2016). One high-throughput in-situ cultivation device called the iChip has led to the discovery of teixobactin, the first representative of a new class of antibiotics with promising activity against Gram-positive bacteria and low potential for resistance (Gunjal et al. 2020; Ling et al. 2015). While innovative "blanket" cultivation approaches such as the iChip have been very successful, there is increasing interest in harnessing metagenomic information to isolate specific community members. For example, Cross et al. used metagenome-derived protein sequences to raise polyclonal antibodies against a predicted membrane protein and then used these antibodies to label, sort and eventually culture hitherto uncultivated Saccharibacteria and SR1 bacteria from the human mouth (Cross et al. 2019). Rubin et al. used transposons to deliver CRISPR machinery into bacteria in synthetic soil communities as well as real-life gut communities, endowing the bacteria with antibiotic resistance and additional metabolic capabilities (Rubin et al. 2022). These marker genes were then used to select for the engineered bacteria, leading to their successful isolation. While this proof-of-principle study only targeted easily cultivable bacteria such as *Klebsiella* and *Escherichia*, it could be used to isolate more elusive bacteria in the future.

1.5 Culture-independent specialised metabolite discovery

Sequencing of environmental 16S rRNA genes was used to explore and quantify the general bacterial diversity, and the same technology was also applied to specialised metabolite biosynthesis genes, such as NRPS A-domains and PKS KS-domains. Studies using degenerate primers for these domains showed remarkable sequence diversity in uncultured bacteria (Borsetto et al. 2019; Schirmer et al. 2005; Charlop-Powers et al. 2015). Amplicon as well as short-read metagenomic studies have shown promising candidates containing a large number

of BGCs in the phyla Verrucomicrobia, Acidobacteria, Chloroflexi and several candidate phyla (Crits-Christoph et al. 2018; Nayfach et al. 2020; Sharrar et al. 2020).

In addition to amplicon sequencing, metagenomic clone libraries have been used to both explore the diversity of metagenomic BGCs and to obtain the products encoded in these BGCs by heterologous expression (Gillespie et al. 2002). The workhorse of metagenomic library approaches has been the fosmid – a plasmid that can be packed into phage particles and then transfected into an *E. coli* host (Brady 2007). Its advantages include high effectiveness due to the high efficiency of phage transfection compared to e.g. chemical transformation, as well an intrinsic size-selection step in the phage particle assembly that ensures average insert sizes of 40 kb on average (De Tomaso and Weissman 2003). However, this size selection is also its limitation, since many BGCs are larger than the maximum insert size and therefore need to be "stitched together" using TAR cloning (J. H. Kim et al. 2010). This limits their application in functional screens compared to BAC vectors that can hold inserts of several hundred kilobases. Construction of BAC libraries, however, is more complicated, necessitating efficient selection for high-molecular weight DNA by techniques such as separation of bacteria from the matrix (e.g. soil), embedding in agarose gel plugs followed by in-plug cell lysis, enzymatic digestion of the agarose and eventually cloning and transformation (Nasrin et al. 2018). Furthermore, the high efficiency of the fosmid approach has led to megalibraries with an estimated $1.5*10^7$ individual fosmid clones, corresponding to roughly six terabases of DNA. BAC approaches have only reached a size of ca. 2 terabases to the author's knowledge (Owen et al. 2013; Nasrin et al. 2018). An important differentiation in library approaches is also the choice of screening method, i.e. functional or sequence-based. Functional screening works by observing activity of library hosts in assays such as antibiotic assays or enzyme inhibition, while sequence-guided screening uses sequence information to retrieve BGC-bearing clones using PCR primers

(L. Robinson, Piel, and Sunagawa 2021). Functional screening of libraries has the advantage of immediate detection of activity, but is constrained by the activity of the BGC in the library host organism, usually *E. coli,* a notoriously unreliable host for heterologous expression of BGCs (Ke and Yoshikuni 2020; Liu et al. 2020). Sequence-guided screening is free from this constraint but necessitates the downstream expression of the recovered BGCs using a suitable expression host and potentially genetic engineering (H.-S. Kang, Charlop-Powers, and Brady 2016). Furthermore, sequence-guided screening is often conducted with only a small fragment of the BGC actually known through e.g. NRPS A-domain amplicon sequencing or the use of degenerate primers. This means that very little information about the detected BGC is known until the plasmids are recovered and sequenced. This problem is partly circumvented in the recent approach of Libis et al., who used NRPS/PKS domain amplicon sequencing to determine co-occurrence patterns of unique A/KS domains within arrayed metagenomic libraries (Libis et al. 2019). Using these patterns, it was possible to infer which domains were present on the same cosmid clone, likely constituting a BGC. These inferred BGCs were compared to the A/KS domains of known BGCs in order to select the inferred BGCs harbouring the most novel A/KS domains combinations. After iterative qPCR screening, cloning of the BGC and heterologous expression, the omnipeptins were discovered. However, no sequence-guided screening approach to date could recover completely novel types of BGCs, since the biosynthetic genes are unknown and therefore cannot be screened for. As currently only 1% of bacterial proteins are experimentally characterised and only about a third of bacterial proteins have any computationally assigned function, it is likely that many novel BGCs go undetected (Hoskisson and Seipke 2020). An overview of culture-based and culture-independent techniques is provided in Figure 1.2. Further advantages and disadvantages of sequence-guided vs functional screening can be found in Table 1.1. A selection of recently reported natural product discoveries from metagenomes can be found in Table 1.2.

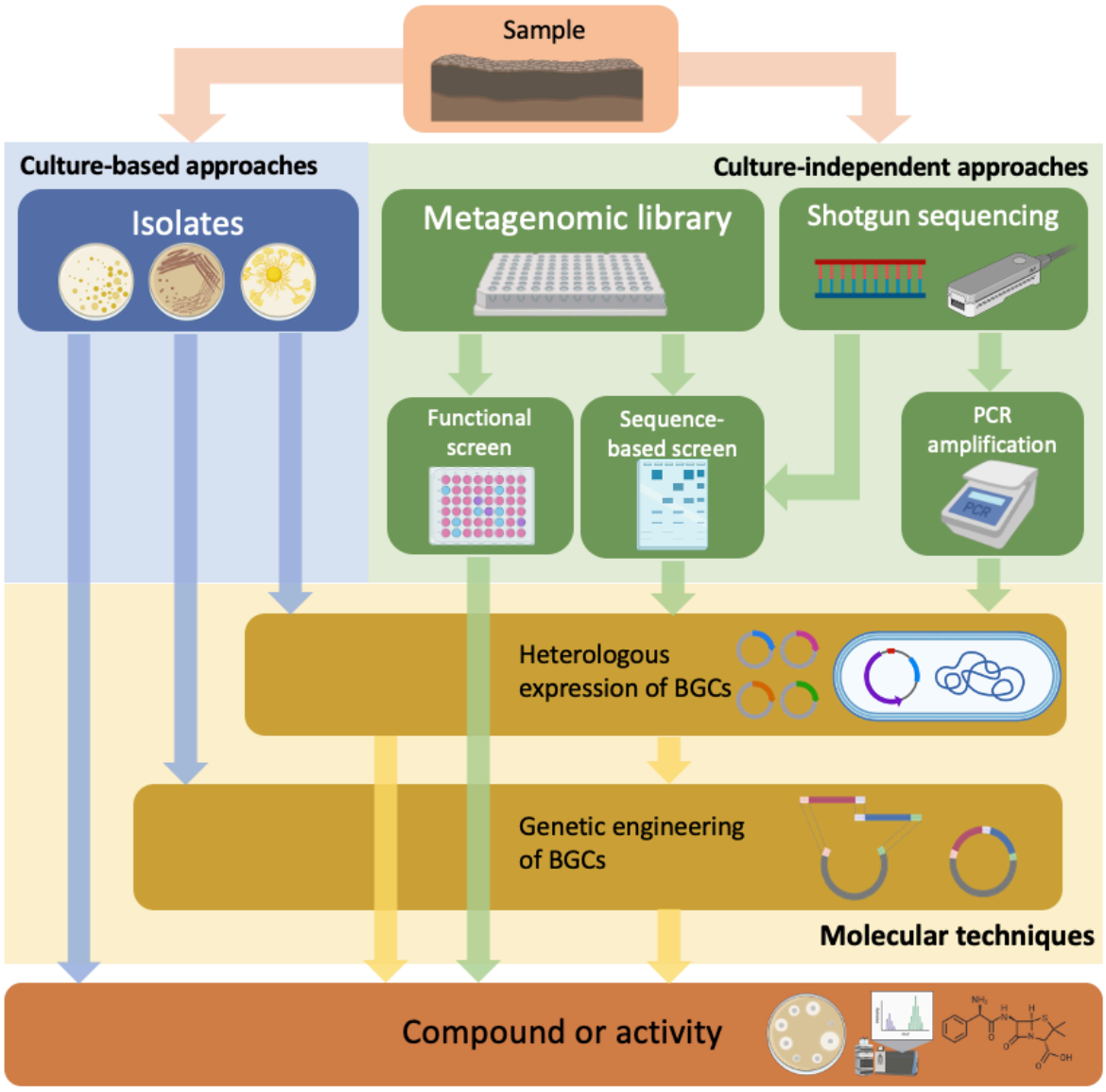


Figure 1.2: Simplified overview of culture-based and culture-independent workflows for natural product discovery.

The falling cost of sequencing has enabled large-scale shotgun metagenomic sequencing of different microbiomes (Nayfach et al. 2020; Desai et al. 2012). Shotgun metagenomic sequencing has the advantage of being able to retrieve a relatively unbiased picture of the sequence diversity of a given sample. This, however, also means that it is untargeted and the sequences of interest – such as BGCs – will only amount to a tiny fraction of the combined

Table 1.1: Advantages and disadvantages of sequence-guided and functional metagenomic library screening

Sequence-guided screening	Functional screening
Activity does not need to be present in library host	Activity needs to be present in detectable levels in library host
Library pooling is possible – even large libraries can be kept in one 96 well plate. However, this requires iterative dilution and screening	Library pooling is not possible, likely necessitating automation to pick, store and assay thousands of colonies in 384-well plates
Activity of recovered BGCs might never discovered due to lack of expression	Each hit is an activity
Can target specific types of BGCs with sequence information	Cannot target BGC type, potentially leading to rediscovery
Limited targeting of specific activity e.g. against a pathogen	Very specific activity selectable as long as an appropriate assay is developed
Can use fosmids – BGCs can be stitched together for large BGCs	Can use fosmids, but BACs needed for large BGCs
Potential activity of BGC product not limited	Activity of BGC product limited by assay employed for screening

Table 1.2: Non-exhaustive list of natural products discovered from metagenomes in recent years, with activity and discovery method.

Product	Activity	Method	Reference
Malacidins	Antibacterial	Metagenomic library, sequence-guided	(Hover et al. 2018)
Omnipeptins	Not reported	Metagenomic library, sequence-guided	(Libis et al. 2019)
Metathramycin	Cytotoxic	Metagenomic library, sequence-guided	(J. Stevenson et al. 2021)
Metatricycloene	Antibacterial	Metagenomic library, functional (in *Streptomyces*)	(Iqbal et al. 2016)
Chloramphenicol derivatives	Antibacterial	Metagenomic library, functional (in *E. coli*)	(Nasrin et al. 2018)
Antibacterial enzymes	Antibacterial	Metagenomic library, functional (in *Ralstonia*)	(Iqbal, Craig, and Brady 2014)
Polybrominated diphenyl ethers	Not reported	Shotgun metagenome and cloning of PCR products	(Agarwal et al. 2017)

data, thereby making it necessary to sequence a large amount of DNA. Early Sanger and 454 clone library approaches as well as NGS read-based approaches mostly directly analysed the sequenced reads (Pearce et al. 2012; Tyson et al. 2004). This maximises the amount of sequences analysed by bypassing the metagenomic assembly step which requires many reads of the same nucleotide position ("coverage") to assemble a sequence (Luo et al. 2012). Read-based analyses, however, are limited by read length, which for Illumina is between 50 and 150 bp. While gene fragments such as NRPS A-domains can be analysed through reads alone, whole BGCs as well the origin of these BGCs or indeed any genomic context cannot. Therefore, metagenomic assembly is highly advantageous to analyse the structure and function of a given microbiome. While metagenomic assembly was extremely expensive and only possible on low-diversity samples as late as 15 years ago, it has been made relatively affordable by the development of short-read NGS (Illumina) sequencing (Desai et al. 2012; Tyson et al. 2004). Metagenomic assembly, like genomic assembly, generally works by constructing graphs based on overlaps between reads, finding the most likely path through the graph and assembling continuous and unambiguous parts of the path into contigs (Compeau, Pevzner, and Tesler 2011). However, the assembly of such contigs is hindered by high diversity in the metagenome, highly conserved or repetitive regions such as 16S rRNA gene sequences, the presence of closely related bacterial species or strains, as well as areas of low read coverage that occur due to the stochastic nature of shotgun sequencing (Quince et al. 2021). Low coverage can be ameliorated by deeper sequencing, but repeat sequences, high diversity and especially the presence of closely related strains sharing significant amounts of sequence identity are not easy to solve (Quince et al. 2021). This problem is partly solved by the usage of binning algorithms that try to sort contigs into "bins" that ideally represent the original genome of the organism the contigs originated from. Binning algorithms work by harnessing intrinsic features such as GC content, k-mer frequency, but also differential coverage

information (Alneberg et al. 2014; D. D. Kang et al. 2019; Wu et al. 2014). This has enabled the development of genome-centric metagenomics and the recovery of hundreds of metagenome-assembled genomes (MAGs), giving access to (fractured) genomes of thousands of uncultured organisms (Nayfach et al. 2020). With regards to BGCs, modular NRPS and PKS BGCs fall into the repeat sequence category, making assemblies "break" in the middle of a BGC (Meleshko et al. 2019). Therefore, Illumina-derived MAGs usually do not contain complete BGCs. Soil bacteria have been the source of a large number of natural products, making uncultured soil bacteria a highly desirable target for drug discovery (Daniel 2004). At the same time, soil is also the most diverse biome on earth, making metagenomic assembly especially challenging (Howe et al. 2014). However, long-read sequencing technologies like Oxford Nanopore and PacBio HiFi sequencing promise to solve this problem by virtue of providing reads that are thousands of basepairs in length, as opposed to 50-150 basepairs in short-read sequencing. First studies have shown that long-read metagenomics result in much larger contig sizes, complete BGCs and even closed genomes from metagenomic samples (Singleton et al. 2020; Van Goethem et al. 2021; Sevim et al. 2019; Moss, Maghini, and Bhatt 2020). These results indicate that long read metagenomics could be game-changing in the discovery and analysis of BGCs in uncultured phyla.

1.6 Activation of specialised metabolite production

Even before the advent of widespread genome sequencing, it was known that one bacterial or fungal strain could often produce many different compounds, depending on the culture conditions employed ("one strain many compounds", Bode et al. 2002). Genome sequencing of actinobacterial producer strains confirmed this by revealing that most BGCs in cultured strains did not have an associated product, and that there could be up to 50 BGCs in a single strain (*Streptomyces hygroscopicus* XM201, RefSeq GCF_002021875.1). Variation of culture

conditions has therefore been used as a relatively straightforward and successful method to discover novel compounds from known and previously screened strains. The list of possible variations of culture conditions is potentially endless and includes, for example, media composition with regards to pH and nutrients, elicitation with small molecules or heavy metals, challenge with live or dead bacterial cells, and many others (Tomm, Ucciferri, and Ross 2019). However, the sheer number of potential variables – many of which are difficult to recreate in the laboratory – also make it unlikely that all BGCs can be activated by this method alone. Therefore, employing molecular methods may be necessary to activate certain BGCs.

Molecular techniques have been used to investigate the genes involved in specialised metabolite production – mostly in actinomycetes – as well as regulation since the 1980s (Horinouchi et al. 1989). Since most specialised metabolites are produced by the concerted action of several or even dozens of genes, they cannot easily be overexpressed by replacing a single promoter. However, BGCs are often controlled by a regulatory cascade comprised of only a few genes acting as transcriptional activators and transcriptional repressors on different levels. Therefore, knocking out or overexpressing a single regulatory gene can lead to abolishing or activating the production of a specialised metabolite encoded in a BGC (Rutledge and Challis 2015). This observation has guided many successful approaches to activating BGCs so far, either by overexpressing transcriptional activators such as SARPs or LALs, or by disrupting transcriptional repressors such as TetRs or GntRs (Koomsiri et al. 2019; Laureti et al. 2011; Sidda et al. 2013; Smanski et al. 2009). These regulators can be pathway-specific and only control one BGC, or global/pleiotropic regulators that affect several BGCs as well as processes such as differentiation and sporulation. While e.g. SARPs are mostly pathway-specific regulators, TetRs can act both on specific pathways and global regulatory networks (Xia et al. 2020). Moreover, regulatory networks can be very complex and involve several

activators and repressors within a single BGC, thereby requiring some a priori knowledge to predictably activate the expression of a BGC (Rigali et al. 2018). Furthermore, some BGCs do not contain obvious regulatory genes, either because they are controlled by global regulators or because they contain an as of yet uncharacterised regulatory gene. One example of a widespread, yet only recently discovered regulator family in *Streptomyces* is the LmbU family (Ju, Zhang, and Elliot 2017). Many more uncharacterised regulatory genes are likely found in less well explored bacterial lineages.

While the "regulatory gene" approach harnesses the native regulatory capacity of the BGC to induce expression, the synthetic biology-inspired "BGC refactoring" approach seeks to be more independent of the native regulatory mechanisms (Z. Shao et al. 2013). A large part of BGC refactoring to date has consisted of promoter engineering, based on the premise that a transcriptionally silent BGC can be activated by replacing the inactive promoters with active ones (L. Li, MacIntyre, and Brady 2021). This approach has been enabled by the increasing sophistication of molecular biology tools, including CRISPR/Cas9 genome editing and has resulted in the successful activation of multiple silent BGCs (H.-S. Kang, Charlop-Powers, and Brady 2016). However, many bacterial strains of interest are not amenable to genetic manipulation. While molecular tools such as CRISPR or engineered promoters might work across many species within the same bacterial genus, introducing DNA into a non-model strain can be difficult (Tong et al. 2019; Q. Yan and Fong 2017). In addition, more unusual bacteria such as e.g. Acidobacteria do not even come with any developed tools that could be introduced. To address this problem, more generalisable approaches for the genetic manipulation of non-model bacteria are being researched and a recent study reported successful BGC activation in "undomesticated" *Photorhabdus* and *Xenorhabdus* isolates (G. Wang et al. 2019). While these successes point to a future when any bacterium can be genetically manipulated to induce

expression of BGCs, as of today heterologous expression is still the technique of choice for engineering BGCs from genetically intractable organisms.

In its simplest form, heterologous expression entails the cloning of a BGC into a plasmid and subsequently transferring it into an expression host. This plasmid can be obtained by generation and screening of a (meta-) genomic library, but can also be constructed using more targeted techniques such as yeast-based transformation-associated recombination (TAR) cloning or PCR amplification and Gibson cloning (Greunke et al. 2018; Kouprina and Larionov 2016). In some cases, removing the BGC from its native environment is enough to relieve transcriptional repression, thereby leading to expression of a BGC silent in its native host (X. Zhang, Hindra, and Elliot 2019). However, this is not always the case, and further genetic engineering of the BGC, either by refactoring or by manipulating regulatory genes, is necessary. Furthermore, heterologous expression adds a whole different set of challenges to expression (Xu and Wright 2019). In the native host, it is likely that the cellular machinery for successful expression of a BGC is in place, and only a regulation mechanism is preventing it. In a heterologous host, the repressive regulatory signal mechanism is absent, but so might be other factors that enable successful expression in the first place. These factors can be tRNA supply based on codon usage, recognition of promoters and RBS, supply of precursors, resistance to toxic products, and others (Xu and Wright 2019). To minimise these problems, a close phylogenetic relationship between native and host organisms is desirable. However, this might not always be possible if no appropriate host organism is available. Furthermore, even if the organism of origin is closely related to the expression host and shares features such as e.g. GC content, there can be challenges. For example, the TTA codon in *Streptomyces* is very rare and production of its corresponding tRNA (encoded by the *bldA* gene) is tightly regulated, since it has an important function in bacterial development. This role has been suggested to be exclusive to

Streptomyces (Ventura et al. 2007). Heterologously expressed BGCs with a higher frequency of TTA codons may therefore not be expressed well unless the TTA codons are replaced (Molnár et al. 2005; Zhu et al. 2011). Moreover, even between members of the same genus, there can be differences in promoter activity (W. Wang et al. 2013). Therefore, heterologous expression often makes further molecular manipulations such as refactoring necessary for successful expression.

1.7 The natural role of natural products

When a specialised metabolite is isolated, it is usually tested for activities that are useful for medicinal, agricultural or other purposes. While properties of a compound such as antibiotic activity can thus be efficiently measured, the actual role that a molecule plays in the microbiome is not always clear. Some specialised metabolites with well-established roles are for example the variety of metallophores that scavenge different metals necessary for enzymatic functions, such as iron, copper, molybdenum and vanadium (Dassama, Kenney, and Rosenzweig 2017; Wichard 2016; Bellenger et al. 2008). Recently, it was shown that the widely produced volatile terpenes geosmin and 2-methylisoborneol attract springtails to sporulating *Streptomyces* colonies, thereby promoting spore dispersal (Becher et al. 2020). The carcinogenic metabolite colibactin produced by *E. coli* has been shown to trigger prophage induction, and thereby induce changes in gut microbiomes (Silpe et al. 2022). Antibiotics, which are often regarded purely in their capacity to kill or arrest growth in bacteria, are being examined in a new light as potential signalling molecules with a variety of antagonistic and mutualistic effects depending on concentration and interaction partner (Romero et al. 2011; Tyc et al. 2017). However, the significance of many specialised metabolites continues to remain unknown. It is likely that one compound will often have many effects that emerge through the complex interactions resulting from the co-evolution of microbiomes, which might

be difficult to analyse in pure culture. In situ metatranscriptomic studies of BGC activity can help shine a light on the timing and conditions of BGC expression and thereby help elucidate the roles of encoded metabolites (Van Goethem et al. 2021). Examining the up- or downregulation of a BGC in manipulation experiments could also help guide prioritisation of BGCs to recover a desired activity.

2 Results 1: Biosynthetic potential of uncultured Antarctic soil bacteria revealed through long-read metagenomic sequencing

2.1 Introduction

Over the last decade, metagenomics has shown that a vast amount of the bacterial diversity on Earth is comprised of uncultured bacterial taxa, with 97.9% of bacterial operational taxonomic units (OTUs) estimated as unsequenced (Z. Zhang et al. 2020). First efforts to characterise and harness the specialised metabolite diversity encoded in metagenomes have shown promising results (Milshteyn, Schneider, and Brady 2014; Katz, Hover, and Brady 2016; Trindade et al. 2015). Metagenomic library screenings have yielded novel compounds, among them antibiotics (Katz, Hover, and Brady 2016; Hover et al. 2018; Libis et al. 2019), while sequence-based studies have documented their diversity. In a study of grasslands with 1.3 Tb of short-read sequence data, Crits-Christoph et al. recovered hundreds of metagenome-assembled genomes (MAGs) obtained through a combination of binning approaches (Crits-Christoph et al. 2018). Analysis of the MAGs revealed a large number of BGCs in Acidobacteria and Verrucomicrobia, widespread but underexplored phyla of soil bacteria. Analysis of nonribosomal peptide synthetase (NRPS) and polyketide synthase (PKS) domains indicated that NRPS and PKS from these groups were highly divergent from known BGCs of these classes. Borsetto et al. also reported a high degree of diversity of NRPS and PKS domains in Verrucomicrobia and other difficult-to-culture phyla (Borsetto et al. 2019). Finding efficient ways to access this treasure trove of diverse and unexplored specialised metabolites will expand our understanding of microbial natural products, yield novel and useful compounds, and be an important step towards the development of much-needed antimicrobials.

Recent advances in long-read sequencing technology have made it possible to recover largely complete genomes from metagenomic sequencing projects. A sequencing effort of 26 Gb returned 20 circular genomes from human stool samples (Moss, Maghini, and Bhatt 2020), while a recent study using 1 Tb of long-read data from wastewater treatment plants recovered thousands of high-quality MAGs, 50 of which were circular (Singleton et al. 2020). Using mock community data, Pérez et al. demonstrated that full-length BGCs could be successfully recovered from long-read metagenomic sequencing (Latorre-Pérez et al. 2019). In light of recent advances in PCR-based cloning techniques that comprise heterologous expression of BGCs based on PCR amplification (Greunke et al. 2018; D'Agostino and Gulder 2018; Qian et al. 2020; R. Duell et al. 2020), the recovery of full-length metagenomic BGC sequences is promising as these sequences would be amenable to PCR amplification. At the time of the study, no BGC discovery from environmental long-read metagenomes had been reported. However, one such study using PacBio SMRT sequencing has been released since then (Van Goethem et al. 2021).

In recent years, a number of tools to explore and understand BGC diversity have been developed. Genomes can be mined for known classes of BGCs using tools such as antiSMASH (Blin, Shaw, et al. 2019), while the MiBiG database (Kautsar, Blin, et al. 2020) links BGCs to known compounds. BGCs can be compared in networking-based tools such as BiG-SCAPE (Navarro-Muñoz et al. 2020) and BiG-SLiCE (Kautsar, Hooft, et al. 2020) to assess relations of BGCs and estimate their novelty relative to extant BGC databases.

The isolated, harsh and unique environments of Antarctica show high degrees of endemism in their bacterial life, but their diversity remains underexplored (Kleinteich et al. 2017). Little is known about the specialised metabolites of Antarctic microorganisms. Few studies have

explored polar, and specifically Antarctic, natural products using functional screening of isolates and metabolomics (T. R. Silva et al. 2018; Shekh et al. 2011; Mojib et al. 2010; Giudice, Bruni, and Michaud 2007; Millán-Aguiñaga et al. 2019). A high number pigmented bacterial isolates indicates that carotenoids and PKS, among other pigments, could be abundant BGC classes (Dieser, Greenwood, and Foreman 2010). One culturing study suggested that Antarctic isolates show a below average potential for antimicrobials (T. R. Silva et al. 2018). On the other hand, a primer-based study showed a promising diversity of NRPS and PKS diversity in soil from Mars Oasis in the southern maritime Antarctic (Borsetto et al. 2019), a site with exceptionally high diversity of micro- and macroorganismal life for its latitude (Yergeau et al. 2007; Pearce et al. 2012). Low-temperature, aerated Antarctic soils have previously also been linked to methanotrophy (Lau et al. 2015; Edwards et al. 2017), and these soils could therefore harbour methanobactins, small ribosomally synthetised peptides that scavenge copper needed for methane monooxygenases.

2.2 Materials and Methods:

2.2.1 Site description

Mars Oasis is situated on the south-eastern coast of Alexander Island in the southern maritime Antarctic at 71° 52' 42" S, 68° 15' 00" W (Figure 2.2A). Mean soil pH is 7.9, with NO_3^--N and NH_4^+-N concentrations of 0.007 mg kg^{-1} and 0.095 mg kg^{-1}, and total organic C, N, phosphorus and potassium concentrations of 0.26%, 0.02%, 8.01% and 0.22%, respectively. Soil moisture concentrations range between 2% and 6% in December–February, when snow or rainfall events are very rare, with the majority of precipitation falling as snow between March and November. Mars Oasis has a continental Antarctic climate, with frequent periods of cloudless skies during summer, when temperatures at soil surfaces reach 19 °C. During midwinter, the temperatures of surface soils decline to -32 °C. Mean annual air temperature is *c*. -10 °C (Misiak et al. 2021).

2.2.2 Soil sample, extraction and sequencing

One sample of surface soil (*c*. 0-50 mm, *c.* 2.5 kg) was collected with clean spades from the lower terrace at Mars Oasis (S71 52.691, W68 14.943) by British Antarctic Survey staff on 8 December 2017 and was kept cool for several hours before being stored at -20 °C. Soil was kept at this temperature until being thawed for DNA extraction. A gentle chemical lysis and DNA extraction of 50 g of soil were performed and the DNA was subjected to size selection to approximately 20 Kb and larger by agarose gel electrophoresis using a protocol previously used for metagenomic library construction (Brady 2007). DNA was sequenced using Oxford Nanopore Technologies (ONT) MinION and Illumina HiSeq 150 bp paired-end reads. For long reads, the DNA was sequenced using three R9.4.1 flow cells and the SQK-LSK109 kit. The nuclease flush protocol was used between each independent library run on a flow cell. Short read DNA library preparation and Illumina sequencing were performed by Novogene

according to their in-house pipeline. In short, one µg of DNA was sheared to 350 bp, then prepared for sequencing using NEBNext DNA Library Prep Kit. The library was enriched by PCR and underwent SPRI-bead purification prior to sequencing on a HiSeq sequencing platform.

2.2.3 Read processing, assembly, polishing and quality control

A flowchart of the processing is provided in Figure 2.1. The long reads fast5 data were basecalled with Guppy v.3.03 (HAC model). Basecalled raw reads were assembled using Flye v2.5 using the --meta flag. The resulting assembly was polished with 4 iterations of Racon (Vaser et al. 2017) v1.4.7 followed by one run of Medaka (*Nanoporetech/Medaka* 2017) v0.7.1. Then, the short reads were used for six rounds of polishing with pilon (Walker et al. 2014) v1.23. The approximate assembly quality was checked at every step using ideel (Watson 2018). Long reads were also classified with kraken2 2.0.7b using the GTDB r89 database. Short reads were used to estimate diversity and predict coverage with nonpareil 3.304. Furthermore, short reads were assembled with SPAdes 3.14.1 using the --bio flag ("biosyntheticSPAdes"). Read and assembly statistics can be found in Table 2.1. Initial assessment of potential indels showed that 82% of all proteins were shorter than 0.9 times the length of the closest reference protein in the UniProt database and 7.2% were longer than 1.1 times the length of the closest reference protein. After polishing using Racon, Medaka and pilon, the proportion of potentially truncated proteins was reduced to 70%, while that of proteins that were potentially too long slightly increased to 7.6%.

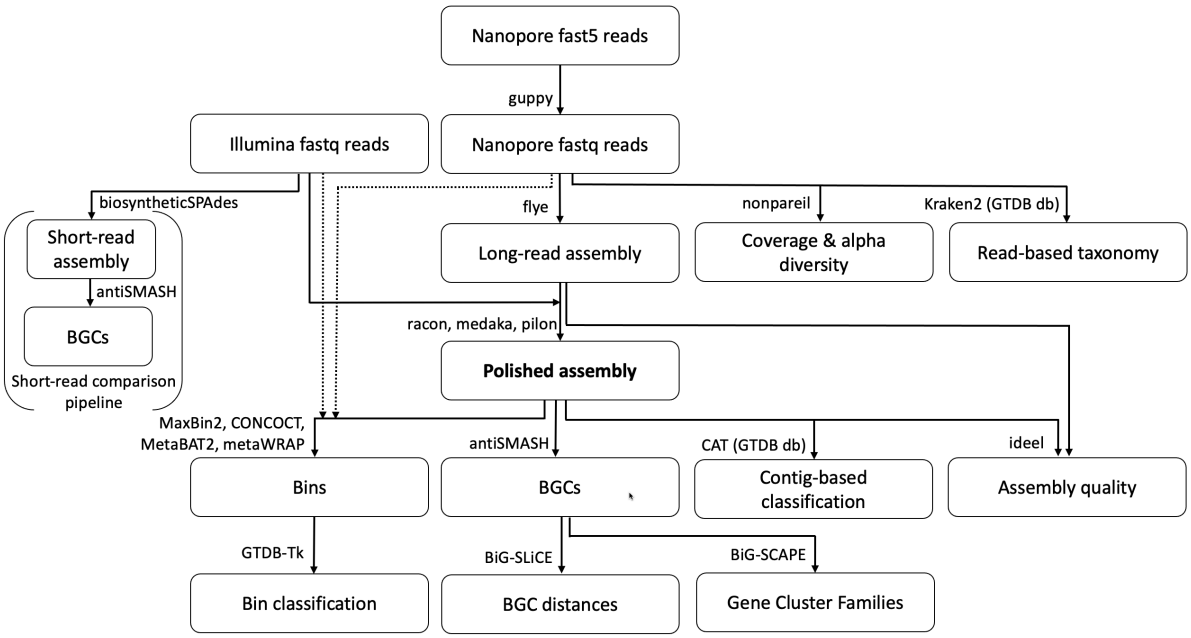


Figure 2.1: Flowchart of the bioinformatic processing of the sequence data.

2.2.4 Genome mining, binning, taxonomic assignment and quality control

For detecting BGCs, the polished assembly was analysed by antiSMASH (Blin, Shaw, et al. 2019) v5.1. For taxonomic assignment of contigs, proteins were predicted using Prodigal (Hyatt et al. 2010), and CAT (von Meijenfeldt et al. 2019) (settings --sensitive -r 10 and -f 0.3) was used with a DIAMOND (Buchfink, Xie, and Huson 2015) database built from proteins in the GTDB_r89_54k database (Parks et al. 2018) as well as the NCBI non-redundant protein database. The contigs were also binned with MetaBAT2 (D. D. Kang et al. 2019), CONCOCT (Alneberg et al. 2014) and MaxBin2 (Wu et al. 2014, 2), using long- and short-read abundance profiles generated with bowtie2 (Langmead and Salzberg 2012) and minimap2 (H. Li 2018, 2) as a proxy for differential coverage. The resulting bins were subjected to metawrap-refine (Uritskiy, DiRuggiero, and Taylor 2018) to produce the final bins and classified using GTDB-Tk 0.3.2 (r89). BiG-SCAPE (Navarro-Muñoz et al. 2020) 1.0.1 was run in --auto mode with --mibig enabled to identify BGCs families. Networks using similarity thresholds of 0.1, 0.3, 0.5 and 0.7 were examined, since higher thresholds led to extensively large proposed BGC

families. In order to calculate BGC novelty, BiG-SLiCE 1.1.0 (Kautsar, Hooft, et al. 2020) was run in --query mode with a previously prepared dataset which had been computed from 1.2 million BGCs using --complete_only and t = 900 as threshold (Kautsar et al. 2021). The resulting distance *d* indicates how closely a given BGC is related to previously computed gene cluster families (GCFs), with a higher *d* indicating higher novelty. For this analysis, the values of $d > t$ and $d > 2t$ (i.e. $d > 900$ and $d > 1800$, respectively) were highlighted, as they were previously suggested as arbitrary cutoffs for "core", "putative" and "orphan" BGCs (Kautsar et al. 2021).

2.2.5 Precursor peptide homology searches and sequence logo construction

ORFs were aligned using Clustal Omega (Sievers et al. 2011) and a HMMER (Finn, Clements, and Eddy 2011) search was performed in the EBI reference proteome database with a cut-off E-value of 1E-10. The resulting protein sequences were aligned using Clustal Omega and a HMM was generated and visualised using skylign.org (Wheeler, Clements, and Finn 2014).

2.3 Results

2.3.1 Soil diversity, taxonomic classification and binning of BGCs

Nonpareil analysis estimated an abundance-weighted coverage of 85.3% for the 44.4 Gb used in the long-read assembly. To achieve 95% and 99% coverage respectively, 250 Gb and 1.6 Tb of sequencing were predicted to be necessary. Alpha diversity was estimated at N_d = 21.6. Contigs were binned using CONCOCT, MaxBin2 and MetaBAT2, consensus bins were generated using metaWRAP refine and classified using GTDB-Tk. This yielded 114 bacterial bins with CheckM completeness > 50% and contamination < 10% containing 278 BGCs (see Table 2.1) Since only 278 BGCs had been binned, an additional contig-based classification approach was adopted. All contigs were classified using CAT with a database based on Genome Taxonomy Database (GTDB) r89 proteins, leading to a classification of 93% of BGC-containing contigs at a phylum level (Figure 2.2B-C). A cross-check of bin-level classification and contig-level classification of the 269 binned and CAT-classified BGC-containing contigs showed three conflicts at different levels in total (phylum: 0, class: 1, order: 1, family: 0, genus: 1, species: 0). Of the 2,892 total binned and CAT-classified contigs, 52 (1.7%) were classified differently at order level using CAT. This

Table 2.1: Raw sequence, polished assembly, BGC mining and binning statistics

Nanopore reads	No. of reads	9.3 million
	Total length	44.4 Gb
	N50	9.4 Kb
150bp PE Illumina reads	No. of reads	186.6 million
	Total length	28 Gb
Nonpareil analysis	Abundance-weighted coverage at 44.4 Gb	85.3%
	Diversity N_d	21.6
Polished assembly	No. of contigs	48,422
	length	2.4 Gb
	N50	84.8 Kb
	Max length	129.6 Kb
antiSMASH BGCs	No. of BGCs	1417
	BGCs on contig edge	564
	Total length	40.5 Mb
	Mean length	28.5 Kb
	Max length	129.6 Kb
metaWRAP 50/10 bins	No. of bins	114
	Mean no. of contigs per bin	18.5
	BGCs in bins	278
	Average bin N50	224 Kb

indicates that the risk of misclassification of BGC-containing contigs by CAT is low but cannot be excluded. Bin-level classification was preferred where available.

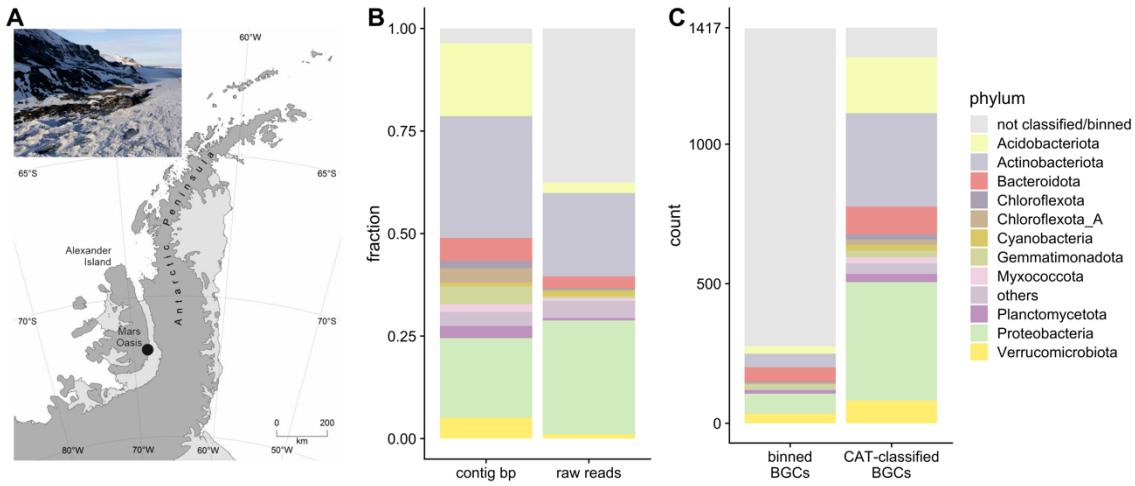


Figure 2.2: Location and phylogenetic classification overview.(A): Map of the Antarctic Peninsula with Mars Oasis indicated. Inset: Aerial photo of the site taken in austral summer (obtained from Kevin Newsham, personal communication); (B): Phylogenetic classification of contigs (by CAT) and long reads (by kraken2); (C) phylogenetic classification of BGC-containing contigs using binning and CAT classification approaches.

2.3.2 Recovery of diverse and full-length BGCs

The polished assembly was analysed using antiSMASH v5.1. A total of 1417 BGCs were identified on 1,350 contigs. A total of 564 BGCs (39.8%) were identified as being on a contig edge and were therefore categorised potentially incomplete, while 853 (60.2%) were full-length. The most abundant classes of BGCs were terpenes (27.2%), followed by NRPS (15.7%) and bacteriocins (10.1%). In particular, terpenes were dominated by few sub-classes. Out of 401 observed terpene BGCs, 321 contained a squalene/phytoene synthase Pfam domain (PF00494). This indicates that the product of these BGCs is a tri- or tetraterpene. Forty-four BGCs also contained a squalene/hopene cyclase (N terminal; PF13249), 39 BGCs contained a carotenoid synthase (PF04240), while 47 contained a lycopene cyclase domain (PF05834).

Approximately half of the ribosomally synthesized and post-translationally modified peptides (RiPPs) identified in the sample contained methanobactin-like DUF692 domains (PF05114). However, no BGCs resembling known methanobactin BGCs were found.

The proportion of proteins identified as too short on BGC-containing contigs was estimated at 63%. It is possible that this measure was influenced by the UniProt reference database not containing representative proteins for the mostly uncultivated strains recovered in this study. However, fragmentation of ORFs through indels was clearly visible, especially in NRPS and PKS BGCs in which whole megasynthase genes were broken up into several fragments.

2.3.3 Long reads and GTDB improve phylogenetic classification of environmental BGCs

The use of GTDB proteins instead of the NCBI non-redundant protein database increased the classification success of BGC-containing contigs from 36.8% classified at order level with the NCBI database to 71.8% with GTDB. The difference was mainly due to BGCs from MAG-derived orders which were not present in the NCBI database, such as UBA7966. However, the GTDB database is also much smaller than the NCBI nr database, and many MAG-derived clades especially at lower taxonomic ranks do not have many representatives in the GTDB database. To avoid misclassifications, analysis was conducted at class and order level, even if contigs were classified at lower taxonomic ranks.

To assess the advantages of long-read sequencing for BGCs detection and classification, the output was compared with BiosyntheticSPAdes, which allows the assembly of NRPS and PKS from short-read sequences by following an ambiguous assembly graph using *a priori* information about their modularity. Using BiosyntheticSPAdes with the 28 Gb of short reads, 228 unambiguous NRPS/PKS BGCs were predicted. Sixty-one of these were above 5 Kb long

and five NRPS were larger than 30 Kb. Furthermore, 202 other BGCs were predicted from other contigs. 96.7% of BGCs were marked as on a contig edge, i.e. not full-length. Indeed, 392 out of 430 BiosyntheticSPAdes BGCs could be aligned to 255 long-read BGCs using blastn (E-value < 1E-90), indicating that mostly the same BGCs were assembled, but they were fragmented in the short-read assembly (Figure 2.3). In the case of NRPS/PKS BGCs, even the BGCs on contigs with the highest coverage (>120x) were fragmented into two or more contigs. Classification success using the same binning and CAT approach was lower (68% at phylum level, 50% at order level; 48 BGCs binned). This could be attributed to the lack of genomic context around the BGCs. While BiosyntheticSPAdes predicted a large number of BGCs in total, the practical usability and interpretability of the output remained low, since completeness, cluster borders and potential modification genes could not be assessed and phylogenetic classification success was reduced.

2.3.4 Highly divergent BGCs found in unusual specialised metabolite producer phyla

Examination of the BGC counts by BGC type and phylum showed that the three well-known producer phyla Actinobacteriota (NCBI taxonomy: Actinobacteria), Proteobacteria and Bacteroidota (NCBI taxonomy: Bacteroidetes) together contributed over 60% of BGCs (Figure 2.4A). BGCs attributed to Acidobacteriota and Verrucomicrobiota represented up to 20% of the total BGCs, while other phyla constituted the remaining 12%, and 7% remained unclassified at phylum level. In particular, 20% of NRPS remained unclassified at phylum level. No archaeal BGCs were found.

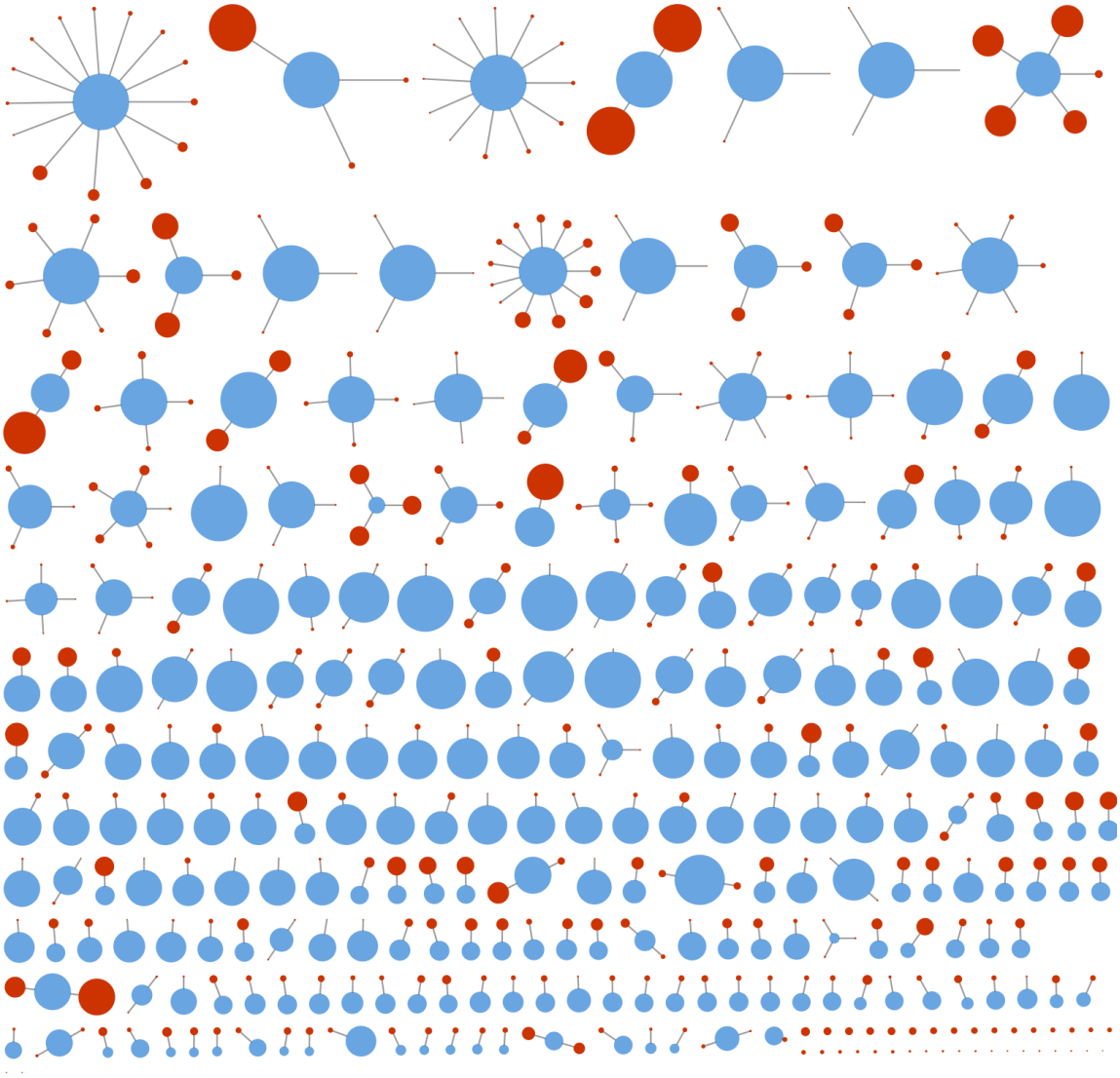


Figure 2.3: Network visualisation of short-read derived BGCs (red) aligning to long-read derived BGCs (blue). The size is approximately proportional to BGC length as defined by antiSMASH borders. It is visible that most short-read derived BGCs are shorter than long-read derived BGCs. For few high-coverage BGCs, biosyntheticSPAdes assembled the same parts of the BGC in the first assembly phase as well as in the NRPS/PKS-specific assembly phase, leading to a longer combined length (and thereby larger dot size) than the long-read derived BGC.

The 1417 BGCs were then analysed with BiG-SLiCE's query mode in order to calculate their distance (*d*) from a set of pre-computed gene cluster families (GCFs) comprised of 1.2 mio known BGCs. The analysis showed that 845 out of 1417 BGCs (59.6%) had a $d > 900$, indicating that they were only distantly related to a GCF. Fifty-five outliers were found with

d > 1800, indicating extremely divergent BGCs. A wide span of distances was present within each phylum which indicates that each phylum contained BGCs that are both closely and distantly related to known BGCs (Figure 2.4B). The median distances showed significant variation between phyla, with Bacteroidota containing the highest novelty (median *d* = 1227) and Planctomycetota the lowest (median *d* = 742). This overall score was, however, influenced by the fact that different classes of BGC scored differently. For example, NRPS/PKS BGCs scored higher than e.g. terpenes or bacteriocins. Rankings of single BGC classes showed that the high Bacteroidota score was partly driven by the large number of NRPS (Figure 2.4C) and the small number of terpenes and bacteriocins (Figure 2.4E-F) in the phylum. This is evidenced by the fact that other phyla scored the highest in individual BGC classes. For NRPS BGCs, Gemmatimonadota, Acidobacteriota and Verrucomicrobiota showed the highest values for *d* (Figure 2.4C). Gemmatimonadota furthermore showed the highest value for *d* when considering terpene BGCs (Figure 2.4E), while Acidobacteriota scored high for lassopeptides, arylpolyenes and PKS (Figure 2.4 G,H,D). Furthermore, BGCs on a contig edge tended to score lower. To check whether low coverage and the resulting insertion and deletion errors in the assembly led to overestimation of *d*, contig coverage as well as percentage of correctly-sized ORFs (as calculated by ideel) were plotted against *d*. There was a positive correlation of *d* values with increased coverage up until a coverage of ca 10, indicating an underestimation of novelty at low coverage. Similarly, for contigs with under 20% correctly-sized ORFs, there was a slight positive correlation between the percentage of correctly-sized ORFs and distance. As expected, coverage showed a strong positive correlation with percentage of correctly-sized ORFs (Figure 2.6).

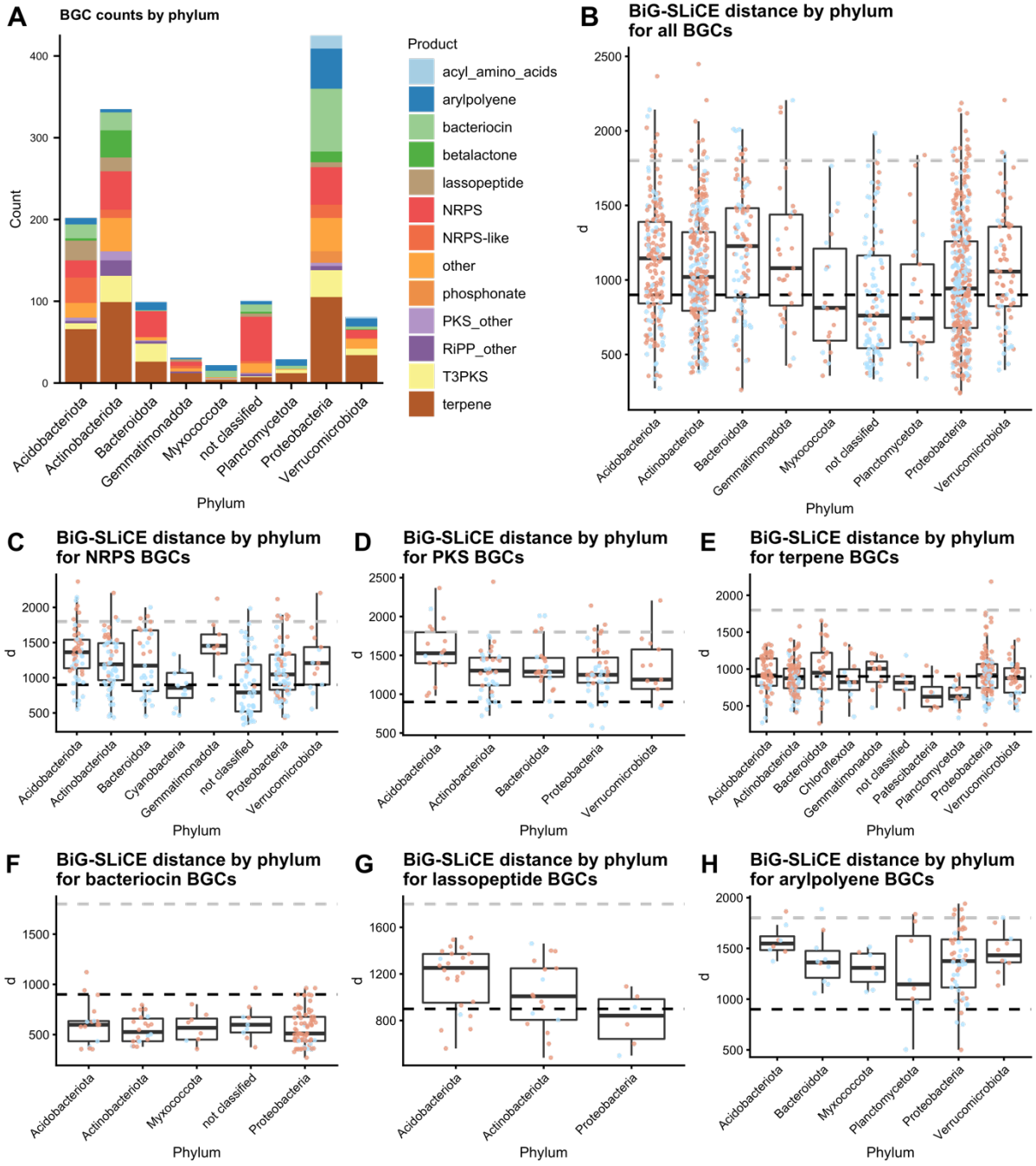


Figure 2.4: BGC distribution and BiG-SLiCE scores. (A) BGCs by phylum and BGC type (phyla with a count <20 removed; products with count <10 under "others", (B) BiG-SLiCE distances of BGCs by phylum, with the black dotted line indicating d = 900 and the grey dotted line d = 1800 (phyla with a count <20 removed); (C-H) BiG-SLiCE distances for different BGC types plotted by phylum (phyla with < 5 BGCs of the type removed; hybrid BGCs counted for both classes). Each point indicates a BGC. Salmon = BGC not on contig edge, Light blue = BGC on contig edge.

2.3.5 Acidobacterial BGCs

Analysis of acidobacterial BGCs by order (Figure 2.5A) showed that terpenes were the most numerous, but with significant contributions from PKS, NRPS, lassopeptide and bacteriocin clusters. The orders of Pyrinomonadales and Vicinamibacterales constituted >60% of BGCs.

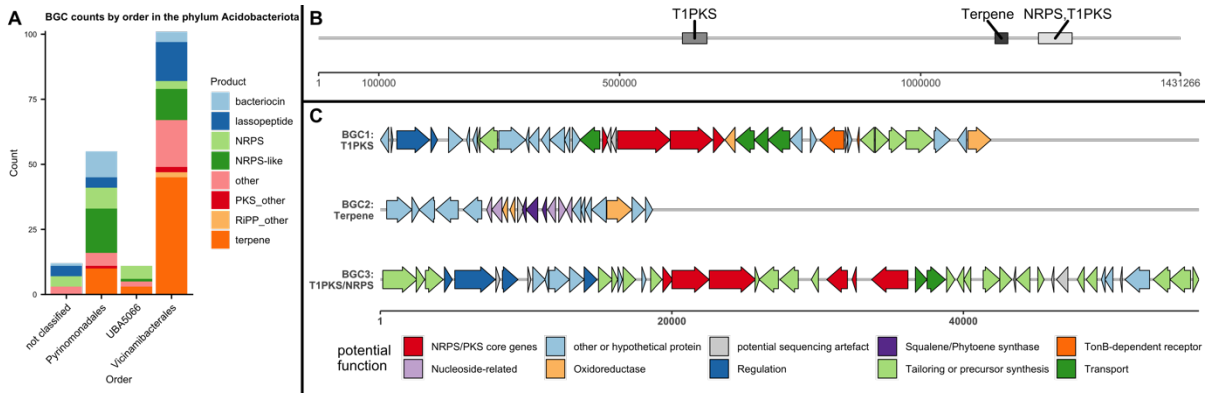


Figure 2.5: Acidobacteria BGCs.(A) BGC counts by BGC type and order in phylum Acidobacteriota; (B) Map of a large Acidobacteriota contig (order Vicinamibacterales) and the BGCs on it (C) Cluster map of proposed functions of genes in BGC1, BGC2 and BGC3. Functions were predicted from BLASTing against NCBI nr database as well as antiSMASH module predictions.

BiG-SCAPE analysis showed that BGCs mainly clustered together within orders. None of the families contained MiBiG clusters at the cut-offs used. Acidobacteriota showed a large number of lassopeptides, 16 of which grouped into two GCFs. NRPS-like BGCs also contributed a large number to the sample. In particular, one NRPS-like family from the order Vicinamibacterales showed homology to the VEPE BGC from *Myxococcus xanthus* (MiBiG BGC0000871). Furthermore, seven NRPS/PKS with a megasynthase gene length of over 20 Kb were found with the largest BGC measuring 89 Kb of NRPS and PKS megasynthase genes. The largest Acidobacteriota (order Vicinamibacterales) contig was 1.5 Mb in size and contained three BGCs: a PKS, a terpene and a NRPS/PKS hybrid cluster (Figure 2.5B,C). BGC1 (d = 1397) contained a partial one-module NRPS followed by a partial PKS module as well as transporter genes and a TonB-dependent receptor protein, suggesting a role as a

siderophore. BGC2 (d = 1103) contained squalene/phytoene synthase genes and several potential tailoring enzymes. BGC3 (d = 1977) contained a complete NRPS and a partial NRPS module and an incomplete PKS domains. Several gaps visible in the BGC make a sequencing error seem possible, leading to truncated genes and therefore missing domains.

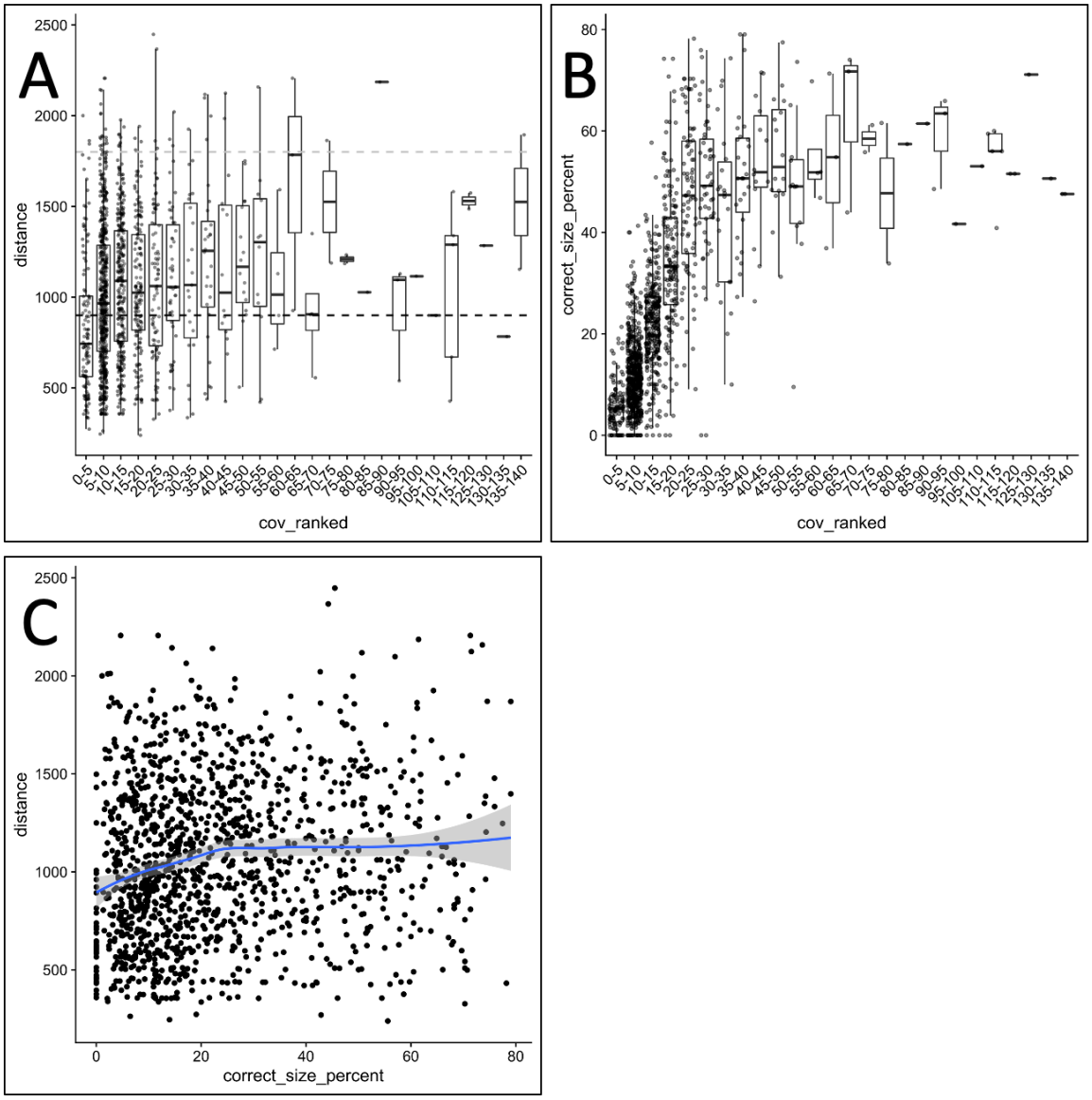


Figure 2.6: Coverage relating to distance and "correct size" ORFs with each point indicating a BGC. (A) Ranked coverage compared to BiG-SLiCE distance. Spearman's correlation using all data points showed a ρ of 0.178 with a highly significant p = 1.606e-11. When coverages 0-5 and 5-10 were removed, ρ was reduced to 0.057 and p increased to a non-significant 0.125. (B) Ranked coverage of BGC-containing contigs compared to percentage of "correct size" ORFs on same contig. "correct size" is defined by being between 0.9 to 1.1-times the length of a reference protein as calculated by ideel. A strong trend between coverage and number of "correct size" ORFs is visible up until coverage 25-30 (Spearman's ρ = 0.82, p = 2.2e-16); (C) Percentage of "correct size" ORFs on a BGC-containing contig compared to bigslice distance of the BGC. A LOESS curve was fitted. Spearman's rank correlation showed a small correlation (ρ = 0.163, p = 1.851e-9). No correlation was observed when data points with correct_size_percent below 20 were removed (ρ = 0.008, p = 0.8355).

2.3.6 Verrucomicrobial BGCs

The analysis of Verrucomicrobial BGCs by order (Figure 2.7A) showed that the vast majority of BGCs were terpenes, followed by arylpolyenes, PKS, NRPS, as well as ladderanes. The most prolific producer orders were Opitutales, Pedosphaerales and Chtoniobacterales.

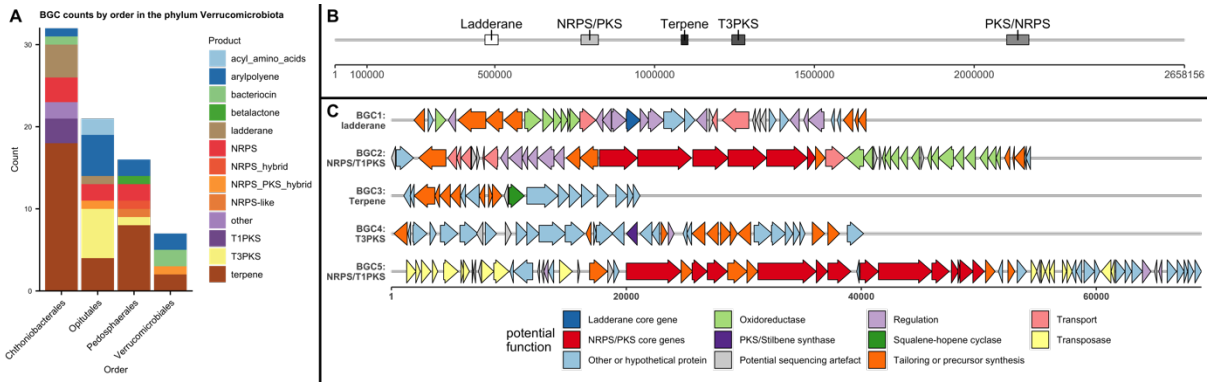


Figure 2.7: Verrucomicrobial BGCs. (A) BGC counts by BGC type and order in phylum Verrucomicrobiota, (B) map of a large Verrucomicrobiota contig (order Opitutales) and the BGCs on it; (C) Cluster map of proposed functions of genes in BGC1 – BGC5. Functions were predicted from BLASTing against NCBI nr database as well as antiSMASH module predictions. X axis represents basepairs.

Verrucomicrobial BGCs did not show strong clustering into conserved GCFs compared to Acidobacteriota. One NRPS and one PKS BGC were the only BGCs that clustered with MiBiG clusters. The largest Verrucomicrobiota contig (order Opitutales) was 2.6 Mb in size and featured five BGCs, two of which were NRPS-PKS hybrids with megasynthase genes above 20 Kb (Figure 2.7B, C). BGC1 (d = 1479) contained a ladderane-type 3-oxoacyl-[acyl-carrier-protein] synthase. BGC2 (d = 1305) contained four NRPS modules interspersed by one PKS module. BGC3 (d = 673) contained a squalene-hopene cyclase, indicating a role in hopanoid biosynthesis. BGC4 (d = 1142) encoded a chalcone/stilbene synthase. BGC5 (d = 1340) contained a PKS module followed by five NRPS modules. The third module, however, showed a truncated A domain, with the antiSMASH HMM NRPS-A_a3 only matching around 50 bp at the end of ORF ctg423_1968. This could be explained by a sequencing error in which an indel lead to a frameshift, causing a premature stop codon. Indeed, nucleotide-level BLAST of

the gap between ctg423_1968 and the PCP-domain containing ctg423_1970 showed a match to known A domains. It is, however, not possible to rule out potential pseudogenisation.

2.3.7 Uncultivated and underexplored classes and orders from Actinobacteriota and Proteobacteria show a large biosynthetic potential

2.3.7.1 Actinobacteriota: Acidimicrobiia and Thermoleophilia

The phylum Actinobacteriota (335 BGCs) featured a large amount of BGCs unclassified at order level. Therefore, they were analysed by class (Figure 2.8A). The class Actinobacteria (114 BGCs) contained BGC-rich genera such as *Streptomyces* and *Pseudonocardia* and accordingly contributed a large amount of BGCs in the sample. The class Acidimicrobiia (90 BGCs) contained the genera *Illumatobacter* and *Microthrix* and several uncultured genera. The class Thermoleophilia (95 BGCs) contained genera such as *Solirubrobacter* and *Patulibacter*, besides uncultured genera, and contributed to a large amount of the bacteriocin and betalactone BGCs. The amount of BGCs in these classes that were not placed into lower taxonomic ranks indicated that there is a large unexplored diversity of uncultured Actinobacteriota containing a great diversity of BGCs.

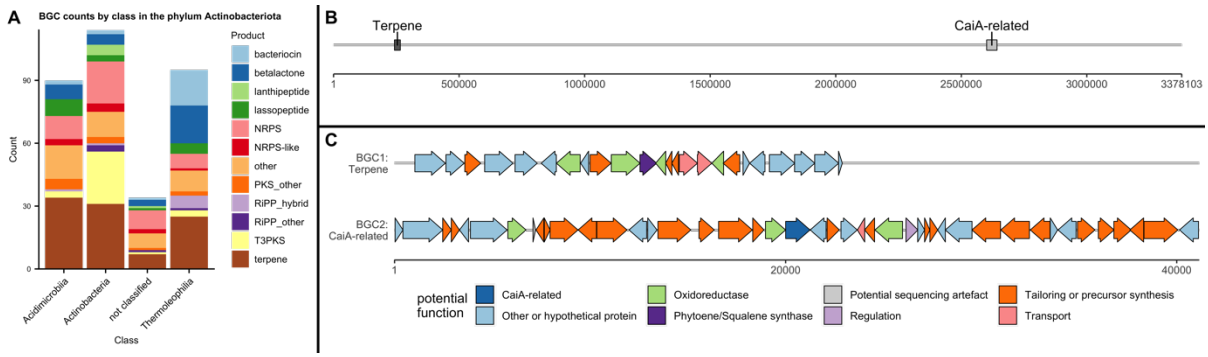


Figure 2.8: Actinobacteriota BGCs. (A) BGC counts by BGC type and class in Actinobacteriota; (B) Map of a large Actinobacteriota contig (order IMCC26256) and number of basepairs; (C) Cluster map of proposed functions of genes in BGC1 and BGC2. Functions were predicted from BLASTing against NCBI nr database as well as antiSMASH module predictions. X axis represents basepairs.

Remarkably, one circular genome from the uncultured order IMCC26256 from the class Acidimicrobiia was recovered in a single contig, measuring 3.3 Mb in size and containing two BGCs (Figure 2.8B-C). The terpene BGC (*d* = 1398) contained a squalene synthase, a lycopene cyclase and polyprenyl synthetases, suggesting a role in pigment formation. The CaiA-related BGC (*d* = 1869) contained an acyl-CoA dehydrogenase related to CaiA (involved in saccharide antibiotic BGCs). BLAST hits indicated other genes related to small organic acids, sugars and nucleoside metabolism.

Two families of terpenes containing terpene cylases, methyltransferases and/or P450s showing similarity to the known geosmin and 2-methylisoborneol BGCs were found, with members belonging to both Acidimicrobiia, Thermoleophilia and unclassified Actinobacteriota. One BGC from a *Streptomyces* spp. was detected, containing a LmbU-like gene on the very edge of the contig. BiG-SCAPE analysis showed that Actinobacteriota BGCs mostly grouped within the classes, and one lanthipeptide BGC grouped with MiBiG BGCs at the cut-off used.

2.3.7.2 Proteobacteria: the uncultured methanotrophic order UBA7966 as a specialised metabolite producer

Analysis at the order level of the proteobacterial BGCs showed that the biggest contributor was the Burkholderiales order with 116 BGCs (Figure 2.9A) followed by order UBA7966 with 96 BGCs. UBA7966 BGCs included a variety BGCs, including terpenes, bacteriocins, phosphonates, NRPS & NRPS hybrids, NRPS-like, and arylpolyenes. In particular, the high abundance of NRPS-like and phosphonate BGCs in UBA7966 contrasted with the lower counts in other proteobacterial orders in the dataset. By order, UBA7966 contigs also showed a high average coverage 26x, compared to the total average of 10.2x, indicating a high abundance. The total length of UBA7966 contigs was 53 Mb, indicating the presence of several genomes.

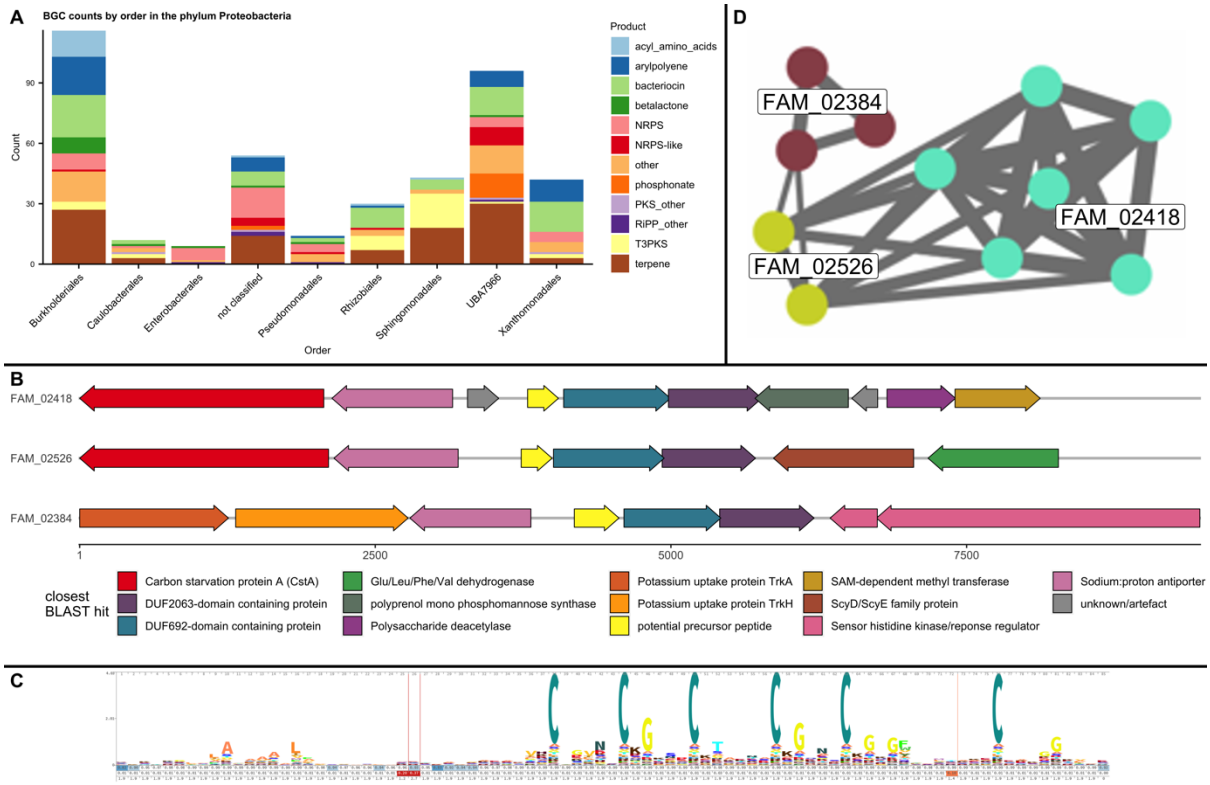


Figure 2.9: Proteobacteria BGCs and DUF692 BGCs. (A) BGC counts by BGC type and order in the phylum Proteobacteria; (B) Cluster layout of three gammaproteobacterial DUF692-containing BGCs representatives: contig_12391 for FAM_02418, contig_14956 for FAM_02526, and scaffold_15362 for FAM_02384; (C) Sequence logo generated from an HMM of 301 potential precursor peptides; (D) Similarity network generated from BiG-SCAPE with brown: FAM_02384, turquoise: FAM_02418, green: FAM_02526.

The order UBA7966 is an uncultured order consisting of one family, UBA7966, which contains two genera, *UBA7966* and *USCγ-Taylor*. UBA7966-family bin bin.3 was assigned no genus, while all CAT-assigned contigs were assigned species *USCγ-Taylor sp002007425*, the only species in the *USCγ-Taylor* genus. The *USCγ-Taylor* genus is based on a putatively methanotrophic MAG extracted from a methane-oxidising soil metagenome from Taylor Valley in Antarctica (Genbank accession GCA_002007425.1) (Edwards et al. 2017). The low number of UBA7966 reference genomes in the GTDB database means, however, that these classifications remain an approximation. The two closest orders to UBA7966 that contain cultured representatives, Beggiatoales and Nitrosococcales, both have members implicated in methanotrophy, sulphur cycling and ammonia oxidation as well as varying degrees of

chemolithotrophy and chemoautotrophy (Zopfi et al. 2001; Sweerts et al. 1990; Klotz et al. 2006; Boden et al. 2010). On all the contigs assigned to order UBA7966 by CAT, four *pmoCAB* operons were found, with *pmoA* showing 92.9% to 96.8% identity with *pmoA* from *USCγ-Taylor*. This indicates that, in addition to the methanotrophy of *USCγ-Taylor*, other members of the order UBA7966 could be involved in similar lifestyles.

When analysed with BiG-SCAPE at cut-off 0.7 phosphonates (median d = 1421), NRPS/NRPS-like (median d = 1262) and bacteriocins seemed to form especially conserved GCFs. Other GCFs were shared with other proteobacterial orders. With 96 BGCs, UBA7966 contributed a similar number of BGCs as the established specialised metabolite producing order Burkholderiales (116 BGCs). However, the BiG-SLiCE distances of UBA7966 were higher than Burkholderiales for all but one BGC class, indicating more novel BGCs (Figure 2.10).

The potential methanotrophy of UBA7966 suggested the potential presence of methanobactins, but no BGCs corresponding to known methanobactins were found in the dataset. On the other hand, an abundance of DUF692-containing BGCs were observed, grouping into three GCFs. DUF692 proteins are a diverse family of proteins with largely unknown functions, although some are known to be involved in methanobactin biosynthesis (Dassama, Kenney, and Rosenzweig 2017). The analysis of three related GCFs containing DUF692 domains (including BGCs from UBA7966 and unclassified gammaproteobacterial contigs) showed that FAM_02526 (two BGCs), FAM_02384 (three BGCs) and FAM_02418 (six BGCs) (Figure 2.9B-D) all contained a short (circa 240 bp) ORF followed by first a DUF692-domain containing protein, then a DUF2063-domain containing protein. Furthermore, a putative cation antiporter was found upstream of the precursor peptide. The three families differed by the genes

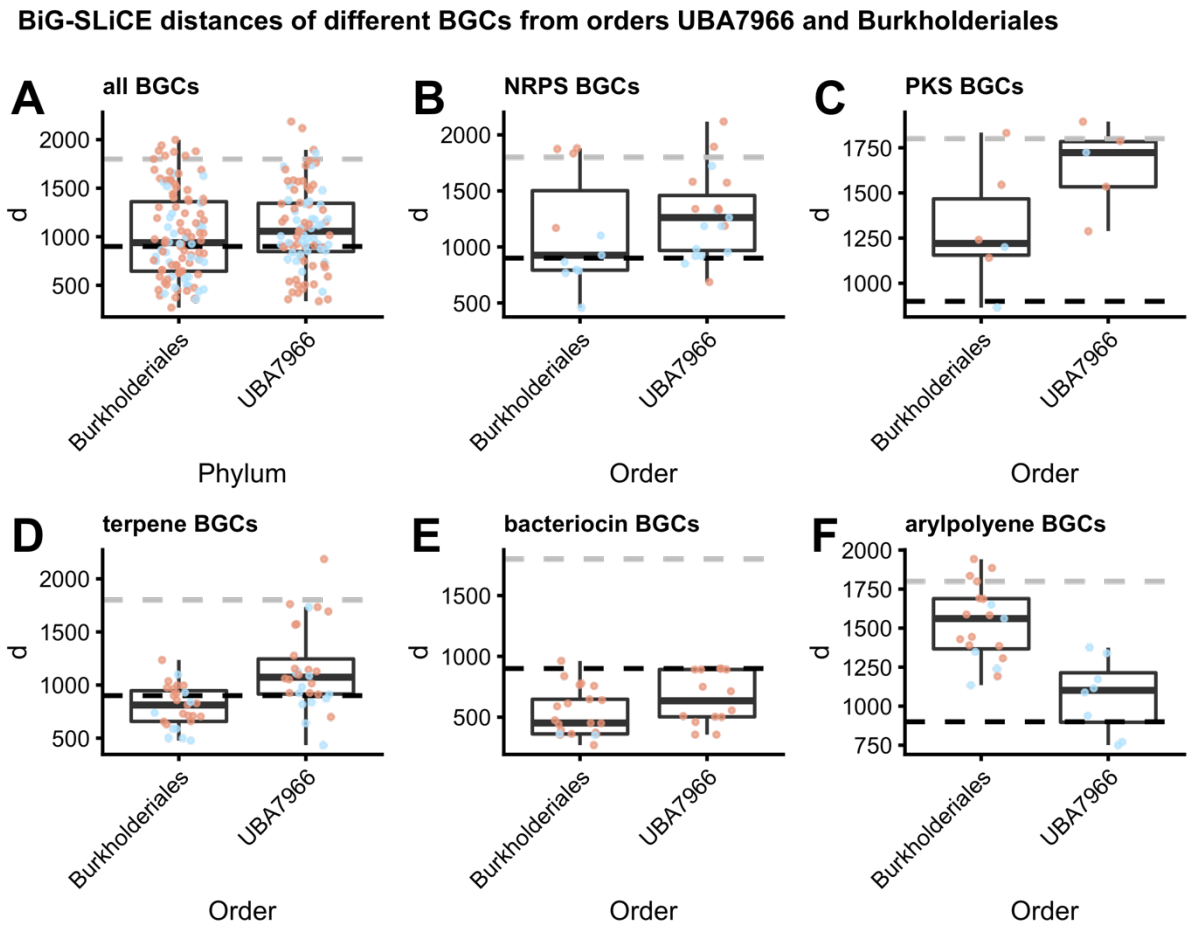


Figure 2.10: Comparison of distances of BGCs of different classes between UBA7966 and Burkholderiales orders. (A) all BGCs; (B-F) specific BGC classes. Each point indicates a BGC. Salmon = BGC not on contig edge, Light blue = BGC on contig edge

surrounding this core cluster (Figure 2.9B). The 11 small translated 240bp ORFs were aligned using Clustal Omega and a HMM search was made in ebi reference proteome database with a cut-off E-value of 1E-10. The resulting 290 protein sequences (almost exclusively from Proteobacteria) plus 11 original sequences were aligned using Clustal Omega and a HMM was generated and visualised using skylign.org. The resulting logo showed a low degree of sequence conservation except for a pattern of six conserved cysteines – some followed by glycines – within forty amino acids towards the N-terminus, and a slightly conserved hydrophobic patch towards the C-terminus (Figure 2.9C). This might represent a potential precursor peptide, with the six cysteines marking the potential core peptide. Additionally, a

Sec signal peptide was detected in the first 25 amino acids of the peptide, indicating export into the periplasm.

The UBA7966 order also contained larger BGCs such as four NRPS/ NRPS-PKS BGCs with megasynthase genes with a length of more than 20 Kb, the largest cluster possessing 56 Kb of PKS (seven modules) along with NRPS (three modules) genes. This latter BGC also formed a BiG-SCAPE GCF with several MiBiG BGCs which shared the presence of a small peptide moiety followed by several malonyl units.

2.3.8 Low numbers of BGC found in other underexplored phyla

Lower numbers of BGCs were detected in the phyla Gemmatimonadota (31 BGCs), Planctomycetota (29), Myxococcota (22), Patescibacteria (9), Methylomirabilota (5), Bdellovibrionota_B (8), Elusimicrobiota (4), Armatimonadota (4) and Binatota (3) (Figure 2.11A)

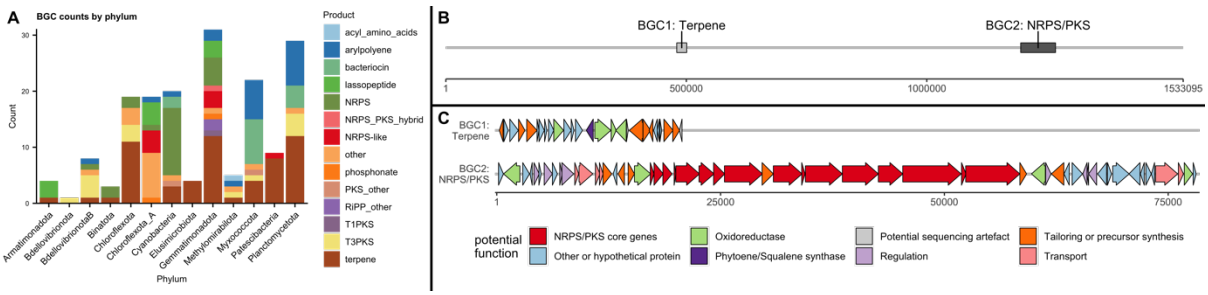


Figure 2.11: BGCs from other phyla. (A) Distribution of BGCs among phyla with 31 or fewer BGCs in the dataset; (B) Map of a large Gemmatimonadota contig (order Gemmatimonadales) and BGCs detected on it; (C) Cluster map of proposed functions of genes in BGC1 and BGC2. Functions were predicted from BLASTing against NCBI nr database as well as antiSMASH module predictions. X axis represents basepairs.

One remarkably long (1.5 Mb, Figure 2.11B,C) Gemmatimonadota contig from the order Gemmatimonadales was found to contain two BGCs: one terpene ($d = 998$) and one NRPS/PKS BGC ($d = 1423$). BGC1 contained a phytoene synthase and several related oxidases. BGC2 contained six PKS modules and two NRPS modules as well as modifying enzymes presence of a TonB receptor indicated that the product could serve as a siderophore.

2.4 Discussion

2.4.1 Metagenomics reveal biosynthetic potential of underexplored bacterial lineages

In the present dataset, a large number of BGCs were found in underexplored phyla not usually associated with specialised metabolites. Two previous studies noted NRPS and PKS novelty and diversity in Acidobacteria and Verrucomicrobia (Crits-Christoph et al. 2018; Borsetto et al. 2019). The present study indicates that these underexplored phyla harbour not only novel NRPS/PKS, but new BGCs from many different classes, such as lassopeptides and bacteriocins. While Crits-Christoph et al. [7] highlighted two promising acidobacterial MAGs from the classes Blastocatellia and the Acidobacteriales, in the present sample the classes Blastocatellia and Vicinamibacteria were the main contributors of acidobacterial BGCs. Furthermore, many BGCs were found in other ubiquitous phyla such as Patescibacteria, Gemmatimonadota and Armatimonadota. Three BGCs (two NRPS and one terpene) were placed in the phylum Binatota. The phylum Binatota was proposed by Chuvochina et al. based on a handful of soil MAGs with no cultured representatives (Parks et al. 2018). To our knowledge, this is the first description of BGCs belonging to the phylum Binatota. Further highly divergent BGCs were found in the underexplored Actinobacteriota classes Acidimicrobiia and Thermoleophilia. This suggests that Actinobacteriota, which contain the heavily exploited genus *Streptomyces*, contain unknown lineages harbouring interesting BGC diversity.

In the present dataset, 845 out of 1417 BGCs (59.6%) had a $d > 900$ and 55 (3.9%) had a $d > 1800$ to the closest GCF. These numbers contrast starkly with the 1.2 million original BGCs in the BiG-SLiCE dataset, of which only 13.9% and 0.2% showed $d > 900$ and $d > 1800$ respectively. While it is necessary to note that sequence diversity does not demonstrate chemical diversity, the striking amount of sequence divergence encountered in just one soil

sample adds to the mounting evidence that uncultured and underexplored phyla – especially Acidobacteriota – are promising candidates for the discovery of novel specialised metabolites. It is furthermore worth noting that the great biosynthetic diversity found at Mars Oasis is under threat from climate change, with the maritime Antarctic warming by 1–3 °C between the 1950s and the turn of the millennium (Adams et al. 2009), and, despite a recent pause in this warming trend (Turner et al. 2016), similar increases in temperature being predicted for later this century as greenhouse gases continue to accumulate in the atmosphere (Turner et al. 2016; Fraser et al. 2018).

The large number of terpene BGCs, most of them putatively C30/C40 carotenoids or hopanoids, could be interpreted with respect to the roles of these compounds in membrane function at extreme temperatures (Belin et al. 2018; Bale et al. 2019; Dieser, Greenwood, and Foreman 2010), as well as UV protection (Dieser, Greenwood, and Foreman 2010; Osmond et al. 2000). A previous study similarly noted a high number of pigmented bacteria among isolates from Antarctic samples (Dieser, Greenwood, and Foreman 2010). Kautsar et al. (Kautsar et al. 2021) recorded only 7.8% terpene BGCs in their large-scale survey of publicly available bacterial genomes, as opposed to the ca. 25% in this survey. Previous short-read metagenomic studies of aquatic and soil environments also reported high numbers of terpene BGCs, with terpenes representing between 15% and 50% of the reported BGCs, respectively (Chen et al. 2020; Cuadrat et al. 2018; Sharrar et al. 2020). However, the representativeness of BGC counts obtained through metagenomic studies remains questionable. In this study, for example, the 85.3% abundance-weighted coverage estimated by nonpareil indicates that many less abundant members of the community are not represented in the dataset. Furthermore, small terpene BGCs are easier to assemble than long and repetitive NRPS/PKS BGCs, therefore leading to bias.

In this study, a large number of BGCs were observed in potentially methanotrophic members of the uncultured order UBA7966. Methanotrophic organisms have not usually been linked to specialised metabolite production, except for siderophore-like RiPPs called methanobactins able to scavenge the copper needed for methane and/or ammonia oxygenase enzymes (Dassama, Kenney, and Rosenzweig 2017). It is likely that the lack of known natural products is related to difficulties associated with cultivation such as specific nutrient requirements and often slow growth, as well as to the amount of energy, carbon and nitrogen available for specialised metabolite production. While no methanobactin BGCs were seen in UBA7966-classified contigs, examining three gammaproteobacterial DUF692-domain containing GCFs revealed the presence of a potential conserved six cysteine precursor peptide. The conserved cysteines in the potential precursor peptides are resemblant of ranthipeptides (formerly known as SCIFFs), which contain six cysteines in forty-five amino acids. Ranthipeptides, however, contain thioethers formed by radical SAM enzymes (Haft and Basu 2011). DUF692 domain proteins are furthermore known to be involved in methanobactin and TglA-thiaGlu biosynthesis (Dassama, Kenney, and Rosenzweig 2017; Ting et al. 2019), and at least one member has been shown to contain two iron atoms potentially acting as cofactors (Ting et al. 2019). All DUF692 protein containing GCFs in the order UBA7966 observed in the present study also contained DUF2063 proteins. DUF2063 family proteins are mostly uncharacterised, though the crystal structure of a member from *Neisseria gonorrhoeae* indicates that DUF2063 might be a DNA-binding domain involved in virulence, and there has been one report of co-occurrence of DUF2063 and DUF692 proteins (Das et al. 2010). Other studies discovered the two neighbouring proteins in operons related to stress response at high calcium concentration (Sarkisova et al. 2014) in *Pseudomonas* as well as responding to gold and copper ions (Jwanoswki et al. 2017) in *Legionella*. The two genes were also found in the atmospheric methane oxidiser *Methylocapsa gorgona* (Tveit et al. 2019). It is therefore possible that these

BGCs could be another form of RiPP involved in chelating metals. While the six cysteines could be involved in forming thioether bonds, disulfide bonds or lanthithionine groups like in many other RiPPs, they could potentially also be directly involved in metal coordination as is the case in the group of small metal-binding proteins called metallothioneins (Ziller and Fraissinet-Tachet 2018). Furthermore, the detected signal peptide indicates translocation using the Sec translocation machinery, which would preclude cyclisation reactions taking place before export.

2.4.2 Long reads make mining and phylogenetic classification of metagenomic BGCs feasible

The advantage of long reads could be observed from comparing the results achieved from long reads vs. short reads, with the short reads providing a lower number of BGCs and a significantly lower taxonomic classification success compared to the BGCs assembled and annotated using long reads. While the number of bases used in the assembly was about a third lower for short reads (28 Gb vs 44 Gb), the number of recovered BGCs was more than two thirds lower (430 BGCs vs 1417 BGCs) and the BGCs assembled from short reads were mostly incomplete. Moreover, this study showed that long-read metagenomes constitute a valuable tool to achieve similar or even improved results to deep short-read metagenomes (Crits-Christoph et al. 2018; Chen et al. 2020; Cuadrat et al. 2018). For example, Cuadrat et al. used 500 million reads (*c.* 50 Gb if read length was 100 bp) for BGC genome mining of a lake community recovering 243 BGCs with a total of 2,200 ORFs, which averages to nine ORFs per BGC indicating small and/or incomplete BGCs (Cuadrat et al. 2018). A larger short-read study of microbial mats recovered 1,477 BGCs (Chen et al. 2020). While this study did not report the number of sequenced bases or BGC completeness, the median BGC length of 103 BGCs from 15 representative and highly complete MAGs was 11.9 Kb, also indicating mostly small and/or incomplete BGCs. Another study by Crits-Christoph et al. (Crits-Christoph et al. 2018) used

1.3 Tb of short-read sequence data of grassland soil to mine selected bins of four phyla, recovering a total of 1,599 BGCs, 240 of which were NRPS/PKS BGCs, including several large and complete ones (Crits-Christoph et al. 2018). The present study indicates that the long-read approach requires a relatively low sequencing input similar to the two smaller studies to provide a result similar to the larger study. While the contigs, MAGs and BGCs produced using shallow ONT sequencing are not as accurate as the ones produced using deep short read sequencing, our results show that they can be used to profile the biosynthetic potential of complex environmental samples, estimate their diversity and could be used to guide isolation and heterologous expression strategies. Lower error rates could be achieved through higher coverage in long and short reads as well as advances in long-read basecalling.

It can furthermore be concluded that contig-level classification using CAT shows advantages compared to genome-resolved metagenomics in single-sample data, where binning is inefficient. Cuadrat et al, Crits-Christoph et al. and Chen et al. used genome-resolved metagenomics (Cuadrat et al. 2018; Crits-Christoph et al. 2018; Chen et al. 2020), in which contigs are binned and bins are mined for BGCs. While it is favourable to attribute BGCs to distinct MAGs, it is viable only when a large number of samples are used, making binning efficient through differential abundance (Albertsen et al. 2013). When using only one sample, binning becomes inefficient and, in our case, missed the vast number of BGCs, with 1,139 of 1417 BGCs not being binned. Contig-based classification approaches offer an alternative, but their accuracy is limited by contig length (von Meijenfeldt et al. 2019) and the classification dependent on the database used. In our data, a contig N50 of >80 Kb provided ample sequence data for accurate classification, leading to >90% classification at phylum level. Usage of GTDB-derived databases ensured improved classification of uncultured taxa, and few conflicts with single-copy core gene-based bin-level classification were detected.

3 Results 2: Development of a novel metagenomic library screen, traditional isolation and comparison to shotgun metagenome for BGC recovery

3.1 Introduction

3.1.1 Metagenomic libraries

Most natural products from metagenomes have been discovered using metagenomic libraries. These have the significant advantage of making DNA easily accessible and therefore amenable to applications such as cloning. However, since most clones in a metagenomic library do not contain the sequences of interest, they need to be screened. Metagenomic library screening strategies include relatively simple, completely functional library screenings, in which a library is screened for a phenotype or an activity (Brady 2007). The disadvantage of this approach is that the libraries need to be transferred into an expression host as well as the necessity of an appropriate activity assay. On the other hand, sequence-guided approaches use sequence information to recover clones of interest from a library. Simple sequence-guided approaches use PCR with degenerate primers to screen libraries for conserved sequences, for example NRPS A-domains and PKS KS-domains (Amos et al. 2015). More refined sequence-guided approaches employ MiSeq sequencing of the amplicons obtained with these degenerate primers (Owen et al. 2013). Sequences of interest are then identified by comparing the amplicon sequences to known sequences. Then, through multiple rounds of qPCR with specific primers followed by dilution, a clone containing the desired BGC (or BGC fragment) is obtained. Examples of natural products (and the target domains) discovered by sequence-guided metagenomic library screening are malacidins (calcium-binding motif), arimetamycin A (anthracycline-like domains), as well as clarepoxcins A–E and landepoxcins A and B (epoxyketone proteasome inhibitors) (Hover et al. 2018; H.-S. Kang and Brady 2013; Owen et

al. 2015). The disadvantage of this approach is that only compounds that are at least related to known compounds can be discovered. Therefore, most sequence-guided metagenomic library screens lead to the recovery of congeners of known compounds or novel members of a known natural product class. This problem is partly circumvented in the recent approach of Libis et al., who used NRPS/PKS domain amplicon sequencing to determine co-occurrence patterns of unique A/KS domains within metagenomic libraries (Libis et al. 2019). However, no screening approach to date could recover novel classes of BGCs – i.e. BGCs utilising types of biosynthetic machinery that show no or little homology to known biosynthetic genes – since the biosynthetic genes are unknown and therefore cannot be screened for.

The recovery of BGCs or BGC fragments using sequence-guided screening of large, pooled metagenomic libraries (containing thousands of distinct clones per well) is a laborious process involving iterative library screening using (q)PCR followed by dilution to eventually obtain single clones (J. H. Kim et al. 2010). However, obtaining a natural product from a cloned BGC is not straightforward, with promoter engineering and refactoring being the most promising methods to activate expression (H.-S. Kang, Charlop-Powers, and Brady 2016; S.-H. Kim et al. 2019). There are no published numbers on the success rate of compounds obtained per cloned metagenomic BGC, but the known challenges of heterologous expression coupled with the phylogenetically diverse origin of the sequences indicate that the attrition rate is likely to be high. Therefore, it would be highly desirable to establish a reliable and convenient method of activating BGCs from metagenomic libraries.

3.1.2 Use of γ-butyrolactone regulatory cassettes for natural product discovery

A commonly used method for activating BGC expression in isolates is the overexpression of activators or the deletion of repressors (Aigle and Corre 2012). This has also been achieved for

isolate-derived BGCs cloned into genetically tractable heterologous hosts (Alberti et al. 2019). The regulatory system controlling the methylenomycin BGC from *Streptomyces coelicolor* A3(2) is particularly well studied. In this BGC, production of the antibiotic methylenomycin is regulated by autoinducer molecules, the methylenomycin furans (MMFs). MMFs are related to γ-butyrolactone signalling molecules, such as A-factor from *S. griseus*. In *S. coelicolor*, MMF biosynthesis involves the action of MmfL, MmfH and MmfP. The resulting molecules bind to the TetR-like transcriptional repressor MmfR, leading to its dissociation from the methylenomycin autoregulatory elements (MAREs) that are present within promoters of the BGC. Since the MMF biosynthetic gene promoters themselves contain MAREs, a positive feedback loop is initiated which leads to the production of methylenomycin. The BGC contains a second TetR-like transcriptional repressor termed MmyR, which also binds to MAREs, but is not deactivated by binding of MMFs. The deletion of *mmyR* leads to overproduction of methylenomycin (O'Rourke et al. 2009). A very similar pattern of overproduction of a specialised metabolite upon deletion of a *mmyR* homologue was also seen in Jadomycin (*S. venezuelae*), Virginiamycin (*S. virginiae*) and Lankamycin (*S. rochei*) (Zou et al. 2014; Lee, Kitani, and Nihira 2010; Arakawa et al. 2007).

The observation that the deletion of *mmyR* led to the overproduction of methylenomycin inspired Sidda et al. to mine other *Streptomyces* genomes for the presence of homologous gene cassettes (*mmfR, mmfLHP, mmyR)* (Sidda et al. 2013). The deletion of the *mmyR*-homologue in *S. venezuelae* KSCC 10712 led to overproduction of a novel class of γ-aminobutyrates termed gaburedins (Sidda et al. 2013). Similarly, Alberti et al. cloned a silent BGC containing the same regulatory cassette from the genetically intractable *S. sclerotialus* into *S. albus* and triggered the expression of the silent BGC by deletion of the *mmyR* homologue, resulting in the discovery of scleric acid (Alberti et al. 2019).

Other γ-butyrolactone regulation systems show different architectures and levels of control. For example, the *S. griseus mmfL* and *mmfR* homologues *afsA* and *arpA* are located more than 100 kb apart on the chromosome and no *mmyR*-like function has been identified (Poon 2015). In the gaburedin BGC, MmfR directly activates gaburedin biosynthesis, while many other characterised BGCs feature an intermediate activator, such as MmyB in the case of methylenomycin (Sidda et al. 2013; O'Rourke et al. 2009). γ-butyrolactone regulation systems can furthermore have far-reaching effects. For example, the *scb* genes in the coelimycin BGC in *S. coelicolor* regulate coelimycin biosynthesis but also have pleiotropic effect, with the deletion of the *mmyR*-homologue *scbR2* leading to an increase in coelimycin production, but also to differential expression of >40% of all genes (Bednarz, Kotowska, and Pawlik 2019).

The prevalence of highly similar γ-butyrolactone regulatory cassettes coupled with the successful deletions of *mmyR* homologues to achieve overproduction of natural products led to the idea that it could be possible to screen metagenomic libraries for the regulatory cassette and thereby discover novel BGCs and subsequently activate them by simply knocking out the *mmyR* homologue (Figure 3.1). This would not only provide a reliable method for activation of recovered BGCs, but also enable discovery of novel BGC classes since the screen would not be directed towards biosynthetic enzymes like usual metagenomic library screens. Most γ-butyrolactone regulation systems are known from members of the *Streptomyces* genus which are readily culturable. However, rare actinomycetes have also been shown to possess γ-butyrolactone signalling systems and can be much harder to isolate, making a metagenomic approach desirable (Arul Jose and Jebakumar 2013; Choi et al. 2003; 2004; Aroonsri et al. 2008; Lazzarini et al. 2000).

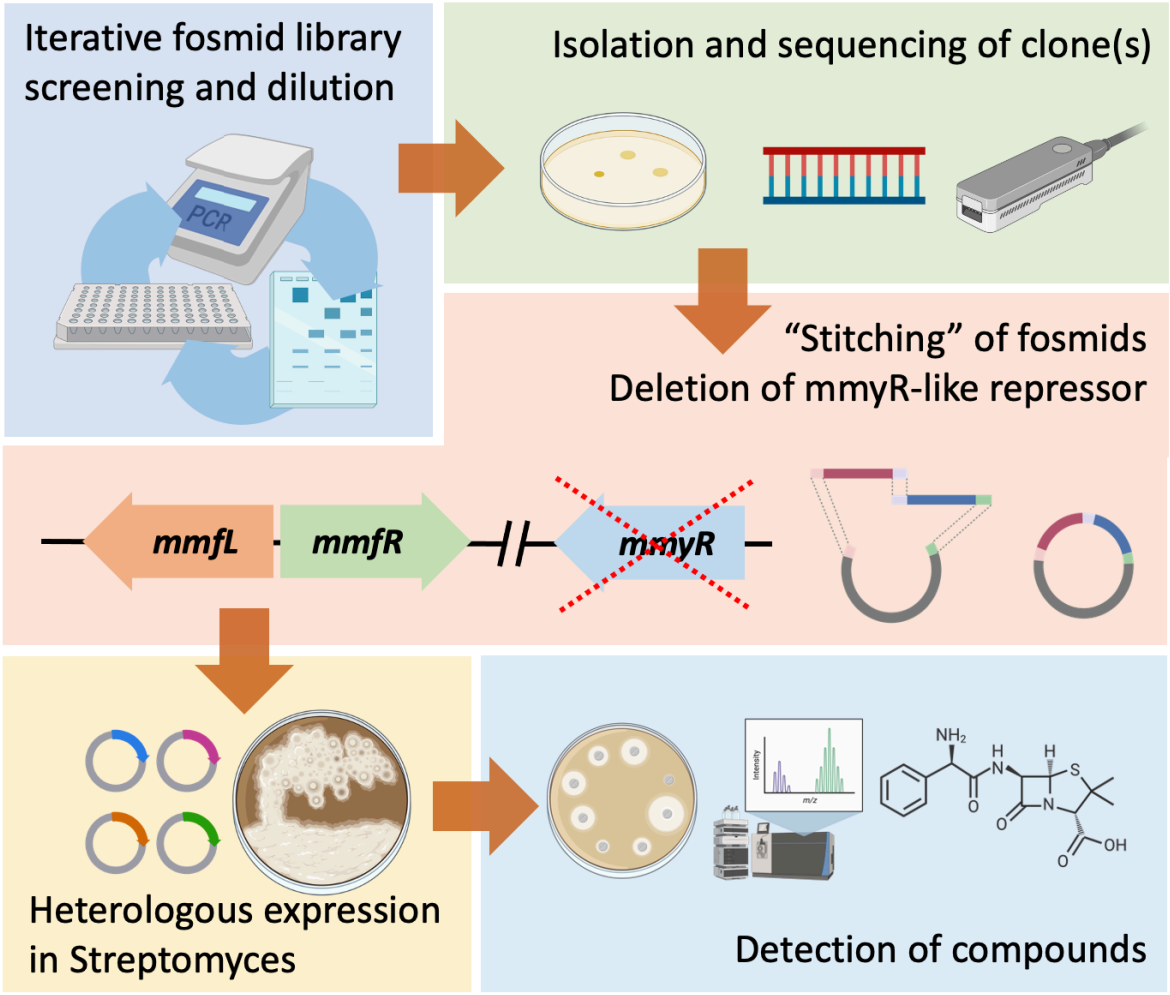


Figure 3.1: Proposed workflow for metagenomic library screening and BGC activation through regulatory gene cassettes.

3.1.3 Isolation of Antarctic soil bacteria

As early as 1985, the discrepancy between observable and cultivable microorganisms had been noted, but until the advent of environmental 16S rRNA gene clone libraries in the 1990s, cultivation had been the only way of assessing bacterial diversity (Staley and Konopka 1985; Schmidt, DeLong, and Pace 1991). With increasing metagenomic sequencing, it became evident that the portion of readily culturable bacteria was just a fraction of the diversity present in the environment. Additionally, biases introduced through culture can enrich low abundance species, leading to little overlap between sequencing studies and cultivation studies in the same environment (Hardoim et al. 2014).

Studies comparing the bacterial diversity obtained through cultivation and 16S rRNA gene amplicon sequencing are plentiful. However, the author is not aware of any studies comparing the specialised metabolite potential of isolates and metagenomic sequences from the same environment. Due to the small fraction of bacteria culturable through traditional methods, many BGCs found in metagenome would most likely not be represented in the isolates. However, it less clear how many isolate-derived BGCs would be detected in the metagenome. Furthermore, overlapping biosynthetic capabilities between different bacteria could occur through horizontal gene transfer (HGT), which is known to occur both within species and genera as well as between phyla and even from bacteria to fungi (Ziemert et al. 2014; Jensen et al. 2007; McDonald and Currie 2017; Cruz-Morales et al. 2017; Kroken et al. 2003).

3.2 Aims and rationale

The aim of the present study was to complement the metagenomic sequencing approach using a metagenomic library approach. Furthermore, traditional isolation work was undertaken and was used for comparative analysis. The metagenomic library and regulatory cassette screening approach would enable the recovery and activation of novel BGCs potentially undetected by antiSMASH genome mining, while the isolation experiment would allow an assessment of BGC diversity in readily cultivable bacteria and enable a comparison with the metagenome-derived BGCs.

In details, the objectives were as follows:

1. Develop a screen for regulatory genes for recovery and activation of metagenomic BGCs
    a. Construction of library
    b. Design and validation of primers

c. Screening and recovery of clones

2. Examine the differences between culturable bacteria and metagenome
    a. Isolate and identify 50-100 bacteria
    b. Sequence a selected set of bacteria
    c. Assess differences in the biosynthetic potential of isolates & metagenome from the same soil

3.3 Materials and Methods

3.3.1 SPRI bead preparation

Custom SPRI beads were prepared by modifying a previously published protocol (Ramawatar and Schwessinger 2018). In short, the buffer was removed from AMPure XP beads and replaced with a custom buffer (11% v/v PEG-8000, 0.25% v/v Tween-20, 10 mM Tris-HCl, 1 mM EDTA pH 8, 0.4% v/v washed beads). Variable concentrations NaCl and cleanup ratios were tested, with 1.6M NaCl and 0.8x providing a good compromise of DNA recovery and size-selection.

3.3.2 Soil DNA extraction & metagenomic library preparation

Soil DNA was first extracted and the library (CopyControl Fosmid Library Production Kit, Lucigen) constructed a previously published protocol (Brady 2007). In short, the steps were:

1) Gentle lysis & DNA extraction using CTAB
2) Isopropanol precipitation
3) Size selection by agarose gel electrophoresis, cutting out the HMW (c. 20 kb+) band an elution using dialysis tubing
4) Concentration of DNA using Amicon centrifugal concentrator

This was followed by the library construction:

5) Blunt-ending of DNA using end-repair enzyme mix
6) Isopropanol precipitation
7) Ligation with linearised vector pCC1FOS
8) Packaging into lambda phage heads. Control reactions carried out with supplied control vector.
9) Transfection of library into *Escherichia coli* Epi-300
10) Titration of library and transferring into 96-well plate

In modified protocol 1, steps 3, 4 and 6 were replaced with 0.8x clean-ups using custom SPRI beads. In modified protocol 2, only steps 4 and 6 were replaced with custom SPRI bead clean-ups.

3.3.3 Primer design for library screening

Degenerate primers were designed for conserved stretches in the *mmfL/mmfR* genes. For an initial search for *mmfL/mmfR*-like genes in Actinobacteria, a set of 352 actinobacterial genomes containing all major cultured actinobacterial orders was manually curated at JGI IMG. This set was used to conduct BLAST searches using the translated ORFs of *mmfL* (E-value 1E-2) and *mmfR* (E-value 1E-5). To specifically detect *mmfL*/*mmfR* homologs directly adjacent to each other as well as to gain information about orientation of the two genes, a ClusterTools 0.2 BLASTp search with an E-value cut-off of 1E-5 and a window size of 2500 bp was conducted on a database of all actinobacterial genomes in RefSeq. The results showed all co-occurrences of *mmfL*/*mmfR* homologs within 2500 bp of the genome. To visualise conserved sequence motifs for designing primers on, the protein sequences of all divergently oriented *mmfL*/*mmfR* homologs were aligned using ClustalO and a hmm logo was generated using skylign.org (Wheeler, Clements, and Finn 2014). The visualised alignments were assessed for conserved sequences of at least 3-4 AA long. Degenerate primers were designed by placing the 3' end at

the most conserved site and taking into account actinobacterial codon usage (Table 3.1). A diagram explaining positions of primers can be found in the Results section (Figure 3.3).

Table 3.1: Codon table with Actinobacteria preferences derived from Lal et al. (2016). Crossed-out codons show codons with little usage; bold and underlined codons show heavily preferred codons.

1st base	2nd base T		2nd base C		2nd base A		2nd base G		3rd base
T	TTT	(Phe/F)	TCT	(Ser/S)	TAT	(Tyr/Y)	TGT	(Cys/C)	**T**
	TTC		TCC		TAC		**TGC**		**C**
	~~TTA~~		TCA		TAA	Stop (*Ochre*)	TGA	Stop (*Opal*)	**A**
	TTG		TCG		TAG	Stop (*Amber*)	TGG	(Trp/W)	**G**
C	CTT	(Leu/L)	CCT	(Pro/P)	CAT	(His/H)	CGT	(Arg/R)	**T**
	CTC		**CCC**		**CAC**		**CGC**		**C**
	~~CTA~~		CCA		CAA	(Gln/Q)	~~CGA~~		**A**
	CTG		**CCG**		**CAG**		CGG		**G**
A	ATT	(Ile/I)	ACT	(Thr/T)	AAT	(Asn/N)	AGT	(Ser/S)	**T**
	ATC		**ACC**		AAC		**AGC**		**C**
	~~ATA~~		ACA		AAA	(Lys/K)	AGA	(Arg/R)	**A**
	ATG	(Met/M)	ACG		**AAG**		AGG		**G**
G	GTT	(Val/V)	GCT	(Ala/A)	GAT	(Asp/D)	GGT	(Gly/G)	**T**
	GTC		GCC		GAC		**GGC**		**C**
	~~GTA~~		GCA		GAA	(Glu/E)	GGA		**A**
	GTG		GCG		GAG		GGG		**G**

3.3.4 PCR

All PCRs were performed with KAPA Taq polymerase (Sigma-Aldrich) with the addition of BSA and DMSO using the calculated annealing temperature or empirically observed optimum annealing temperature. Since fragments were small, extension time was set at 30 seconds per kilobase.

3.3.5 Primer List

The primers used in the study are found in Table 3.2.

Table 3.2: The primers used in this study.

No	Sequence	Notes
1_020	CCGCTCCTTGCTSGGRAARTGRAARTA	*mmfR* screening; 32x degenerate
1_025	CCCAGCCGCTCCTTGCTSGGRAARTGRAARTA	*mmfR* screening; 16x degenerate
1_021	GCGCCAGTCGGTCAGVARVACNTC	*mmfL* screening; 72x degenerate
1_024	GTGGCTGCGSGGCCASYGNGC	*mmfL* screening; 32x degenerate
1_026	CTCGCTGACGCTGCTNYKRTGNAC	*mmfL* screening; 128x degenerate
1_027	GCCGCTCTGGCGVABSGTYTC	*mmfL* screening; 36x degenerate
1_028	TTCTGCCGCGGCAGCCTTGTGTACG	*mmfL* in *S. coelicolor* A3(2) in same position as 1_026; non degenerate
1_029	GGCCAGTGATTCCTTGCTGGGGAAGTG	*mmfR* in *S. coelicolor* A3(2) in same position as 1_020; non degenerate
2_005	TGATGTTCGACCACACCTCG	Screening for T1PKS from contig_7544; Gemmatimonadota
2_006	ACGAGGACCTGGCTTCCAA	
2_009	CGCACCATTTCCTATTGCCG	Screening for NRPS/PKS from contig_2148; Verrucomicrobiota
2_010	GGTGTACTTCCGTTCCGGTT	
27F	AGAGTTTGATCCTGGCTCAG	Universal 16S rRNA gene primers for amplification and sequencing (Frank et al. 2008)
1492R	GGTTACCTTGTTACGACTT	
NRPS_F	CGCGCGCATGTACTGGACNGGNGAYYT	Degenerate NRPS primers (Amos et al. 2015)
NRPS_R	GGAGTGGCCGCCCARNYBRAARAA	
16S_ill_F	TCGTCGGCAGCGTCAGATGTGTATAAGAGAC AGCCTACGGGNGGCWGCAG	Primers targeting the 16S rRNA gene V3 and V4 regions, with Illumina MiSeq adapters, adapted from Klindworth et al. ((Klindworth et al. 2013)
16S_ill_R	GTCTCGTGGGCTCGGAGATGTGTATAAGAGAC AGGACTACHVGGGTATCTAATCC	

3.3.6 Library screening

To screen the library by PCR, a "working copy" of the original library was induced with 10 mM arabinose (Epicentre induction solution), grown overnight in LB in deep 96 well plates, and plasmid DNA was purified using the Qiagen 96 Turbo kit. PCR was performed using the appropriate primer pair and the result was assessed using 96 well agarose gel electrophoresis. When a well showed a positive signal, another iteration of the screening process was conducted. First, the concentration of bacteria in the well was assessed by plating a serial dilution. According to the number of colony forming units in the well and the estimated number of unique clones in it, the positive well was then diluted into another 96 well plate. The dilution had to be chosen to ensure that unique clones were separated into different wells, but not diluted to extinction. Then, this second plate was processed like the original plate. After a third iteration, the positive well was diluted and plated and colonies picked for colony PCR.

3.3.7 16S rRNA gene amplicon sequencing

16S rRNA gene amplicon sequencing was carried out for eDNA extracted using gentle chemical lysis, FastDNA Spin Kit for Soil (MP Bio) as well as plasmid DNA purified by miniprep from the Mars Oasis library. The library preparation was carried out by Dr. Chiara Borsetto. In short, The V3-V4 regions were amplified using primers 16S_ill_F and 16S_ill_R targeting the V3-V4 region and containing MiSeq adapters. After confirmation of amplification by gel electrophoresis and purification of PCR products, the amplicons were indexed by PCR and the concentration normalised with a SequalPrep Normalisation Plate. 2x 300 bp paired-end sequencing was carried out by the Warwick Genomics Facility. Quality control was done by fastqc (Babraham, 2019) and sequence analysis was conducted using QIIME 2 v2020.8.0 (Bolyen et al. 2019), employing the dada2 (Callahan et al. 2016, 2) classifier with the SILVA

138 database (Quast et al. 2013) for taxonomy assignment. *E. coli* reads were removed from the Mars Oasis library plasmids since they made up about 50% of all reads.

3.3.8 Media

Bacteria were grown in the following media:

- Luria-Bertani solid (LBA; 10 g/L tryptone, 5 g/L yeast extract, 10 g/L sodium chloride, 15g/L agar)
- Luria-Bertani liquid medium (LB; 10 g/L tryptone, 5 g/L yeast extract, 10 g/L sodium chloride)
- Nutrient Broth (10 g/L peptone, 1g/L beef extract, 2g/L yeast extract, 5g/L sodium chloride
- Nutrient Agar (10 g/L peptone, 1g/L beef extract, 2g/L yeast extract, 5g/L sodium chloride, 15g/L agar)
- R2A (0.5l/L yeast extract, 0.5g/L Proteose Peptone No. 3, 0.5g/L Casamino Acids, 0.5g/L Dextrose, 0.5g/L Soluble Starch, 0.3/L Sodium Pyruvate, 0.3g/L Dipotassium Phosphate, 0.05g/L Magnesium Sulfate, 15g/L agar)
- Soil extract/Nutrient agar (SENA): 500g of soil obtained a site in Cryfield, Coventry (52.3762, -1.5694, previously characterised as sandy silt loam(Borsetto 2017)) was extracted with 1L 50 mM NaOH by stirring overnight at RT. After centrifugation and filtering, pH was adjusted to 7.5 using HCl. Soil extract was autoclaved and stored at 4°C. For SENA preparation, 200 mL of soil extract were mixed with 400 mL of dilute Nutrient agar (0.08g/L Nutrient Broth, 30g/L agar) and 200mL of sterile water.

3.3.9 Isolation of bacteria

1 g of Mars Oasis soil (see Chapter 1 for sample description) was weighed and vortexed with 25 mL sterile 0.9% saline for one minute. The resulting solution was serially diluted up to 10^{-6} using 0.9% saline. 100 µL of each dilution were plated on four plates of SENA each. The plates were incubated at 16°C for three weeks, then single colonies were picked, trying to avoid similar morphologies from the same plate. The bacteria were streaked on SENA containing 50 µg/mL of cycloheximide to discourage the fungal growth that had appeared on several plates. If possible, the isolates were subsequently grown on standard media (NA/R2A/LBA). The isolates were named after the origin (MA), dilution level (0-6), plate number (I, II, III, IV) and number of colony picked (1+), resulting in e.g. MA-2IV3, or just 2IV3 for the purposes of this study. For cryostocks, bacterial biomass was scraped off a freshly grown plate into 15% glycerol, flash frozen in liquid nitrogen and stored at -80°C

3.3.10 Identification of bacteria

Single colonies were picked, resuspended in 50 µL lysis buffer (10 mM Tris-HCl, 1 mM EDTA, 0.1% Triton-X100, pH 8) and heated to 100°C for 15 minutes. After centrifugation, 1 µL of lysate was added to a PCR reaction using primers 27F and 1492R (Frank et al. 2008). After confirmation using agarose gel electrophoresis, the PCR product was sent for Sanger sequencing using the same primers. The resulting reads were aligned using SnapGene to form a full sequence, which was then searched against the NCBI nucleotide database using BLASTn to establish taxonomy. A tree was built using SILVA (Quast et al. 2013) (FastTree de-novo including neighbours) and visualised with iTOL (Letunic and Bork 2007). To assess duplicate isolates, an all-vs-all BLASTn search was conducted using a 99.5% cutoff and the results visualised using Cytoscape.

3.3.11 Sequencing and sequence processing

Isolates were grown on agarose plates with different media for several days at room temperature. To obtain short reads, cells were scraped into tubes containing DNA/RNA shield (Zymo Research) and sent for Illumina sequencing at microbesNG. Where long reads were available, the short reads were used for hybrid assembly. Where no long reads were available, the assembly provided by microbesNG was used.

To obtain long reads, the biomass was processed using the ProMega Wizard Genomic DNA Extraction kit. A barcoded library was built using the ONT SQK-LSK109 kit as well as the native barcoding expansion (EXP-NBD104) and sequenced on a MinION using R9.4.1 flow cells. The raw reads were basecalled using guppy v.3.03 (HAC model). Where short reads were available, a hybrid assembly was done using Unicycler v0.4.8 (Wick et al. 2017). Where no short reads were available, a long-read-only assembly was done using flye v2.5 (Kolmogorov et al. 2019), polished 4x using racon v1.4.7 (Vaser et al. 2017) and 1x using medaka v0.7.1 (*Nanoporetech/Medaka*, 2017).

Isolate assemblies were taxonomically classified using GTDB-Tk 0.3.2 r89 (Chaumeil et al., 2019) and assembly quality was determined using CheckM (Parks et al. 2015). BGCs were mined using antiSMASH v5.1 (Blin, Shaw, et al. 2019) and analysed using BiG-SLiCE 1.1.0 (Kautsar, Hooft, et al. 2020). Assemblies were also submitted to autoMLST (Alanjary, Steinke, and Ziemert 2019) to obtain an average nucleotide identity (ANI) estimate to the nearest genome in RefSeq.

3.3.12 Comparing isolates and metagenomes

To test the overlap of isolate 16S rRNA gene sequences and the metagenomic 16S rRNA gene sequences, first barrnap v0.9 (Seemann 2013) was used to extract 16S rRNA gene sequences

from the metagenomic assembly. The isolate-derived sequences were then aligned against the metagenome-derived sequences using BLASTn with an identity cutoff of 97%, an E-value of 10E-10 and only counting alignments with 1000 or more bp to ensure matching of most of the 16S rRNA gene sequence. Additionally, both isolate and metagenome 16S rRNA gene sequences were classified using dada2 assignTaxonomy with a classifier derived from the GTDB r89 database (Alishum 2019, 2). The resulting network was visualised using R package igraph as well as Cytoscape.

To assess the representation of the sequenced isolates (and BGCs) in the metagenome reads, the metagenomic nanopore reads were mapped to the isolate assemblies (and BGC fastas) using minimap2 (H. Li 2018), the reads were quality filtered with samtools (H. Li et al. 2009) to only retain alignments with a quality score over 30, coverage was calculated using bedtools (Quinlan and Hall 2010) genomecov, and the median coverage calculated using R.

3.4 Results

3.4.1 Metagenomic library construction and screening

3.4.1.1 Metagenomic library construction

A metagenomic library was constructed from Mars Oasis soil using an established protocol involving several precipitation steps, agarose gel electrophoresis for size selection, and concentration using a centrifugal filter tube (see Methods). However, a large amount of DNA was lost from the initial purification step until the ligation step (36 µg in crude extract to 2 µg in ligation step; ca 5.6% yield). Furthermore, the total library size was only ca. 300 clones, corresponding to ca. one megabase at an average insert size of 35 kb. Running the blunt-ended, precipitated DNA on a gel revealed that the DNA had degraded in the preparation process.

To avoid DNA degradation during processing, a faster and less complicated alternative protocol (modified protocol 1) was tested which replaced the size selection (step 3), centrifugal filtering (step 4) and precipitation (step 6) with size-selective binding of DNA to paramagnetic SPRI beads. A SPRI buffer aimed at eliminating fragments below 3000 bp was prepared and tested on NEB 1 kb DNA ladder. This showed successful elimination of DNA fragments below 3000 kb using a buffer/bead volume of 0.8x the DNA volume (Figure 3.2A). For library construction, two of these SPRI clean-ups were employed: the first for size selection and concentration of eDNA from the first isopropanol precipitation, the second for buffer exchange after blunt-ending. The SPRI method showed efficient reduction of fragments < 3-4 kb (Figure 3.2B) as well as a slightly better yield of 8.9%. One transfection reaction gave a library the total size of ca. 122,000 clones with an insert size of approximately 35 kb, leading to a total size of ca. 4.3 Gb. Compared to the control library, efficiency was calculated to be ca 6.6%. The low efficiency was attributed to the presence of fragments above 3-4 kb which were retained by the SPRI beads and readily ligated with the vector, but were not packaged into phage heads. Therefore, a third attempt (modified protocol 2) re-introduced the size selection

by agarose gel electrophoresis and only replaced the centrifugal filtering (step 4) and precipitation (step 6) with SPRI bead clean-ups. This led to the production of a 436,000-membered library, corresponding to ca 15 Gb and an efficiency of 23.3% versus the control. This library was pooled with the 122,000-membered library, giving a total size of ca. 19.4 Gb. This size was deemed sufficient for detection and recovery of BGCs in high to mid abundance species.

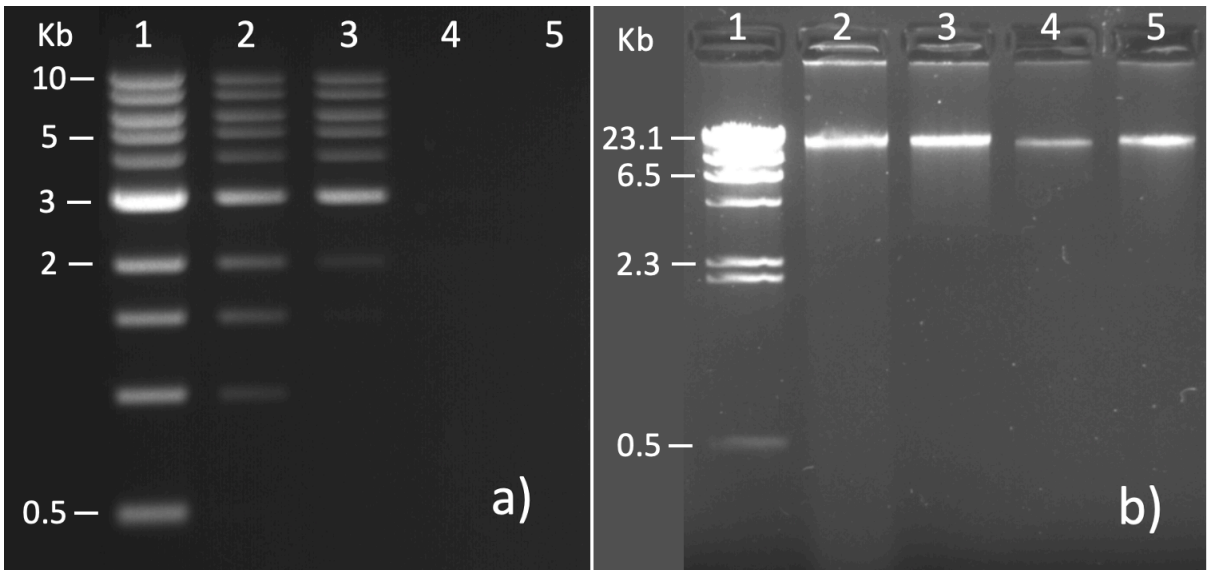


Figure 3.2: SPRI-cleanup of a) a 1 kb ladder and b) eDNA. In a) lane 1 is the umodified ladder, lane 2 is a 1.0x cleanup, lane 3 is a 0.8x cleanup, lane 4 is a 0.6x cleanup and lane 5 is a 0.4x cleanup. Fragments above ca 3 kb are retained at 0.8x. In b) lane 1 is a Lambda HindIII digest. Lane 2 is eDNA after CTAB extraction and precipitation, the smear of small fragments is visible. Lane 3 is the same DNA after a 0.8x SPRI cleanup. Lane 4 is an error, lane 5 is the same DNA after blunt-ending and a second cleanup. Fragments above 3-4 kb are retained along with the HMW DNA.

3.4.1.2 Primer design and testing

The genes selected for screening were *mmfL* and *mmfR*, as they are usually co-located in the regulatory cassettes. An initial search for *mmfL*- and *mmfR*-like genes revealed that *mmfL*- and *mmfR*-like genes were distributed differently. In the representative 352 genome database, there were 89 hits for *mmfL* in 50 genomes, with the most hits coming from *Streptomyces*, *Rhodococcus* and *Amylocatopsis*. For *mmfR*, the same search showed 5391 hits in 340 genomes, the three top genera being *Corynebacterium*, *Mycobacterium* and *Bifidobacterium*.

The results indicate that not only are *mmfL* homologs much less widespread than *mmfR* homologs, but also that there are on average >15 *mmfR* homologs per genome, while for *mmfL* this number is <2. This did not come as a surprise, given that *mmfR* belongs to the TetR-family transcriptional regulators which are ubiquitous among Actinobacteria. It also suggested that it would be advantageous to not design primers based on an alignment of all *mmfL*/*mmfR* genes found through BLAST, but to only use *mmfR* homologs adjacent to *mmfL* homologs and vice versa.

A search using ClusterTools 0.2 against the RefSeq actinobacterial genomes revealed differences in orientation of the genes in different genera. Analysis of orientation and taxonomy showed that most hits came from *Streptomyces* species as well as rare actinomycete genera such as *Amycolatopsis* and always featured the two genes in a divergent orientation. Less frequent was a convergent orientation which occured mostly in *Rhodococcus*. A small number of hits were observed in tandem orientation (5'-*mmfR*-*mmfL*-3'). Due to a previously demonstrated role in specialised metabolite production, the divergently oriented genes from *Streptomyces* and rare actinomycetes were deemed the most likely candidates for BGC expression and therefore only divergently oriented genes were chosen as a target. Alignment and hmm logo generation of the translated amino acid sequences of all divergently oriented *mmfL/mmfR* homologues revealed one highly conserved region in *mmfR* and several conserved regions in *mmfL* homologues which were then used to design degenerate primers (Figure 3.3). Two primers for *mmfR* and four primers for *mmfL* were generated, leading to products between 548 and 692 bp in length. (Table 3.3)

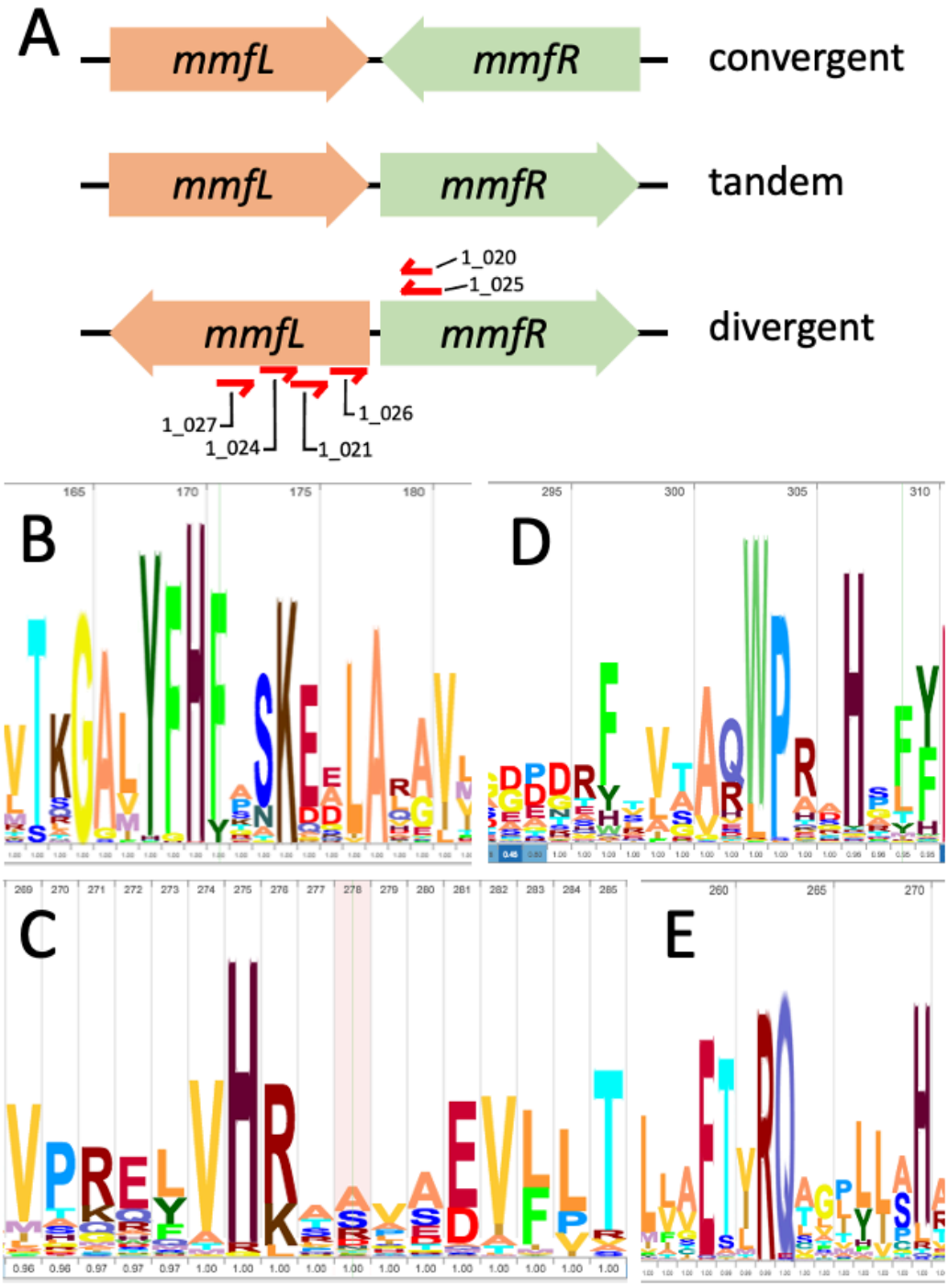


Figure 3.3: Orientation of *mmfL/mmfR* genes, primer positions and conserved motifs. (A) The three possible orientations of any two neighbouring genes shown with *mmfL* and *mmfR*. Primers were designed for the divergent orientation found in *Streptomyces* and rare actinomycetes. Primer binding locations on the two genes are indicated, but not to scale. (B) Conserved mmfR stretch for 1_020 & 1_025; (C) conserved mmfL stretch for 1_026 (VHR) and 1_021 (EVLLT); (D) conserved mmfL stretch for 1_024; (E) conserved mmfL stretch for 1_027.

Table 3.3: Primers designed for metagenomic library screening to detect adjacent *mmfL* and *mmfR* homologues

Target binding	Primer	Sequence	Tm range	Degeneracy
mmfR	1_020	CCGCTCCTTGCTSGGRAARTGRAARTA	58 - 64	32x
	1_025	CCCAGCCGCTCCTTGCTSGGRAARTGRAARTA	64 - 70	16x
mmfL	1_021	GCGCCAGTCGGTCAGVARVACNTC	59 - 66	72x
	1_024	GTGGCTGCGSGGCCASYGNGC	64 - 68	32x
	1_026	CTCGCTGACGCTGCTNYKRTGNAC	57 - 66	128x
	1_027	GCCGCTCTGGCGVABSGTYTC	58 - 64	36x

Initial screening of primer combinations was conducted at three temperatures using the plasmid C73_797 containing the methylenomycin gene cluster, and thereby *mmfL* and *mmfR* (Figure 3.4). Strong amplification without unspecific bands was observed in primer combinations 1_020/1_027 and 1_025/1_027. The two primer combinations were further tested on *S. coelicolor* M145 which contains a different gamma-butyrolactone BGC (*scbA* and *scbR* as homologs for *mmfL* and *mmfR*), *S. coelicolor* A3(2) containing both the methylenomycin BGC as well as the *scbA/scbR* as targets, as well as *S. venezuelae* containing the homologues *jadW* and *jadR*. Screening showed that both combinations gave results in all targets, and amplification was stronger with 1_020/1_027. In S. *coelicolor* A3(2), all primer combinations preferentially amplified *scbA/scbR* over *mmfL* and *mmfR* as could be seen by the size of the band. The temperature of 63°C was chosen for further testing since it showed highest specificity.

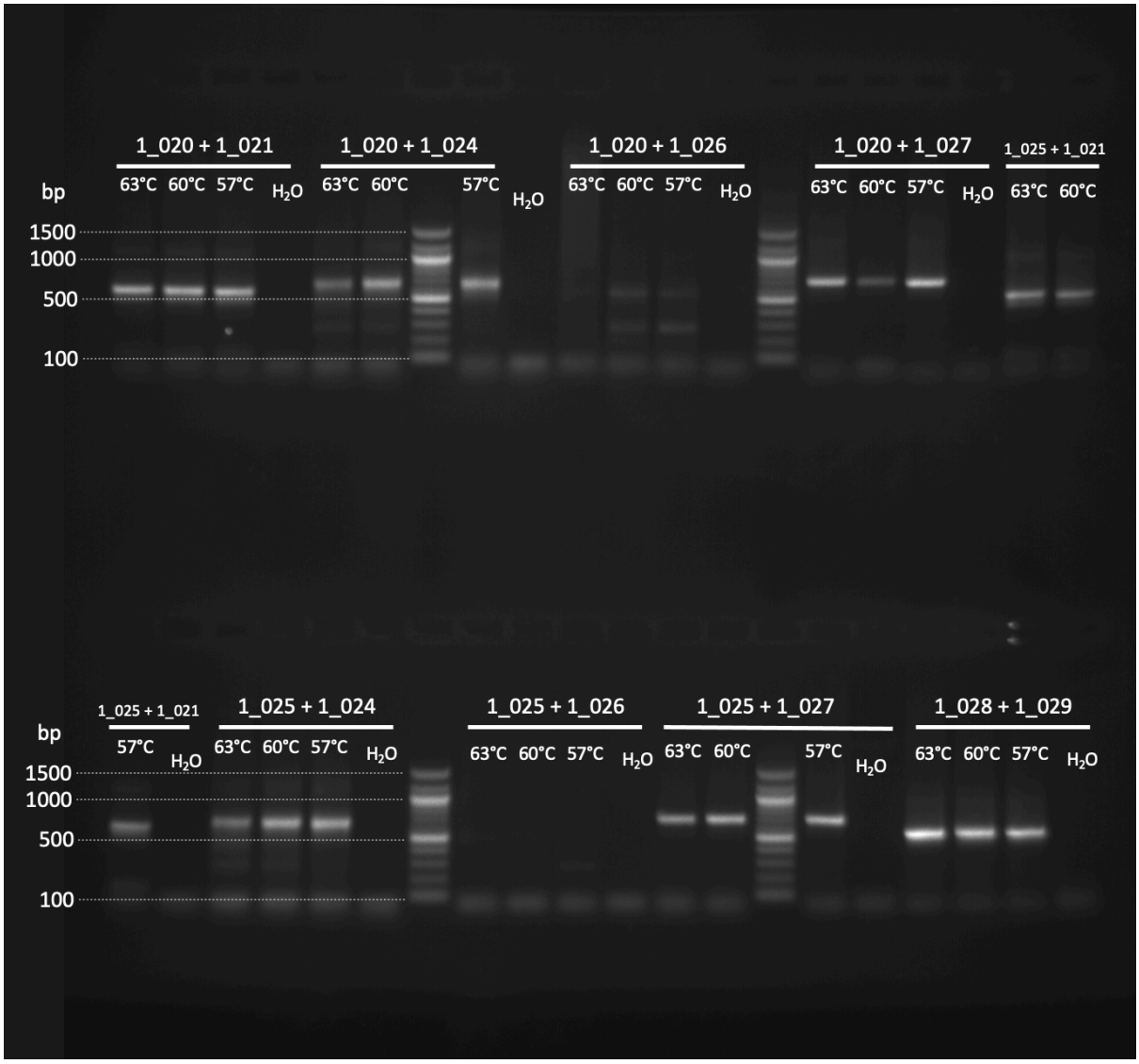


Figure 3.4: Agarose gel electrophoresis of PCR products of different degenerate primer combinations tested on *mmfL/mmfR*-containing plasmid C37_737 at different temperatures. As a benchmark, non-degenerate primers 1_028 and 1_029 were used. Water was used as the negative control.

Primer pair 1_020/1_027 was further tested using different strains of *Streptomyces* strains as well as unrelated bacteria (Figure 3.5). The results showed that the primers picked up other homologs in *S. avermitilis, S. violaceoniger, S. hygroscopicus, Streptomyces sp.* BTG678 and the Antarctic *Streptomyces* isolates 3I4 and 2III1. However, several samples showed two bands, and only three of the samples could be confirmed using Sanger sequencing. This could be related to the promiscuous nature of degenerate primers, the abundance of *tetR* genes in

actinobacterial genomes as well as the issues arising from using degenerate primers for Sanger sequencing. To test for false positives, bacteria which did not contain target genes were screened (genera *Streptomyces*, *Nocardioides, Hymenobacter, Kocuria, Rhodococcus, Ralstonia, Escherichia, Pseudomonas* and *Flavobacterium*). Of all these samples, only one of the three *Rhodococcus* strains showed a signal. Subsequent Sanger sequencing revealed that this arose from a *tetR* gene adjacent to a membrane protein. Importantly, the *Streptomyces* strain known not to contain a *mmfL* homologue (*S. albus* DSM 41398) did not show a signal.

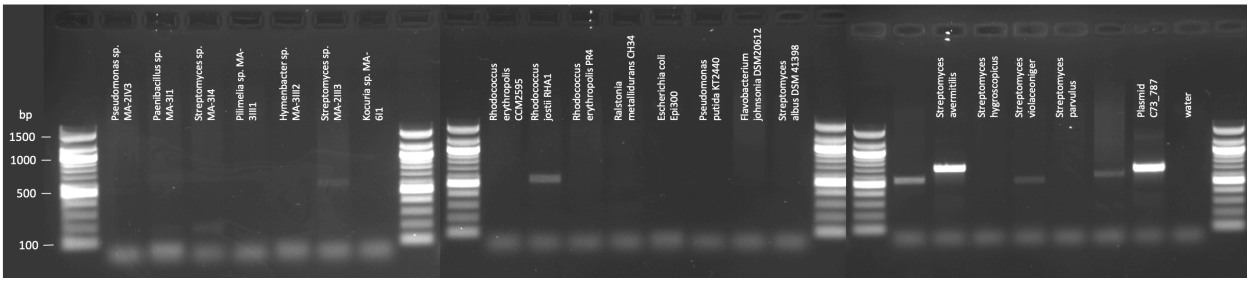


Figure 3.5: Testing of primers 1_020/1_027 on genomic DNA of different bacteria.

Taken together, the primers 1_020/1_027 were able to detect many different *mmfL/mmfR* homologs, picked up signals from two strains that were isolated from the Antarctic soil sample, and showed only one false positive when tested on a panel of phylogenetically diverse bacteria. Therefore, they were chosen for screening of the Antarctic metagenomic library.

3.4.1.3 Screening of metagenomic library

The metagenomic library was PCR screened with primer pair 1_020/1_027. However, no hits could be observed after repeated attempts. To check if there was a problem with the eDNA or library, a previously published set of degenerate primers targeting NRPS A-domains was tested on eDNA and the pooled library. While the NRPS primers showed clear bands, none were observed for 1_020/1_027 (Figure 3.6A). Furthermore, primers for specific BGCs identified from the sequenced metagenome also showed clear bands and the clones could be successfully

isolated and verified using Sanger sequencing (Figure 3.6B). This indicated that the library was functional and contained metagenome-derived BGCs.

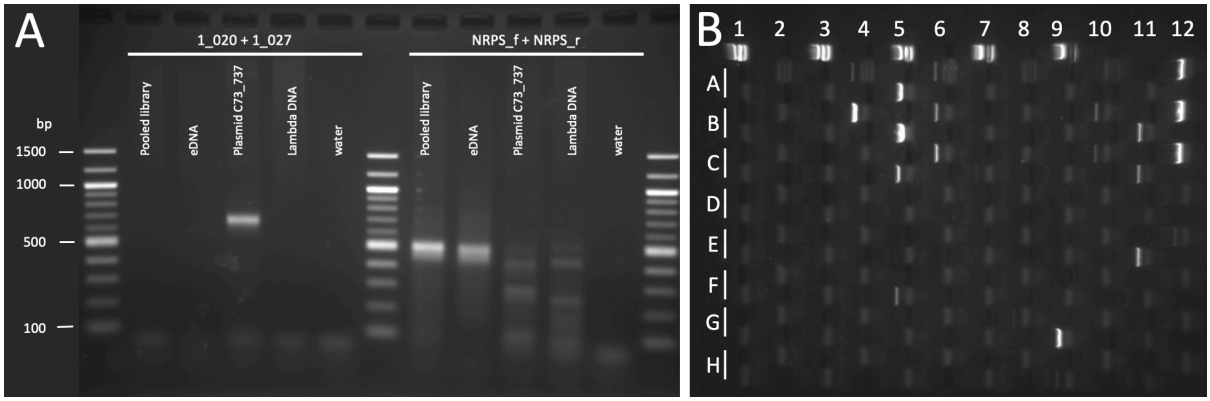


Figure 3.6: PCR screening using different primers on metagenomic library DNA. A) Comparison of *mmfL/mmfR*-targeting primers and NRPS-targeting primers on eDNA, library DNA and controls; b) Screening of the metagenomic library with primers 2_005/2_006 specific for a metagenome-derived NRPS/PKS hybrid

The total lack of hits for *mmfL*/*mmfR* was puzzling, since the primers had been shown to work on *Streptomyces* isolates from the same soil. Furthermore, the taxonomic assessment of the metagenome with kraken2 showed that more than 10% of reads were classified as *Streptomyces*. Since the DNA extraction method was the same for metagenomic library construction and sequencing, *Streptomyces* DNA should have been abundant in the library. The high reported abundance furthermore matched the observation from isolation plates, where there were many colonies matching *Streptomyces* morphology. However, taxonomic assignment of assembled metagenomic BGCs using contig-level classification and binning showed that only one out of >1400 BGCs and only 41 out of 49,262 contigs were assigned to *Streptomyces*, both corresponding to ca. 0.1% of the total. Agreeing with this, a BLASTn search of *mmfL* against the assembled metagenome did not produce any significant hits. This discrepancy between the read taxonomy and the contig taxonomy could be explained by two hypotheses:

1. The kraken2 read assessment was wrong – no significant amount of *Streptomyces* was present in the extracted DNA. This could be explained by the known problem of database-reliant methods like kraken2 that are prone to misclassify unknown sequences.
2. The *Streptomyces* reads did not assemble into contigs. This could be explained by the known difficulty of assembling *Streptomyces* genomes into continuous contigs, and by the high frequency of genomic rearrangements in natural *Streptomyces* populations, which could lead to assembly problems.

3.4.2 16S rRNA gene amplicon sequencing

To understand whether 16S rRNA gene amplicon sequencing would detect any significant amount of *Streptomyces* in the soil DNA or the library, V3 and V4 regions were amplified and sequenced from three different samples:

A. High molecular weight soil DNA extracted by gentle chemical lysis using CTAB. This reflected the DNA used for metagenomic library construction and sequencing.
B. Lower molecular weight soil DNA extracted using the FastPrep soil kit. This more destructive extraction method would lead to a better representation of hard-to-lyse organisms.
C. Pooled metagenomic library plasmids that reflected the DNA present in the metagenomic library.

The results of the 16S rRNA gene sequencing showed that very few reads were assigned to the genus *Streptomyces* (Table 3.4). The FastDNA Soil Kit (B) showed a slightly higher proportion of *Streptomyces* reads compared to the CTAB sample (A) and the library plasmids (C), which might be caused by the harsher lysis by bead-beating that is more efficient at lysing spores.

Since no duplicates were run, the variability of these counts remained unknown. Still, it is evident that the genus assignment based on 16S rRNA gene sequencing strongly conflicted with the kraken2 read assignment. In fact, the values around 0.1% agreed with the assignment of ca. 0.1% of assembled contigs and 0.1% of BGCs to *Streptomyces*. Since this confirmed the suspicion that that the target DNA was not present in the library, no further screenings were conducted.

Table 3.4: 16S rRNA gene sequencing of different DNA sources and the contribution of Streptomyces reads.

Sample	Extraction Method	Total reads	Streptomyces reads	% of total reads
A	CTAB + beads	19,442	11	0.06
B	FastDNA Soil Kit	17,248	41	0.24
C	Library miniprep	8,255*	11	0.13

*After removing *E. coli* reads, the library host organism

3.4.3 Isolation of Antarctic soil bacteria

An isolation experiment using soil extract/nutrient agar was conducted to complement the metagenomic approaches. 66 isolates were recovered and successfully identified using 16S rRNA gene sequencing and BLAST alignment to the NBCI database (Figure 3.7A). Out of these 66 isolates, 34 (51.5%) were classified as Actinobacteria, 17 as Alphaproteobacteria (25.8%), 8 as Gammaproteobacteria (12.1%), 5 as Bacteroidetes (7.6%), and 3 as Firmicutes (4.5%). Out of 34 actinobacterial isolates, 18 were classified as belonging to the order of Streptomycetales, 6 as Micrococcales, 5 as Propionibacteriales and 4 as Frankiales. Three Firmicutes were isolated, all belonging to the Bacillales order. All five Bacteroidetes isolates were classified as Cytophagales (*Hymenobacter* genus). Out of 17 Alphaproteobacteria, 11 were Sphingomonadales, 5 Rhizobiales and 1 Acetobacterales. Out of 8 Gammaproteobacteria,

6 were classified as Betaproteobacteriales (all Burkholderiaceae), and two were assigned Pseudomonadales. While *Streptomyces* species were abundant and easily isolated from the lower dilutions, they were notably less common in the higher dilutions. Members of the Frankiales and Propionibacteriales were only isolated from dilutions of 10^{-2} or more. Fifteen isolates (ca. 23%) showed a similarity equal or lower than 97% to isolates in the NCBI 16S rRNA gene database, indicating novel species.

To assess how many identical or closely related strains were isolated, all full-length 16S rRNA gene sequences were aligned to each other and sequence pairs with 99.5% identity were visualised in a network graph (Figure 3.7B). This revealed three clusters of closely related *Streptomyces* as well as two *Streptomyces* isolates connected to these clusters by only a single edge. Furthermore, three *Sphingomonas* isolates as well as a pair each of *Nocardioides, Massilia_A, Pseudarthrobacter_A* and *Pseudomonas_E* showed >99.5% sequence identity. When using this cut-off for sequence identity, there was only a small amount of redundancy in the isolates, with only 15 of 66 samples being duplicates of already sampled strains, indicating a great diversity of readily cultivable bacteria that were not sampled.

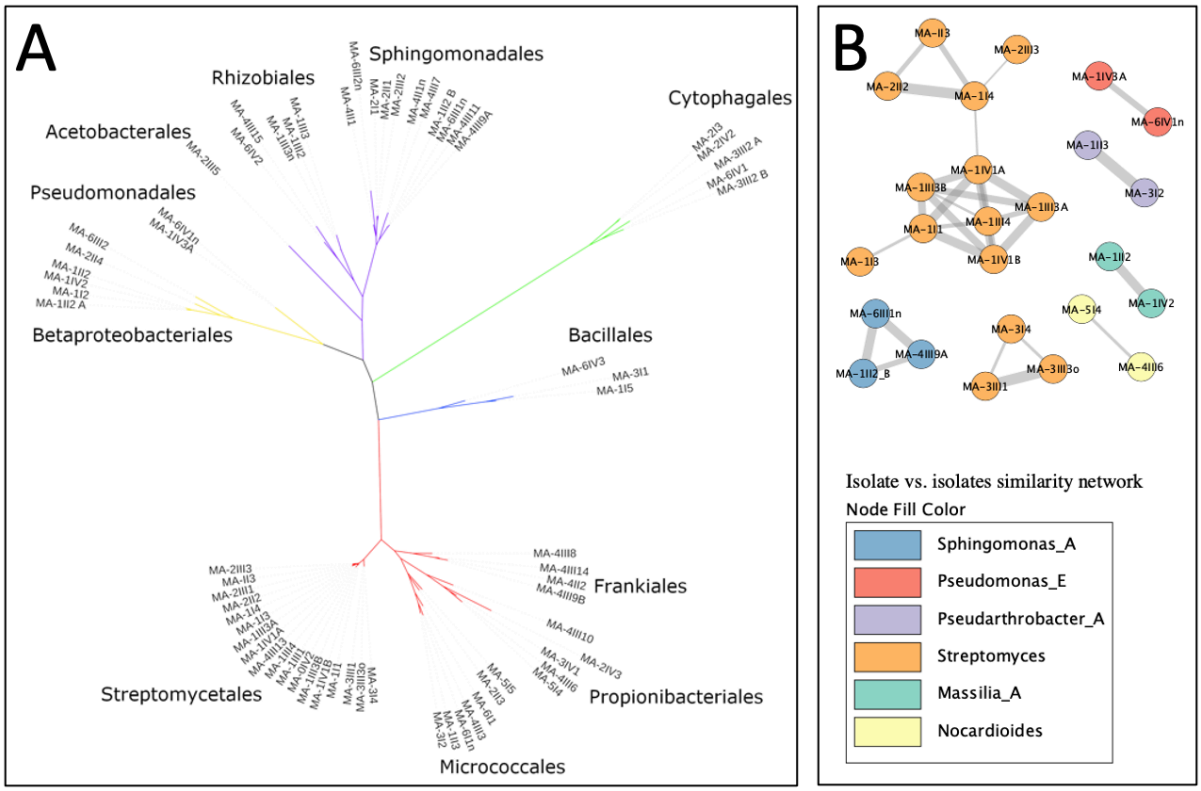


Figure 3.7: Isolate diversity recovered from Mars Oasis soil. A) Tree based on alignment of all Sanger 16S rRNA gene sequences; B) Network graph showing clusters of isolates with ≥99.5% identity in BLASTn, line thickness corresponding to identity.

3.4.4 Sequencing and analysis of isolates

A small but diverse set of isolates was sequenced using Illumina and Oxford Nanopore (Table 3.5). CheckM completeness was ≥98.6% for all hybrid and short read assemblies, with the long-read assemblies trailing behind with 92.8% (2I3) and 75.6% (6III1). The low score for 6III1 can be explained by the relatively low coverage, making error correction less efficient. The estimated ANI to the nearest genome in RefSeq was below 95% for all isolates except for 2III2, indicating novel species. Supporting this, GTDB-Tk classified all isolate at genus level, but not species level.

Table 3.5: Isolate classification, assembly type, CheckM completeness, BGC count as detected by antiSMASH, closes RefSeq match, ANI to RefSeq match and isolate details of RefSeq match

No	Phylum (GTDB)	Genus (GTDB)	Assembly	percent complete	BGCs	Match RefSeq ID	Match name	Match ANI	Match Origin/notes
2IV3	Actinobacteriota	*Nocardioides*	hybrid	98.96	3	GCF_900103935	*Nocardioides szechwanensis*	0.829	Glacieo, China
1III3	Actinobacteriota	*Knoellia*	hybrid	99.73	5	GCF_000152705	*Janibacter* sp. HTCC2649	0.872	Ocean water, Bermuda
2I3 white	Actinobacteriota	*Tsukamurella*	SR only	99.64	14	GCF_001575225	*Tsukamurella tyrosinosolvens*	0.902	Originally from blood but found in many environments
3I4	Actinobacteriota	*Streptomyces*	hybrid	99.23	26	GCF_000429085	*Streptomyces flavidovirens* DSM 40150	0.868	Soil (USSR)
4III3	Actinobacteriota	*Arthrobacter*	hybrid	99.31	4	GCF_000816565	*Arthrobacter* sp. L77	0.806	Pangong Lake, a subglacial lake in the Himalayas
2III1	Actinobacteriota	*Streptomyces*	hybrid	99.05	25	GCF_000720115	*Streptomyces* sp. NRRL F-2580	0.879	Soil (USA)
3I1	Firmicutes	*Fictibacillus*	SR only	98.68	4	GCF_001999465	*Fictibacillus arsenicus*	0.922	Siderite concretion in West Bengal; high arsenic resistance
6IV3	Firmicutes	*Bacillus*	SR only	99.45	4	GCF_000517385	*Bacillus boroniphilus* JCM 21738	0.804	Boron-rich soil in Hisarcik, Turkey; Requires boron
1I5	Firmicutes_I	*Paenibacillus*	SR only	99.85	10	GCF_000787385	*Paenibacillus* sp. P1XP2	0.814	Food Waste Bioreactor (Australia)
2I3	Bacteroidota	*Hymenobacter*	LR only	92.82	3	GCF_900115775	*Siccationidurans arizonensis*	0.869	Biological soil crusts from arid areas, southwestern USA
2III2	Proteobacteria	*Sphingomonas*	SR only	99.66	3	GCF_001421625	*Sphingomonas* sp. Leaf226	0.994	Leaf of Arabidopsis thaliana, Switzerland
6III1	Proteobacteria	*Sphingomonas*	LR only	75.55	3	GCF_000711715	*Sphingomonas astaxanthinifaciens* DSM 22298	0.789	Radioactive water, Tottori, Japan; gamma ray tolerance
1IV3A	Proteobacteria	*Pseudomonas*	hybrid	100	10	GCF_001424925	*Pseudomonas* sp. Root329	0.931	Roots of Arabidopsis thaliana, Potsdam

3.4.4.1 *Comparing isolate genomes to the metagenomic assembly*

The presence of both isolates as well as a long-read metagenomic assembly enabled a multi-level comparison based on 16S rRNA gene sequences, sequenced genomes and BGCs.

16S rRNA gene gene comparison indicated a small overlap between metagenome assembly and isolates (Figure 3.8). Using BLASTn alignment with 97% identity cutoff, 57 matches between the two datasets were found, stemming from a combination of 26 assembly-derived sequences aligning to 12 isolate-derived sequences. 35 of the matches were *Sphingomonas* and *Sphingomonas*_A, indicating the presence of several closely related strains in the sample. No isolate-derived and metagenome-derived sequences showed 100% identity between them, and only two *Nocardioides* isolates (MA-5I4 and MA-4III6) showed >99.5% identity to a metagenome-derived sequence.

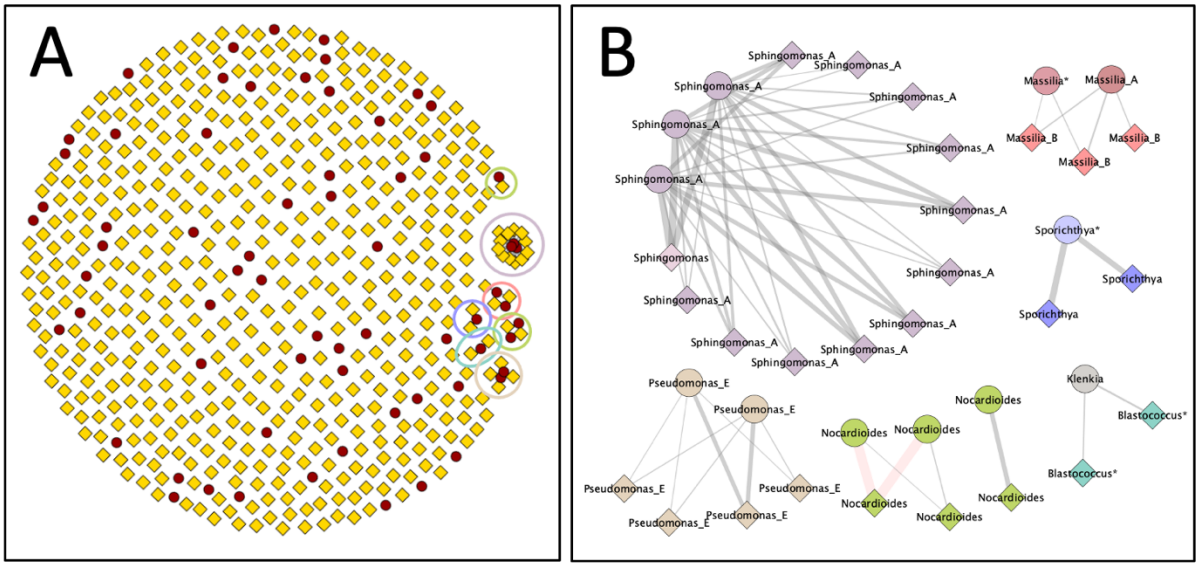


Figure 3.8: BLAST-derived network of isolate and metagenomic assembly 16S rRNA gene sequences. A) Network graph of all isolate (red circles) and metagenomic assembly (gold squares) 16S rRNA gene sequences with a 97% identity cutoff for clustering. Circles show the clusters further detailed in B. B) Network graph containing only those 16S rRNA gene sequences that cluster together in A. Diamond = metagenome 16S, circle = isolate 16S, a. thicker edge corresponds to a higher % identity, with ≥99.5% identity marked in red. Different classification because NCBI blast of sequences vs GTDB classification of contigs

Sequenced isolates were not well represented in the metagenome. Mapping of the metagenomic nanopore reads to the isolate genomes revealed 10 out of 13 isolates were barely present in the metagenomic reads with median coverages from 0x to 2x (Figure 3.9A). 1IV3A (*Pseudomonas*) and 2IV3 (*Nocardioides*) showed 8x to 9x median coverage, putting them into the low end of coverages observed in the assembled metagenome. However, the fraction of each of the genome that had any metagenomic reads mapped to it was well below one, reaching a maximum of 0.75. The *Sphingomonas* isolate 6III1, however, stood out as an exception. It showed a high median coverage of 90x, and 97% of its genome were mapped. This would most likely allow for a contiguous assembly from the metagenome. The two sequenced *Streptomyces* isolates 2III1 and 3I4 had a median coverage of 0, with <25% of the genome mapped, indicating extremely low abundance in the metagenomic DNA.

3.4.4.2 Isolate and metagenome BGC comparison

The isolate BGC coverage obtained from mapping the metagenomic nanopore reads to the BGCs reflected the total genome coverage (Figure 3.9B). BiG-SCAPE assessment networking at cut-off of 0.7 revealed no shared families between the metagenome and the isolate BGCs.

The phylogenetic distribution of BGCs detected in the sequenced isolates was markedly different from the distribution in the metagenome (Figure 3.9C). While in both cases Actinobacteriota contributed a large number of BGCs, these were mostly from *Streptomyces* in the isolates. In the metagenome, only one BGC was assigned to the genus *Streptomyces*. Proteobacteria came second in the isolates, but compared to *Streptomyces* did not contribute many BGCs. No difficult-to-isolate phyla such as Planctomycetota, Acidobacteriota or Verrucomicrobiota or Chloroflexota were isolated either.

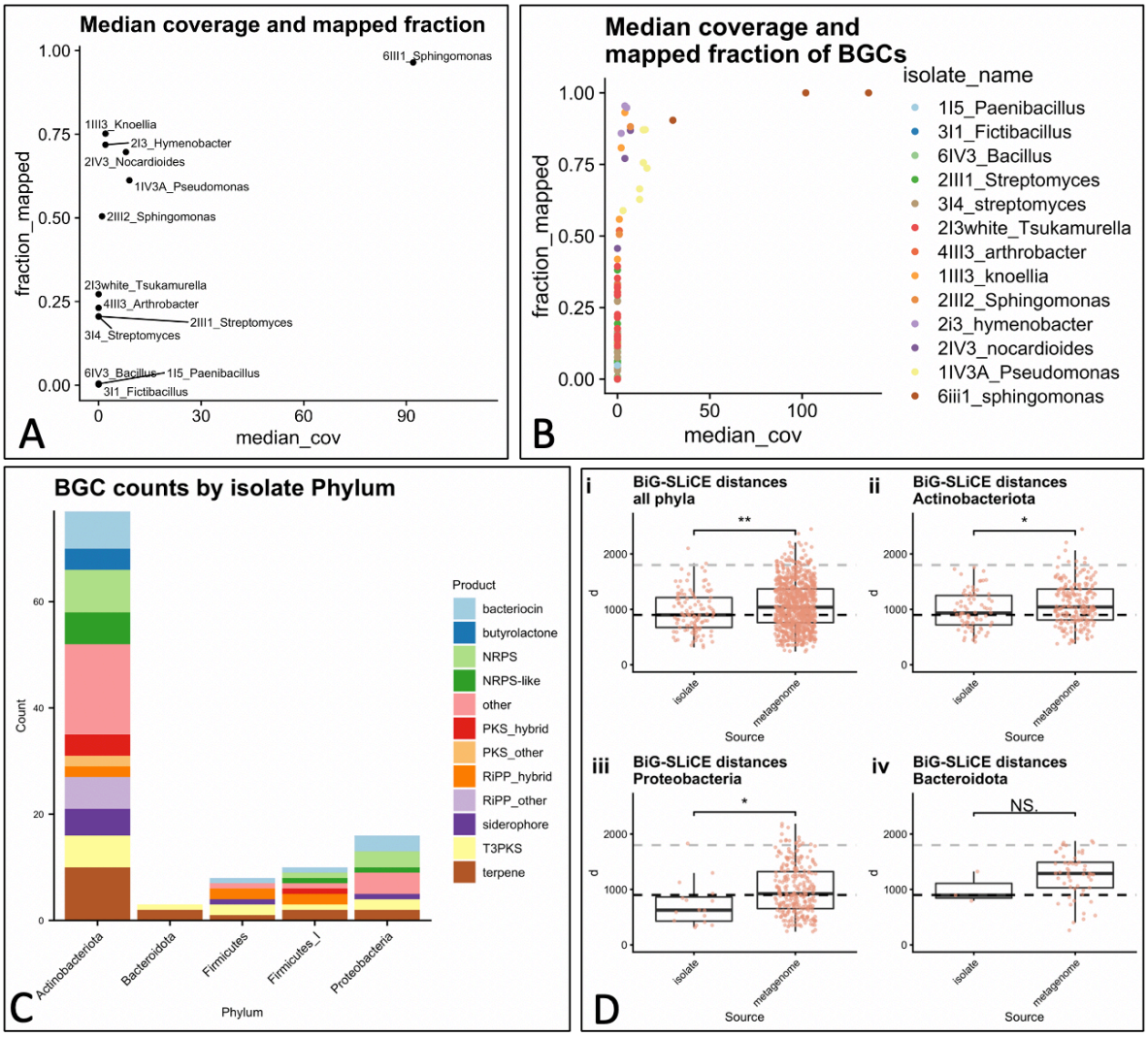


Figure 3.9: Comparison between isolates and the metagenome: A) Median coverage and mapped fraction of isolate genomes in the metagenomic nanopore reads; B) Median coverage and mapped fraction of isolate BGCs in the metagenomic nanopore reads; C) Number and type of BGCs by isolate phylum; D) Comparison of BiG-SLiCE distance scores between metagenomic BGCs and isolate BGCs: i) All phyla p = 0.006; ii) Actinobacteriota p = 0.014; iii) Proteobacteria p = 0.041; iv) Bacteroidota p = 0.252 (not significant)

To compare the sequence novelty of BGCs contained in the isolate genomes and the BGCs obtained from the metagenomic assembly, the isolate and metagenome BGCs were scored using BiG-SLiCE and compared using the Wilxocon Rank Sum test (Figure 3.9D). The results showed a highly significant difference when comparing all BGCs (p = 0.006). Differences within the actinobacterial and proteobacterial BGCs were also significant (p = 0.014 & p =

0.04). Bacteroidota BGCs did not show any significant differences, likely due to the small sample size in the isolate data. No Firmicute BGC were present in the metagenomic BGCs, making a comparison impossible.

3.5 Discussion

3.5.1 Metagenomic library screening using regulatory genes

The purpose of this study was to recover BGCs from a metagenomic library by screening for BGC-associated regulatory cassettes, which would then provide a convenient way for activation of the recovered BGCs. While metagenomic library construction, primer design and assay verification were successful, the objective of recovering BGCs from the metagenomic library using *mmfL*/*mmfR* degenerate primers was not accomplished due to the lack of target sequences in the library.

The absence of *Streptomyces* DNA in the library came as a surprise after having observed many filamentous, streptomycete-like colonies on isolation plates and kraken2 having identified 10% of metagenomic reads as *Streptomyces*. The abundance in *Streptomyces* isolates, yet lack of *Streptomyces* DNA can be seen an example of the Great Plate Count Anomaly. This term describes the discrepancy between the amount of bacterial cells present in soil and the colonies obtained on agar plates (Staley and Konopka 1985). *Streptomyces*, as heterotrophic generalists known to grow on a large variety of media and produce copious amounts of dormant spores, were likely to outcompete other bacteria and thereby be overrepresented on isolation plates (Schlatter et al. 2013). However, the relative absence of *Streptomyces* from the higher dilution plates gave a hint about the true abundance. The misclassification of reads as *Streptomyces* by kraken2 can be explained by the absence of closely related sequences in the GTDB database, as well as the low confidence setting employed. Since the abundances of *Streptomyces* and other *mmfL*/*mmfR*-containing actinomycetes can be higher in other soils and especially in the rhizosphere of plants, the approach might have some use there, or in bigger libraries (Viaene et al. 2016). The present results underline the need for critical examination of classification results and awareness of biases.

It would be desirable to extend the "discovery by regulatory genes" approach to the bacteria whose DNA is more abundant in the library. While regulatory genes involved in specialised metabolite biosynthesis have been discovered in Proteobacteria, no regulation cassettes associated as closely with secondary metabolism as *mmfL*/*mmfR* have been characterised to the author's knowledge (Brotherton, Medema, and Greenberg 2018; Mao et al. 2017; Wallenstein et al. 2020). Therefore, screening for transcriptional regulators such as LAL family proteins would likely produce many hits outside of BGCs (Schrijver and Mot, 1999). It would be possible to screen for colocalization of biosynthetic genes and regulatory genes by using one primer binding to e.g. an NRPS A-domain and one primer binding to a regulatory gene. This would, however, only discover known biosynthetic classes that can easily be screened for, thereby removing a hypothesised key advantage of the *mmfL*/*mmfR* screen. Furthermore, there is no research on the regulators of secondary metabolism of mostly uncultured phyla such as Acidobacteriota which were abundant in the sample.

Another application for the identified conserved sequences in the *mmfL*/*mmfR* genes could be the targeted interruption of the genes using CRISPR. While some *mmfL*-homologues (such as *mmfL* itself) are involved in activation of specific BGCs, others such as *afsA* are responsible for a multitude of effects. For example, disruption of *mmfL*-homologue *farX* in *S. lavendulae* FRI-5 led to abolishing the production of the blue pigment, but increased production of D-cycloserine, which is usually only produced in the first hours of growth (Kitani et al. 2010). Culp et al. exploited the conservation of genes in the streptothricin and streptomycin BGCs by targeting and deactivating them simultaneously in a dozen of previously characterised *Streptomyces* isolates, thereby shifting their specialised metabolite profile and enabling discovery of novel compounds produced by the bacteria (Culp et al. 2019). In a similar fashion,

the simultaneous deactivation of pleiotropic regulators in multiple strains could enable the discovery of otherwise “hidden” natural products.

3.5.2 Isolation

The genera isolated are comparable to isolation studies conducted on other soils in Antarctica, with *Arthrobacter, Pseudarthrobacter, Nocardioides, Microlunatus, Hymenobacter, Pseudomonas, Streptomyces, Sphingomonas, Sphingobium, Massilia, Variovorax, Fictibacillus, Paenibacillus* and *Bacillus* being commonly isolated (Pudasaini et al. 2017; Vander Schaaf et al. 2015; Aislabie et al. 2013; Tomova et al. 2015; Siebert and Hirsch 1988; Smith et al. 2006). Isolates of the genera *Beijerinckia, Caballeronia, Kocuria, Rhizobium, Clavibacter, Knoellia, Tsukamurella* and *Roseomonas* are less common in literature (Pulschen et al. 2017; A. V. da Silva et al. 2021; Kuhn et al. 2014; Nakai et al. 2013; P. Yan et al. 2012; Selbmann et al. 2010; Yi Pan et al. 2013; Wong et al. 2019). No reports were found on *Blastococcus* and *Sporichthya* isolates from Antarctica.

A high species-level novelty was observed from the sequenced isolates, with 12 out of 13 isolates having an ANI lower than the commonly used interspecies threshold of 95% (Jain et al. 2018). This indicates a large amount of novel bacterial diversity also in readily culturable bacteria. However, no new genera were found. With regards to the potential for discovering novel natural products, a recent global analysis of BGCs in MAGs and isolate genomes showed that species within a genus had a much more homogenous BGC diversity than genera within a family (Gavriilidou et al. 2021). However, speciation also has been shown to go hand in hand with specialised metabolite diversification at least in the genus *Salinispora*. (Chase et al. 2020). This indicates that novel species within well-known genera could potentially contain novel

congeners of known products, making readily culturable bacteria a worthwhile target to explore to discover compounds related to known natural products.

The closest related Refseq isolates give indications of the physiology of the Mars Oasis isolates. These were mostly isolated from unusual or extreme environments, including traits such as low temperature, aridity, and presence of stressors like arsenic, boron or radiation. They were of soil, aquatic or plant-associated origin, which reflects the characteristics of Mars Oasis as an arid soil with seasonal meltwater ponds as well as stands of bryophytes (Convey and Smith 1997). Previous studies did not show particularly high levels of copper, lead, zinc, iron or nickel, other heavy metals such as arsenic or mercury were not examined (Chong et al. 2012). It has been reported that other Antarctic soils and sediments show high heavy metal contents as a result of anthropogenic pollution and natural processes such as food chain accumulation, and as a consequence harbour heavy metal resistant bacteria (Romaniuk et al. 2018; Magesh et al. 2021; Tomova et al. 2015; Stoilova-Disheva, Vasileva-Tonkova, and Tomova 2014). It is possible that the present isolates hold similar capabilities. Furthermore, siderophore production has been implicated in heavy metal resistance, indicating that part of the biosynthetic arsenal of these organisms could be devoted to metal detoxification (Schalk, Hannauer, and Braud 2011; Hussein and Joo 2014; Hesse et al. 2018).

The overlap between sequences from the metagenome and sequences obtained from isolates was relatively small, with only some 16S rRNA gene sequences showing overlap and one *Sphingomonas* isolate showing a high coverage in the metagenome reads. This stark contrast between bacterial diversity observed using culture-dependent and culture-independent methods is well-documented and can be attributed to the inherent limitations of culture-dependent approaches (Prakash et al. 2021; Tytgat et al. 2014). Many abundant bacteria are

hard to culture due to factors such as slow growth, the absence of suitable energy sources, unsuitability of agar for the isolation as well as sensitivity to hydrogen peroxide generated by autoclaving of phosphate buffered medium (Vartoukian, Palmer, and Wade 2010; Kato et al. 2018; Tamaki et al. 2009). In the present study, soil extract was used to provide nutrients, but the use of agar and the short incubation time made the isolation of e.g. Acidobacteria or Verrucomicrobia unlikely, as their successful isolation has usually been reported using gellan gum and incubation times of up to three months (Kielak et al. 2016; Janssen et al. 2002; George et al. 2011). Additionally, isolating the highly abundant, potential atmospheric methane oxidisers UBA7966/*USCy* would most likely not have been achieved on agar plates with the incubation time used. The successful isolation of the slow-growing atmospheric methane oxidiser *Methylocapsa gorgona* required several months of incubation in liquid medium (Tveit et al. 2019).

The overlap in BGCs between sequenced isolates and the metagenome was also very limited, and the isolate-derived BGCs were markedly more similar to BGCs in databases. This finding indicates that the chance of discovering the same BGCs in uncultured bacteria and culturable bacteria from the same soil is low. This agrees with the findings that despite prominent examples of horizontal gene transfer between distantly related organisms, vertical inheritance is an important driver in defining BGC diversity of bacterial genera (Chase et al. 2021; Jensen et al. 2007).

4 Results 3: Cloning and expression of metagenomic BGCs

4.1 Introduction

4.1.1 Expression of metagenomic BGCs so far

With the advent of genome sequencing twenty years ago, it became clear that many BGCs in cultivable bacteria are silent under laboratory conditions. The increasing amount of sequence data has therefore led to a mounting number of BGCs without an associated product (cryptic BGCs). There are, however, many ways to "awaken" these BGCs and obtain products. While the manipulation of culture conditions is an established approach that has yielded a great number of compounds (Lincke et al. 2010; Rateb et al. 2011; Onaka et al. 2011; Akhter et al. 2018), molecular techniques like overexpression or knockouts of transcriptional regulators (Sidda et al. 2013; Bunet et al. 2011; Laureti et al. 2011; Alberti et al. 2019) as well as heterologous expression (Alberti et al. 2019; Gomez-Escribano et al. 2019; Qian et al. 2020; Lin, Hopson, and Cane 2006; Saleh et al. 2012) have become routine.

As a consequence of falling sequencing costs, the study of environmental and host-associated metagenomes became common, revealing an even greater diversity of BGCs present in the uncultivated majority of bacteria (Chen et al. 2020; Crits-Christoph et al. 2018; Sharrar et al. 2020; Borsetto et al. 2019). However, accessing these BGCs and producing the compounds encoded in the BGCs has been a major challenge (L. Robinson, Piel, and Sunagawa 2021). The only way to obtain a product from a BGC found in an unculturable bacterium is by heterologous expression. In order to perform heterologous expression, the DNA encoding the BGC must be cloned into a plasmid that can be introduced into an expression host. Until recently, metagenomic libraries were the only feasible way to recover metagenomic BGC DNA. In libraries, metagenomic DNA can be stored in an easily accessible way, and whole BGCs can be recovered even if only a fraction of their sequence is known (e.g. from amplicon

sequencing). This enables recovery of BGCs even if their full sequence is not known, which avoids the need for ultra-deep sequencing of complex metagenomes (Hover et al. 2018). It has been estimated that screening of large libraries can recover BGCs from soil that would only be assembled at sequencing depths of multiple terabases of short-read data (Libis et al. 2019). A key drawback of metagenomic libraries is the labour-intensive, iterative (q)PCR screening and dilution process required to recover a single BGC fragment from a large library.

The increase in metagenomic sequencing has enabled "metagenome mining" approaches for exploration of BGCs in shotgun metagenomes. Targeted recovery of BGCs identified through metagenome mining has resulted in production of natural products through heterologous expression in several examples from sponge metagenomes (Agarwal et al. 2017; Nakashima et al. 2016; Freeman et al. 2012). In these organisms, the concentration of specific metabolites produced by bacterial symbionts can be high enough for purification and structural elucidation, thereby providing a structure and often a biological activity. Their relatively simple microbiome enables assembly of BGCs without ultra-deep sequencing. Most metagenome mining approaches still employ targeted screening of metagenomic libraries. However, a library-free approach was taken by Agarwal et al., who used synthetic genes cloned into a cyanobacterial expression host (Agarwal et al. 2017).

Despite the falling prices of DNA synthesis, gene synthesis is still an expensive undertaking and only viable for small BGCs. Furthermore, any errors present in the sequence – introduced for example through sequencing errors – are propagated. These errors might not be obvious but can be detrimental to the function of an enzyme and are especially common in nanopore sequencing (Y. Wang et al. 2021). Recently, DiPaC has been demonstrated as a viable technique for heterologous expression of BGCs from isolates (D'Agostino and Gulder 2018;

Greunke et al. 2018). DiPaC relies on PCR-amplifying BGC operons using high-fidelity polymerase and cloning them into an expression vector using homology-based assembly such as HiFi assembly or SLIC. Large BGCs consisting of several tens of kilobases have been cloned in this manner (D'Agostino and Gulder 2018; Greunke et al. 2018).

PCR-based cloning of BGCs as exemplified by DiPaC is compatible with error-prone sequences such as those derived from nanopore sequencing: as long as the sequence is correct for the short stretch where primer binding occurs (ca. 20 bp), successful amplification will lead to a faithful copy of the native DNA. The estimated error rate of 5.3×10^{-7} for Q5 polymerase suggests that amplicons the size of large BGCs (ca. 100,000 bp) are unlikely to contain errors introduced by amplification (Potapov and Ong 2017). While the chance of point mutations or frameshift errors is low, there are other drawbacks to this technique. Through the use of native DNA, there is a chance that the codon usage will not match the tRNA supply in the expression host, thereby preventing translation of the mRNA and making successful expression impossible (L. Robinson, Piel, and Sunagawa 2021). This chance can be minimised by using a phylogenetically related expression host – which is, however, not possible with BGCs from phyla such as Acidobacteriota, which have no established expression hosts.

4.1.1.1 Factors affecting the success of heterologous expression of a BGC

There are many factors affecting the success of heterologous expression of a BGC. Starting at the very beginning, the genes necessary for biosynthesis must be identified and chosen for expression. BGCs are often surrounded by other biosynthetic genes involved in primary metabolism. In absence of a known BGC product, this can make the definition of BGC boundaries difficult, unless the BGC is demarcated by e.g. transposases (W. Li et al. 2009; Blin, Kim, et al. 2019).

When the BGC containing all the genes necessary for biosynthesis of the specialised metabolite is successfully recovered, cloned and transferred into an expression host, the next barrier is sufficient transcription. This can be negatively affected by the inability of native promoters to induce transcription in the heterologous host, by transcriptional repressors in the BGC or by terminator sequences. Employing promoters known to work in the heterologous host and avoiding terminator sequences as well as removing potential negative regulators can all ensure transcription of the BGC (Alberti et al. 2019; Saleh et al. 2012; S.-H. Kim et al. 2019; D'Agostino and Gulder 2018). After an mRNA has been produced through the process of transcription, it needs to be translated into proteins by ribosomes. This can be affected by the compatibility of native ribosome binding sites (RBS) present in the BGC and the ribosomes provided by the host (M. M. Zhang et al. 2016). If ribosomes do not bind to the mRNA, translation will not be initiated. Once translation is initiated, the presence of rare codons in the mRNA can lead to ribosome stalling and termination of translation (Keiler 2015). If all of the above conditions have been met and a biosynthetic enzyme has been produced, the right cofactors and precursors (e.g. cobalamine, non-proteinogenic amino acids or methylmalonyl-CoA) must be present for the enzyme to produce the specialised metabolite (Lanz et al. 2018; Jiang and Pfeifer 2013).

Given that a metabolite is produced in sufficient quantities, it must be detected. This can be achieved either by simple observation through e.g. a pigmented phenotype, through analytical methods such as LC/MS or through activity assays. This in turn depends on the properties of the compound such as stability, solubility, absorption spectrum, ionizability and biological activity.

4.2 Aims and rationale

The overall aim of the work reported here was to achieve sequence-guided cloning and heterologous expression of BGCs from the complex Mars Oasis metagenome without the use of metagenomic libraries. Since the BGCs did not have known products associated with them, this was an exploratory approach akin to those previously taken with BGCs from isolates.

The intermediary goals were to:

1) construct expression plasmids
2) amplify BGCs from the metagenome by PCR
3) clone BGCs into the plasmids using SLIC
4) transfer the plasmids into expression hosts via transformation and conjugation
5) identify and characterise products

4.3 Materials and Methods

4.3.1 Bacterial strains and media

The bacterial strains used in this study can be found in Table 4.1.

Table 4.1: Bacterial strains used in this study

Strain name	Genotype and comments	Incubation T (°C)	Reference
Escherichia coli JM109	*endA1, recA1, gyrA96, thi, hsdR17 (rk–, mk+), relA1, supE44, Δ(lac-proAB), [F´ traD36, proAB, laqIqZΔM15]*	37	Promega
Escherichia coli ET12567/pR9406	*dam-13:: Tn9 dcm-6 hsdM* Chl^R with helper plasmid pR9406	37	(Widdick et al. 2018)
Streptomyces coelicolor M1154	M145 derivative Δ*act* Δ*red* Δ*cpk* Δ*cda rpoB*(C1298T) *rpsL*(A262G)	30	(Gomez-Escribano and Bibb 2011)
Streptomyces albus J1074		30	
Pseudomonas putida KT2440 trfA	*trfA*, rmo-, mod+	30	(Borsetto 2017)
Escherichia coli BL21	F^- *ompT gal dcm lon hsdS*$_B$(r_B^- m_B^-) [*malB*$^+$]$_{K-12}$(λ^S)	37	Promega
Micrococcus luteus		30	

Bacterial strains were kept as cryostocks with 15% v/v glycerol prepared from spore suspensions (*Streptomyces*) or liquid cultures (all other bacteria) flash frozen in liquid nitrogen and stored at -80°C. Spore suspensions were prepared from plates grown on SFM at 30°C for five to eight days as previously described (Kieser et al. 2000).

Streptomyces strains were grown on the following media with and without antibiotics for selection:

- Soybean Flour Mannitol medium (SFM; 20 g/L Soybean flour, 20 g/L D-mannitol, 20 g/L agar)
- Bennet's Glucose Medium (BGM; 1g/L yeast extract; 1g/L beef extract, 2g/L N-Z Amine Type A, 10g/L glucose, 15g/L agar)
- Supplemented Minimal Medium Solid (SMMS; 2 g L^{-1} casaminoacids, 8.68 g L^{-1} TES buffer, 15 g L^{-1} agar, with 10 mL of 50mM NaH_2PO_4 + K_2HPO_4, 5 mL of 1M $MgSO_4$, 18 mL of 50% w/v glucose as well as 1 mL of trace element solution [0.1 g L^{-1} each of $ZnSO_4.7H_2O$, $FeSO_4.7H_2O$, $MnCl_2.4H_2O$, $CaCl_2.6H_2O$ and NaCl] added just before use)

All other bacterial strains were grown with and without antibiotics on the following media:

- Luria-Bertani liquid medium (LB; 10 g/L tryptone, 5 g/L yeast extract, 10 g/L sodium chloride)
- Luria-Bertani solid (LBA; 10 g/L tryptone, 5 g/L yeast extract, 10 g/L sodium chloride, 15g/L agar)
- Terrific Broth (TB; 24g/L yeast extract, 20g/L tryptone 0.4% v/v glycerol, 10% v/v phosphate buffer [0.17M KH_2PO_4, 0.72 M K_2HPO_4])

4.3.2 Vectors

All vectors used in this study (see Table 4.2) were maintained in *E. coli* JM109 and *E. coli* TOP10. Plasmids were extracted with GeneJET Plasmid Miniprep kit (Thermo Fisher Scientific) and stored at -20°C.

Table 4.2: Vectors used in this study.

Vector name	Description	Reference
pBCaBAC	Ap^R; *aac(3)IV*, *oriT*, *ΦC31 attP*, *int ΦC31*	(Borsetto 2017)
pBCkBAC	Km^R; *aphI, oriT*, *ΦC31 attP*, *int ΦC31*	(Borsetto 2017)
pBCaBAC-g1	Ap^R; *aac(3)IV*, *oriT*, *ΦC31 attP*, *int ΦC31*	This study
pBCaBAC-g2	Ap^R; *aac(3)IV*, *oriT*, *ΦC31 attP*, *int ΦC31*	This study
pBCkBAC-g2	Km^R; *aphI, oriT*, *ΦC31 attP*, *int ΦC31*	This study
pBCkBAC-g1	Km^R; *aphI, oriT*, *ΦC31 attP*, *int ΦC31*	This study
TX-TL_SP44_RFP	Amp^R, *ampR*	Contains mScarlet (*Streptomyces* codon-optimised). Patrick Capel, personal communication

4.3.3 gBlocks

All gBlocks used in this study (see Table 4.3) were synthetised by IDT.

Table 4.3: gBlocks used in this study.

Name	Sequence	Length (bp)	Notes
g1	TGCCACCTGACGTCTAAGAATGTGCGGGCTCTAACACG TCCTAGTATGGTAGGATGAGCAAGTTTAAACTTAATTA AATGCATCCTTAGGAGTACTGTGCACGCTAGCATTTAA ATTGGCCACGACTTTACATTAGATGTGCCTTGGTTGTC AAAGCAGAGACGGTTCGAATGTGAACAGCTCACTCAAA GGCGGTAAT	199	P21 and sp24 promoters (convergent) with MCS in middle. Flanked by VF2 and VR primers
g2	TGCCACCTGACGTCTAAGAAGAGAGCGTTCACCGACAA ACAACAGATAAAACGAAAGGCCCAGTCTTTCGACTGAG CCTTTCGTTTTATTTGATGCCTGGATACAATTAAAGGC TCCTTTTGGAGCCTTTTTTTTTGGAGATTTTCAACGTA GGTTCTGTTAAGTAACTGAGTTTAAACTTAATTAAATG CATCCTTAGGTACACCAGACTTTACAACACCGCACAGC ATGTTGTCAAAGCAGAGACGGTTCGAATGTGAACAACC CAATGTCGTTAGTGTGTGCGGGCTCTAACACGTCCTAG TATGGTAGGATGAGCAAAGTACTGTGCACGCTAGCATT TAAATACGCTTACCTCTTAAGAGGTTGCAGATCTGGTA ATCATGGTCATAGCTGTTTCCTGTGTGCTCGAGGCCTG ATGACTCCTGTTGATAGATCCAGTAATGACCTCAGAAC TCCATCTGGATTTGTTCAGAACGCTCGGTTGCCGCCGG GCGTTTTTTATTGGTGAGAATGCTCACTCAAAGGCGGT AAT	535	p21 and sp44 promoters (divergent) with 10bp spacer between it. Around promoters: MCS. Around MCS: Double terminators. Flanked by VF2 and VR primers

4.3.4 Primers

4.3.4.1 Primers for vector construction and verification

The primers used in the process of vector construction, verification of constructs and verification of inserts can be found in Table 4.4

Table 4.4: Primers for vector construction and verification

No	Sequence (5' to 3')	Product length (bp)	Notes
3_001	GAGCTGGTTGCCCTCGCC	9500	Linearisation of pBCk/aBAC vector with removal of KanR/ApraR and flanking sequences
3_002	CATGGGGACGTGCTTGGCAATC		
VF2	TGCCACCTGACGTCTAAGAA	variable	PCR primers for amplifying expression cassette (with insert) and sequencing inserts
VR	ATTACCGCCTTTGAGTGAGC		
3_017	CTTAGGTACACCAGACTTTACAACAC	c. 11400	Linearisation of pBCk/aBAC-g2 for insertion of gene for sp44
3_018	GAACCTACGTTGAAAATCTCCA		
3_019	TTTGCTCATCCTACCATACTAGGAC	c. 11400	Linearisation of pBCk/aBAC-g2 for insertion of gene for p21
3_020	TCTTAAGAGGTTGCAGATCTGG		
3_021	AACTTGCTCATCCTACCATACTAG	c. 11100	Linearisation of pBCk/aBAC-g1 for insertion of gene for p21
3_022	AGCATTTAAATTGGCCACGAC		
3_023	TAAACTTGCTCATCCTACCATACTAG	c. 11000	Linearisation of pBCk/aBAC-g1 for insertion of gene for sp24 OR p21
3_024	AAATTGGCCACGACTTTACATTAG		
3_025	TAGTATGGTAGGATGAGCAAAGTACT TTAACTTTAAGAAGGAGATATACAC	769	Amplification of mScarlet (*Streptomyces* CO) from TX-TL_SP44_RFP for cloning into pBCk/aBAC-g2 under p21
3_026	AGATCTGCAACCTCTTAAGACTTGTA CAGCTCGTCCATGCC		
3_027	CTGTGCGGTGTTGTAAAGTCTGGTG	769	Amplification of mScarlet (*Streptomyces* CO) from TX-TL_SP44_RFP for cloning into pBCk/aBAC-g2 under sp44
3_028	GAGATTTTCAACGTAGGTTCCTTGTA CAGCTCGTCCATGC		
3_029	TGTAAAGTCGTGGCCAATTTGTACTT TAACTTTAAGAAGGAGATATACACAT ATGGTG	1075	Amplification of mScarlet (*Streptomyces* CO) + terminator from TX-TL_SP44_RFP for cloning into pBCk/aBAC-g1 under sp24
3_030	TGGTAGGATGAGCAAGTTTACTCTGG CAAACATATAAACGCAGAAAG		
3_031	TGTAAAGTCGTGGCCAATTTCTCTGG CAAACATATAAACGCAGAAAG	1075	Amplification of mScarlet (*Streptomyces* CO) + terminator from TX-TL_SP44_RFP for cloning into pBCk/aBAC-g1 under p21
3_032	TGGTAGGATGAGCAAGTTTAGTACTT TAACTTTAAGAAGGAGATATACACAT ATGGTG		
3_047	AGCTCATCGCTAATAACTTCG	variable	pBCa/kBAC: Verify gblock cloning
3_048	TTTTAAGGCAGTTATTGGTGC		
3_005	CTTATTCAGGCGTAGCAACCAG	731	Verifying plasmid integration in *Streptomyces* (binds in backbone)
int_seq _RV*	AAGGACTCTTACCGCTGCC		

* from (Borsetto 2017)

4.3.4.2 *Primers for BGC amplification and cloning*

The primers for BGC amplification and cloning can be found in Appendix A, Supplementary Table 1.

4.3.5 Primer design

Primers for the amplification of fragments from BGCs were designed in several steps.

1. The antiSMASH output was visually examined to find BGCs in which all putative biosynthetic and transport genes were organised in one or two operons. ARNold was used to predict rho-independent terminators (Naville et al. 2011).
2. The putative edges of the operons were searched against the nr database with blastx to check for potential indel errors affecting the start and stop coordinates of ORFs in order to avoid cloning partial genes.
3. After the likely start and end of operons were determined, primers were generated using primerBLAST to find the best primers with the least unspecific hits to other bacteria. The sequences were submitted to primerBLAST with the desired coordinates of the primers and the desired T_m. The coordinates were given so that 8 to 30 bp before the putative start codon were included and up to 200 bp after the stop codon were allowed. To avoid unspecific amplification, primerBLAST was set to filter hits to RefSeq representative bacterial genomes.
4. The best primer pair (T_m close to 60°C, lowest potential for secondary structure formation, fewest unspecific hits) was selected and 20 bp of vector homology arms were added, leading to a total length of c. 40-45 bp. Primers were confirmed in silico using SnapGene (yielding in-silico assembled plasmids) and then ordered as lyophilised powder (Integrated DNA Technologies).

4.3.6 PCR

PCRs for amplification of fragments to be cloned was carried out using Q5 polymerase (New England Biolabs) with GC enhancer. PCR for verifying constructs and colonies was done using KAPA Taq (Sigma-Aldrich) polymerase with addition of BSA and DMSO. The first attempt for amplification of a fragment was always carried out using a touchdown protocol. This protocol consisted of a touchdown phase (10 cycles) and an amplification phase (20-25 cycles). The touchdown phase started with an annealing temperature 5-7 °C above the calculated optimal annealing temperature and decreased by 1 °C each cycle, ending 3-5 °C below the optimal annealing temperature. This temperature was then used as annealing temperature of the remaining run. If this did not yield satisfactory results, a range of constant annealing temperatures were tried, though results were mostly the same. Extension time was set at 60 seconds per kilobase.

4.3.7 Agarose Gel Electrophoresis and Gel Purification

Amplification success was assessed on agarose gels (0.4% to 1%) with the addition of 5 µL GelRed (Biotium) per 100 mL gel. 5 µL of each PCR product mixed with 1 µL of purple loading dye (NEB) were loaded into wells. Additionally, 1kb DNA ladder (NEB) or GenerRuler High Range ladder (Thermo Fisher Scientific) were loaded into a well and the gels were run at 3-5 V/cm for one to three hours in TAE buffer and then visualised using UV light. For gel purification of a DNA fragment from a successfully amplified sample, the remainder of the PCR product (45 µL) was run on a gel in the same manner. Fragments were quickly cut out with a razor blade under minimal exposure to UV light. For large fragments, UV exposure was eliminated by running a small part of the sample in a separate well which was exposed to UV light, while the large part of the sample was not. DNA was extracted from agarose chunks with

the Monarch DNA Gel Extraction Kit (NEB), eluted in 6-10 µL and quantified using Qubit (Thermo Fisher Scientific).

4.3.8 Sequence and ligation independent cloning (SLIC)

SLIC (Jeong et al. 2012) was carried out by mixing the following reagents:

PCR-linearised and gel-purified vector	20 fmol
PCR-amplified and gel-purified insert	40 fmol (i.e. twice as much as vector)
Buffer 2.1 (NEB)	1 µL
T4 DNA polymerase (NEB)	0.5 µL
Molecular grade H_2O	to 10 µL

After addition of the enzyme, the reaction was mixed and incubated at RT for 2.5 minutes, then on ice for 10 minutes. Afterwards, the mixture was transformed into chemically competent *E. coli* JM109. Clones were verified using colony PCR using VF2/VR and 3_017/VR primers, leading to the amplification of the inserted fragment. Amplicons were additionally sent for Sanger sequencing with the same primers, verifying the first several hundred bp of insert.

If the concentration to the fragment was too low to achieve the desired amount and vector-to-insert ratio within the 10 µL reaction, both the total amount of DNA as well as the ratio were reduced. No successful cloning was observed below 10 fmol of both vector and insert.

4.3.9 Preparation and transformation of competent cells

Chemically competent *E. coli* JM109 and ET12567/pR9406 were prepared using the TSS method (Chung, Niemela, and Miller 1989). In short, an exponential phase culture was spun down, washed and mixed with ice-cold LB broth containing 10% w/v PEG 8000, 5% v/v DMSO, and 50 mM Mg^{2+}, pH 6.5. The resulting cells were aliquoted and flash frozen in liquid nitrogen. For transformation of *E. coli*, aliquots were thawed on ice until liquid. Ice-cold 5x

KCM (500 mM KCl, 150 mM CaCl2, 250 mM MgCl2) was added, followed by plasmid DNA or the cloning reaction. After incubation on ice for 30 minutes, the cells were heat shocked at 42 °C for 30 seconds, placed back on ice for 5 minutes and then incubated with SOC at 37 °C with shaking for an hour. Serial dilutions were plated and incubated overnight.

Electrocompetent *P. putida* KT-2440 were prepared according to established protocols (New England Biolabs 2015). In short, an exponential phase culture was spun down and washed repeatedly in ice-cold 10% glycerol, resuspended in the same, aliquoted and flash-frozen in liquid nitrogen. For transformation, aliquots were thawed, transferred into a 1 mm electroporation cuvette, DNA was added and transformed through electroporation at 2.5 kV and 200 ohms. SOC was added and the cells were incubated at 37 °C with shaking for an hour. Serial dilutions were plated and incubated overnight.

4.3.10 Conjugation

For conjugating plasmids into *Streptomyces*, they were initially transformed into the methylation-deficient *E. coli* ET12567/pR9406 which also contained the helper plasmid. Conjugations were then conducted as described previously (Kieser et al. 2000). In short, an early exponential phase culture of *E. coli* was washed and plated with germinated *Streptomyces* spores onto SFM with 10 mM $MgCl_2$. After overnight incubation, the plates were overlaid and lightly scrubbed with 1mL of water containing 20µl of 25 mg/mL nalidixic acid to kill off *E. coli*, as well as 150 µL of 50mg/mL kanamycin to select for exconjugants. After several days of growth, colonies were picked and passaged twice on plates containing 25 µg/mL nalidixic acid to kill off *E. coli*, as well as 50 µg/mL kanamycin. Exconjugants were confirmed by PCR with primers 3_005 and int_seq_RV which bind in the vector backbone.

4.3.11 LC-MS

P. putida transformants were grown for 48h at 30 °C with shaking in 50 mL Terrific Broth in 250 mL spring loaded flasks. Supernatant and cells were separated by centrifugation and frozen. For analysis, supernatant was filtered using 0.2 µm spin columns (Thermo Fisher Scientific). Pelleted cells were lysed by adding 25 mL of methanol and sonicating. After centrifugation, the supernatant was filtered with 0.2 µm spin columns.

Streptomyces exconjugants except for the ones harbouring carotenoid BGCs were grown on SMMS. Exconjugants with carotenoid BGCs and control strains were grown on BGM and with exposure to light in an illuminated incubator. After seven days of incubation at 30 °C, plates were extracted using acidified methanol. After solvent evaporation (GeneVac, maximum temperature 37 °C), the residue was resuspended in a 1:1 methanol:water mixture.

All samples were run by Prof. Lijiang Song on a Dionex 3000RS UHPLC with an Agilent Zorbax Eclipse plus column (C18, 100x2.1mm, 1.8µm), which was coupled to a Bruker MaXis Impact Q-TOF mass spectrometer. Mobile phase A consisted of water with 0.1% formic acid, while mobile phase B consisted of acetonitrile with 0.1% formic acid. The flow rate was constant at 0.2 ml/min. Runs were started with 5 minutes of 5% B (isocratic). Thereafter, a gradient from 5% to 100% B within 15 minutes was employed. Ionisation was achieved through electrospray ionisation and the scan range was set to 50 – 3,000 m/z. For each sample, positive and negative ion mode were employed in separate runs. Source conditions were as follows: nebulizer gas (N_2) at 1.4bar; dry gas (N_2) at 8L/min; end plate offset at -500V; capillary at -4500V; dry temperature at 180 °C. Ion transfer conditions were: ion funnel 1 RF at 200Vpp; ion funnel 2 RF at 200vpp; quadrupole ion energy at 5ev; quadrupole low mass set at 55m/z; collision energy at 5.0ev; hexapole RF at 200Vpp; collision RF ramping from 800 to

1500 Vpp; pre-Pulse storage time set at 5 µs; transfer time set from 100 to 155µs. For calibration, 20 µL of 10 mM sodium formate were injected at the beginning of each run. For analysis, ion and UV-VIS chromatograms were compared to the control samples and checked for new peaks using DataAnalysis (Bruker). Ion chromatograms were checked in steps of 100 m/z.

4.3.12 Antimicrobial assays

Agar plug diffusion was used for assessing antibiotic activity against *Escherichia coli* JM21 and *Micrococcus luteus. Streptomyces* exconjugants were grown on plates as for LC-MS. Since *P. putida* transformants are dependent on kanamycin to maintain the plasmid over many generations, and kanamycin is active against the target organisms, a different method had to be used. *P. putida* exconjugants were grown in 5mL LB with kanamycin overnight. The resulting cells were pelleted, washed twice with PBS to remove kanamycin, resuspended in 100 µL PBS and plated onto an LB agar plate without antibiotics. This plate was incubated at 30 °C overnight and then used for the assay. This method assumes that the high copy number plasmid is only lost through cell division over many generations, so reducing generation number through plating a large amount of cells at once would allow the plasmid to be kept in most cells. This was verified using an mScarlet-expressing transformant that showed slight red coloration when the pelleted cells were plated, but very strong red coloration the next day, indicating that mScarlet was still being produced despite the lack of antibiotic. For the assay, 100 µL of overnight culture of either *E. coli* or *M. luteus* were added to 50 mL of molten LB agar, cooled down to just above solidification temperature. The thus inoculated agar was poured into square plastic petri dishes and left to solidify. The back side of a sterile 1000 µL pipette tip was used to cut agar plugs from *Streptomyces* and *P. putida* plates that were subsequently transferred onto the *E. coli* and *M. luteus* plates in upside-down orientation. The

plates were then incubated overnight at 30 °C for *M. luteus* and 37 °C for *E. coli*. Plugs from empty kanamycin and ampicillin-containing LB agar plates were used as positive controls. Empty plasmid containing transformants and exconjugants were used as negative controls. Activity was assessed by observing inhibition zones around plugs.

4.3.13 Sequencing

Streptomyces exconjugants were grown for seven days at 30 °C on SFM medium. Biomass was scraped off using a plastic spreader and transferred into tubes containing DNA/RNA shield (Zymo Research). The tubes were then sent for Illumina sequencing at microbesNG.

4.3.14 Read mapping

To check for integration and coverage of the plasmid, the trimmed reads received from microbesNG were mapped onto the plasmids using bowtie2 (Langmead and Salzberg 2012). After processing with samtools 1.9 (H. Li et al. 2009), the mappings were visualised using BamView 1.1.8 (Carver et al. 2013).

To detect integration sites, it was necessary to find those reads that contained both part of the plasmid as well as part of the chromosome (split read mapping). During plasmid integration, the attP site present on the plasmid recombines with the attB site present in the chromosome, giving rise to the attL and attR sites which then flank the integrated plasmid. Consequently, many of the reads containing the left or right part of the attP sequence (both c. 50 bp) in an exconjugant will also map to the chromosomal integration site. Therefore, exconjugant reads were mapped to the 50 bp left/right parts of the attP sequence using bowtie2 (--local flag). Using samtools 1.9, unmapped reads were removed and the SAM files converted back into a FASTQ file. This FASTQ file, now containing only the exconjugant reads mapping to the 50 bp attP left/right site, was mapped onto the *Streptomyces* reference genome (*Streptomyces*

albus J1074, RefSeq GCA_000359525.1; *Streptomyces coelicolor* A3(2), RefSeq GCA_000203835.1) using bowtie2 (--local flag). Since the reference genome contained no attP sequences, only the part of the read containing the flanking sequences was mapped. The mappings were visualised with BamView, revealing the integration sites.

To discover whether there were any reads spanning the intact integration sites present in the wild type, exconjugant reads were mapped against the previously defined integration sites in the wild-type genome (*Streptomyces albus*: Site 1: 3168219-3168653, Site 2: 3505643-3506077; *Streptomyces coelicolor*: Site 1: 4171075-4171396, Site 2: 4178509-4178904) using bowtie2 (--local flag). Only mapped reads were kept and converted to FASTQ with samtools 1.9. The resulting reads were mapped onto the wild-type genome using bowtie2 (--local flag) and visualised using BamView, revealing reads stretching over the integration site.

4.4 Results

4.4.1 Vector construction

4.4.1.1 Cloning and Expression Strategy

As explained above, there are many pitfalls on the way from cloning a gene cluster to detecting the natural product it encodes. Many of these risks can be reduced by the investment of time and resources and the thorough investigation of any failures at each step. One of the main problems is the difference in codon usage in donor and host organism which can lead to failure in the translation step. However, due to the error-prone nanopore sequence, the synthesis of codon-optimised genes was not possible. This would in turn make other optimisation steps, such as using qPCR to test for transcription of the different genes, systematically knocking out potential regulators, or feeding of precursors an endeavour likely to waste resources. It was therefore decided to clone and express a large and phylogenetically diverse set of BGCs using their native sequences, the rationale being that some of them would yield a detectable product.

Due to the limited amount of optimisation devoted to each BGC, a relatively high attrition rate was expected.

However, a set of measures were taken to maximise the likelihood of expression:

1. Selection of BGCs diverse in phylogenetic distance to expression hosts and diverse in biosynthetic pathway to mitigate the risk
2. Subset of BGCs with potentially visible phenotypes (production of carotenoids) to aid detection
3. Use of strong promoters known to work in the expression host to ensure transcription of operons
4. Introduction of BGCs in different heterologous hosts to increase chances of expression. *Streptomyces coelicolor* M1154 and *Streptomyces albus* J1074 were used as the target hosts for which promoters were chosen. However, *Pseudomonas putida* KT2440 *trfA* was also tested.

4.4.1.2 Gblock and vector design

Two different gBlocks were designed to allow for the expression of BGCs featuring a maximum of two operons in both possible orientation – convergent and divergent (Figure 4.1). Both gBlocks contained strong constitutive promoters, an MCS with several restriction sites after the promoters, and flanking sequences containing the primer binding sites for VF2 and VR to allow for PCR amplification and Sanger sequencing. gBlock g1 was designed for the insertion of a single fragment containing two convergent operons. It therefore contained the promoters p21 and sp24 in convergent orientation with an MCS between them. gBlock g2 was designed for the insertion of two fragments in divergent orientation. It contained the sp44 and p21 promoters in a back-to-back orientation. To ensure that transcription was contained to the

cloned BGC and did not extend to the rest of the plasmid, strong terminators (fd & rrnB T1, ermE & lambda t0) were added downstream of the promoters.

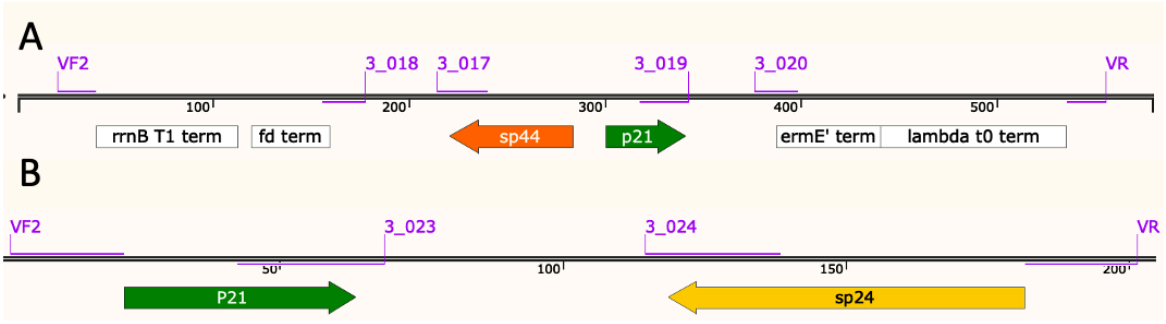


Figure 4.1: Map of the gblocks that were inserted into pBCkBAC. Promoters (colourful), terminators (white), primers for linearising the plasmids (pink) shown. Restriction sites in MCS not shown. (A) Map of g1; (B) Map of g2

The vectors to be endowed with the gBlocks were pBCaBAC and pBCkBAC. These are a set of BAC-derived vectors with apramycin or kanamycin resistance created by Chiara Borsetto (Borsetto 2017). They are able to replicate in *E. coli* (low copy number) and the specifically engineered *Pseudomonas putida* KT2440 *trfA* (high copy number), as well as to integrate into *Streptomyces* hosts. The latter is achieved by the presence of a phiC31 recombinase and a phiC31 attP site, and therefore needs a phiC31 attB site for successful integration. Both plasmids feature an MCS within a *lacZ* gene.

4.4.1.3 Gblock cloning and vector construction

The gblocks were cloned into the vectors pBCkBAC and pBCaBAC. Linear copies of the vectors without *lacZ* were amplified by PCR using primers 3_013 and 3_034 which also added overhangs for cloning. The gblocks were cloned into the vectors using SLIC and verified by Sanger sequencing. For the purpose of amplification of linear vector copies for cloning in BGC genes, several primer combinations were tested and primer pairs were chosen based on specific amplification of the linear plasmid (Figure 4.2). For all further experiments, kanamycin resistance was chosen over apramycin resistance since it worked more efficiently with *P. putida* (C. Borsetto, pers. comm.).

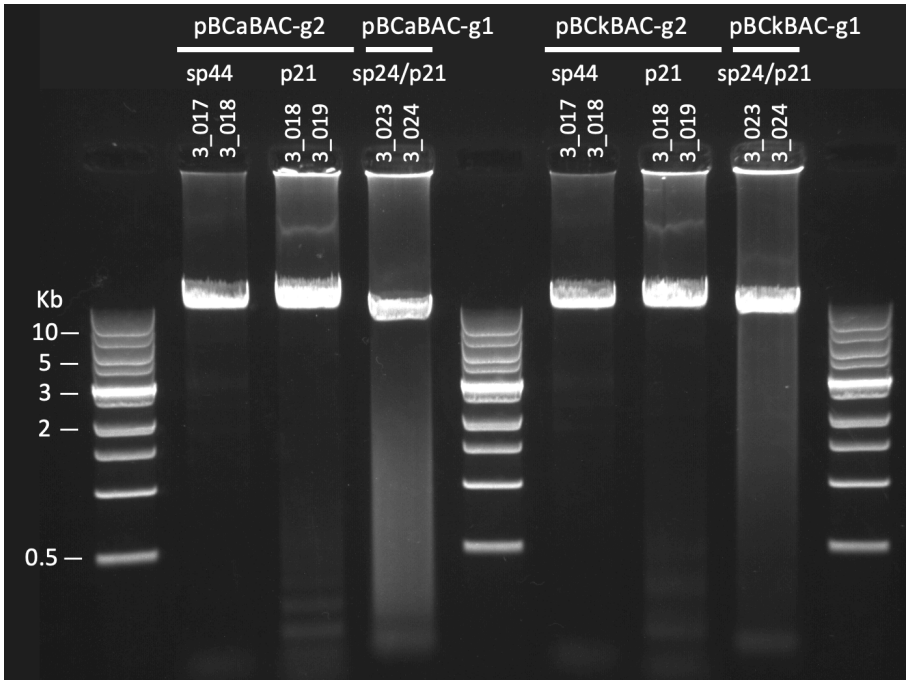


Figure 4.2: Gel photo of pBCaBAC/pBCkBAC g1 and g2 opened with primers specific for the cloning behind a given promoter (sp44, p21, or sp24/p21).

4.4.1.4 *Verifying cassette expression in E. coli, Streptomyces and Pseudomonas using mScarlet red fluorescent protein*

The functionality of the promoter cassettes was assessed by cloning the red fluorescent protein mScarlet (*Streptomyces* codon optimised) downstream of each promoter in both plasmids, leading to four plasmids (Table 4.5). When transformed into *E. coli*, all plasmids showed weak red fluorescence, in agreement with the low copy number of the plasmid. The plasmids were conjugated into *S. coelicolor* and *S. albus* as well as transformed into *P. putida*. Exposure to UV light led to red fluorescence for all plasmids in both *Streptomyces* species. Solid LB medium was found to show the highest intensity, likely due to vigorous growth, little sporulation and transparency of the medium. Fluorescence was also observed in SFM, SMMS and BGM, but not compared or measured. Consistently, sp44 showed the strongest signal (Figure 4.3A, B). In *P. putida*, red fluorescence was strongly visible in mScarlet 3 & 8 (sp24 and sp44 promoters, respectively), while it was weak in mScarlet 4 (P21 in pBCkBAC-g1) and absent in mScarlet 7 (Figure 4.3 C).

Table 4.5: mScarlet-containing plasmids and fluorescence of transformants/exconjugants under UV on LB agar, as assessed by eye. *Streptomyces* were incubated at 30°C for one week, while *P. putida* was incubated at 30°C for two days.

			Fluorescence		
Plasmid name	Vector	Promoter	*S. coelicolor*	*S. albus*	*P. putida*
mScarlet 3	pbCkBAC-g1	sp24	+	++	+++
mScarlet 4	pbCkBAC-g1	P21	++	+++	+
mScarlet 7	pBCkBAC-g2	P21	+++	++	-
mScarlet 8	pBCkBAC-g2	sp44	+++	+++	++

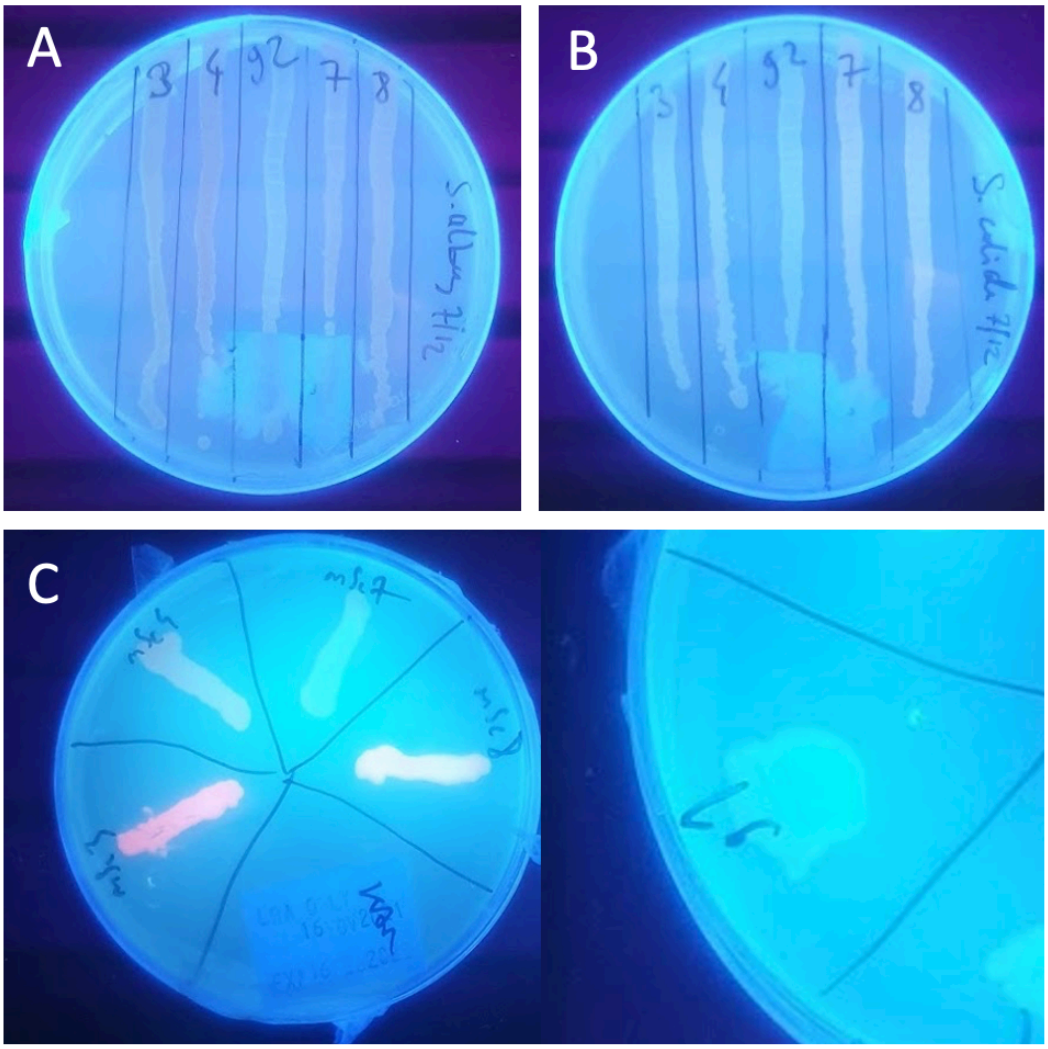


Figure 4.3: Red fluorescence of mScarlet-containing exconjugants/tranformants on LB agar compared with empty plasmid exconjugants under UV light in (A) *S. albus*, (B) *S. coelicolor*, (C) *P. putida*. *Streptomyces* were incubated at 30°C for one week, while *P. putida* was incubated at 30°C for two days.

With a base level of expression of all promoters in both *Streptomyces* species confirmed, it was decided to proceed with the cloning. Since *P. putida* was not the primary choice host, and several BGCs only contained one fragment under the promoter of sp44, the lack of expression under the P21 promoter was deemed acceptable.

4.4.2 BGC selection, cloning, transformation and conjugation into expression hosts

4.4.2.1 BGC selection

BGCs were selected according to several considerations:

1. The BGCs should reflect a diversity of phylogeny and biosynthetic pathway
2. The biosynthetic genes and possible transporters should be in maximum two putative operons
3. There should not be a predicted rho-independent terminator within a putative operon
4. The amplicons should not be larger than a reasonable PCR size, i.e. up to 20kb.

41 BGCs were selected for amplification, leading to 58 fragments to be amplified. Five large fragments were also selected for an alternative approach involving the amplification of two smaller, overlapping fragments to be cloned into a vector in a three-way reaction. All in all, primers for 68 PCR reactions were designed with 20bp overhangs for cloning them into the gblocks g1 and g2.

The selected BGCs, their characteristics and taxonomic annotation can be found in Table 4.6. A more detailed description of successfully cloned BGCs can be found below.

Table 4.6: BGCs selected for cloning.

contig name	Product	Vector	Fragment 1 length (bp)	Fragment 2 length (bp)	BGC length (bp)	Phylum	Class	Order	Family	Genus	BiG-SLiCE d	Plasmid
contig_15892	lassopeptide	g2	7263		22672	Acidobacteriota	Acidobacteriae	20CM-2-55-15	not classified	not classified	1429	11
contig_4314	lassopeptide	g2	7909		59125	Acidobacteriota	Acidobacteriae	20CM-2-55-15	not classified	not classified	1406	6
contig_736	NRPS	g2	23158		50068	Acidobacteriota	Blastocatellia	Pyrinomonadales	Pyrinomonadaceae	not classified	1541	
contig_11857	NRPS	g2	25812		63382	Acidobacteriota	Thermoanaerobaculia	UBA5066	Gp7-AA6	*Gp7-AA6*	1048	
contig_23551	T3PKS	g1	5372		40504	Acidobacteriota	Vicinamibacteria	Fen-336	Fen-336	not classified	1563	
contig_4	terpene	g2	5754		21750	Acidobacteriota	Vicinamibacteria	Vicinamibacterales	2-12-FULL-66-21	not classified	1239	57
contig_6313	lassopeptide	g1	12791		23416	Acidobacteriota	Vicinamibacteria	Vicinamibacterales	2-12-FULL-66-21	not classified	838	
contig_10632	NRPS PKS	g1	18477		41245	Acidobacteriota	Vicinamibacteria	Vicinamibacterales	not classified	not classified	1538	
contig_2313	terpene	g2	10425		21857	Acidobacteriota	Vicinamibacteria	Vicinamibacterales	not classified	not classified	557	
contig_25828	lassopeptide	g1	13505		23193	Acidobacteriota	Vicinamibacteria	Vicinamibacterales	not classified	not classified	1437	
contig_36584	T1PKS	g2	20181	6877	49762	Acidobacteriota	Vicinamibacteria	Vicinamibacterales	not classified	not classified	1516	
contig_6994	NRPS PKS	g1	18248		55908	Acidobacteriota	Vicinamibacteria	Vicinamibacterales	not classified	not classified	1977	
contig_6994	T1PKS	g2	8390	5198	40856	Acidobacteriota	Vicinamibacteria	Vicinamibacterales	not classified	not classified	1397	
contig_795	terpene	g2	8272		20630	Acidobacteriota	Vicinamibacteria	Vicinamibacterales	not classified	not classified	934	
contig_13679	lassopeptide	g1	15013		23403	Acidobacteriota	Vicinamibacteria	Vicinamibacterales	UBA2999	not classified	1512	
contig_228	lassopeptide	g1	11886		22247	Actinobacteriota	Acidimicrobiia	UBA5794	not classified	not classified	1008	
contig_10649	T3PKS	g1	9273		41101	Actinobacteriota	Actinobacteria	not classified	not classified	not classified	1456	33
contig_1186	lanthipeptide	g1	6994		24265	Actinobacteriota	Actinobacteria	not classified	not classified	not classified	936	
contig_13147	NRPS PKS	g2	18946		49401	Actinobacteriota	Actinobacteria	not classified	not classified	not classified	1559	
contig_9172	NRPS	g1	15354		52197	Actinobacteriota	Thermoleophilia	not classified	not classified	not classified	1030	
contig_13589	terpene	g1	7651		20687	Actinobacteriota	Thermoleophilia	Solirubrobacterales	Solirubrobacteraceae	not classified	1251	3
contig_3134	NRPS	g2	6255	14863	46004	Actinobacteriota	Thermoleophilia	Solirubrobacterales	Solirubrobacteraceae	not classified	2206	
contig_11044	NRPS	g2	13558	9007	44436	Gemmatimonadota	Gemmatimonadetes	Gemmatimonadales	Gemmatimonadacea	not classified	1454	
contig_4743	NRPS	g2	7498	8490	50340	Gemmatimonadota	Gemmatimonadetes	Gemmatimonadales	Gemmatimonadacea	not classified	2124	
contig_7544	T1PKS	g2	17320		50885	Gemmatimonadota	Gemmatimonadetes	not classified	not classified	not classified	1496	
scaffold_35893	lassopeptide	g1	10189		22501	Gemmatimonadota	Gemmatimonadetes	not classified	not classified	not classified	425	
contig_5955	T3PKS	g2	13125	2166	33729	Planctomycetota	UBA1135	UBA1135	GCA-002686595	not classified	1304	
contig_115	lassopeptide	g2	9229		22314	Proteobacteria	Alphaproteobacteria	Caulobacterales	Caulobacteraceae	*Brevundimonas*	601	
contig_13212	terpene	g2	11062		20627	Proteobacteria	Gammaproteobacteria	not classified	not classified	not classified	689	62
contig_13212	terpene	g2	4355		20402	Proteobacteria	Gammaproteobacteria	not classified	not classified	not classified	1042	22
contig_14956	bacteriocin	g2	2042	3267	10936	Proteobacteria	Gammaproteobacteria	UBA7966	UBA7966	*USCγ-Taylor*	504	32
contig_24847	NRPS	g2	12403		50278	Proteobacteria	Gammaproteobacteria	UBA7966	UBA7966	*USCγ-Taylor*	687	
contig_2807	NRPS	g2	16666	17420	73790	Proteobacteria	Gammaproteobacteria	UBA7966	UBA7966	*USCγ-Taylor*	1339	
contig_291	bacteriocin	g2	2163	3447	10936	Proteobacteria	Gammaproteobacteria	UBA7966	UBA7966	*USCγ-Taylor*	899	30
contig_414	bacteriocin	g2	2183	3597	10897	Proteobacteria	Gammaproteobacteria	UBA7966	UBA7966	*USCγ-Taylor*	892	28
contig_665	bacteriocin	g2	2202	1178	10903	Proteobacteria	Gammaproteobacteria	UBA7966	UBA7966	*USCγ-Taylor*	899	26
scaffold_45328	T1PKS	g2	9579	4613	43393	Verrucomicrobiota	Verrucomicrobiae	Chthoniobacterales	not classified	not classified	1375	8
scaffold_13961	T1PKS	g2	13520	2053	47647	Verrucomicrobiota	Verrucomicrobiae	Chthoniobacterales	UBA10450	*Palsa-1392*	1067	
contig_10669	T3PKS	g2	5231		40984	Verrucomicrobiota	Verrucomicrobiae	Opitutales	Opitutaceae	*Opitutus*	1171	34
contig_2148	NRPS	g2	21000		63452	Verrucomicrobiota	Verrucomicrobiae	Opitutales	Opitutaceae	*Opitutus*	1436	
scaffold_11847	NRPS	g2	23750		55766	Verrucomicrobiota	Verrucomicrobiae	Pedosphaerales	AV2	not classified	1208	

4.4.2.2 Amplification and cloning success

Through the pipeline from amplification to successfully cloned plasmid, the number of BGCs was significantly reduced. Out of 68 PCR reactions that were attempted, 48 (71%) successfully amplified as determined by a band of the right size appearing on the gel. While a number of large fragments could be successfully amplified, Figure 4.4A shows that amplification success decreases with fragment size. After gel excision and DNA purification, a significant trend of decreasing DNA concentration with size could be observed (Pearson's R = -0.37, p = 0.007). The combined effect of decreased DNA concentration and increased molecular weight led to an even stronger negative association between molarity and fragment size (Figure 4.4B). Therefore, only 32 of the 48 amplified fragments (67%) showed a molarity of 4 fmol/µL or more (as measured by qubit), which emerged as the minimum concentration necessary for cloning. This is because the 10 µL SLIC reaction needs to accommodate both vector and fragment in sufficiently high concentrations (combined amount > 30 fmol) as well as reagents. For five of the sufficiently concentrated fragments, no cloning was attempted because the second part of the BGC had not amplified sufficiently. Four fragments were attempted as a three-way cloning reaction involving two fragments and the vector, which did not work. In the end, out of 23 single fragment cloning reactions attempted with the minimum amount of DNA needed, 18 (78%) were successful and led to the construction of 13 plasmids (with intermediate plasmid opening steps for two-fragment BGCs). The total attrition rate from primer design to cloned BGC was 75%, with a strong bias favouring small BGCs at every step (Figure 4.5, Figure 4.6). For example, the median fragment size at the primer design stage was 9kb, while the median cloned fragment size was only 4.3kb. The largest fragment amplified was 20kb long, but the largest cloned fragment measured only 11kb. These results show that fragment size and its impact on amplification and purification success is a major limitation in amplification and cloning of BGCs from metagenomic soil DNA. The plasmids were numbered after the last fragment that was cloned into them – e.g. 33 for the plasmid containing fragment

33, and 28 for the plasmid containing fragment 27 (first reaction) and fragment 28 (second reaction). In addition, the number of the colony picked for further processing was added after a hyphen, giving rise to e.g. 33-1 and 28-2.

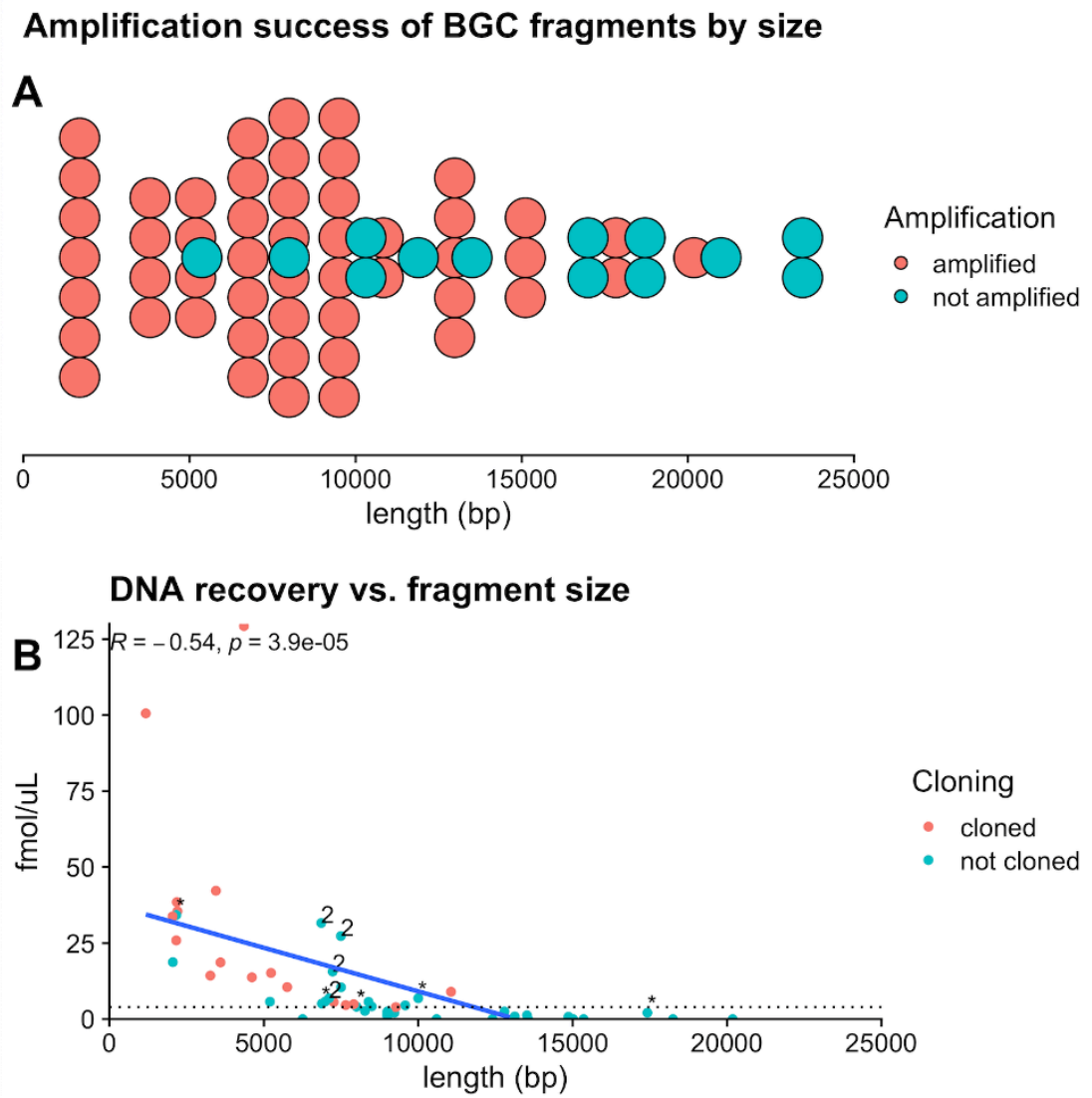


Figure 4.4: Amplification success of targeted BGCs. (A) Amplification success by fragment size. (B) DNA recovery plotted by fragment size, with colour indicating cloning success. Asterisks indicate fragments that were not cloned because the second part of the BGC was not amplified. "2" indicates fragments that were cloned as a three-part assembly, which did not work. The dotted line indicates a fragment concentration of 4 fmol/µL

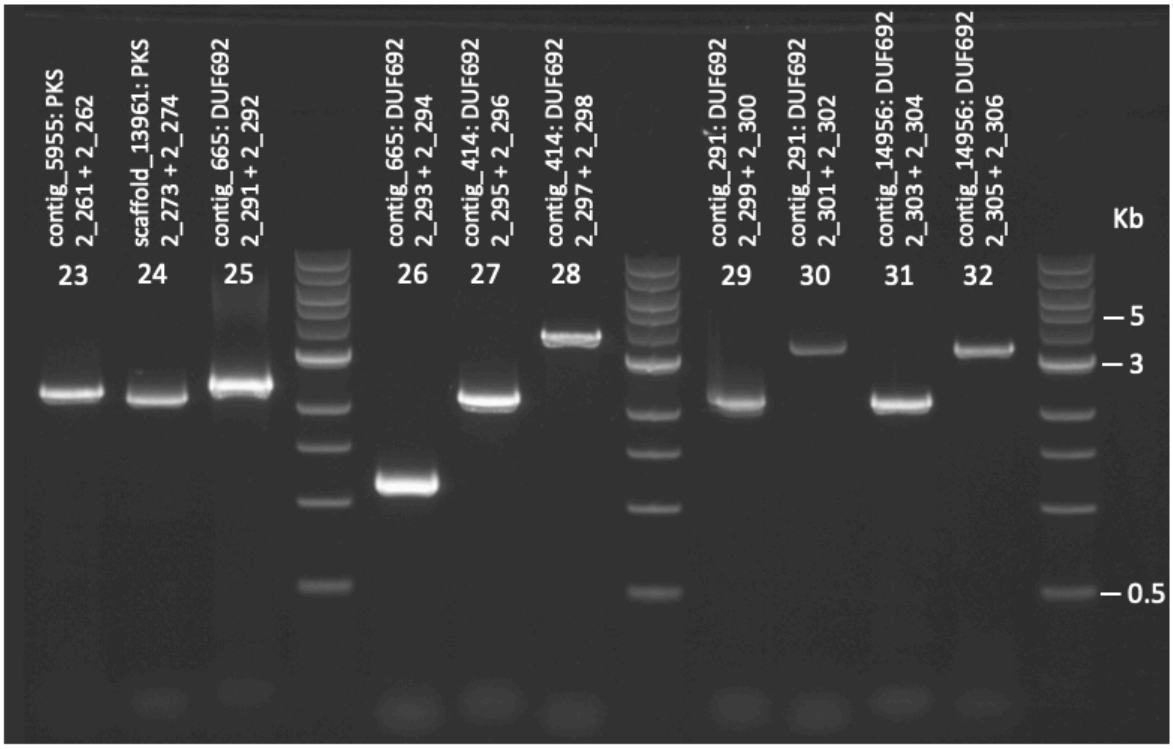


Figure 4.5: Agarose gel of selected small BGC fragments with reaction number, target contig and primer pair annotated. The successful specific amplification of fragments 23-32 with sizes from 1178 to 3597 bp shows the high success rate of small fragment amplification.

4.4.2.3 Cloning

Sanger sequencing with primers flanking the insertion site revealed some differences between the nanopore assembly and the cloned fragments. Alignments of partly sequenced colony PCR products to the *in silico* plasmid sequence mostly showed single nucleotide mismatches and differences in homopolymer length, likely due to the inaccuracy of the nanopore assembly. However, cloning reaction number 30 (the putative sodium-calcium exchanger from the DUF692-containing BGC on contig_291) yielded two slightly different fragments recovered from three positive colonies (30-1, 30-2, 30-3). Intriguingly, none of the sequenced colonies matched the nanopore assembly (Figure 4.7). 30-1 and 30-2 were identical and differed from 30-3 by set of deletions and insertions as well as several single nucleotide differences. While 30-3 was closer to the sequence assembled from the nanopore data, it also contained additional

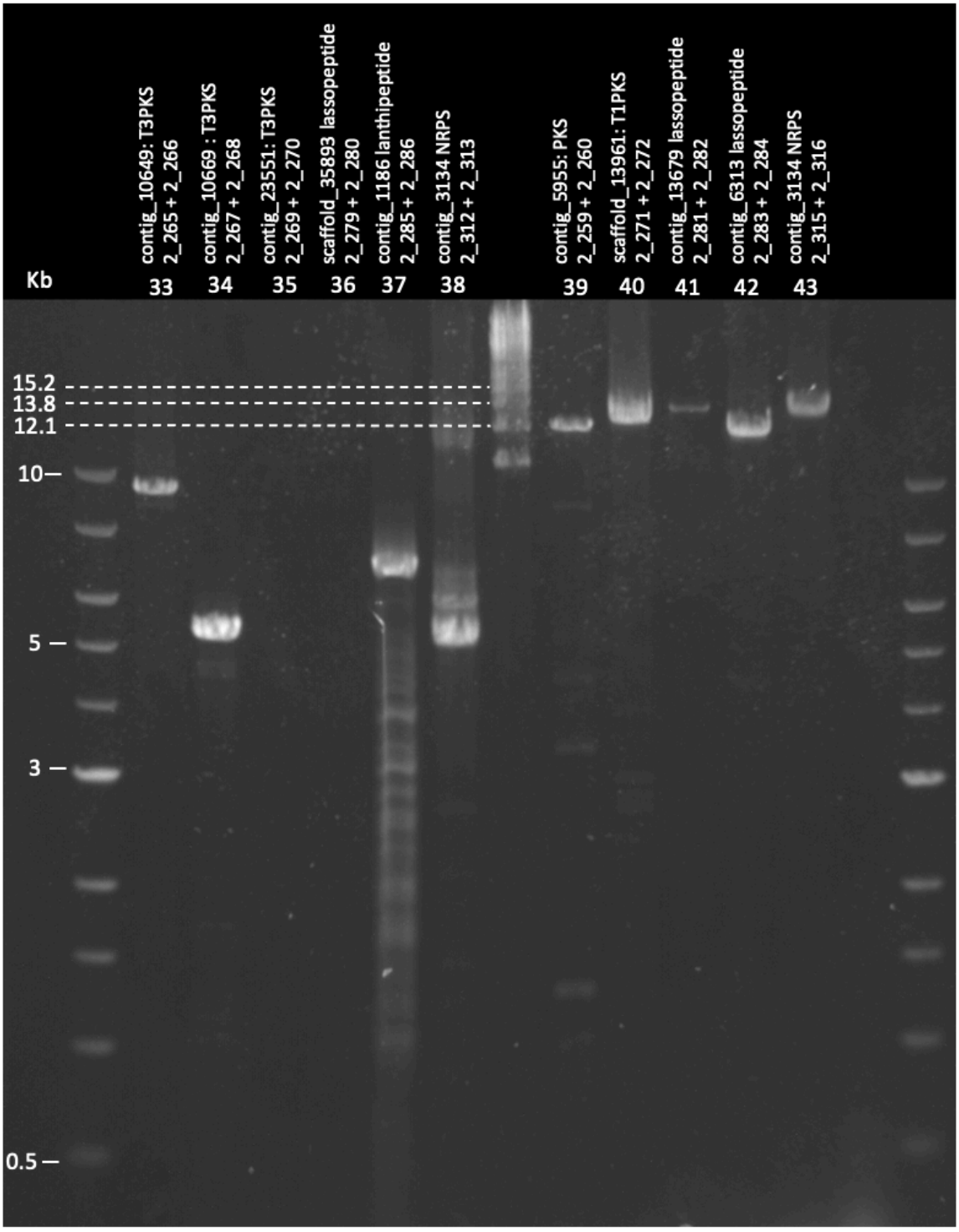


Figure 4.6: Agarose gel of selected large BGC fragments with reaction number, target contig and primer pair annotated. The partially successful, but often weak and unspecific amplification of fragments 33-43 with sizes from 5231 bp to 15013 bp shows the reduced reliability of long-range amplification.

insertions not found in the assembly. This indicates that two different fragments were amplified, and none of them was the fragment from the nanopore assembly. Whether the assembly fragment for which the primers were designed was just not amplified as well as the other two fragments, or whether the BGC was a chimera generated by the long-read assembly through merging of two closely related strains remains unclear.

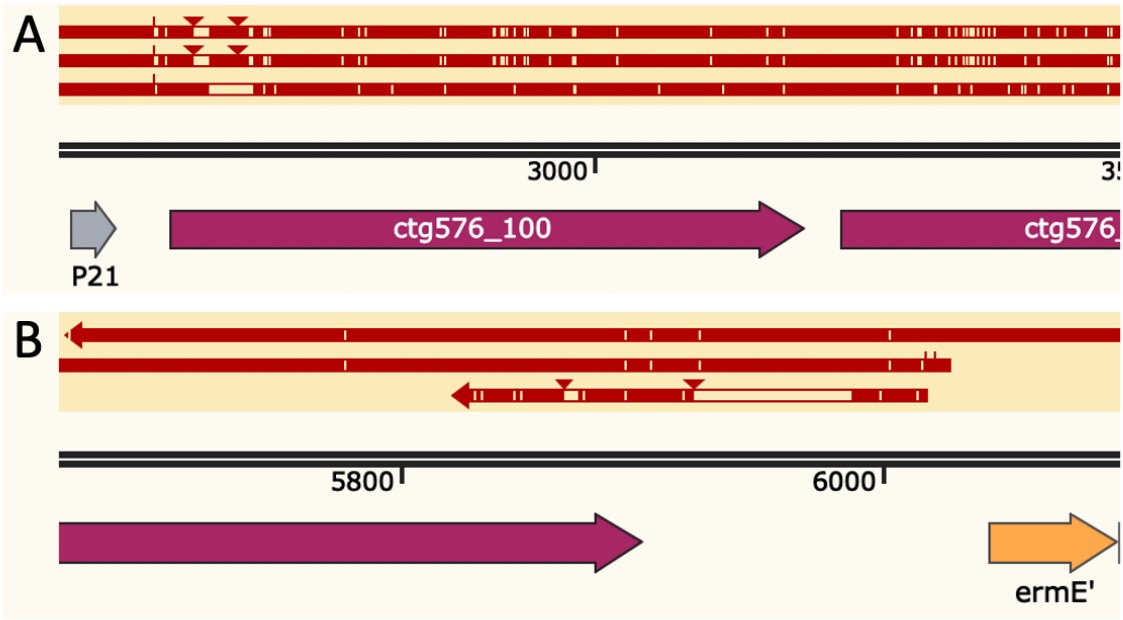


Figure 4.7: Alignment of three Sanger sequences (from top to bottom: 30-1, 30-2, 30-3) to the in silico constructed plasmid containing the DUF692 BGC from contig_291 (order UBA7966). Coloured line represents matching stretches, empty line represents deletions in the Sanger sequence, arrowheads represent insertions in the Sanger sequence relative to the in silico plasmid. (A) Beginning of insert, (B) end of insert.

It should be noted that Sanger sequencing could not be used to check for single nucleotide errors introduced through the PCR or cloning process, owing to the inaccuracies introduced by the nanopore sequencing.

4.4.2.4 Description of the successfully cloned BGCs

Thirteen BGCs were successfully cloned.

4.4.2.4.1 Lassopeptide BGCs

Two lassopeptide BGCs (Figure 4.8, Table 4.7) were successfully cloned into pBCkBAC-g2 under the sp44 promoter, giving rise to plasmids 11-4 (contig_15892) and 6-1 (contig_4314). Both BGCs were predicted to come from the uncultured order 20CM-2-55-15 within the Acidobacteriae class. They both showed a relatively high BiG-SLiCE distance (d), indicating novelty compared to BGCs in databases. No lassopeptides were predicted by antiSMASH in 11-4. However, a small peptide (3rd ORF in Figure 4.8 on plasmid 11-4) showed similarity to two MAG-derived ORFs (MCA9670917.1 from a Myxococcales MAG and TNE48647.1 from a Deltaproteobacteria MAG). These ORFs were adjacent to other lassopeptide biosynthesis genes, indicating that this could indeed be a precursor peptide. However, since it was not picked up by lassopeptide precursor prediction algorithms, it is likely to significantly diverge from known lassopeptides. In 6-1, no lassopeptides were predicted in the sequence polished with exconjugant short reads. In the unpolished sequence, a sequencing artefact had been picked up as a potential precursor peptide by antiSMASH but disappeared after polishing.

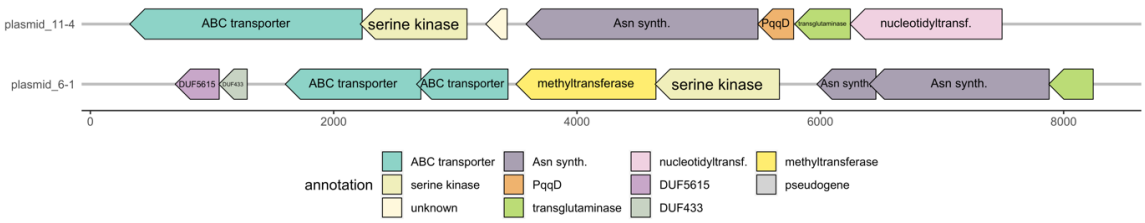


Figure 4.8: Map of lassopeptide BGCs cloned into plasmids.

Table 4.7: Taxonomic classification of the cloned lassopeptide BGCs.

No	P	C	O	F	G	BiG-SLiCE d
11	Acidobacteriota	Acidobacteriae	20CM-2-55-15	NA	NA	1429
6	Acidobacteriota	Acidobacteriae	20CM-2-55-15	NA	NA	1406

4.4.2.4.2 PKS BGCs

Three PKS BGCs were successfully cloned (Figure 4.9A, Table 4.8).

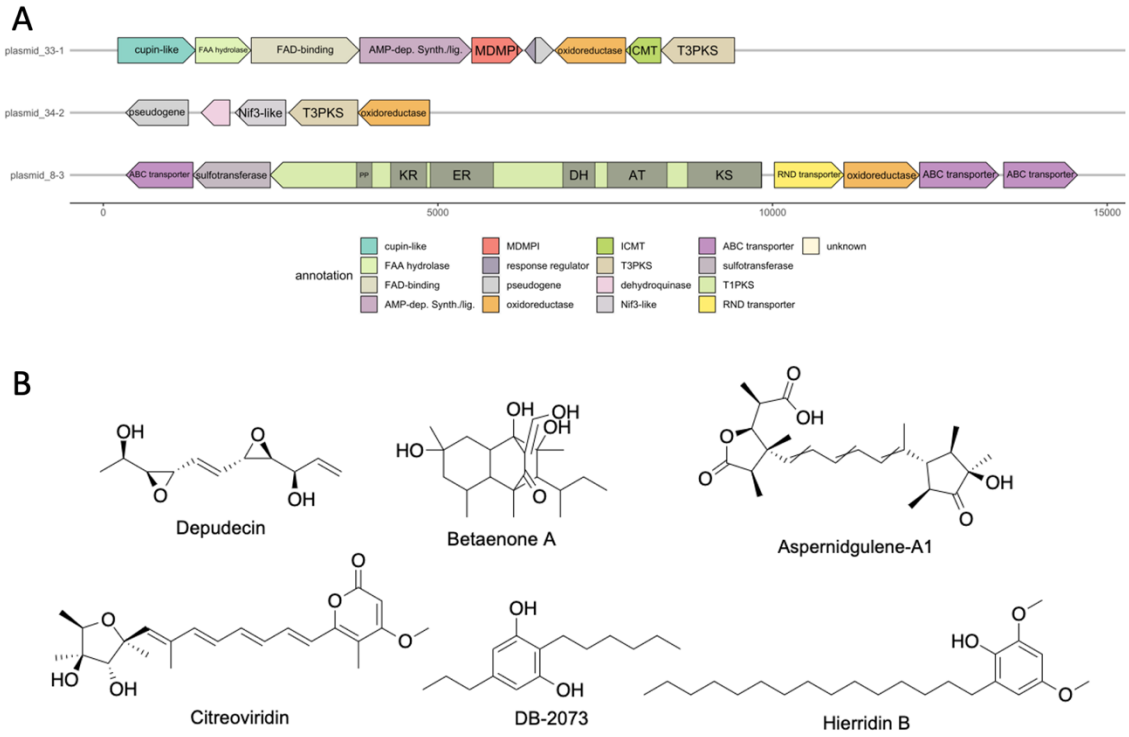


Figure 4.9: Cloned PKS BGCs and products of related BGCs. (A) Map of BGCs as cloned into plasmids. (B) structures of compounds from related BGCs: depudecin, betaenone A, aspernidgulene A1, citreoviridin, DB-2073 (an alkylresorcinol antibiotic from Pseudomonas) and hierridin B

Table 4.8: Taxonomic classification of the cloned PKS BGCs

No	P	C	O	F	G	BiG-SLiCE d
8	Verrucomicrobiota	Verrucomicrobiae	Chthoniobacterales	NA	NA	1375
33	Actinobacteriota	Actinobacteria	NA	NA	NA	1456
34	Verrucomicrobiota	Verrucomicrobiae	Opitutales	Opitutaceae	Opitutus	1171

A type 1 PKS assigned to the Verrucomicrobial order Chthoniobacterales (scaffold_45328) was cloned into pBCkBAC-g2, with the putative core biosynthesis genes under promoter sp44 and putative transporter genes under P21, giving rise to the plasmid 8-1. AT specificity was predicted to be malonyl-CoA. The PKS gene showed 20%-25% sequence similarity to the PKS

genes of the MiBiG BGCs for depudecin from *Alternaria brassicicola*, betaenones A, B & C from *Phoma betae* as well as aspernidgulenes A1, A2, B1 and citreoviridin from *Aspergillus* species (Figure 4.9B). Based on these similarities, it was concluded that BGC 8 most likely encoded an iterative PKS.

The type 3 PKS from contig_10649 (lowest taxonomic classification: class Actinobacteria) was cloned into pBCkBAC-g2, with the core biosynthetic genes under sp44 and other putative biosynthetic genes under P21, resulting in plasmid 33-1. The core biosynthetic genes (oxidoreductase, ICMT, T3PKS) showed 50%-64% sequence similarity to the MiBiG entry for alkylresorcinol from *Streptomyces* griseus, indicating that a type of phenolic lipid might be produced by this BGC (Figure 4.9B). Another type 3 PKS from contig_10669, assigned verrucomicrobial genus *Opitutus* was cloned into pBCkBAC-g2 under promoter sp44, giving rise to plasmid 34-2. The T3PKS gene showed similarity to the MiBiG entries hierridins B & C from Cyanobium sp. LEGE 06113 (Figure 4.9B). Similar to plasmid 33-1, a phenolic lipid might be produced by this BGC.

4.4.2.4.3 Terpene BGCs

Four terpene BGCs were successfully cloned (Figure 4.10A, Table 4.9)

Table 4.9: Taxonomic classification of the cloned terpene BGCs. Genera were not assigned.

No	P	C	O	F	BiG-SLiCE d
3	Actinobacteriota	Thermoleophilia	Solirubrobacterales	Solirubrobacteraceae	1251
22	Proteobacteria	Gammaproteobacteria	NA	NA	689
57	Acidobacteriota	Vicinamibacteria	Vicinamibacterales	2-12-FULL-66-21	1239
62	Proteobacteria	Gammaproteobacteria	NA	NA	689

A terpene synthase was cloned from the gammaproteobacterial contig_13212 into pBCkBAC-g2, resulting in plasmid 22-1. A gene assigned a regulatory function (diguanylate cyclase) was included as well, given the chance that it might be involved in biosynthesis. BLAST showed

sequence similarity to germacrene A synthases as well as 5-epi-alpha-selinene synthases (Figure 4.10B). Based on this homology, and since no modification enzymes were present, the product was expected to be an unoxidized, and thereby volatile, sesquiterpene (15 carbons).

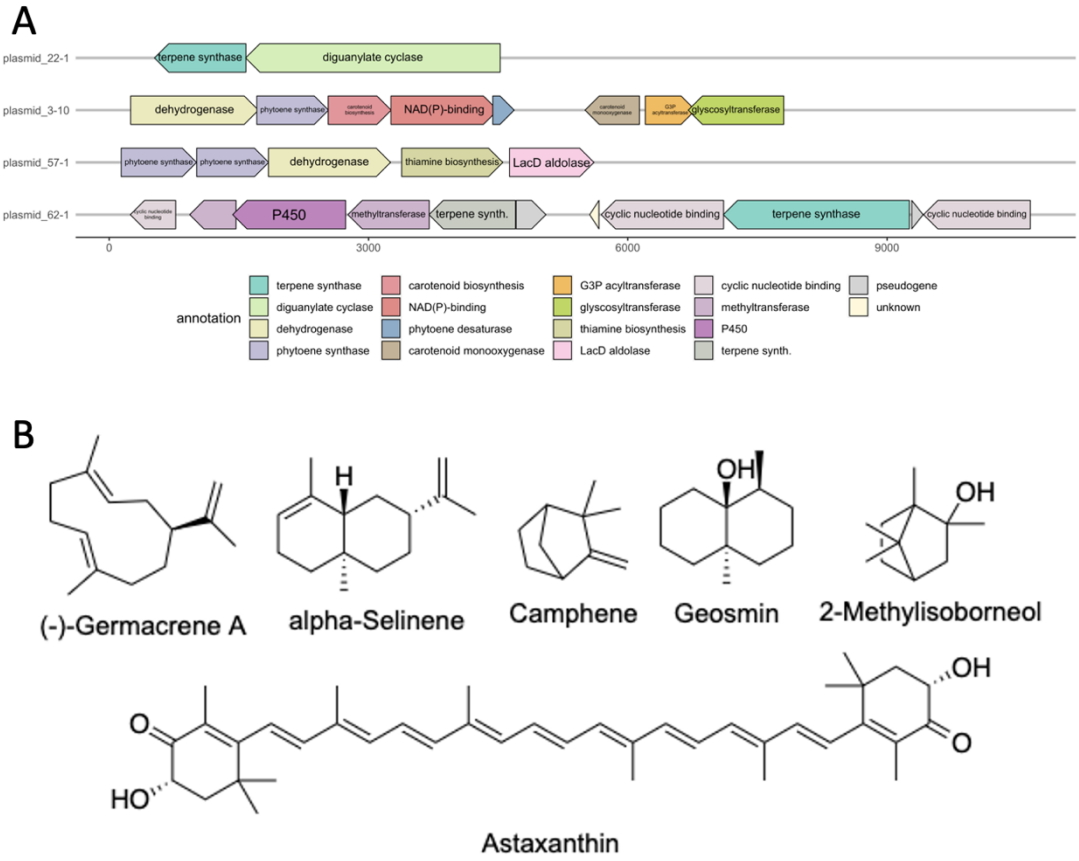


Figure 4.10: Cloned terpene BGCs and products of related BGCs. (A) Map of BGCs as cloned into plasmids. (B) Structures of Germacrene A, alpha-selinene, 2-methylisoborneol, camphene, geosmin and astaxanthin (a carotenoid occurring in bacteria)

A putative carotenoid BGC from the actinobacterial Solirubrobacteraceae family was cloned into pBCkBAC-g1, leading to plasmid 3-10 containing a phytoene synthase and several enzymes involved in carotenoid biosynthesis. However, in the unpolished contig, only two operons were visible. After polishing with exconjugant short reads, a potential third operon appeared, containing a glycerol-3-phosphate-acyltransferase. This might hinder the expression of the carotenoid monooxygenase. However, a coloured carotenoid pigment was expected to be produced by the remaining genes.

Another carotenoid BGC from the uncultured acidobacterial family 2-12-FULL-66-21 was cloned into pBCkBAC-g2 (plasmid 57-1) containing two phytoene synthases as well as an oxidoreductase. Two subsequent genes were included as well, at the chance that they contribute to biosynthesis. The product expected was an oxidised phytoene pigment.

A further terpene BGC from contig_13212, assigned to the class Gammaproteobacteria was cloned into pBCkBAC-g2 under promoter sp44. The BGC contained two terpene synthases, potentially indicating the production of two different compounds. However, since no rho-independent terminator could be detected between the two synthases, it was treated as one BGC. The first terpene synthase showed sequence similiarity to camphene synthases was followed by two methyltransferases and a P450 (Figure 4.10B). antiSMASH also detected similarity to the 2-methylisoborneol BGC, which however features only one methyltransferase and no P450.

The second terpene synthase showed homology to germacradienol/geosmin synthases and did not show any oxidases (Figure 4.10B). Three cyclic-nucleotide binding proteins, likely regulatory, were present in the BGC as well. The products expected from this BGC were a potentially bicyclic and oxidised monoterpene and a sesquiterpene without further oxidation reactions.

Both *Streptomyces coelicolor* and *Streptomyces albus* contain a carotenoid biosynthesis pathway, indicating the potential availability of precursors. However, both pathways are inactive under normal laboratory conditions. In *Streptomyces coelicolor*, carotenoid production can be activated by culturing it on BGM and exposing it to light. For *S. albus*, no such strategy is known, and no carotenoid production was observed even under illumination.

Both *Streptomyces* species as well as *P. putida* contain a farnesyl pyrophosphate synthase and both *Streptomyces* strains are known to produce the sesquiterpene geosmin, indicating the availability of precursors.

4.4.2.4.4 DUF692 BGCs

Four DUF692 BGCs (Figure 4.11, Table 4.10) were cloned into pBCkBAC-g2 with the DUF692-containing operon under sp44 and the other conserved, putative transport-related genes under P21, thereby giving rise to plasmids 26-1, 28-2, 30-1/30-3 and 32-4. All BGCs were assigned to the genus *USCγ*. Based on the analysis of conserved sequences (see Chapter 1), the product was expected to be a ca. 40-50 amino acid peptide containing six cysteines that would either be free for metal coordination or forming intramolecular bridges.

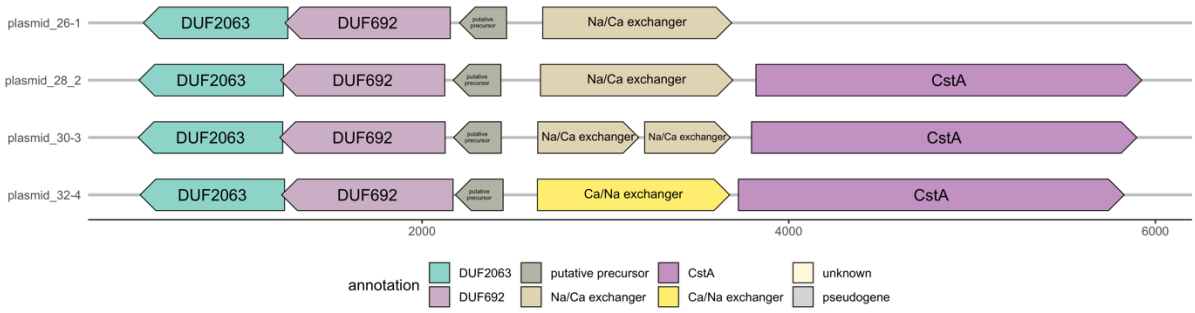


Figure 4.11: Map of the DUF692 BGCs cloned into the plasmids

Table 4.10: Taxonomic classification of cloned DUF692 BGCs

No	P	C	O	F	G	BiG-SLiCE d
26	Proteobacteria	Gammaproteobacteria	UBA7966	UBA7966	*USCγ-Taylor*	899
28	Proteobacteria	Gammaproteobacteria	UBA7966	UBA7966	*USCγ-Taylor*	892
30	Proteobacteria	Gammaproteobacteria	UBA7966	UBA7966	*USCγ-Taylor*	899
32	Proteobacteria	Gammaproteobacteria	UBA7966	UBA7966	*USCγ-Taylor*	504

4.4.2.5 Transformation into P. putida

All plasmids were transformed through electroporation into *P. putida* successfully. A phenotype was observed with the DUF692-derived BGCs 28-2, 26-1, 30-1 and 30-3, with PCR-positive colonies being a much smaller size and growing more slowly when streaked (Figure 4.12). Furthermore, colony growth seemed to negatively correlate with the band size observed after colony PCR, indicating that the more plasmid was present, the lower the growth rate. However, this phenotype disappeared after cryostock preparation and re-streaking, indicating that the BGC-containing plasmid was a burden to the host fitness. DUF692 BGC 32-4 did not show a small colony phenotype, though it is unclear why, as the encoded genes are highly similar on a protein level.

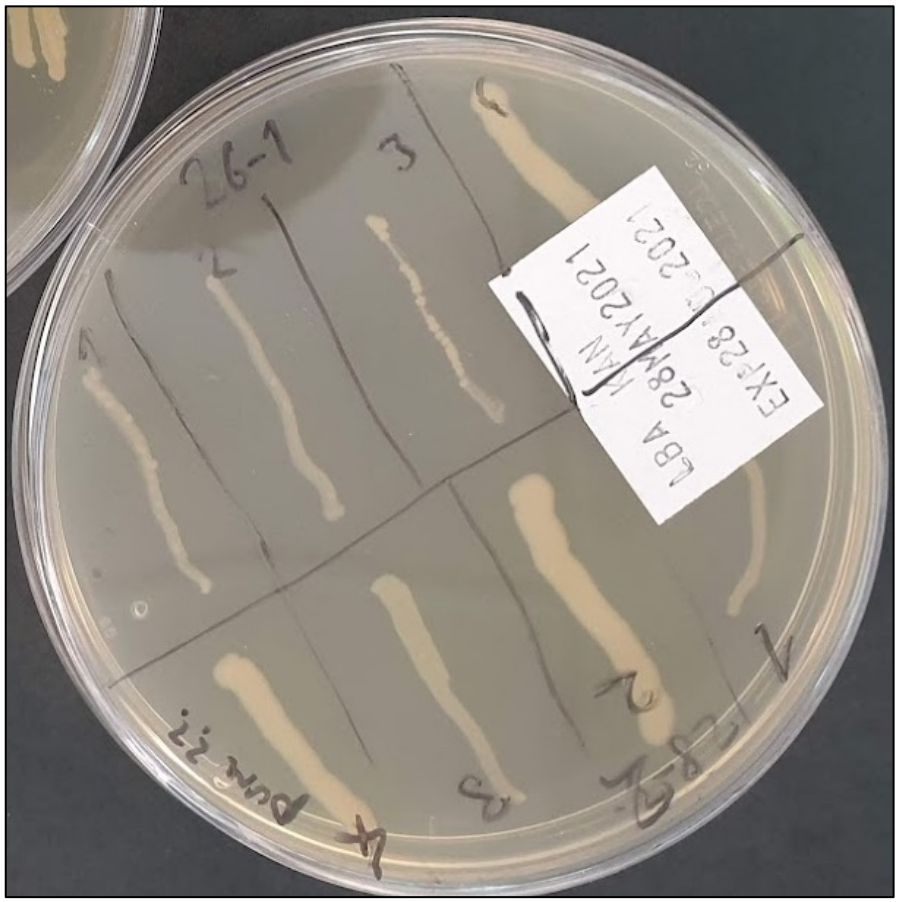


Figure 4.12: Photo of colonies picked from *P. putida* transformation plates and streaked. The thinner streaks originated from small colonies and were PCR-positive (28-2: 1, 3; 26-1: 1, 2, 3), the thicker originated from regular sized colonies and were PCR-negative (28-2: 2, 4; 26-1: 4).

4.4.2.6 Conjugation into Streptomyces coelicolor M1154 & Streptomyces albus and LC-MS

The 12 BGC-containing plasmids were transformed into non-methylating *E. coli* ET12567 containing the helper plasmid pR9604. Conjugation was performed and exconjugants were selected with kanamycin and passaged. Several exconjugants could not be achieved in one or both of the strains. Successful exconjugants were sequenced using Illumina 150bp paired-end sequencing. Exconjugants of empty plasmids were obtained for both plasmids in *S. albus* and for g2 in *S. coelicolor*, but none of them were sequenced. Furthermore, it was observed that kanamycin, even at the high dosage used, did not suppress the growth of negative colonies as well as hygromycin which is used in integrative plasmid pOSV556 (data not shown).

To evaluate the integration success of the plasmids, the trimmed reads from each exconjugant were mapped onto the in-silico plasmid sequence and visualised using BamView (Figure 4.13, Table 4.11). In *Streptomyces albus*, seven out of nine exconjugants showed successful integration of the complete insert sequence (Figure 4.13A). Two exconjugants were not successful, with 28-2 showing no plasmid integration at all and 30-1 showing a partial integration with most of the insert missing. In *S. coelicolor*, only four out of eight exconjugants showed successful insertion as reflected by reads mapping onto the full plasmid sequence. The others (6-1, 11-4, 32-4, 62-1) were only partial insertions (Figure 4.13B, C). The exconjugants 3-10 and 33-1 furthermore showed uneven coverage across the plasmid sequence, potentially indicating that multiple, but not equally successful integration events had taken place (Figure 4.13D).

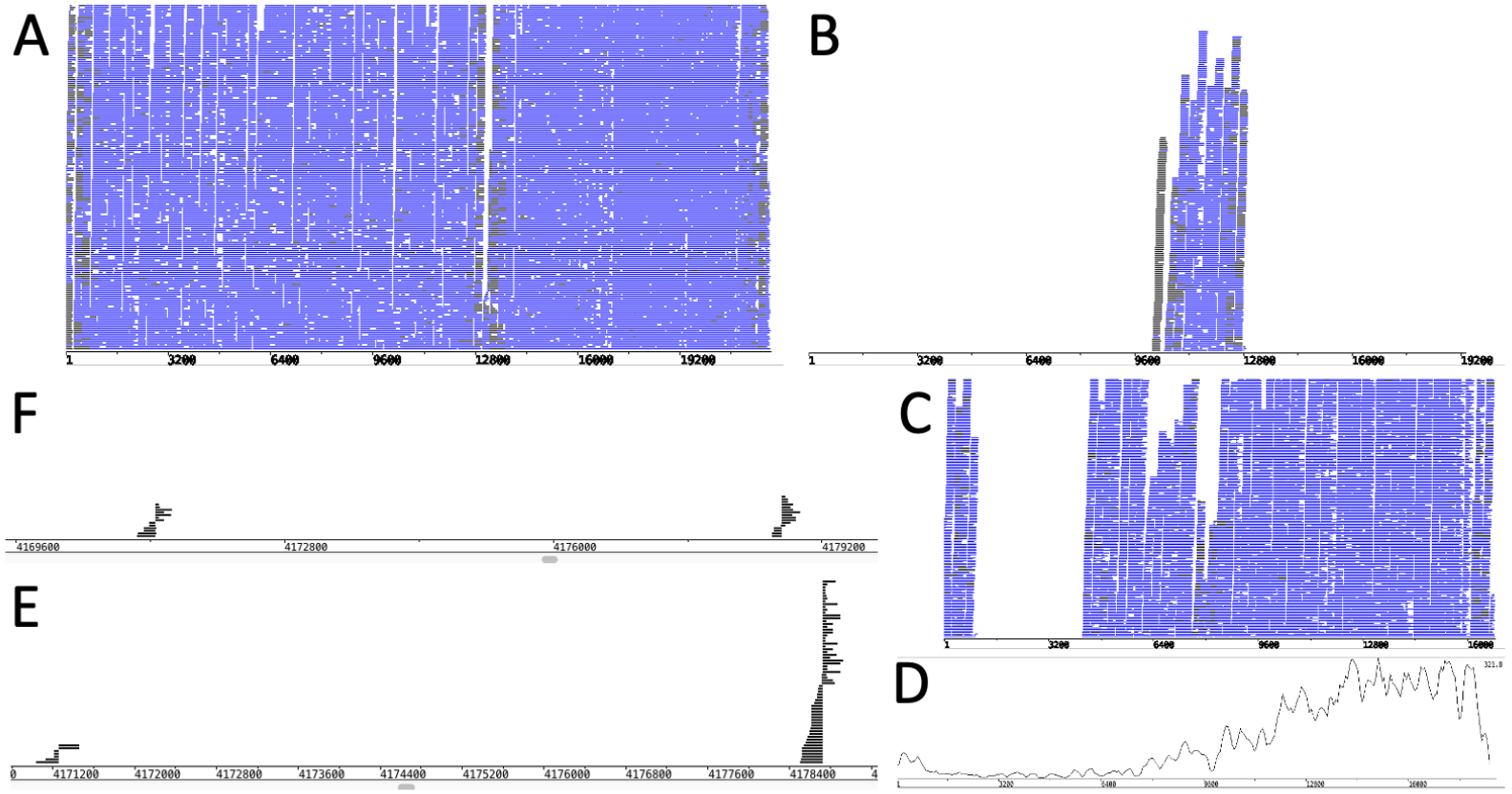


Figure 4.13: BamView line/stack visualisations of exconjugants reads mapped onto in-silico plasmids and reference genomes. The x axis are sequence coordinates while the Y axis shows coverage via read stacking or line chart. (A) Mapping of S. albus 62-1 exconjugant reads onto the in silico plasmid shows full and even coverage. The visible gap is the middle of attP, which splits upon integration, (B) Mapping *S. coelicolor* 6-1 excojugant reads onto the in silico plasmid shows deletion of most of the plasmid apart from KanR and attP (C), Mapping of *S. coelicolor* 32-4 exconjugant reads onto the in silico plasmid shows deletion of the promoter cassette and surrounding genes, (D) Mapping *S. coelicolor* 3-10 exconjugant reads onto the in silico plasmid shows full, but highly uneven coverage of the plasmid (grey line = maximal coverage = 321.8x), (E) Mapping the attP-matching reads from the *S. coelicolor* 32-4 exconjugant reveals two insertion sites with highly uneven coverage, (F) Mapping the attP-

Table 4.11: Attachment sites derived from mapping in *S. coelicolor* and *S. albus*, aligned to show conserved motifs of GGnG, Tn (integration site) and CnCC.

S. coelicolor	site1	AAACGCGGAGGCCCGGGAGAAGCT**TC**TGCCTCTCCCGGGCCTCCGCTGTA
	site2	CGGTGCGGGTGCCAGGGCGTGCCC**TT**GGGCTCCCCGGGCGCGTACTCCAC
S. albus	site1	CGCGACGGGTGCCAGGGAGAGCCG**TA**CGTCTCGCCCTGGCACCCCGCCGC
	site2	CGGTGCGGGTGCCAGGGCGTCCCC**TT**CGGCTCCCCGGCCGCGTACTCCAC

Read mapping was conducted to determine integration sites. When the plasmid integrated into the genome, the attP site recombined with the chromosomal attB site, generating attL and attR regions flanking the integrated plasmid. The attL and attR site both contain part of the chromosomal attB and the plasmid-derived attP site. Therefore, many of the 150bp reads containing the left or right part of the attP site would also contain the chromosomal part of the attL/attR site. In order to map the integration events, all reads of each exconjugant were mapped onto the left and right parts of the attP site present in the plasmid. The resulting exconjugant reads that mapped onto attP halves were then in turn mapped onto the chromosome of the wild type, thereby revealing the position of the integration sites.

In *S. coelicolor*, the read mapping revealed two distinct integration sites (site1, site2, Figure 4.13E,F, Table 4.11). While all exconjugants showed evidence of integration into site2, only some of them showed evidence of integration into site1, indicating a preferential integration into site2. This could explain the difference in coverage along some of the integrated plasmids, indicating that one plasmid had integrated fully, the other one only partially. However, some of the integration sites showed a much higher coverage than others when examined by mapping exconjugant reads onto the flanking sites. This could not be explained by multiple integration events in one single genome, but rather indicated a mixture of at least two different exconjugants with varying abundances in the sequenced DNA. This means that one exconjugant would have integrated the plasmid at site1 and site2, while another exconjugant

only integrated it at site 2. The presence of both exconjugants in the DNA would lead to a difference in coverage at the sites, with site2 having a higher coverage.

The presence of different exconjugants was tested by checking for the presence or absence of exconjugant-derived reads stretching across intact integration sites, therefore indicating the presence of an exconjugant with only one integration event. Any number of reads spanning the integration site would indicate a mixture of at least two exconjugant strains. Indeed, all *S. coelicolor* strains showing integration at two sites featured several reads spanning site1 when mapped to the wild type. However, no reads spanning site2 were found. This indicates that all exconjugants had integrated the plasmid into site2, but only some of them had integrated it into site1.

In *S. albus*, two integration sites (site1, site2) were observed as well (Table 4.11). The plasmids were integrated into both sites in all exconjugants. When looking for different integration events, only 57-1 showed reads spanning site1 when mapped to the wild type. This indicates that only 57-1 featured a discernible mixture of strains.

It is noteworthy that none of the DUF692 BGCs were successfully conjugated into *Streptomyces*. While exconjugants could not be generated for 26-1 and 30-3, the putative exconjugant 28-2 was a false positive that had acquired kanamycin resistance in another way, while 30-1 and 32-4 showed incomplete insertion.

Table 4.12: Summary of mapping results from *S. coelicolor* and *S. albus* exconjugants

Species	BGC	Plasmid integration	Site1	Site2	Putative strains
S. coelicolor	3-10	full	+	+	2+
	6-1	incomplete		+	
	11-4	incomplete		+	
	22-1	full		+	
	32-4	incomplete	+	+	2+
	33-1	full		+	
	34-2	full	+	+	2+
	62-1	incomplete		+	
S. albus	6-1	full	+	+	
	8-3	full	+	+	
	11-4	full	+	+	
	33-1	full	+	+	
	34-2	full	+	+	
	57-1	full	+	+	2+
	62-1	full	+	+	
	30-1	incomplete	+	+	
	28-1	absent			

4.4.3 Cultivation, LC-MS analysis and antibiotic activity of strains

S. coelicolor and *S. albus* were grown on solid supplemented minimal medium at 30°C for 7 days and then extracted with methanol. Extracts were run on a Bruker maXis and spectra of empty plasmid exconjugants were compared with BGC-containing exconjugants. However, no novel ions could be found that were unique to a BGC-containing strain. Base peak chromatograms can be found in Appendix B.

No antibiotic activity of any exconjugants or transformants was observed in assays against *E. coli* BL21, *E. coli* JM109 or *Micrococcus luteus* (see Appendix B). Upon visual inspection, no additional or different pigment production compared to the control was observed in the carotenoid-expressing *Streptomyces* exconjugants 3-10 and 57-1. No smell could be detected in 22-1 and 61-2 that could point towards production of a volatile terpene.

4.5 Discussion

4.5.1 Amplification and cloning

A clear association between amplicon length and amplification success was observed, with ca. 20kb appearing to be the upper limit. PCR success could be influenced by the DNA fragment length affected by the DNA extraction as well as freeze-thaw cycles, or by impurities in the DNA acting as PCR inhibitors (W. Shao, Khin, and Kopp 2012; Wnuk et al. 2020). Indeed, the extracted DNA had a brown tint, indicating presence of co-extracted humic acids. In previous examples of direct pathway cloning, the largest PCR amplicon successfully cloned with SLIC was 23 kb in length (D'Agostino and Gulder 2018). In the same study, the authors stressed the importance of HMW DNA free from impurities for successful amplification of long amplicons. Consequently, they used a very gentle lysis followed by phenol-chloroform extraction, which is known to produce extremely high molecular weight DNA (Trigodet et al. 2021). Phenol-chloroform extraction has been used to extract DNA from soils, but shows limited efficacy of removal of humic acids without further, soil-specific extraction approaches (Wnuk et al. 2020; Técher et al. 2010; Sagova-Mareckova et al. 2008). For successful amplification of long DNA fragments from soil DNA, further optimisation of DNA extraction is necessary or another approach, such as the separation of bacterial cells from the soil matrix has to be taken (Liles et al. 2009).

Furthermore, long amplicons showed weaker bands, making it difficult to obtain the amount of DNA required for cloning. The minimum amount of DNA necessary for successful cloning (c. 4 fmol/µL, leading to c. 20 fmol in a 10 µL SLIC reaction that includes other reactants) roughly matched previous reports (Jeong et al. 2012). The unsuccessful cloning of several fragments with sufficient concentrations might have been caused by other factors such as secondary structure formation or toxicity to *E. coli*.

Plasmid 30 contained two different fragments, none matching the metagenomic assembly. This could have been caused by a misassembly. Metagenomic assemblers are known to have difficulties resolving multiple closely related strains and producing contigs that are chimeras between two or more strains (Sevim et al. 2019; Kolmogorov et al. 2019). In the metagenome, the large number of contigs classified as UBA7966, the multiple copies of pmmoABC on them as well as the fact that no high-quality bins were produced all indicate the presence of several closely related strains, potentially leading to misassembled contigs.

4.5.1.1 *Pseudomonas putida* transformations

A transient small-colony/reduced-growth phenotype was observed for several DUF692 BGCs when expressed in *Pseudomonas putida* KT2440 trfA. This indicates that the BGC product, or possibly a gene in the BGC, is detrimental to the fitness of *P. putida*.

It is well known that, given enough generations, a bacterial population will accumulate mutations that reduce the fitness cost of a plasmid, and the less fit cells will be outcompeted by the fitter ones. This can occur through mutations rendering the expressed genes non-functional, reducing expression or reducing the plasmid copy number (Carroll and Wong 2018). It can, however, also occur through compensatory mutations in chromosomal genes or pathways, leading to an adaptation of the bacteria to the plasmid (Hall et al. 2021). It is unclear which of these occurred in the case of DUF692 BGCs. However, the absence of any detectable product suggests that production has been abolished.

Since *P. putida* did not show expression of mScarlet under the P21 promoter in pBCkBAC-g2, the genes responsible for the phenotype are most likely proposed biosynthetic genes (DUF692, precursor peptide, DUF2063). On the other hand, there is a chance that the absence of the

transport-related genes might lead to the phenotype. The plasmid did not lead to a phenotype in *E. coli*, a similarly related Gammaproteobacterium. In *E. coli*, however, the plasmid is present only in 1-2 copies per cell, unless induced (using L-arabinose in strain EPI300) (Aakvik et al. 2009, 2). This indicates that the toxicity of the product might be dose-dependent and not lethal. This observation is compatible with a proposed role in metal chelating.

The genes responsible for the reduced fitness could potentially be detected by single nucleotide polymorphism (SNP) analysis of mutated genes, or by knocking out the genes one by one. Employing inducible promoters to uncouple growth from production would be a way to make product detection more likely.

4.5.1.2 Streptomyces conjugation

The conjugation of several plasmids was unsuccessful or partial, especially in *Streptomyces coelicolor*. This could be due to toxicity of the expressed products for the cell. However, the fact that the empty pBCkBAC-g2 plasmid exconjugant was not achieved in *S. coelicolor* hints at other potential factors such as random recombination events. Full integration, partial integration and inability to achieve exconjugants were observed in both g1- and g2-based plasmids. The insertion into multiple sites in the genome is consistent with previous reports of simultaneous integration of attP-containing plasmids into attB sites and pseudo-attB sites in the same genome in *Streptomyces coelicolor* and other *Streptomyces* species through the phi-C31 integrase (Combes et al. 2002). Indeed, the preferred site2 proved identical with the intended target attB, while site1 was identified as pseB2, a pseudo-attB site with lower integration efficiency (Combes et al. 2002). To the author's knowledge, the integration into pseudo-attB sites has not been reported for *S. albus*, but does not come as a surprise. It should be noted that the number of copies of the inserted plasmid cannot easily be determined by

comparing the coverage to the rest of the genome, since the differing GC contents of genome, plasmid and insert would lead to bias in sequencing depth (Benjamini and Speed 2012).

The uneven coverage observed along the plasmid sequences of many of the integrated plasmids indicate that either the integration into one of the sites was not complete, or it could indicate the presence of several strains with different integration success in the same sites. Long-read sequencing could be employed to determine this. Since seemingly single colonies were picked, it is possible that recombination and excision of sequences took place after the integration event, leading to different amounts of BGC left in different individuals of the same population. *Streptomyces* are known for genomic instability and frequent genome rearrangements, but these arrangements occur mostly at the distal arms of the chromosome (Hoff et al. 2018; Choulet et al. 2006; Tidjani et al. 2019). In the central regions of the chromosome, however, double strand breaks (DSBs) have been shown to lead to deletions via non-homologous end joining (NHEJ) (Hoff et al. 2018). The phiC31 recombinase that is responsible for attP x attB recombination induces DSBs, but also immediately joins the ends together (Merrick, Zhao, and Rosser 2018). It is conceivable that the deletions resulted from a failed attP x attB recombination event, and that due to toxicity of the intact BGC, only the exconjugants containing failed integration events survived.

The conjugation of DUF692 BGCs was unsuccessful in all cases. Either no exconjugants could be generated, or the insertion was incomplete. In light of the transient phenotype (reduced growth) observed in *Pseudomonas* transformants of DUF692 BGCs, it seems plausible that the BGCs could be lethal to *Streptomyces*, making integration of a functional expression plasmid impossible. The fact that the DUF692 BGCs 30-1 (*S. albus*) and 32-4 (*S. coelicolor*) were the only BGCs that were present in both integration sites, but incomplete in both of them reinforces

this view. A more successful approach might be to employ inducible promoters instead of constitutive promoters, as strong constitutive promoters have been known to hamper successful heterologous expression of secondary metabolites (R. Duell et al. 2020; C.-H. Ji, Kim, and Kang 2019). Another approach could be to use cell-free TX-TL systems to circumvent problems of toxicity (X. Ji, Liu, and Li 2022).

Streptomyces coelicolor showed a much higher frequency of incorrect insertion. If the failed insertions are indeed related to BGC toxicity, this could be explained by the fact that the heavily mutated *Streptomyces coelicolor* M1154 shows a reduced tolerance to toxic products. Indeed, the slow growth of its parent strain M1152 has been explained by increased oxidative stress in the cell (Sulheim et al. 2020). Oxidative stress can increase susceptibility to antibiotics in *E. coli* (X. Wang and Zhao 2009). Alternatively, the high frequency of incorrect insertion could also be related to the fact that the *Streptomyces coelicolor* integration sites are only ca. 7000 bp apart, potentially making recombination events between the inserts likely due to proximity of identical DNA sequences.

The fact that some exconjugants integrated one, while others integrated two copies of the plasmid could also affect the fluorescence observed by mScarlet exconjugants. For example, promoter activity of sp24 (mScarlet 1) has been reported to surpass P21 (mScarlet4, 7) (Myronovskyi and Luzhetskyy 2016). However, as can be seen in Figure 4.3 above, the mScarlet 4 and 7 exconjugants both show greater fluorescence than mScarlet 1. This could be due a difference in copy number. Taken together, the differences in copy number, difficulty to achieve exconjugants, and truncation of plasmids indicate that the pBCkBAC-g1/g2 expression plasmids need redesign and optimisation if they are to be routinely used.

4.5.1.3 Experimental design

A premise of this work was the idea that the unlikely event of successful expression could be made possible by ensuring that enough BGCs made it through the whole pipeline. However, this was not the case, and no novel compounds were found. This was in part due to the unsuitability of many of the procedures for medium to high throughput experiments, with large amounts of manual labour, long incubation times and unreliable outcomes. Thereby, only 13 BGCs were examined in the end – not enough to overcome the barriers to successful expression by sheer numbers. However, experimental design considerations such as the use of constitutive instead of inducible promoters, or the use of GC-MS for the detection of potentially volatile terpenes are likely responsible as well. It is likely that more advances in synthetic biology, such as cross-species inducible promoters, tRNA complemented expression hosts or higher throughput methods for conjugation are necessary for this type of pipeline to be successfully carried out in a reasonable time. Until then, a more focused approach on single, low-risk BGCs is likely necessary.

5 General Discussion

Three approaches for evaluating and accessing biosynthetic diversity of Mars Oasis soil bacteria were taken: a long-read shotgun metagenome with subsequent heterologous expression, traditional isolation work, and a novel metagenomic library screen. While the long read metagenome revealed a large number of BGCs, accessing them using PCR-based cloning and heterologous expression proved difficult. Sequencing of bacterial isolates obtained through traditional isolation methods showed the limited overlap between BGCs observed in the isolates and in the metagenomic dataset. A novel metagenomic library screening approach for regulatory genes was validated in isolates but did not yield success due to the absence of the target organism from the metagenomic library. The trade-offs between potential novelty and chance of success are clear: while the metagenome and heterologous expression work promise the expression of highly novel clusters from uncultured lineages, the difficulties of cloning and expression give this approach a high chance of failure to discover anything at all. Traditional isolation approaches yield many commonly cultivated genera with BGCs more closely related to BGCs in databases, indicating that any detected specialised metabolites might be closely related to known compounds. Though the metagenomic library screening approach theoretically has the potential to discover novel BGCs with convenient mechanism of activation, no BGCs were recovered through it. It is clear that in natural product discovery, there are no low-hanging fruits.

5.1 Metagenomics for BGC discovery

In the present metagenomic sequencing work, the use of nanopore metagenomic sequencing, binning and contig-based classification approaches using GTDB combined with BGC genome mining enabled the identification of 1417 BGCs, 65% of which were complete, from a wide range of soil bacteria. This confirms and further expands our knowledge of the biosynthetic

potential of difficult-to-culture phyla such as Verrucomicrobiota, Acidobacteriota and Gemmatimonadota. In addition, uncultured and underexplored lineages of the well-known producer phyla Actinobacteriota (classes Thermoleophilia and Acidimicrobiia) and Proteobacteria (order UBA7966) show a previously undetected biosynthetic potential. The present work furthermore demonstrates that ONT long-read sequencing enables the assembly, detection and taxonomic classification of full-length BGCs on large contigs from a highly complex environment such as soil using only 72 Gb of sequencing data, which presents a >10-fold reduction compared to studies using short reads to recover large and complete BGCs (Crits-Christoph et al. 2018). However, while the low amount of bases sequenced demonstrates the great efficacy of nanopore long reads for metagenomic assembly, it also leads to a high error rate in the assembly. While it could be demonstrated that the high error rate did not lead to an overestimation of BGC novelty with BiG-SLiCE, the errors nevertheless strongly affected gene prediction and thereby have a negative effect of most downstream applications. For example, synthesising DNA based on the genes of any but the highest coverage organisms would likely incorporate the sequencing errors, rendering the synthetic DNA useless. Increasing short read coverage to at least the same coverage as the nanopore reads would likely have improved polishing outcomes. However, rapid advances in ONT sequencing technology might render short-read polishing obsolete soon. A recent benchmark of metagenomic sequencing using ONT's R10.4 flow cells combined with Q20+ chemistry led to the near-perfect assembly of the Zymo mock community using only nanopore reads (Sereika et al. 2021). Most importantly, homopolymers of up to 10 bases can be resolved with relative confidence. The second weakness of the present work is the usage of only one soil DNA sample, which rendered binning inefficient. Using the short read and long read coverages to simulate differential abundance led to a small improvement in binning efficiency. However, using several different DNA extraction methods would likely have led to improved binning

(Albertsen et al. 2013). The alternative CAT-based approach enabled classification of >60% of BGCs at order level. While a reference database approach like CAT is likely to be more error-prone than binning, the large size of the contigs – often in the range of megabases – potentially makes up for that.

With nanopore sequencing becoming more widespread and increasing both in accuracy and throughput, it will soon be commonplace to profile the biosynthetic potential of uncultured microbes from diverse environments without enormous sequencing efforts. This will lead to a great improvement in completeness and taxonomic classification of BGCs recovered from metagenomes, which in turn will let us pinpoint the most promising BGC-rich clades from Acidobacteriota as well as other phyla. BGCs can then be recovered by targeted library screening or PCR-based cloning as in the present work. To translate these genomic discoveries to actual novel compounds, however, hurdles to the heterologous expression of BGCs from distant phyla need to be overcome through e.g. refactoring or host optimisation. Luckily, detecting specific BGC-rich genera would also allow for more targeted isolation, opening another avenue for novel natural product discovery. Furthermore, metagenomes with a good BGC resolution can be coupled with transcriptomics to analyse the transcriptional response of the BGCs to different stimuli (Van Goethem et al. 2021; Crits-Christoph et al. 2018). This allows prediction of the functions and natural roles of unknown BGCs in absence of characterised homologs or tell-tale genes such as siderophore receptors. For remote and endangered environments such as the Antarctic Peninsula, which is warming rapidly due to climate change, metagenomic strategies will prove especially valuable to document the microbial biodiversity for future studies.

Genome mining is the only method available for predicting the biosynthetic potential of uncultured bacteria. However, the limitations of this technology must be considered when labelling lineages as having a high or low potential for specialised metabolite production. Hoskisson et al. applied the helpful concept of "(un)known (un)knowns" to BGCs and their products, emphasising that the "unknown unknowns" – i.e. BGCs not detected by genome mining algorithms and not detected from cultures – are likely to contain the greatest untapped novelty (see Table 5.1; Hoskisson and Seipke 2020). Some of these BGCs are most likely present in cultured and well-exploited bacteria, but have thus far escaped detection due to limitations in activity or analytical detection methods. However, they are likely to be much more numerous in uncultured bacterial lineages, where activity- or analytical chemistry-guided approaches are usually impossible (with the exception of certain symbiont communities). Therefore, uncultured lineages, especially if they are only distally related to characterised isolates, might very possibly be more biosynthetically talented than is currently detectable with genome mining tools. However, evolutionary mining approaches that use e.g. duplicated household genes likely to be involved in specialised metabolite biosynthesis or self-resistance, as well as "black box" machine learning approaches are opening the door towards assessing and exploiting the "unknown unknowns" from both isolates and MAGs (Sélem-Mojica et al. 2019; Alanjary et al. 2017; Hannigan et al. 2019). Another important factor when estimating the unknowns of specialised metabolism in the environment is the level of difficulty associated with obtaining sufficient data about the organism (Table 5.2). Easily culturable bacteria, regardless of abundance, are readily isolated and sequenced. Difficult-to-culture bacteria that are abundant in the environment, such as the different lineages of the Acidobacteriota phylum found in the present study, are characterizable through long-read metagenomic sequencing. This characterisation can lead to predictions about their biosynthetic potential as well as aid in targeted cultivation efforts. Low-abundance hard-to-culture lineages however are less likely to

be characterised even by long-read metagenomic sequencing because of the high coverage necessary. This makes appraisal of their BGC content as well as isolation more challenging. However, targeted enrichment cultures as well as techniques such as nanopore adaptive sequencing and single-cell genomics can help overcome these challenges (Payne et al. 2021; Doud et al. 2020; Kogawa et al. 2022).

Table 5.1: Known knowns, unknown knowns, known unknowns and unknown unknowns in specialised metabolite discovery. Adapted from Hoskisson et al. (2020).

		BGC: known	BGC: unknown
Product: known		**Known knowns** BGCs and product linked (e.g. all MiBiG entries)	**Unknown knowns** Product without linked BGC (e.g. many NPAtlas entries)
Product: unknown		**Known unknowns** BGC detected, but no product linked (e.g. most of the BiG-FAM or antiSMASH databases)	**Unknown unknowns** BGC not detected, no product observed

Table 5.2: Consequences of cultivability and abundance on ease of characterisation

	High abundance	Low abundance
Easy to isolate	Readily characterised through isolates. E.g. *Sphingomonas* in the present study	Readily characterised through isolates. E.g. *Streptomyces* in the present study
Difficult to isolate	Good potential for metagenomic characterisation. E.g. Acidobacteria in the present study	Low potential for metagenomic characterisation or isolation. E.g. different candidate phyla

5.2 Heterologous expression of metagenome-derived BGCs

In the present work, the heterologous expression study did not lead to discovery of any novel compounds. A transient phenotype was observed in DUF692-expressing *P. putida* and the same BGCs did not produce any viable exconjugants in *Streptomyces*. This suggests toxicity of the BGC product. The signal peptide detected in the potential precursor peptide indicates that the peptide is exported across the cell membrane into the periplasm via the Sec pathway. Sec signal peptides have been reported to be interchangeable across bacterial lineages, but the export efficiency varies greatly depending on the signal peptide and protein (Hemmerich et al. 2016). Therefore, it could be the case that the cysteine-rich precursor peptide is toxic in large quantities, and while *P. putida* can efficiently export it and thereby reduce the damage, *Streptomyces* exconjugants cannot and therefore do not survive. Furthermore, while it is uncommon for RiPPs to be exported using Sec pathway, it has been documented for a small number of bacteriocins (Worobo et al. 1995; Herranz and Driessen 2005). The Sec translocation pathway has also been shown to be able to transport modified peptides, but not fully modified and cyclised nisin (Kuipers et al. 2006; Kuipers, Rink, and Moll 2009). Therefore, it is unlikely that the potential precursor peptide is heavily modified or cyclised before export. However, the peptide might be modified in the periplasmatic space. For example, it is possible that some or all of the cysteines could be oxidised to disulfide bonds by periplasmatic enzymes, as is the case with many exported proteins (Denoncin and Collet 2013). Furthermore, it is unknown whether the peptide is exported further across the outer membrane by a secretion system. The improbability of cyclisation catalysed by DUF692 enzymes before export firstly raises the question of their function, and secondly makes an involvement of some or all of the cysteines in metal coordination seem plausible. It is well-known that particulate methane monooxygenases (pMMOs) depend on copper as a cofactor (Dassama, Kenney, and Rosenzweig 2017). Moreover, several methanotroph methanol dehydrogenases have been

shown to depend on lanthanides, and selectively lanthanide-binding proteins have been discovered (Huang, Yu, and Chistoserdova 2018; Kato et al. 2020; Cotruvo et al. 2018). Metal-binding peptides have many potential applications ranging from mining to bioremediation to medical applications (Mejáre and Bülow 2001; Semrau et al. 2020). Furthermore, improving our understanding of the mechanisms for metal uptake and homeostasis in methanotrophs would be important for our understanding of their role in climate change.

5.3 Isolation work, biogeography and BGC novelty

In culture-based studies, many factors affect the novelty of the recovered bacterial strains, their BGCs and therefore also the detected specialised metabolites. It is unlikely to discover novel natural products using well-established methods in well-researched bacterial genera from well-explored environments (Figure 5.1). While the difference between outcomes of isolation or detection methods, as well as the "exploitation level" of a genus are relatively straightforward to define, it is less obvious where the line can be drawn to identify the most promising environments, since the biogeography of specialised metabolite production remains understudied. For example, it could be argued that the physical distance of Antarctica from everywhere else makes Mars Oasis soils a good candidate for yielding novel natural products. However, it is unclear to what extent geographic distance is an important factor for BGC novelty. A global, amplicon-based study of NRPS and PKS domains in soil showed an influence of both biome type and geographic distance on domain similarity (Charlop-Powers et al. 2015). A similar study in Australia singled out latitude as the main factor in domain composition (Lemetre et al. 2017). Yet another study that was limited to US soils showed a geographic split between the southwestern and northeastern coast samples. However, more granular analysis of dozens of soil factors indicated that these geographic differences also correlated with changes in moisture, organic matter and content of different minerals,

potentially explaining a large part of the difference (Charlop-Powers et al. 2014). Chase et al. (2021) determined that within a set of 118 closely related marine *Salinispora* isolates (>99% 16S rRNA gene identity), geographic location only explained 11% of variation in gene cluster family distribution, and only 3% of all gene clusters found in the set were unique. This indicates that, if the findings of Chase et al. are generalisable for BGC evolution in other lineages, the geographic distance of Antarctica to other land masses might not contribute greatly to BGC novelty at Mars Oasis. However, in addition to sheer distance, the dispersal of organisms to Antarctica is limited further through the Antarctic circumpolar current and atmospheric circumpolar vortex that started about 30 million years ago, concurrent with the start of glaciation (DeConto and Pollard 2003). This isolation combined with an extreme environment could impede the colonisation by microorganisms and thereby lead to a more pronounced phylogenetic and functional divergence than in between e.g. soils on different continents. In support of this, aerobiological studies have indicated that the majority of aerosolised bacteria over Antarctica originates from the continent (Bottos et al. 2014). Furthermore, studies on bacterial isolates indicate that dispersal is an important factor in prokaryotic biogeography (Andam et al., 2016). However, it is still likely that most bacterial lineages on Antarctica have close relatives elsewhere, as bacterial 16S rRNA genes only diverge an estimated 1% per 50 million years (Ochman, Elwyn, and Moran 1999). Furthermore, an environment such as Mars Oasis, with its meltwater pools and stands of bryophytes, could be more hospitable to colonising organisms than e.g. the hyper-arid deserts of the McMurdo Dry Valleys, where hydration has been suggested to be sustained for a large part through oxidation of hydrogen to H_2O (Ortiz et al. 2021). Mars Oasis has been shown to differ in soil chemistry from other polar desert soils in the vicinity (Chong et al. 2012). However, respective influences of differences in soil properties, climate as well as the isolation provided by the oceanic and atmospheric currents on the community composition have not been determined. The combination of

undersampling of the Antarctic continent and the general inability of 16S rRNA gene amplicon sequencing studies to distinguish between ecotypes or to determine metabolic capacities means that basic questions about biogeography in the Antarctic and elsewhere remain open. Furthermore, the high taxonomic rank that 16S rRNA gene sequencing studies are often interpreted at, coupled with a complete absence of isolate genome or physiological data, leads to potentially confusing findings. For example, a general trend in community composition of many soils is an increase in the abundance of Acidobacteria at low pH. However, it has been observed that several clades of Acidobacteria actually prefer neutral or alkaline soil, such as the soil used in the present work (Lladó, López-Mondéjar, and Baldrian 2018; Ivanova et al. 2020). However, as techniques are evolving, we might come closer to giving more informed answers to questions of biogeography even in absence of isolates. For example, an analysis of global catalogue of MAGs showed that most unique gene variants are rare and biome-specific, as well as less likely to be adaptive than more prevalent and cosmopolitan genes, demonstrating the potential of large MAG catalogues for deducing evolutionary patterns of prokaryotic genes and genomes (Coelho et al. 2022). Furthermore, a recent paper by Ortiz et al. (2021) combined genome-resolved metagenomics, metabolic modelling, isolation and microcosm work to study trace gas metabolism in Antarctic soils. Applying these tools across a range of biomes might lead to fruitful meta-analyses that could shine a light on the determinants of biogeography of bacterial communities in Antarctica and beyond.

The isolates obtained by conventional isolation in the present study showed a relatively high level of novelty at species level based on 16S rRNA gene similarity, which is concordant with a previous metagenomic analysis as well as the geographical isolation of Antarctica for 30 million years and the fact that it is a relatively unexplored environment (Pearce et al. 2012). However, they also confirm that untargeted isolation approaches using standard laboratory

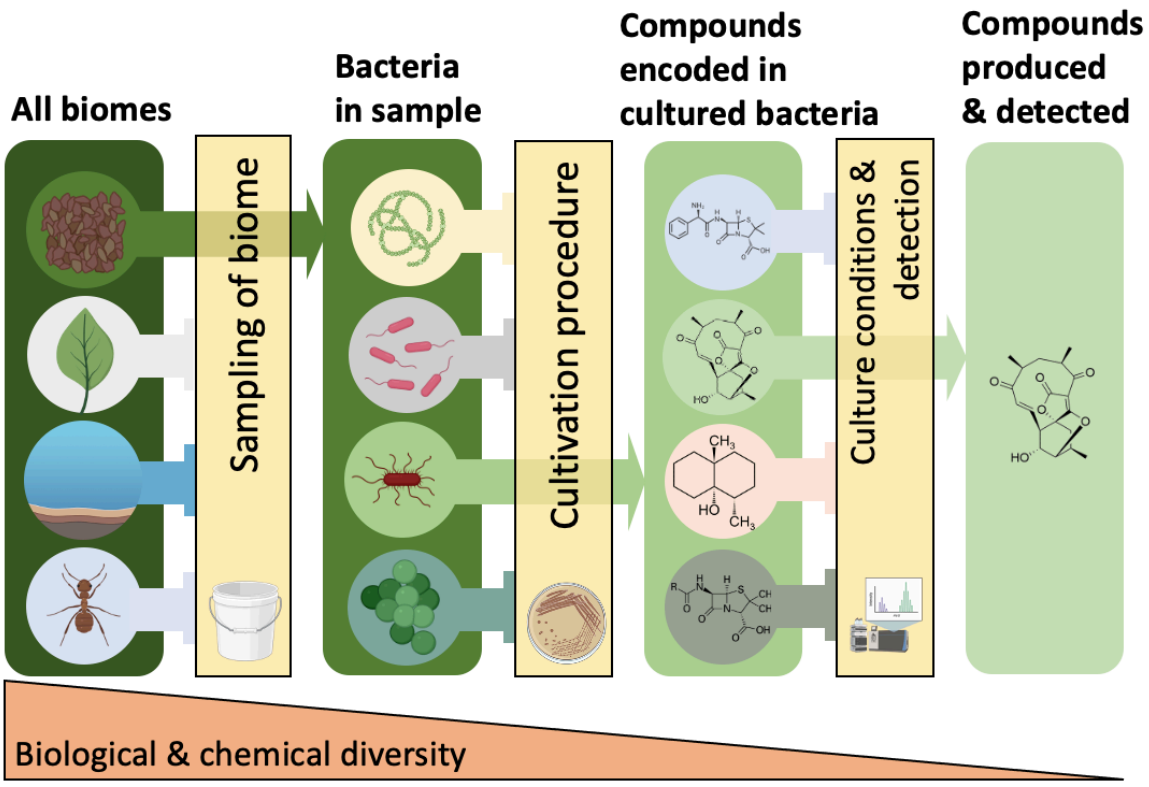


Figure 5.1: There is a loss of diversity with each selection step in the natural products discovery pipeline, from choosing the biome to sample, to isolation procedures, culture conditions and detection of compounds. Implementing novel methods at each of the selection steps can increase the chances of discovery of novel compounds.

media are of little use to expand the range of isolated genera or higher taxonomic orders when working with a well-studied environment such as soil. When taking into account findings that many BGCs are relatively conserved within genera, it is likely that the isolates obtained here would not produce significantly different compounds than isolates of the same genera obtained elsewhere (Gavriilidou et al. 2021; Chase et al. 2021). However, it is possible that the geographic isolation coupled with climactic changes through millions of years could have yielded unique adaptations for cold temperatures, which could potentially be reflected in the types of specialised metabolites produced. For example, compounds that decay quickly at room temperature might be stable enough at sub-zero temperatures; enzymes catalysing novel types of reactions might have evolved through cold adaptation. Notwithstanding the potential novelty

induced by these adaptations, standard isolation procedures could have higher chances of yielding highly novel compounds when applied to underexplored types of environments. A computational study on a large set of MAGs from different environments showed that 74% of BGC families are specific for a given biome type such as soil, freshwater, gut or thermal vents (Gavriilidou et al. 2021). Therefore, exploring unusual biomes such as the diverse host-associated microbiomes or the deep subsurface might be the most promising way forward when using standard isolation and screening procedures.

In addition to the sampled biome, the isolation procedure is the other factor of prime importance for specialised metabolite discovery since it holds the potential to significantly alter the taxonomic composition of recovered bacteria and therefore BGC diversity. It has been shown in previous studies as well as in the present work that several uncultured lineages in various underexplored phyla contain a large amount of BGCs with a high degree of divergence from known sequences (Borsetto et al. 2019; Crits-Christoph et al. 2018). A method of reliably isolating BGC-rich acidobacterial genera would likely constitute a breakthrough for natural product discovery, giving access to a treasure trove of novel BGCs with undiscovered products. The developments in high-throughput cultivation as well as targeted, metagenome-guided isolation techniques indicate that this breakthrough could occur soon. However, searches in PubMed indicate that the number of novel species and genera described per year is decreasing, while publications about metagenomes are steadily increasing (Figure 5.2). While there are likely to be other factors involved, such as the impact of the Covid19 pandemic in 2020, it seems that interest in isolation and description of novel species might be waning in the face of culture-independent techniques. Possibly, the quantity is counterbalanced by quality, i.e. targeted isolation of specific hard-to-culture bacteria, but this cannot easily be verified.

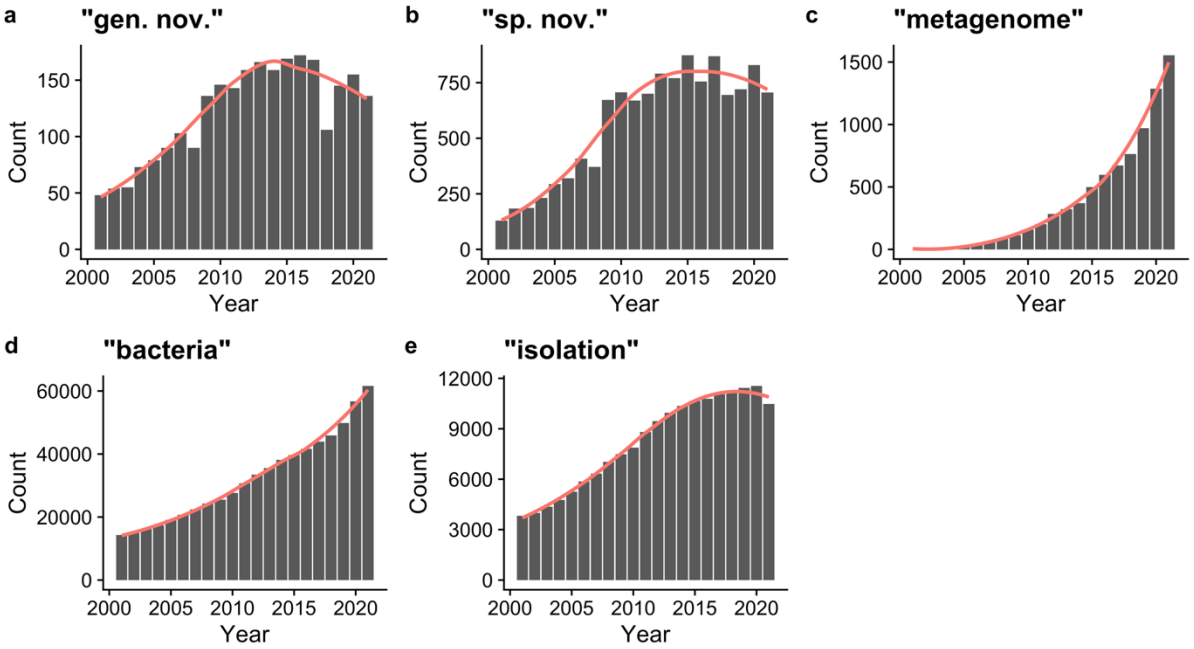


Figure 5.2: Numbers of PubMed results per year from 2001 to 2021 using the following search terms: a) ("gen. nov."[Title/Abstract]) AND (bacteria[Title/Abstract] OR bacterial[Title/Abstract] OR bacterium[Title/Abstract]); b) ("sp. nov."[Title/Abstract]) AND (bacteria[Title/Abstract] OR bacterial[Title/Abstract] OR bacterium[Title/Abstract]); c) (metagenome[Title/Abstract] OR metagenomic[Title/Abstract]) AND (bacteria[Title/Abstract] OR bacterial[Title/Abstract] OR bacterium[Title/Abstract]); d) (bacteria[Title/Abstract] OR bacterial[Title/Abstract] OR bacterium[Title/Abstract]); e) ("isolation" OR "isolated") AND (bacteria[Title/Abstract] OR bacterial[Title/Abstract] OR bacterium[Title/Abstract])

It is worth mentioning that an innovative isolation approach was planned in this work that would have combined the previously reported approaches of droplet microfluidics, matrix encapsulation and in situ incubation in hollow fibre chambers (Mahler et al. 2021; Ben-Dov, Kramarsky-Winter, and Kushmaro 2009; Aoi et al. 2009). Alginate droplets containing single bacterial cells would be generated using a microfluidic device and fed into a hollow fibre which would be placed back into the soil. Following incubation, the droplets would be analysed by metagenomic sequencing as well as deposited on solid media to obtain colonies (Figure 5.3). However, it was not possible to obtain hollow fibres, so the project could not proceed.

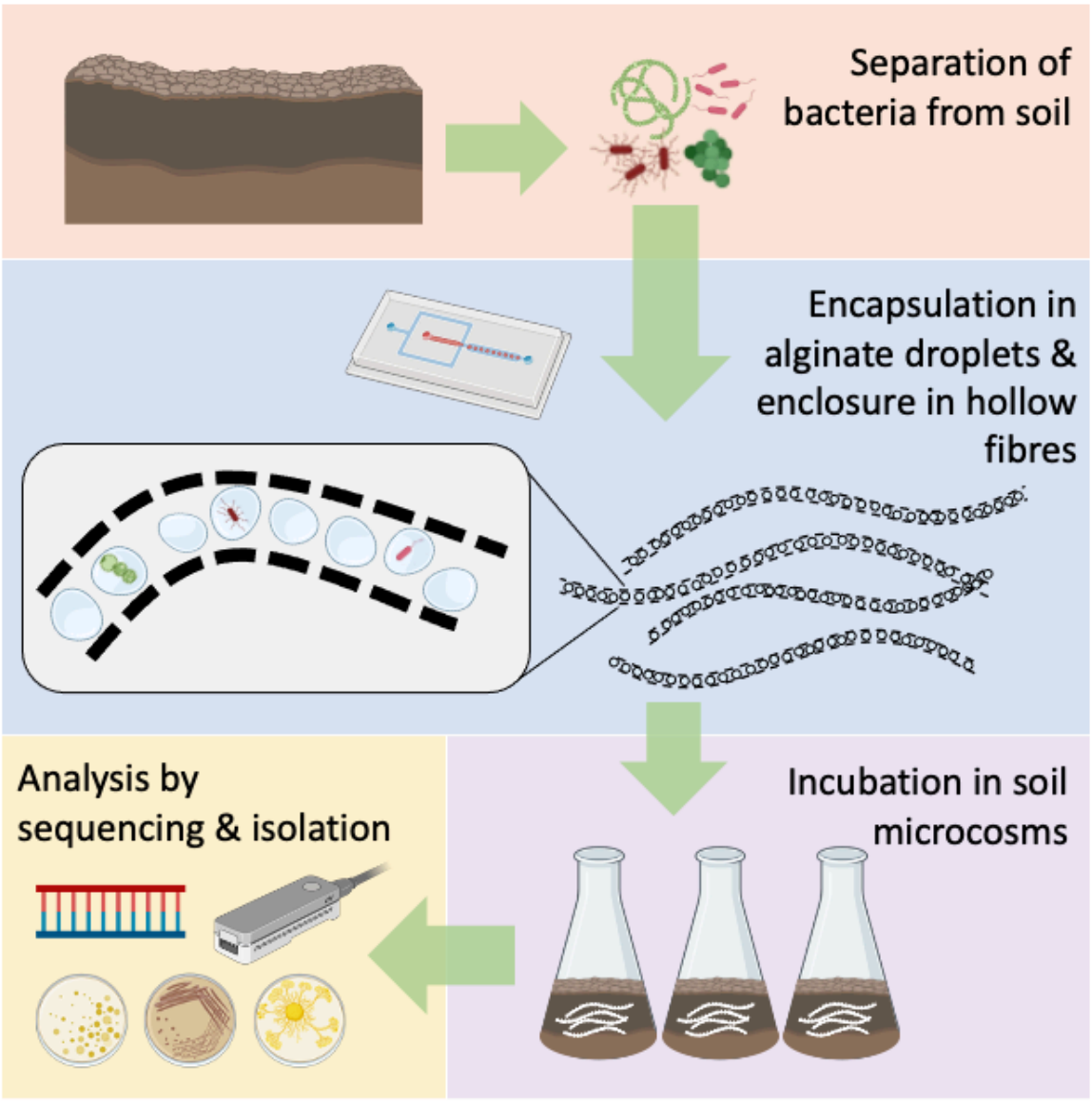


Figure 5.3: Outline of the proposed microfluidics-based in-situ incubation strategy.

5.4 Outlook

It must needs be remarked that existing culture collections of Actinobacteria such as *Streptomyces* should not be overlooked in the hunt for novelty. They still contain a vast repertoire of cryptic BGCs encoding an immense number of potentially useful natural products. While it has been suggested that BGC families are mostly specific to genera, it has also been demonstrated that many BGCs are rare within genera (Gavriilidou et al. 2021; Watve et al.

2001; Baltz 2006; Chase et al. 2021). Traditional screening platforms are significantly hampered by the rediscovery of highly common BGCs that make discovery of rarer BGCs less likely (Culp et al. 2019). However, thanks to the constantly evolving approaches aimed at tackling the issue of rediscovery, strain libraries will likely continue to yield novel compounds. These approaches include targeted activation of BGCs of interest as well as improved dereplication of spectroscopy signals by methods such as GNPS (Wang et al. 2016). Furthermore, the increasing automatic integration of genomic and metabolomic data (metabologenomics) is a promising development. Hitherto, metabologenomics has consisted of large-scale presence/absence correlations as well as specific feature-based predictions based on e.g. NRPS peptide sequence (Goering et al. 2016; Kersten et al. 2011). These two approaches are still at the heart of metabologenomics, but are being improved, expanded and further automated in programs such as NPLinker (Eldjárn et al. 2021). However, as of today, connecting an observed MS signal to a specific BGC in an isolate is still often a puzzle that requires a large amount of expertise and time to solve. A way of reliably and automatically linking MS signals to BGCs by means of rules-based or machine learning algorithms has the potential to revolutionise natural product discovery. This would not only improve the prioritisation of strains and BGCs for study, but also enable complementary analyses of BGC sequence diversity and chemical compound diversity in isolates, culture collections and the environment.

5.4.1 Future work

There are many ways in which the work presented here could be improved, e.g. by deeper short-read sequencing with different DNA extraction protocols to reduce indel errors and improve MAG quality to obtain a set of high-quality MAGs. Furthermore, some aspects of the heterologous expression study could be repeated using a lower growth temperature as well as

different extraction regimes. However, there are also ways in which to not only improve what has been done, but also develop on it.

Firstly, the UBA7966 DUF692 BGCs that showed a phenotype in *P. putida* and proved lethal in *Streptomyces* could be studied further. Since a phenotype has been observed, tweaking expression might realistically lead to isolation of the peptide that could then be assessed for structure and function. Furthermore, the gene functions in the small BGC would be straightforward to determine by knockouts. To examine the hypothesis of the proposed role of the peptide for metal binding, microcosms with addition of metals such as copper as well as EDTA could be set up and the transcriptional response measured. The high abundance of the UBA7966 order also makes the Mars Oasis soil a promising candidate for cultivation of the lineage. This might take a long time, however, since the previous isolation of an atmospheric methane oxidiser took over two years (Tveit et al. 2019).

Secondly, a metagenome-guided cultivation approach could be employed to isolate target groups. While binning was not highly efficient, taxonomic classification was successful for a large number of BGCs and the contig size means that many BGCs are on the same contigs as large parts of the genome. This means that through analysis of classified bins and contigs, the nutritional requirements and potential antibiotic resistances of specific lineages in the metagenome could be identified and harnessed for a metagenome-guided cultivation approach. The prime target for this would be BGC-rich acidobacterial groups.

Thirdly, the relatively low diversity (for soil) and presence of many interesting producer phyla could make a promising target for an integrated “omics” approach: Soil extracts could be analysed for specialised metabolites which could then be tied to BGCs obtained through long-

read metagenomic sequencing and correlated to BGC expression levels obtained through metatranscriptomics. This could vastly improve prediction of structure and function of specialised metabolites from uncultured bacteria at Mars Oasis.

6 Bibliography

Aakvik, Trine, Kristin Fløgstad Degnes, Rannveig Dahlsrud, Frank Schmidt, Ragnar Dam, Lihua Yu, Uwe Völker, Trond Erling Ellingsen, and Svein Valla. 2009. 'A Plasmid RK2-Based Broad-Host-Range Cloning Vector Useful for Transfer of Metagenomic Libraries to a Variety of Bacterial Species'. *FEMS Microbiology Letters* 296 (2): 149–58. https://doi.org/10.1111/j.1574-6968.2009.01639.x.

Adams, Byron, Rob Arthern, Angus Atkinson, Carlo Barbante, Roberto Bargagli, Dana Bergstrom, Nancy A. N. Bertler, et al. 2009. In *Antarctic Climate Change and the Environment. A Contribution to the International Polar Year 2007-2008*, edited by Turner, J., Bindschadler, R.A., Convey, P., Di Prisco, et al., 183–298. Cambridge: Scientific Committee on Antarctic Research, Scott Polar Research Institute. https://hal.archives-ouvertes.fr/hal-01205939.

Agarwal, Vinayak, Jessica M. Blanton, Sheila Podell, Arnaud Taton, Michelle A. Schorn, Julia Busch, Zhenjian Lin, et al. 2017. 'Metagenomic Discovery of Polybrominated Diphenyl Ether Biosynthesis by Marine Sponges'. *Nature Chemical Biology* 13 (5): 537–43. https://doi.org/10.1038/nchembio.2330.

Aigle, Bertrand, and Christophe Corre. 2012. 'Chapter Seventeen - Waking up Streptomyces Secondary Metabolism by Constitutive Expression of Activators or Genetic Disruption of Repressors'. In *Methods in Enzymology*, edited by David A. Hopwood, 517:343–66. Natural Product Biosynthesis by Microorganisms and Plants, Part C. Academic Press. https://doi.org/10.1016/B978-0-12-404634-4.00017-6.

Aislabie, Jackie M., Anna Lau, Melissa Dsouza, Charis Shepherd, Phillippa Rhodes, and Susan J. Turner. 2013. 'Bacterial Composition of Soils of the Lake Wellman Area, Darwin Mountains, Antarctica'. *Extremophiles* 17 (5): 775–86. https://doi.org/10.1007/s00792-013-0560-6.

Akhter, Najeeb, Yaqin Liu, Bibi Nazia Auckloo, Yutong Shi, Kuiwu Wang, Juanjuan Chen, Xiaodan Wu, and Bin Wu. 2018. 'Stress-Driven Discovery of New Angucycline-Type Antibiotics from a Marine Streptomyces Pratensis NA-ZhouS1'. *Marine Drugs* 16 (9): 331. https://doi.org/10.3390/md16090331.

Alanjary, Mohammad, Brent Kronmiller, Martina Adamek, Kai Blin, Tilmann Weber, Daniel Huson, Benjamin Philmus, and Nadine Ziemert. 2017. 'The Antibiotic Resistant Target Seeker (ARTS), an Exploration Engine for Antibiotic Cluster Prioritization and Novel Drug Target Discovery'. *Nucleic Acids Research* 45 (W1): W42–48. https://doi.org/10.1093/nar/gkx360.

Alanjary, Mohammad, Katharina Steinke, and Nadine Ziemert. 2019. 'AutoMLST: An Automated Web Server for Generating Multi-Locus Species Trees Highlighting Natural Product Potential'. *Nucleic Acids Research* 47 (W1): W276–82. https://doi.org/10.1093/nar/gkz282.

Alberti, Fabrizio, Daniel J. Leng, Ina Wilkening, Lijiang Song, Manuela Tosin, and Christophe Corre. 2019. 'Triggering the Expression of a Silent Gene Cluster from Genetically Intractable Bacteria Results in Scleric Acid Discovery'. *Chemical Science* 10 (2): 453–63. https://doi.org/10.1039/C8SC03814G.

Albertsen, Mads, Philip Hugenholtz, Adam Skarshewski, Kåre L. Nielsen, Gene W. Tyson, and Per H. Nielsen. 2013. 'Genome Sequences of Rare, Uncultured Bacteria Obtained by Differential Coverage Binning of Multiple Metagenomes'. *Nature Biotechnology* 31 (6): 533–38. https://doi.org/10.1038/nbt.2579.

Alishum, Ali. 2019. 'DADA2 formatted 16S rRNA gene sequences for both bacteria & archaea'. Zenodo. https://doi.org/10.5281/zenodo.3266798.

Alneberg, Johannes, Brynjar Smári Bjarnason, Ino de Bruijn, Melanie Schirmer, Joshua Quick, Umer Z. Ijaz, Leo Lahti, Nicholas J. Loman, Anders F. Andersson, and Christopher Quince. 2014. 'Binning Metagenomic Contigs by Coverage and Composition'. *Nature Methods* 11 (11): 1144–46. https://doi.org/10.1038/nmeth.3103.

Amos, Gregory C. A., Chiara Borsetto, Paris Laskaris, Martin Krsek, Andrew E. Berry, Kevin K. Newsham, Leo Calvo-Bado, David A. Pearce, Carlos Vallin, and Elizabeth M. H. Wellington. 2015. '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 (9). https://doi.org/10.1371/journal.pone.0138327.

Andam, Cheryl P., James R. Doroghazi, Ashley N. Campbell, Peter J. Kelly, Mallory J. Choudoir, and Daniel H. Buckley. n.d. 'A Latitudinal Diversity Gradient in Terrestrial Bacteria of the Genus Streptomyces'. *MBio* 7 (2): e02200-15. https://doi.org/10.1128/mBio.02200-15.

Aoi, Yoshiteru, Tomoyuki Kinoshita, Toru Hata, Hiroaki Ohta, Haruko Obokata, and Satoshi Tsuneda. 2009. 'Hollow-Fiber Membrane Chamber as a Device for In Situ Environmental Cultivation'. *Appl. Environ. Microbiol.* 75 (11): 3826–33. https://doi.org/10.1128/AEM.02542-08.

Arakawa, Kenji., Susumu. Mochizuki, Kohei. Yamada, Takenori. Noma, and Haruyasu.YR Kinashi. 2007. 'γ-Butyrolactone Autoregulator-Receptor System Involved in Lankacidin and Lankamycin Production and Morphological Differentiation in Streptomyces Rochei'. *Microbiology* 153 (6): 1817–27. https://doi.org/10.1099/mic.0.2006/002170-0.

Årdal, Christine, Manica Balasegaram, Ramanan Laxminarayan, David McAdams, Kevin Outterson, John H. Rex, and Nithima Sumpradit. 2020. 'Antibiotic Development — Economic, Regulatory and Societal Challenges'. *Nature Reviews Microbiology* 18 (5): 267–74. https://doi.org/10.1038/s41579-019-0293-3.

Aroonsri, Aiyada, Shigeru Kitani, Sun-Uk Choi, and Takuya Nihira. 2008. 'Isolation and Characterization of BamA Genes, Homologues of the γ-Butyrolactone Autoregulator-Receptor Gene in Amycolatopsis Mediterranei, a Rifamycin Producer'. *Biotechnology Letters* 30 (11): 2019–24. https://doi.org/10.1007/s10529-008-9794-2.

Arul Jose, Polpass, and Solomon Robinson David Jebakumar. 2013. 'Non-Streptomycete Actinomycetes Nourish the Current Microbial Antibiotic Drug Discovery'. *Frontiers in Microbiology* 4. https://doi.org/10.3389/fmicb.2013.00240.

Atanasov, Atanas G., Sergey B. Zotchev, Verena M. Dirsch, and Claudiu T. Supuran. 2021. 'Natural Products in Drug Discovery: Advances and Opportunities'. *Nature Reviews Drug Discovery* 20 (3): 200–216. https://doi.org/10.1038/s41573-020-00114-z.

'Babraham Bioinformatics - FastQC A Quality Control Tool for High Throughput Sequence Data'. n.d. Accessed 31 March 2022. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Bachmann, Brian O., and Jacques Ravel. 2009. 'Chapter 8 Methods for In Silico Prediction of Microbial Polyketide and Nonribosomal Peptide Biosynthetic Pathways from DNA Sequence Data'. In *Methods in Enzymology*, 458:181–217. Complex Enzymes in Microbial Natural Product Biosynthesis, Part A: Overview Articles and Peptides. Academic Press. https://doi.org/10.1016/S0076-6879(09)04808-3.

Bachmann, Brian O, Steven G Van Lanen, and Richard H Baltz. 2014. 'Microbial Genome Mining for Accelerated Natural Products Discovery: Is a Renaissance in the Making?' *Journal of Industrial Microbiology and Biotechnology* 41 (2): 175–84. https://doi.org/10.1007/s10295-013-1389-9.

Bale, Nicole J., W. Irene C. Rijpstra, Diana X. Sahonero-Canavesi, Igor Y. Oshkin, Svetlana E. Belova, Svetlana N. Dedysh, and Jaap S. Sinninghe Damsté. 2019. 'Fatty Acid and

Hopanoid Adaption to Cold in the Methanotroph Methylovulum Psychrotolerans'. *Frontiers in Microbiology* 10 (April). https://doi.org/10.3389/fmicb.2019.00589.

Baltz, Richard H. 2006. 'Marcel Faber Roundtable: Is Our Antibiotic Pipeline Unproductive Because of Starvation, Constipation or Lack of Inspiration?' *Journal of Industrial Microbiology and Biotechnology* 33 (7): 507–13. https://doi.org/10.1007/s10295-005-0077-9.

Becher, Paul G., Vasiliki Verschut, Maureen J. Bibb, Matthew J. Bush, Béla P. Molnár, Elisabeth Barane, Mahmoud M. Al-Bassam, et al. 2020. 'Developmentally Regulated Volatiles Geosmin and 2-Methylisoborneol Attract a Soil Arthropod to Streptomyces Bacteria Promoting Spore Dispersal'. *Nature Microbiology* 5 (6): 821–29. https://doi.org/10.1038/s41564-020-0697-x.

Bednarz, Bartosz, Magdalena Kotowska, and Krzysztof J. Pawlik. 2019. 'Multi-Level Regulation of Coelimycin Synthesis in Streptomyces Coelicolor A3(2)'. *Applied Microbiology and Biotechnology* 103 (16): 6423–34. https://doi.org/10.1007/s00253-019-09975-w.

Belin, Brittany J., Nicolas Busset, Eric Giraud, Antonio Molinaro, Alba Silipo, and Dianne K. Newman. 2018. 'Hopanoid Lipids: From Membranes to Plant–Bacteria Interactions'. *Nature Reviews. Microbiology* 16 (5): 304–15. https://doi.org/10.1038/nrmicro.2017.173.

Bellenger, J. P., T. Wichard, A. B. Kustka, and A. M. L. Kraepiel. 2008. 'Uptake of Molybdenum and Vanadium by a Nitrogen-Fixing Soil Bacterium Using Siderophores'. *Nature Geoscience* 1 (4): 243–46. https://doi.org/10.1038/ngeo161.

Ben-Dov, Eitan, Esti Kramarsky-Winter, and Ariel Kushmaro. 2009. 'An in Situ Method for Cultivating Microorganisms Using a Double Encapsulation Technique'. *FEMS Microbiology Ecology* 68 (3): 363–71. https://doi.org/10.1111/j.1574-6941.2009.00682.x.

Benjamini, Yuval, and Terence P. Speed. 2012. 'Summarizing and Correcting the GC Content Bias in High-Throughput Sequencing'. *Nucleic Acids Research* 40 (10): e72. https://doi.org/10.1093/nar/gks001.

Bentley, S. D., K. F. Chater, A.-M. Cerdeño-Tárraga, G. L. Challis, N. R. Thomson, K. D. James, D. E. Harris, et al. 2002. 'Complete Genome Sequence of the Model Actinomycete Streptomyces Coelicolor A3(2)'. *Nature* 417 (6885): 141–47. https://doi.org/10.1038/417141a.

Biolabs, New England. 2015. 'Making Your Own Electrocompetent Cells'. Protocols.Io. 5 February 2015. https://www.protocols.io/view/Making-your-own-electrocompetent-cells-imsv6m.

Blin, Kai, Hyun Uk Kim, Marnix H Medema, and Tilmann Weber. 2019. 'Recent Development of AntiSMASH and Other Computational Approaches to Mine Secondary Metabolite Biosynthetic Gene Clusters'. *Briefings in Bioinformatics* 20 (4): 1103–13. https://doi.org/10.1093/bib/bbx146.

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

Bode, Helge Björn, Barbara Bethe, Regina Höfs, and Axel Zeeck. 2002. 'Big Effects from Small Changes: Possible Ways to Explore Nature's Chemical Diversity'. *ChemBioChem* 3 (7): 619–27. https://doi.org/10.1002/1439-7633(20020703)3:7<619::AID-CBIC619>3.0.CO;2-9.

Boden, Rich, Donovan P. Kelly, J. Colin Murrell, and Hendrik Schäfer. 2010. 'Oxidation of Dimethylsulfide to Tetrathionate by Methylophaga Thiooxidans Sp. Nov.: A New Link

in the Sulfur Cycle'. *Environmental Microbiology* 12 (10): 2688–99. https://doi.org/10.1111/j.1462-2920.2010.02238.x.

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

Borsetto, Chiara. 2017. 'Study and Exploitation of Diverse Soil Environments for Novel Natural Product Discovery Using Metagenomic Approaches'. Phd, University of Warwick. http://webcat.warwick.ac.uk/record=b3140926~S15.

Borsetto, Chiara, Gregory C. A. Amos, Ulisses Nunes da Rocha, Alex L. Mitchell, Robert D. Finn, Rabah Forar Laidi, Carlos Vallin, David A. Pearce, Kevin K. Newsham, and Elizabeth M. H. Wellington. 2019. 'Microbial Community Drivers of PK/NRP Gene Diversity in Selected Global Soils'. *Microbiome* 7 (1): 78. https://doi.org/10.1186/s40168-019-0692-8.

Bottos, Eric M., Anthony C. Woo, Peyman Zawar-Reza, Stephen B. Pointing, and Stephen C. Cary. 2014. 'Airborne Bacterial Populations above Desert Soils of the McMurdo Dry Valleys, Antarctica'. *Microbial Ecology* 67 (1): 120–28. https://doi.org/10.1007/s00248-013-0296-y.

Brady, Sean F. 2007. 'Construction of Soil Environmental DNA Cosmid Libraries and Screening for Clones That Produce Biologically Active Small Molecules'. *Nature Protocols* 2 (5): 1297–1305. https://doi.org/10.1038/nprot.2007.195.

Brotherton, Carolyn A., Marnix H. Medema, and E. Peter Greenberg. 2018. 'LuxR Homolog-Linked Biosynthetic Gene Clusters in Proteobacteria'. *MSystems* 3 (3): e00208-17. https://doi.org/10.1128/mSystems.00208-17.

Buchfink, Benjamin, Chao Xie, and Daniel H. Huson. 2015. 'Fast and Sensitive Protein Alignment Using DIAMOND'. *Nature Methods* 12 (1): 59–60. https://doi.org/10.1038/nmeth.3176.

Bunet, Robert, Lijiang Song, Marta Vaz Mendes, Christophe Corre, Laurence Hotel, Nicolas Rouhier, Xavier Framboisier, Pierre Leblond, Gregory L. Challis, and Bertrand Aigle. 2011. 'Characterization and Manipulation of the Pathway-Specific Late Regulator AlpW Reveals Streptomyces Ambofaciens as a New Producer of Kinamycins'. *Journal of Bacteriology* 193 (5): 1142–53. https://doi.org/10.1128/JB.01269-10.

Callahan, Benjamin J., Paul J. McMurdie, Michael J. Rosen, Andrew W. Han, Amy Jo A. Johnson, and Susan P. Holmes. 2016. 'DADA2: High-Resolution Sample Inference from Illumina Amplicon Data'. *Nature Methods* 13 (7): 581–83. https://doi.org/10.1038/nmeth.3869.

Carroll, Amanda C., and Alex Wong. 2018. 'Plasmid Persistence: Costs, Benefits, and the Plasmid Paradox'. *Canadian Journal of Microbiology*, March. https://doi.org/10.1139/cjm-2017-0609.

Carver, Tim, Simon R. Harris, Thomas D. Otto, Matthew Berriman, Julian Parkhill, and Jacqueline A. McQuillan. 2013. 'BamView: Visualizing and Interpretation of next-Generation Sequencing Read Alignments'. *Briefings in Bioinformatics* 14 (2): 203–12. https://doi.org/10.1093/bib/bbr073.

Charlop-Powers, Zachary, Jeremy G. Owen, Boojala Vijay B. Reddy, Melinda A. Ternei, and Sean F. Brady. 2014. 'Chemical-Biogeographic Survey of Secondary Metabolism in Soil'. *Proceedings of the National Academy of Sciences* 111 (10): 3757–62. https://doi.org/10.1073/pnas.1318021111.

Charlop-Powers, Zachary, Jeremy G Owen, Boojala Vijay B Reddy, Melinda A Ternei, Denise O Guimarães, Ulysses A de Frias, Monica T Pupo, Prudy Seepe, Zhiyang Feng, and Sean F Brady. 2015. 'Global Biogeographic Sampling of Bacterial Secondary

Metabolism'. Edited by Jon Clardy. *ELife* 4 (January): e05048. https://doi.org/10.7554/eLife.05048.

Chase, Alexander B., Douglas Sweeney, Mitchell N. Muskat, Dulce G. Guillén-Matus, and Paul R. Jensen. 2021. 'Vertical Inheritance Facilitates Interspecies Diversification in Biosynthetic Gene Clusters and Specialized Metabolites'. *MBio*, November. https://doi.org/10.1128/mBio.02700-21.

Chase, Alexander B., Douglas Sweeney, Mitchell N. Muskat, Dulce Guillén-Matus, and Paul R. Jensen. 2020. 'Complex Evolutionary Dynamics Govern the Diversity and Distribution of Biosynthetic Gene Clusters and Their Encoded Specialized Metabolites'. bioRxiv. https://doi.org/10.1101/2020.12.19.423547.

Chaumeil, Pierre-Alain, Aaron J. Mussig, Philip Hugenholtz, and Donovan H. Parks. n.d. 'GTDB-Tk: A Toolkit to Classify Genomes with the Genome Taxonomy Database'. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btz848.

Chen, Ray, Hon Lun Wong, Gareth S. Kindler, Fraser Iain MacLeod, Nicole Benaud, Belinda C. Ferrari, and Brendan P. Burns. 2020. 'Discovery of an Abundance of Biosynthetic Gene Clusters in Shark Bay Microbial Mats'. *Frontiers in Microbiology* 11 (August). https://doi.org/10.3389/fmicb.2020.01950.

Choi, Sun-Uk, Chang-Kwon Lee, Yong-Il Hwang, Hiroshi Kinoshita, and Takuya Nihira. 2004. 'Cloning and Functional Analysis by Gene Disruption of a Gene Encoding a γ-Butyrolactone Autoregulator Receptor from Kitasatospora Setae'. *Journal of Bacteriology*, June. https://doi.org/10.1128/JB.186.11.3423-3430.2004.

Choi, Sun-Uk, Chang-Kwon Lee, Yong-Il Hwang, Hiroshi Kinosita, and Takuya Nihira. 2003. 'γ-Butyrolactone Autoregulators and Receptor Proteins in Non-Streptomyces Actinomycetes Producing Commercially Important Secondary Metabolites'. *Archives of Microbiology* 180 (4): 303–7. https://doi.org/10.1007/s00203-003-0591-y.

Chong, C. W., P. Convey, D. A. Pearce, and I. K. P. Tan. 2012. 'Assessment of Soil Bacterial Communities on Alexander Island (in the Maritime and Continental Antarctic Transitional Zone)'. *Polar Biology* 35 (3): 387–99. https://doi.org/10.1007/s00300-011-1084-0.

Choulet, Frédéric, Bertrand Aigle, Alexandre Gallois, Sophie Mangenot, Claude Gerbaud, Chantal Truong, François-Xavier Francou, et al. 2006. 'Evolution of the Terminal Regions of the Streptomyces Linear Chromosome'. *Molecular Biology and Evolution* 23 (12): 2361–69. https://doi.org/10.1093/molbev/msl108.

Chung, C T, S L Niemela, and R H Miller. 1989. 'One-Step Preparation of Competent Escherichia Coli: Transformation and Storage of Bacterial Cells in the Same Solution.' *Proceedings of the National Academy of Sciences of the United States of America* 86 (7): 2172–75.

Cimermancic, Peter, Marnix H. Medema, Jan Claesen, Kenji Kurita, Laura C. Wieland Brown, Konstantinos Mavrommatis, Amrita Pati, et al. 2014. 'Insights into Secondary Metabolism from a Global Analysis of Prokaryotic Biosynthetic Gene Clusters'. *Cell* 158 (2): 412–21. https://doi.org/10.1016/j.cell.2014.06.034.

Coelho, Luis Pedro, Renato Alves, Álvaro Rodríguez del Río, Pernille Neve Myers, Carlos P. Cantalapiedra, Joaquín Giner-Lamia, Thomas Sebastian Schmidt, et al. 2022. 'Towards the Biogeography of Prokaryotic Genes'. *Nature* 601 (7892): 252–56. https://doi.org/10.1038/s41586-021-04233-4.

Combes, Patricia, Rob Till, Sally Bee, and Margaret C. M. Smith. 2002. 'The Streptomyces Genome Contains Multiple Pseudo-AttB Sites for the (Phi)C31-Encoded Site-Specific Recombination System'. *Journal of Bacteriology* 184 (20): 5746–52. https://doi.org/10.1128/JB.184.20.5746-5752.2002.

Compeau, Phillip E. C., Pavel A. Pevzner, and Glenn Tesler. 2011. 'How to Apply de Bruijn Graphs to Genome Assembly'. *Nature Biotechnology* 29 (11): 987–91. https://doi.org/10.1038/nbt.2023.

Convey, Peter, and Ronald I. Lewis Smith. 1997. 'The Terrestrial Arthropod Fauna and Its Habitats in Northern Marguerite Bay and Alexander Island, Maritime Antarctic'. *Antarctic Science* 9 (1): 12–26. https://doi.org/10.1017/S0954102097000035.

Cotruvo, Joseph A., Emily R. Featherston, Joseph A. Mattocks, Jackson V. Ho, and Tatiana N. Laremore. 2018. 'Lanmodulin: A Highly Selective Lanthanide-Binding Protein from a Lanthanide-Utilizing Bacterium'. *Journal of the American Chemical Society* 140 (44): 15056–61. https://doi.org/10.1021/jacs.8b09842.

Crits-Christoph, Alexander, Spencer Diamond, Cristina N. Butterfield, Brian C. Thomas, and Jillian F. Banfield. 2018. 'Novel Soil Bacteria Possess Diverse Genes for Secondary Metabolite Biosynthesis'. *Nature* 558 (7710): 440–44. https://doi.org/10.1038/s41586-018-0207-y.

Cross, Karissa L., James H. Campbell, Manasi Balachandran, Alisha G. Campbell, Connor J. Cooper, Ann Griffen, Matthew Heaton, et al. 2019. 'Targeted Isolation and Cultivation of Uncultivated Bacteria by Reverse Genomics'. *Nature Biotechnology* 37 (11): 1314–21. https://doi.org/10.1038/s41587-019-0260-6.

Cruz-Morales, Pablo, Hilda E. Ramos-Aboites, Cuauhtémoc Licona-Cassani, Nelly Selem-Mójica, Paulina M. Mejía-Ponce, Valeria Souza-Saldívar, and Francisco Barona-Gómez. 2017. 'Actinobacteria Phylogenomics, Selective Isolation from an Iron Oligotrophic Environment and Siderophore Functional Characterization, Unveil New Desferrioxamine Traits'. *FEMS Microbiology Ecology* 93 (9): fix086. https://doi.org/10.1093/femsec/fix086.

Cuadrat, Rafael R. C., Danny Ionescu, Alberto M. R. Dávila, and Hans-Peter Grossart. 2018. 'Recovering Genomics Clusters of Secondary Metabolites from Lakes Using Genome-Resolved Metagenomics'. *Frontiers in Microbiology* 9 (February). https://doi.org/10.3389/fmicb.2018.00251.

Culp, Elizabeth J., Grace Yim, Nicholas Waglechner, Wenliang Wang, Andrew C. Pawlowski, and Gerard D. Wright. 2019. 'Hidden Antibiotics in Actinomycetes Can Be Identified by Inactivation of Gene Clusters for Common Antibiotics'. *Nature Biotechnology* 37 (10): 1149–54. https://doi.org/10.1038/s41587-019-0241-9.

D'Agostino, Paul M., and Tobias A. M. Gulder. 2018. 'Direct Pathway Cloning Combined with Sequence- and Ligation-Independent Cloning for Fast Biosynthetic Gene Cluster Refactoring and Heterologous Expression'. *ACS Synthetic Biology* 7 (7): 1702–8. https://doi.org/10.1021/acssynbio.8b00151.

Daniel, Rolf. 2004. 'The Soil Metagenome – a Rich Resource for the Discovery of Novel Natural Products'. *Current Opinion in Biotechnology* 15 (3): 199–204. https://doi.org/10.1016/j.copbio.2004.04.005.

Das, Debanu, Nick V. Grishin, Abhinav Kumar, Dennis Carlton, Constantina Bakolitsa, Mitchell D. Miller, Polat Abdubek, et al. 2010. 'The Structure of the First Representative of Pfam Family PF09836 Reveals a Two-Domain Organization and Suggests Involvement in Transcriptional Regulation'. *Acta Crystallographica Section F: Structural Biology and Crystallization Communications* 66 (Pt 10): 1174. https://doi.org/10.1107/S1744309109022672.

Dassama, Laura M. K., Grace E. Kenney, and Amy C. Rosenzweig. 2017. 'Methanobactins: From Genome to Function'. *Metallomics : Integrated Biometal Science* 9 (1): 7–20. https://doi.org/10.1039/c6mt00208k.

De Tomaso, Anthony W., and Irving L. Weissman. 2003. 'Construction and Characterization of Large-Insert Genomic Libraries (BAC and Fosmid) from the Ascidian Botryllus

Schlosseri and Initial Physical Mapping of a Histocompatibility Locus'. *Marine Biotechnology* 5 (2): 103–15. https://doi.org/10.1007/s10126-002-0071-4.
DeConto, Robert M., and David Pollard. 2003. 'Rapid Cenozoic Glaciation of Antarctica Induced by Declining Atmospheric CO2'. *Nature* 421 (6920): 245–49. https://doi.org/10.1038/nature01290.
Demain, Arnold L. 2002. 'Prescription for an Ailing Pharmaceutical Industry'. *Nature Biotechnology* 20 (4): 331–331. https://doi.org/10.1038/nbt0402-331.
Denoncin, Katleen, and Jean-François Collet. 2013. 'Disulfide Bond Formation in the Bacterial Periplasm: Major Achievements and Challenges Ahead'. *Antioxidants & Redox Signaling* 19 (1): 63–71. https://doi.org/10.1089/ars.2012.4864.
Desai, Narayan, Dion Antonopoulos, Jack A Gilbert, Elizabeth M Glass, and Folker Meyer. 2012. 'From Genomics to Metagenomics'. *Current Opinion in Biotechnology*, Analytical biotechnology, 23 (1): 72–76. https://doi.org/10.1016/j.copbio.2011.12.017.
Dieser, Markus, Mark Greenwood, and Christine M. Foreman. 2010. 'Carotenoid Pigmentation in Antarctic Heterotrophic Bacteria as a Strategy to Withstand Environmental Stresses'. *Arctic, Antarctic, and Alpine Research* 42 (4): 396–405. https://doi.org/10.1657/1938-4246-42.4.396.
Doud, Devin F. R., Robert M. Bowers, Frederik Schulz, Markus De Raad, Kai Deng, Angela Tarver, Evan Glasgow, et al. 2020. 'Function-Driven Single-Cell Genomics Uncovers Cellulose-Degrading Bacteria from the Rare Biosphere'. *The ISME Journal* 14 (3): 659–75. https://doi.org/10.1038/s41396-019-0557-y.
Edwards, Collin R., Tullis C. Onstott, Jennifer M. Miller, Jessica B. Wiggins, Wei Wang, Charles K. Lee, S. Craig Cary, Stephen B. Pointing, and Maggie C. Y. Lau. 2017. 'Draft Genome Sequence of Uncultured Upland Soil Cluster Gammaproteobacteria Gives Molecular Insights into High-Affinity Methanotrophy'. *Genome Announcements* 5 (17). https://doi.org/10.1128/genomeA.00047-17.
Eldjárn, Grímur Hjörleifsson, Andrew Ramsay, Justin J. J. van der Hooft, Katherine R. Duncan, Sylvia Soldatou, Juho Rousu, Rónán Daly, Joe Wandy, and Simon Rogers. 2021. 'Ranking Microbial Metabolomic and Genomic Links in the NPLinker Framework Using Complementary Scoring Functions'. *PLOS Computational Biology* 17 (5): e1008920. https://doi.org/10.1371/journal.pcbi.1008920.
Finland, M., P. F. Frank, and Clare Wilcox. 1950. 'In Vitro Susceptibility of Pathogenic Staphylococci to Seven Antibiotics. (Penicillin, Streptomycin, Bacitracin, Polymyxin, Aerosporin, Aureomyein and Chloromycetin). With a Note on the Changing Resistance of Staphylococci to Penicillin.' *American Journal of Clinical Pathology* 20 (4): 325–34.
Finn, Robert D., Jody Clements, and Sean R. Eddy. 2011. 'HMMER Web Server: Interactive Sequence Similarity Searching'. *Nucleic Acids Research* 39 (suppl_2): W29–37. https://doi.org/10.1093/nar/gkr367.
Fleming, Alexander. 1929. 'On the Antibacterial Action of Cultures of a Penicillium, with Special Reference to Their Use in the Isolation of B. Influenzæ'. *British Journal of Experimental Pathology* 10 (3): 226–36.
Frank, Jeremy A., Claudia I. Reich, Shobha Sharma, Jon S. Weisbaum, Brenda A. Wilson, and Gary J. Olsen. 2008. 'Critical Evaluation of Two Primers Commonly Used for Amplification of Bacterial 16S RRNA Genes'. *Applied and Environmental Microbiology* 74 (8): 2461–70. https://doi.org/10.1128/AEM.02272-07.
Fraser, Ceridwen I., Adele K. Morrison, Andrew McC Hogg, Erasmo C. Macaya, Erik van Sebille, Peter G. Ryan, Amanda Padovan, Cameron Jack, Nelson Valdivia, and Jonathan M. Waters. 2018. 'Antarctica's Ecological Isolation Will Be Broken by

Storm-Driven Dispersal and Warming'. *Nature Climate Change* 8 (8): 704–8. https://doi.org/10.1038/s41558-018-0209-7.

Freeman, Michael F., Cristian Gurgui, Maximilian J. Helf, Brandon I. Morinaka, Agustinus R. Uria, Neil J. Oldham, Hans-Georg Sahl, Shigeki Matsunaga, and Jörn Piel. 2012. 'Metagenome Mining Reveals Polytheonamides as Posttranslationally Modified Ribosomal Peptides'. *Science*, October. https://doi.org/10.1126/science.1226121.

Gavriilidou, Athina, Satria A. Kautsar, Nestor Zaburannyi, Daniel Krug, Rolf Müller, Marnix H. Medema, and Nadine Ziemert. 2021. 'A Global Survey of Specialized Metabolic Diversity Encoded in Bacterial Genomes'. https://doi.org/10.1101/2021.08.11.455920.

Gavrish, Ekaterina, Annette Bollmann, Slava Epstein, and Kim Lewis. 2008. 'A Trap for in Situ Cultivation of Filamentous Actinobacteria'. *Journal of Microbiological Methods* 72 (3): 257–62. https://doi.org/10.1016/j.mimet.2007.12.009.

'GCF_002021875.1 - 1 Record(s) - 51 Region(s)'. n.d. Accessed 4 March 2022. https://antismash.secondarymetabolites.org/upload/GCF_002021875.1/index.html.

George, Isabelle F., Manuela Hartmann, Mark R. Liles, and Spiros N. Agathos. 2011. 'Recovery of As-Yet-Uncultured Soil Acidobacteria on Dilute Solid Media'. *Applied and Environmental Microbiology*, November.

Gillespie, Doreen E., Sean F. Brady, Alan D. Bettermann, Nicholas P. Cianciotto, Mark R. Liles, Michelle R. Rondon, Jon Clardy, Robert M. Goodman, and Jo Handelsman. 2002. 'Isolation of Antibiotics Turbomycin a and B from a Metagenomic Library of Soil Microbial DNA'. *Applied and Environmental Microbiology* 68 (9): 4301–6. https://doi.org/10.1128/AEM.68.9.4301-4306.2002.

Giudice, Angelina Lo, Vivia Bruni, and Luigi Michaud. 2007. 'Characterization of Antarctic Psychrotrophic Bacteria with Antibacterial Activities against Terrestrial Microorganisms'. *Journal of Basic Microbiology* 47 (6): 496–505. https://doi.org/10.1002/jobm.200700227.

Goering, Anthony W., Ryan A. McClure, James R. Doroghazi, Jessica C. Albright, Nicole A. Haverland, Yongbo Zhang, Kou-San Ju, Regan J. Thomson, William W. Metcalf, and Neil L. Kelleher. 2016. 'Metabologenomics: Correlation of Microbial Gene Clusters with Metabolites Drives Discovery of a Nonribosomal Peptide with an Unusual Amino Acid Monomer'. *ACS Central Science* 2 (2): 99–108. https://doi.org/10.1021/acscentsci.5b00331.

Gomez-Escribano, Juan Pablo, and Mervyn J. Bibb. 2011. 'Engineering Streptomyces Coelicolor for Heterologous Expression of Secondary Metabolite Gene Clusters'. *Microbial Biotechnology* 4 (2): 207–15. https://doi.org/10.1111/j.1751-7915.2010.00219.x.

Gomez-Escribano, Juan Pablo, Jean Franco Castro, Valeria Razmilic, Scott A. Jarmusch, Gerhard Saalbach, Rainer Ebel, Marcel Jaspars, Barbara Andrews, Juan A. Asenjo, and Mervyn J. Bibb. 2019. 'Heterologous Expression of a Cryptic Gene Cluster from Streptomyces Leeuwenhoekii C34T Yields a Novel Lasso Peptide, Leepeptin'. *Applied and Environmental Microbiology*, September. https://doi.org/10.1128/AEM.01752-19.

Greunke, Christian, Elke Regina Duell, Paul Michael D'Agostino, Anna Glöckle, Katharina Lamm, and Tobias Alexander Marius Gulder. 2018. 'Direct Pathway Cloning (DiPaC) to Unlock Natural Product Biosynthetic Potential'. *Metabolic Engineering* 47 (May): 334–45. https://doi.org/10.1016/j.ymben.2018.03.010.

Gunjal, Vidya B., Ritesh Thakare, Sidharth Chopra, and D. Srinivasa Reddy. 2020. 'Teixobactin: A Paving Stone toward a New Class of Antibiotics?' *Journal of Medicinal Chemistry* 63 (21): 12171–95. https://doi.org/10.1021/acs.jmedchem.0c00173.

Haft, Daniel H., and Malay Kumar Basu. 2011. 'Biological Systems Discovery In Silico: Radical S-Adenosylmethionine Protein Families and Their Target Peptides for Posttranslational Modification ▿ '. *Journal of Bacteriology* 193 (11): 2745–55. https://doi.org/10.1128/JB.00040-11.

Hall, James P. J., Rosanna C. T. Wright, Ellie Harrison, Katie J. Muddiman, A. Jamie Wood, Steve Paterson, and Michael A. Brockhurst. 2021. 'Plasmid Fitness Costs Are Caused by Specific Genetic Conflicts Enabling Resolution by Compensatory Mutation'. *PLOS Biology* 19 (10): e3001225. https://doi.org/10.1371/journal.pbio.3001225.

Hannigan, Geoffrey D, David Prihoda, Andrej Palicka, Jindrich Soukup, Ondrej Klempir, Lena Rampula, Jindrich Durcak, et al. 2019. 'A Deep Learning Genome-Mining Strategy for Biosynthetic Gene Cluster Prediction'. *Nucleic Acids Research* 47 (18): e110. https://doi.org/10.1093/nar/gkz654.

Hardoim, Cristiane C. P., Massimiliano Cardinale, Ana C. B. Cúcio, Ana I. S. Esteves, Gabriele Berg, Joana R. Xavier, Cymon J. Cox, and Rodrigo Costa. 2014. 'Effects of Sample Handling and Cultivation Bias on the Specificity of Bacterial Communities in Keratose Marine Sponges'. *Frontiers in Microbiology* 5 (November): 611. https://doi.org/10.3389/fmicb.2014.00611.

Hemmerich, Johannes, Peter Rohe, Britta Kleine, Sarah Jurischka, Wolfgang Wiechert, Roland Freudl, and Marco Oldiges. 2016. 'Use of a Sec Signal Peptide Library from Bacillus Subtilis for the Optimization of Cutinase Secretion in Corynebacterium Glutamicum'. *Microbial Cell Factories* 15 (1): 208. https://doi.org/10.1186/s12934-016-0604-6.

Herranz, Carmen, and Arnold J. M. Driessen. 2005. 'Sec-Mediated Secretion of Bacteriocin Enterocin P by Lactococcus Lactis'. *Applied and Environmental Microbiology* 71 (4): 1959–63. https://doi.org/10.1128/AEM.71.4.1959-1963.2005.

Hesse, Elze, Siobhán O'Brien, Nicolas Tromas, Florian Bayer, Adela M. Luján, Eleanor M. van Veen, Dave J. Hodgson, and Angus Buckling. 2018. 'Ecological Selection of Siderophore-Producing Microbial Taxa in Response to Heavy Metal Contamination'. *Ecology Letters* 21 (1): 117–27. https://doi.org/10.1111/ele.12878.

Hoff, Grégory, Claire Bertrand, Emilie Piotrowski, Annabelle Thibessard, and Pierre Leblond. 2018. 'Genome Plasticity Is Governed by Double Strand Break DNA Repair in Streptomyces'. *Scientific Reports* 8 (1): 5272. https://doi.org/10.1038/s41598-018-23622-w.

Horinouchi, S., H. Suzuki, M. Nishiyama, and T. Beppu. 1989. 'Nucleotide Sequence and Transcriptional Analysis of the Streptomyces Griseus Gene (AfsA) Responsible for A-Factor Biosynthesis'. *Journal of Bacteriology* 171 (2): 1206–10. https://doi.org/10.1128/jb.171.2.1206-1210.1989.

Hoskisson, Paul A., and Ryan F. Seipke. 2020. 'Cryptic or Silent? The Known Unknowns, Unknown Knowns, and Unknown Unknowns of Secondary Metabolism'. *MBio* 11 (5): e02642-20. https://doi.org/10.1128/mBio.02642-20.

Hover, Bradley M., Seong-Hwan Kim, Micah Katz, Zachary Charlop-Powers, Jeremy G. Owen, Melinda A. Ternei, Jeffrey Maniko, et al. 2018. 'Culture-Independent Discovery of the Malacidins as Calcium-Dependent Antibiotics with Activity against Multidrug-Resistant Gram-Positive Pathogens'. *Nature Microbiology* 3 (4): 415–22. https://doi.org/10.1038/s41564-018-0110-1.

Howe, Adina Chuang, Janet K. Jansson, Stephanie A. Malfatti, Susannah G. Tringe, James M. Tiedje, and C. Titus Brown. 2014. 'Tackling Soil Diversity with the Assembly of Large, Complex Metagenomes'. *Proceedings of the National Academy of Sciences of the United States of America* 111 (13): 4904–9. https://doi.org/10.1073/pnas.1402564111.

Huang, Jing, Zheng Yu, and Ludmila Chistoserdova. 2018. 'Lanthanide-Dependent Methanol Dehydrogenases of XoxF4 and XoxF5 Clades Are Differentially Distributed Among

Methylotrophic Bacteria and They Reveal Different Biochemical Properties'. *Frontiers in Microbiology* 9. https://www.frontiersin.org/article/10.3389/fmicb.2018.01366.

Hussein, Khalid A., and Jin Ho Joo. 2014. 'Potential of Siderophore Production by Bacteria Isolated from Heavy Metal: Polluted and Rhizosphere Soils'. *Current Microbiology* 68 (6): 717–23. https://doi.org/10.1007/s00284-014-0530-y.

Hyatt, Doug, Gwo-Liang Chen, Philip F. LoCascio, Miriam L. Land, Frank W. Larimer, and Loren J. Hauser. 2010. 'Prodigal: Prokaryotic Gene Recognition and Translation Initiation Site Identification'. *BMC Bioinformatics* 11 (1): 119. https://doi.org/10.1186/1471-2105-11-119.

Ivanova, Anastasia A., Alena D. Zhelezova, Timofey I. Chernov, and Svetlana N. Dedysh. 2020. 'Linking Ecology and Systematics of Acidobacteria: Distinct Habitat Preferences of the Acidobacteriia and Blastocatellia in Tundra Soils'. *PLoS ONE* 15 (3). https://doi.org/10.1371/journal.pone.0230157.

Jain, Chirag, Luis M. Rodriguez-R, Adam M. Phillippy, Konstantinos T. Konstantinidis, and Srinivas Aluru. 2018. 'High Throughput ANI Analysis of 90K Prokaryotic Genomes Reveals Clear Species Boundaries'. *Nature Communications* 9 (1): 5114. https://doi.org/10.1038/s41467-018-07641-9.

Janssen, Peter H., Penelope S. Yates, Bronwyn E. Grinton, Paul M. Taylor, and Michelle Sait. 2002. 'Improved Culturability of Soil Bacteria and Isolation in Pure Culture of Novel Members of the Divisions Acidobacteria, Actinobacteria, Proteobacteria, and Verrucomicrobia'. *Applied and Environmental Microbiology* 68 (5): 2391–96. https://doi.org/10.1128/AEM.68.5.2391-2396.2002.

Jensen, Paul R., Philip G. Williams, Dong-Chan Oh, Lisa Zeigler, and William Fenical. 2007. 'Species-Specific Secondary Metabolite Production in Marine Actinomycetes of the Genus Salinispora'. *Applied and Environmental Microbiology* 73 (4): 1146–52. https://doi.org/10.1128/AEM.01891-06.

Jeong, Jae-Yeon, Hyung-Soon Yim, Ji-Young Ryu, Hyun Sook Lee, Jung-Hyun Lee, Dong-Seung Seen, and Sung Gyun Kang. 2012. 'One-Step Sequence- and Ligation-Independent Cloning as a Rapid and Versatile Cloning Method for Functional Genomics Studies'. *Applied and Environmental Microbiology* 78 (15): 5440–43. https://doi.org/10.1128/AEM.00844-12.

Ji, Chang-Hun, Hiyoung Kim, and Hahk-Soo Kang. 2019. 'Synthetic Inducible Regulatory Systems Optimized for the Modulation of Secondary Metabolite Production in Streptomyces'. *ACS Synthetic Biology* 8 (3): 577–86. https://doi.org/10.1021/acssynbio.9b00001.

Ji, Xiangyang, Wan-Qiu Liu, and Jian Li. 2022. 'Recent Advances in Applying Cell-Free Systems for High-Value and Complex Natural Product Biosynthesis'. *Current Opinion in Microbiology* 67 (June): 102142. https://doi.org/10.1016/j.mib.2022.102142.

Jiang, Ming, and Blaine A. Pfeifer. 2013. 'Metabolic and Pathway Engineering to Influence Native and Altered Erythromycin Production through E. Coli'. *Metabolic Engineering* 19 (September): 42–49. https://doi.org/10.1016/j.ymben.2013.05.005.

Jong, Anne de, Sacha A. F. T. van Hijum, Jetta J. E. Bijlsma, Jan Kok, and Oscar P. Kuipers. 2006. 'BAGEL: A Web-Based Bacteriocin Genome Mining Tool'. *Nucleic Acids Research* 34 (Web Server issue): W273–79. https://doi.org/10.1093/nar/gkl237.

Ju, Kou-San, Xiafei Zhang, and Marie A. Elliot. 2017. 'New Kid on the Block: LmbU Expands the Repertoire of Specialized Metabolic Regulators in Streptomyces'. *Journal of Bacteriology*, October. https://doi.org/10.1128/JB.00559-17.

Jung, Dawoon, Yoshiteru Aoi, and Slava S. Epstein. 2016. 'In Situ Cultivation Allows for Recovery of Bacterial Types Competitive in Their Natural Environment'. *Microbes and Environments* 31 (4): 456–59. https://doi.org/10.1264/jsme2.ME16079.

Jwanoswki, Kathleen, Christina Wells, Terri Bruce, Jennifer Rutt, Tabitha Banks, and Tamara L. McNealy. 2017. 'The Legionella Pneumophila GIG Operon Responds to Gold and Copper in Planktonic and Biofilm Cultures'. *PLOS ONE* 12 (5): e0174245. https://doi.org/10.1371/journal.pone.0174245.

Kang, Dongwan D., Feng Li, Edward Kirton, Ashleigh Thomas, Rob Egan, Hong An, and Zhong Wang. 2019. 'MetaBAT 2: An Adaptive Binning Algorithm for Robust and Efficient Genome Reconstruction from Metagenome Assemblies'. *PeerJ* 7 (July): e7359. https://doi.org/10.7717/peerj.7359.

Kang, Hahk-Soo, and Sean F. Brady. 2013. 'Arimetamycin A: Improving Clinically Relevant Families of Natural Products through Sequence-Guided Screening of Soil Metagenomes'. *Angewandte Chemie International Edition* 52 (42): 11063–67. https://doi.org/10.1002/anie.201305109.

Kang, Hahk-Soo, Zachary Charlop-Powers, and Sean F. Brady. 2016. 'Multiplexed CRISPR/Cas9- and TAR-Mediated Promoter Engineering of Natural Product Biosynthetic Gene Clusters in Yeast'. *ACS Synthetic Biology* 5 (9): 1002–10. https://doi.org/10.1021/acssynbio.6b00080.

Kato, Souichiro, Motoko Takashino, Kensuke Igarashi, and Wataru Kitagawa. 2020. 'Isolation and Genomic Characterization of a Proteobacterial Methanotroph Requiring Lanthanides'. *Microbes and Environments* 35 (1): ME19128. https://doi.org/10.1264/jsme2.ME19128.

Kato, Souichiro, Ayasa Yamagishi, Serina Daimon, Kosei Kawasaki, Hideyuki Tamaki, Wataru Kitagawa, Ayumi Abe, et al. 2018. 'Isolation of Previously Uncultured Slow-Growing Bacteria by Using a Simple Modification in the Preparation of Agar Media'. *Applied and Environmental Microbiology*, July. https://doi.org/10.1128/AEM.00807-18.

Katz, Micah, Bradley M. Hover, and Sean F. Brady. 2016. 'Culture-Independent Discovery of Natural Products from Soil Metagenomes'. *Journal of Industrial Microbiology & Biotechnology* 43 (2): 129–41. https://doi.org/10.1007/s10295-015-1706-6.

Kautsar, Satria A., Kai Blin, Simon Shaw, Jorge C. Navarro-Muñoz, Barbara R. Terlouw, Justin J. J. van der Hooft, Jeffrey A. van Santen, et al. 2020. 'MIBiG 2.0: A Repository for Biosynthetic Gene Clusters of Known Function'. *Nucleic Acids Research* 48 (D1): D454–58. https://doi.org/10.1093/nar/gkz882.

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

Kautsar, Satria A., Justin J. J. van der Hooft, Dick de Ridder, and Marnix H. Medema. 2020. 'BiG-SLiCE: A Highly Scalable Tool Maps the Diversity of 1.2 Million Biosynthetic Gene Clusters'. *BioRxiv*, August, 2020.08.17.240838. https://doi.org/10.1101/2020.08.17.240838.

Ke, Jing, and Yasuo Yoshikuni. 2020. 'Multi-Chassis Engineering for Heterologous Production of Microbial Natural Products'. *Current Opinion in Biotechnology*, Energy Biotechnologyy ● Environmental Biotechnology, 62 (April): 88–97. https://doi.org/10.1016/j.copbio.2019.09.005.

Keiler, Kenneth C. 2015. 'Mechanisms of Ribosome Rescue in Bacteria'. *Nature Reviews Microbiology* 13 (5): 285–97. https://doi.org/10.1038/nrmicro3438.

Kersten, Roland D., Yu-Liang Yang, Yuquan Xu, Peter Cimermancic, Sang-Jip Nam, William Fenical, Michael A. Fischbach, Bradley S. Moore, and Pieter C. Dorrestein. 2011. 'A Mass Spectrometry-Guided Genome Mining Approach for Natural Product Peptidogenomics'. *Nature Chemical Biology* 7 (11): 794–802. https://doi.org/10.1038/nchembio.684.

Kielak, Anna M., Cristine C. Barreto, George A. Kowalchuk, Johannes A. van Veen, and Eiko E. Kuramae. 2016. 'The Ecology of Acidobacteria: Moving beyond Genes and Genomes'. *Frontiers in Microbiology* 7.

Kieser, T., M. J. Bibb, K. F. Chater, M. J. Butter, D. A. Hopwood, K. F. Chater, M. J. Bittner, et al. 2000. 'Practical Streptomyces Genetics: A Laboratory Manual', January. https://www.scienceopen.com/document?vid=558cee38-7e4d-4feb-b037-0cc912882ba5.

Kim, Jeffrey H., Zhiyang Feng, John D. Bauer, Dimitris Kallifidas, Paula Y. Calle, and Sean F. Brady. 2010. 'Cloning Large Natural Product Gene Clusters from the Environment: Piecing Environmental DNA Gene Clusters Back Together with TAR'. *Biopolymers* 93 (9): 833–44. https://doi.org/10.1002/bip.21450.

Kim, Seong-Hwan, Wanli Lu, Mahmoud Kamal Ahmadi, Daniel Montiel, Melinda A. Ternei, and Sean F. Brady. 2019. 'Atolypenes, Tricyclic Bacterial Sesterterpenes Discovered Using a Multiplexed In Vitro Cas9-TAR Gene Cluster Refactoring Approach'. *ACS Synthetic Biology* 8 (1): 109–18. https://doi.org/10.1021/acssynbio.8b00361.

Kitani, Shigeru, Masashi Doi, Tomohito Shimizu, Asa Maeda, and Takuya Nihira. 2010. 'Control of Secondary Metabolism by FarX, Which Is Involved in the γ-Butyrolactone Biosynthesis of Streptomyces Lavendulae FRI-5'. *Archives of Microbiology* 192 (3): 211–20. https://doi.org/10.1007/s00203-010-0550-3.

Kleinteich, Julia, Falk Hildebrand, Mohammad Bahram, Anita Y. Voigt, Susanna A. Wood, Anne D. Jungblut, Frithjof C. Küpper, et al. 2017. 'Pole-to-Pole Connections: Similarities between Arctic and Antarctic Microbiomes and Their Vulnerability to Environmental Change'. *Frontiers in Ecology and Evolution* 5. https://doi.org/10.3389/fevo.2017.00137.

Klindworth, Anna, Elmar Pruesse, Timmy Schweer, Jörg Peplies, Christian Quast, Matthias Horn, and Frank Oliver Glöckner. 2013. 'Evaluation of General 16S Ribosomal RNA Gene PCR Primers for Classical and Next-Generation Sequencing-Based Diversity Studies'. *Nucleic Acids Research* 41 (1): e1. https://doi.org/10.1093/nar/gks808.

Klotz, Martin G., Daniel J. Arp, Patrick S. G. Chain, Amal F. El-Sheikh, Loren J. Hauser, Norman G. Hommes, Frank W. Larimer, et al. 2006. 'Complete Genome Sequence of the Marine, Chemolithoautotrophic, Ammonia-Oxidizing Bacterium Nitrosococcus Oceani ATCC 19707'. *Applied and Environmental Microbiology* 72 (9): 6299–6315. https://doi.org/10.1128/AEM.00463-06.

Kogawa, Masato, Rimi Miyaoka, Franziska Hemmerling, Masahiro Ando, Kei Yura, Keigo Ide, Yohei Nishikawa, et al. 2022. 'Single-Cell Metabolite Detection and Genomics Reveals Uncultivated Talented Producer'. *PNAS Nexus* 1 (1): pgab007. https://doi.org/10.1093/pnasnexus/pgab007.

Kolmogorov, Mikhail, Mikhail Rayko, Jeffrey Yuan, Evgeny Polevikov, and Pavel Pevzner. 2019. 'MetaFlye: Scalable Long-Read Metagenome Assembly Using Repeat Graphs'. *BioRxiv*, May, 637637. https://doi.org/10.1101/637637.

Könneke, Martin, Anne E. Bernhard, José R. de la Torre, Christopher B. Walker, John B. Waterbury, and David A. Stahl. 2005. 'Isolation of an Autotrophic Ammonia-Oxidizing Marine Archaeon'. *Nature* 437 (7058): 543–46. https://doi.org/10.1038/nature03911.

Koomsiri, Wilaiwan, Yuki Inahashi, Kantinan Leetanasaksakul, Kazuro Shiomi, Yōko Takahashi, Satoshi Ōmura, Markiyan Samborskyy, et al. 2019. 'Sarpeptins A and B, Lipopeptides Produced by Streptomyces Sp. KO-7888 Overexpressing a Specific SARP Regulator'. *Journal of Natural Products* 82 (8): 2144–51. https://doi.org/10.1021/acs.jnatprod.9b00074.

Kouprina, Natalay, and Vladimir Larionov. 2016. 'Transformation-Associated Recombination (TAR) Cloning for Genomics Studies and Synthetic Biology'. *Chromosoma* 125 (4): 621–32. https://doi.org/10.1007/s00412-016-0588-3.

Kroken, Scott, N. Louise Glass, John W. Taylor, O. C. Yoder, and B. Gillian Turgeon. 2003. 'Phylogenomic Analysis of Type I Polyketide Synthase Genes in Pathogenic and Saprobic Ascomycetes'. *Proceedings of the National Academy of Sciences of the United States of America* 100 (26): 15670–75. https://doi.org/10.1073/pnas.2532165100.

Kuhn, Emanuele, Andrew S. Ichimura, Vivian Peng, Christian H. Fritsen, Gareth Trubl, Peter T. Doran, and Alison E. Murray. 2014. 'Brine Assemblages of Ultrasmall Microbial Cells within the Ice Cover of Lake Vida, Antarctica'. *Applied and Environmental Microbiology* 80 (12): 3687–98. https://doi.org/10.1128/AEM.00276-14.

Kuipers, Anneke, Rick Rink, and Gert N. Moll. 2009. 'Translocation of a Thioether-Bridged Azurin Peptide Fragment via the Sec Pathway in Lactococcus Lactis'. *Applied and Environmental Microbiology* 75 (11): 3800–3802. https://doi.org/10.1128/AEM.00341-09.

Kuipers, Anneke, Jenny Wierenga, Rick Rink, Leon D. Kluskens, Arnold J. M. Driessen, Oscar P. Kuipers, and Gert N. Moll. 2006. 'Sec-Mediated Transport of Posttranslationally Dehydrated Peptides in Lactococcus Lactis'. *Applied and Environmental Microbiology* 72 (12): 7626–33. https://doi.org/10.1128/AEM.01802-06.

Lal, Devi, Mansi Verma, Susanta K. Behura, and Rup Lal. 2016. 'Codon Usage Bias in Phylum Actinobacteria: Relevance to Environmental Adaptation and Host Pathogenicity'. *Research in Microbiology* 167 (8): 669–77. https://doi.org/10.1016/j.resmic.2016.06.003.

Langmead, Ben, and Steven L Salzberg. 2012. 'Fast Gapped-Read Alignment with Bowtie 2'. *Nature Methods* 9 (4): 357–59. https://doi.org/10.1038/nmeth.1923.

Lanz, Nicholas D., Anthony J. Blaszczyk, Erin L. McCarthy, Bo Wang, Roy X. Wang, Brianne S. Jones, and Squire J. Booker. 2018. 'Enhanced Solubilization of Class B Radical S-Adenosylmethionine Methylases by Improved Cobalamin Uptake in Escherichia Coli'. *Biochemistry* 57 (9): 1475–90. https://doi.org/10.1021/acs.biochem.7b01205.

Latorre-Pérez, Adriel, Pascual Villalba-Bermell, Javier Pascual, Manuel Porcar, and Cristina Vilanova. 2019. 'Assembly Methods for Nanopore-Based Metagenomic Sequencing: A Comparative Study'. *BioRxiv*, August, 722405. https://doi.org/10.1101/722405.

Lau, M C Y, B T Stackhouse, A C Layton, A Chauhan, T A Vishnivetskaya, K Chourey, J Ronholm, et al. 2015. 'An Active Atmospheric Methane Sink in High Arctic Mineral Cryosols'. *The ISME Journal* 9 (8): 1880–91. https://doi.org/10.1038/ismej.2015.13.

Laureti, Luisa, Lijiang Song, Sheng Huang, Christophe Corre, Pierre Leblond, Gregory L. Challis, and Bertrand Aigle. 2011. 'Identification of a Bioactive 51-Membered Macrolide Complex by Activation of a Silent Polyketide Synthase in Streptomyces Ambofaciens'. *Proceedings of the National Academy of Sciences of the United States of America* 108 (15): 6258–63. https://doi.org/10.1073/pnas.1019077108.

Lazzarini, Ameriga, Linda Cavaletti, Giorgio Toppo, and Flavia Marinelli. 2000. 'Rare Genera of Actinomycetes as Potential Producers of New Antibiotics'. *Antonie van Leeuwenhoek* 78 (3): 399–405. https://doi.org/10.1023/A:1010287600557.

Lee, Yong Jik, Shigeru Kitani, and Takuya Nihira. 2010. 'Null Mutation Analysis of an AfsA-Family Gene, BarX, That Is Involved in Biosynthesis of the {gamma}-Butyrolactone Autoregulator in Streptomyces Virginiae'. *Microbiology (Reading, England)* 156 (Pt 1): 206–10. https://doi.org/10.1099/mic.0.032003-0.

Lemetre, Christophe, Jeffrey Maniko, Zachary Charlop-Powers, Ben Sparrow, Andrew J. Lowe, and Sean F. Brady. 2017. 'Bacterial Natural Product Biosynthetic Domain

Composition in Soil Correlates with Changes in Latitude on a Continent-Wide Scale'. *Proceedings of the National Academy of Sciences* 114 (44): 11615–20. https://doi.org/10.1073/pnas.1710262114.

Letunic, Ivica, and Peer Bork. 2007. 'Interactive Tree Of Life (ITOL): An Online Tool for Phylogenetic Tree Display and Annotation'. *Bioinformatics* 23 (1): 127–28. https://doi.org/10.1093/bioinformatics/btl529.

Lewis, Kim. 2013. 'Platforms for Antibiotic Discovery'. *Nature Reviews Drug Discovery* 12 (5): 371–87. https://doi.org/10.1038/nrd3975.

Li, Heng. 2018. 'Minimap2: Pairwise Alignment for Nucleotide Sequences'. *Bioinformatics* 34 (18): 3094–3100. https://doi.org/10.1093/bioinformatics/bty191.

Li, Heng, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. 2009. 'The Sequence Alignment/Map Format and SAMtools'. *Bioinformatics* 25 (16): 2078–79. https://doi.org/10.1093/bioinformatics/btp352.

Li, Jesse W.-H., and John C. Vederas. 2009. 'Drug Discovery and Natural Products: End of an Era or an Endless Frontier?' *Science* 325 (5937): 161–65. https://doi.org/10.1126/science.1168243.

Li, Lei, Logan W Maclntyre, and Sean F Brady. 2021. 'Refactoring Biosynthetic Gene Clusters for Heterologous Production of Microbial Natural Products'. *Current Opinion in Biotechnology*, Chemical Biotechnology ● Pharmaceutical Biotechnology, 69 (June): 145–52. https://doi.org/10.1016/j.copbio.2020.12.011.

Li, Wenli, Yinggang Luo, Jianhua Ju, Scott R. Rajski, Hiroyuki Osada, and Ben Shen. 2009. 'Characterization of the Tautomycetin Biosynthetic Gene Cluster from Streptomyces Griseochromogenes Provides New Insight into Dialkylmaleic Anhydride Biosynthesis'. *Journal of Natural Products* 72 (3): 450–59. https://doi.org/10.1021/np8007478.

Libis, Vincent, Niv Antonovsky, Mengyin Zhang, Zhuo Shang, Daniel Montiel, Jeffrey Maniko, Melinda A. Ternei, et al. 2019. 'Uncovering the Biosynthetic Potential of Rare Metagenomic DNA Using Co-Occurrence Network Analysis of Targeted Sequences'. *Nature Communications* 10 (1): 3848. https://doi.org/10.1038/s41467-019-11658-z.

Liles, Mark R., Lynn L. Williamson, Jitsupang Rodbumrer, Vigdis Torsvik, Larissa C. Parsley, Robert M. Goodman, and Jo Handelsman. 2009. 'Isolation and Cloning of High-Molecular-Weight Metagenomic DNA from Soil Microorganisms'. *Cold Spring Harbor Protocols* 2009 (8): pdb.prot5271. https://doi.org/10.1101/pdb.prot5271.

Lin, Xin, Russell Hopson, and David E. Cane. 2006. 'Genome Mining in Streptomyces Coelicolor: Molecular Cloning and Characterization of a New Sesquiterpene Synthase'. *Journal of the American Chemical Society* 128 (18): 6022–23. https://doi.org/10.1021/ja061292s.

Lincke, Thorger, Swantje Behnken, Keishi Ishida, Martin Roth, and Christian Hertweck. 2010. 'Closthioamide: An Unprecedented Polythioamide Antibiotic from the Strictly Anaerobic Bacterium Clostridium Cellulolyticum'. *Angewandte Chemie (International Ed. in English)* 49 (11): 2011–13. https://doi.org/10.1002/anie.200906114.

Ling, Losee L., Tanja Schneider, Aaron J. Peoples, Amy L. Spoering, Ina Engels, Brian P. Conlon, Anna Mueller, et al. 2015. 'A New Antibiotic Kills Pathogens without Detectable Resistance'. *Nature* 517 (7535): 455–59. https://doi.org/10.1038/nature14098.

Liu, Xiangyang, Kangmin Hua, Dongxu Liu, Zhen-Long Wu, Ying Wang, Haoran Zhang, Zixin Deng, Blaine A. Pfeifer, and Ming Jiang. 2020. 'Heterologous Biosynthesis of Type II Polyketide Products Using E. Coli'. *ACS Chemical Biology* 15 (5): 1177–83. https://doi.org/10.1021/acschembio.9b00827.

Lladó, Salvador, Rubén López-Mondéjar, and Petr Baldrian. 2018. 'Drivers of Microbial Community Structure in Forest Soils'. *Applied Microbiology and Biotechnology* 102 (10): 4331–38. https://doi.org/10.1007/s00253-018-8950-4.

L. Robinson, Serina, Jörn Piel, and Shinichi Sunagawa. 2021. 'A Roadmap for Metagenomic Enzyme Discovery'. *Natural Product Reports* 38 (11): 1994–2023. https://doi.org/10.1039/D1NP00006C.

Luo, Chengwei, Despina Tsementzi, Nikos C. Kyrpides, and Konstantinos T. Konstantinidis. 2012. 'Individual Genome Assembly from Complex Community Short-Read Metagenomic Datasets'. *The ISME Journal* 6 (4): 898–901. https://doi.org/10.1038/ismej.2011.147.

Magesh, N. S., Anoop Tiwari, Sathish Mohan Botsa, and Tara da Lima Leitao. 2021. 'Hazardous Heavy Metals in the Pristine Lacustrine Systems of Antarctica: Insights from PMF Model and ERA Techniques'. *Journal of Hazardous Materials* 412 (June): 125263. https://doi.org/10.1016/j.jhazmat.2021.125263.

Mahler, Lisa, Sarah P Niehs, Karin Martin, Thomas Weber, Kirstin Scherlach, Christian Hertweck, Martin Roth, and Miriam A Rosenbaum. 2021. 'Highly Parallelized Droplet Cultivation and Prioritization of Antibiotic Producers from Natural Microbial Communities'. *ELife* 10: e64774. https://doi.org/10.7554/eLife.64774.

Mao, Dainan, Leah B. Bushin, Kyuho Moon, Yihan Wu, and Mohammad R. Seyedsayamdost. 2017. 'Discovery of ScmR as a Global Regulator of Secondary Metabolism and Virulence in Burkholderia Thailandensis E264'. *Proceedings of the National Academy of Sciences* 114 (14): E2920–28. https://doi.org/10.1073/pnas.1619529114.

McDonald, Bradon R., and Cameron R. Currie. 2017. 'Lateral Gene Transfer Dynamics in the Ancient Bacterial Genus Streptomyces'. *MBio* 8 (3): e00644-17. https://doi.org/10.1128/mBio.00644-17.

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

Meijenfeldt, F. A. Bastiaan von, Ksenia Arkhipova, Diego D. Cambuy, Felipe H. Coutinho, and Bas E. Dutilh. 2019. 'Robust Taxonomic Classification of Uncharted Microbial Sequences and Bins with CAT and BAT'. *Genome Biology* 20 (1): 217. https://doi.org/10.1186/s13059-019-1817-x.

Mejáre, Malin, and Leif Bülow. 2001. 'Metal-Binding Proteins and Peptides in Bioremediation and Phytoremediation of Heavy Metals'. *Trends in Biotechnology* 19 (2): 67–73. https://doi.org/10.1016/S0167-7799(00)01534-1.

Meleshko, Dmitry, Hosein Mohimani, Vittorio Tracanna, Iman Hajirasouliha, Marnix H. Medema, Anton Korobeynikov, and Pavel A. Pevzner. 2019. 'BiosyntheticSPAdes: Reconstructing Biosynthetic Gene Clusters from Assembly Graphs'. *Genome Research* 29 (8): 1352–62. https://doi.org/10.1101/gr.243477.118.

Merrick, Christine A., Jia Zhao, and Susan J. Rosser. 2018. 'Serine Integrases: Advancing Synthetic Biology'. *ACS Synthetic Biology* 7 (2): 299–310. https://doi.org/10.1021/acssynbio.7b00308.

Millán-Aguiñaga, Natalie, Sylvia Soldatou, Sarah Brozio, John T. Munnoch, John Howe, Paul A. Hoskisson, and Katherine R. Duncan. 2019. 'Awakening Ancient Polar Actinobacteria: Diversity, Evolution and Specialized Metabolite Potential'. *Microbiology,* 165 (11): 1169–80. https://doi.org/10.1099/mic.0.000845.

Milshteyn, Aleksandr, Jessica S. Schneider, and Sean F. Brady. 2014. 'Mining the Metabiome: Identifying Novel Natural Products from Microbial Communities'. *Chemistry & Biology* 21 (9): 1211–23. https://doi.org/10.1016/j.chembiol.2014.08.006.

Misiak, Marta, William P. Goodall-Copestake, Tim H. Sparks, M. Roger Worland, Lynne Boddy, Naresh Magan, Peter Convey, David W. Hopkins, and Kevin K. Newsham. 2021. 'Inhibitory Effects of Climate Change on the Growth and Extracellular Enzyme Activities of a Widespread Antarctic Soil Fungus'. *Global Change Biology* 27 (n/a): 1111–25. https://doi.org/10.1111/gcb.15456.

Mojib, Nazia, Rachel Philpott, Jonathan P. Huang, Michael Niederweis, and Asim K. Bej. 2010. 'Antimycobacterial Activity in Vitro of Pigments Isolated from Antarctic Bacteria'. *Antonie van Leeuwenhoek* 98 (4): 531–40. https://doi.org/10.1007/s10482-010-9470-0.

Molnár, István, D. Steven Hill, Ross Zirkle, Philip E. Hammer, Frank Gross, Thomas G. Buckel, Volker Jungmann, Johannes Paul Pachlatko, and James M. Ligon. 2005. 'Biocatalytic Conversion of Avermectin to 4"-Oxo-Avermectin: Heterologous Expression of the Ema1 Cytochrome P450 Monooxygenase'. *Applied and Environmental Microbiology* 71 (11): 6977–85. https://doi.org/10.1128/AEM.71.11.6977-6985.2005.

Moss, Eli L., Dylan G. Maghini, and Ami S. Bhatt. 2020. 'Complete, Closed Bacterial Genomes from Microbiomes Using Nanopore Sequencing'. *Nature Biotechnology* 38 (6): 701–7. https://doi.org/10.1038/s41587-020-0422-6.

Myronovskyi, Maksym, and Andriy Luzhetskyy. 2016. 'Native and Engineered Promoters in Natural Product Discovery'. *Natural Product Reports* 33 (8): 1006–19. https://doi.org/10.1039/C6NP00002A.

Nakai, Ryosuke, Eri Shibuya, Ana Justel, Eugenio Rico, Antonio Quesada, Fumihisa Kobayashi, Yasunobu Iwasaka, et al. 2013. 'Phylogeographic Analysis of Filterable Bacteria with Special Reference to Rhizobiales Strains That Occur in Cryospheric Habitats'. *Antarctic Science* 25 (2): 219–28. https://doi.org/10.1017/S0954102012000831.

Nakashima, Yu, Yoko Egami, Miki Kimura, Toshiyuki Wakimoto, and Ikuro Abe. 2016. 'Metagenomic Analysis of the Sponge Discodermia Reveals the Production of the Cyanobacterial Natural Product Kasumigamide by "Entotheonella"'. *PLOS ONE* 11 (10): e0164468. https://doi.org/10.1371/journal.pone.0164468.

'Nanoporetech/Medaka'. (2017) 2020. Python. Oxford Nanopore Technologies. https://github.com/nanoporetech/medaka.

Nasrin, Shamima, Suresh Ganji, Kavita S. Kakirde, Melissa R. Jacob, Mei Wang, Ranga Rao Ravu, Paul A. Cobine, et al. 2018. 'Chloramphenicol Derivatives with Antibacterial Activity Identified by Functional Metagenomics'. *Journal of Natural Products* 81 (6): 1321–32. https://doi.org/10.1021/acs.jnatprod.7b00903.

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

Naville, Magali, Adrien Ghuillot-Gaudeffroy, Antonin Marchais, and Daniel Gautheret. 2011. 'ARNold: A Web Tool for the Prediction of Rho-Independent Transcription Terminators'. *RNA Biology* 8 (1): 11–13. https://doi.org/10.4161/rna.8.1.13346.

Nayfach, Stephen, Simon Roux, Rekha Seshadri, Daniel Udwary, Neha Varghese, Frederik Schulz, Dongying Wu, et al. 2020. 'A Genomic Catalog of Earth's Microbiomes'. *Nature Biotechnology*, November, 1–11. https://doi.org/10.1038/s41587-020-0718-6.

Neil, JO. 2016. 'Report on Antimicrobial Resistance'.

Ochman, Howard, Susannah Elwyn, and Nancy A. Moran. 1999. 'Calibrating Bacterial Evolution'. *Proceedings of the National Academy of Sciences* 96 (22): 12638–43. https://doi.org/10.1073/pnas.96.22.12638.

Onaka, Hiroyasu, Yukiko Mori, Yasuhiro Igarashi, and Tamotsu Furumai. 2011. 'Mycolic Acid-Containing Bacteria Induce Natural-Product Biosynthesis in Streptomyces Species'. *Applied and Environmental Microbiology* 77 (2): 400–406. https://doi.org/10.1128/AEM.01337-10.

O'Rourke, Sean, Andreas Wietzorrek, Kay Fowler, Christophe Corre, Greg L. Challis, and Keith F. Chater. 2009. 'Extracellular Signalling, Translational Control, Two Repressors and an Activator All Contribute to the Regulation of Methylenomycin Production in Streptomyces Coelicolor'. *Molecular Microbiology* 71 (3): 763–78. https://doi.org/10.1111/j.1365-2958.2008.06560.x.

Ortiz, Maximiliano, Pok Man Leung, Guy Shelley, Thanavit Jirapanjawat, Philipp A. Nauer, Marc W. Van Goethem, Sean K. Bay, et al. 2021. 'Multiple Energy Sources and Metabolic Strategies Sustain Microbial Diversity in Antarctic Desert Soils'. *Proceedings of the National Academy of Sciences* 118 (45): e2025322118. https://doi.org/10.1073/pnas.2025322118.

Osmond, C. B., C. H. Foyer, G. Bock, Richard J. Cogdell, Tina D. Howard, Robert Bittl, Erberhard Schlodder, Irene Geisenheimer, and Wolfgang Lubitz. 2000. 'How Carotenoids Protect Bacterial Photosynthesis'. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* 355 (1402): 1345–49. https://doi.org/10.1098/rstb.2000.0696.

Owen, Jeremy G., Zachary Charlop-Powers, Alexandra G. Smith, Melinda A. Ternei, Paula Y. Calle, Boojala Vijay B. Reddy, Daniel Montiel, and Sean F. Brady. 2015. 'Multiplexed Metagenome Mining Using Short DNA Sequence Tags Facilitates Targeted Discovery of Epoxyketone Proteasome Inhibitors'. *Proceedings of the National Academy of Sciences* 112 (14): 4221–26. https://doi.org/10.1073/pnas.1501124112.

Owen, Jeremy G., Boojala Vijay B. Reddy, Melinda A. Ternei, Zachary Charlop-Powers, Paula Y. Calle, Jeffrey H. Kim, and Sean F. Brady. 2013. 'Mapping Gene Clusters within Arrayed Metagenomic Libraries to Expand the Structural Diversity of Biomedically Relevant Natural Products'. *Proceedings of the National Academy of Sciences* 110 (29): 11797–802. https://doi.org/10.1073/pnas.1222159110.

Parks, Donovan H., Maria Chuvochina, David W. Waite, Christian Rinke, Adam Skarshewski, Pierre-Alain Chaumeil, and Philip Hugenholtz. 2018. 'A Proposal for a Standardized Bacterial Taxonomy Based on Genome Phylogeny'. Preprint. Microbiology. https://doi.org/10.1101/256800.

Parks, Donovan H., Michael Imelfort, Connor T. Skennerton, Philip Hugenholtz, and Gene W. Tyson. 2015. 'CheckM: Assessing the Quality of Microbial Genomes Recovered from Isolates, Single Cells, and Metagenomes'. *Genome Research* 25 (7): 1043–55. https://doi.org/10.1101/gr.186072.114.

Payne, Alexander, Nadine Holmes, Thomas Clarke, Rory Munro, Bisrat J. Debebe, and Matthew Loose. 2021. 'Readfish Enables Targeted Nanopore Sequencing of Gigabase-Sized Genomes'. *Nature Biotechnology* 39 (4): 442–50. https://doi.org/10.1038/s41587-020-00746-x.

Pearce, David Anthony, Kevin Newsham, Michael Thorne, Leo Calvo-Bado, Martin Krsek, Paris Laskaris, Andy Hodson, and Elizabeth M. H. Wellington. 2012. 'Metagenomic Analysis of a Southern Maritime Antarctic Soil'. *Frontiers in Microbiology* 3. https://doi.org/10.3389/fmicb.2012.00403.

Poon, Vincent. 2015. 'Analysis and Exploitation of AHFCA-Dependent Signalling Systems in Streptomyces Bacteria'. Phd, University of Warwick. http://webcat.warwick.ac.uk/record=b2870443~S1.

Potapov, Vladimir, and Jennifer L. Ong. 2017. 'Examining Sources of Error in PCR by Single-Molecule Sequencing'. *PLoS ONE* 12 (1): e0169774. https://doi.org/10.1371/journal.pone.0169774.

Prakash, Om, Mrinalini Parmar, Manali Vaijanapurkar, Vinay Rale, and Yogesh S Shouche. 2021. 'Recent Trend, Biases and Limitations of Cultivation-Based Diversity Studies of Microbes'. *FEMS Microbiology Letters* 368 (17): fnab118. https://doi.org/10.1093/femsle/fnab118.

Pudasaini, Sarita, John Wilson, Mukan Ji, Josie van Dorst, Ian Snape, Anne S. Palmer, Brendan P. Burns, and Belinda C. Ferrari. 2017. 'Microbial Diversity of Browning Peninsula, Eastern Antarctica Revealed Using Molecular and Cultivation Methods'. *Frontiers in Microbiology* 8.

Pulschen, Andre A., Amanda G. Bendia, Ashwana D. Fricker, Vivian H. Pellizari, Douglas Galante, and Fabio Rodrigues. 2017. 'Isolation of Uncultured Bacteria from Antarctica Using Long Incubation Periods and Low Nutritional Media'. *Frontiers in Microbiology* 8.

Qian, Zhengyi, Torsten Bruhn, Paul M. D'Agostino, Alexander Herrmann, Martin Haslbeck, Noémi Antal, Hans-Peter Fiedler, Ruth Brack-Werner, and Tobias A. M. Gulder. 2020. 'Discovery of the Streptoketides by Direct Cloning and Rapid Heterologous Expression of a Cryptic PKS II Gene Cluster from Streptomyces Sp. Tü 6314'. *The Journal of Organic Chemistry* 85 (2): 664–73. https://doi.org/10.1021/acs.joc.9b02741.

Quast, Christian, Elmar Pruesse, Pelin Yilmaz, Jan Gerken, Timmy Schweer, Pablo Yarza, Jörg Peplies, and Frank Oliver Glöckner. 2013. 'The SILVA Ribosomal RNA Gene Database Project: Improved Data Processing and Web-Based Tools'. *Nucleic Acids Research* 41 (D1): D590–96. https://doi.org/10.1093/nar/gks1219.

Quince, Christopher, Sergey Nurk, Sebastien Raguideau, Robert James, Orkun S. Soyer, J. Kimberly Summers, Antoine Limasset, A. Murat Eren, Rayan Chikhi, and Aaron E. Darling. 2021. 'STRONG: Metagenomics Strain Resolution on Assembly Graphs'. *Genome Biology* 22 (1): 214. https://doi.org/10.1186/s13059-021-02419-7.

Quinlan, Aaron R., and Ira M. Hall. 2010. 'BEDTools: A Flexible Suite of Utilities for Comparing Genomic Features'. *Bioinformatics* 26 (6): 841–42. https://doi.org/10.1093/bioinformatics/btq033.

Ramawatar, and Benjamin Schwessinger. 2018. 'DNA Size Selection (>3-4kb) and Purification of DNA Using an Improved Homemade SPRI Beads Solution.' Protocols.Io. 9 April 2018. https://www.protocols.io/view/dna-size-selection-3-4kb-and-purification-of-dna-u-n7hdhj6.

Rateb, Mostafa E., Wael E. Houssen, William T. A. Harrison, Hai Deng, Chinyere K. Okoro, Juan A. Asenjo, Barbara A. Andrews, et al. 2011. 'Diverse Metabolic Profiles of a Streptomyces Strain Isolated from a Hyper-Arid Environment'. *Journal of Natural Products* 74 (9): 1965–71. https://doi.org/10.1021/np200470u.

R. Duell, Elke, Tobias M. Milzarek, Mustafa El Omari, Luis J. Linares-Otoya, Till F. Schäberle, Gabrielle M. König, and Tobias A. M. Gulder. 2020. 'Identification, Cloning, Expression and Functional Interrogation of the Biosynthetic Pathway of the Polychlorinated Triphenyls Ambigol A–C from Fischerella Ambigua 108b'. *Organic Chemistry Frontiers* 7 (20): 3193–3201. https://doi.org/10.1039/D0QO00707B.

Rigali, Sébastien, Sinaeda Anderssen, Aymeric Naômé, and Gilles P. van Wezel. 2018. 'Cracking the Regulatory Code of Biosynthetic Gene Clusters as a Strategy for Natural Product Discovery'. *Biochemical Pharmacology*, Diamond Jubilee Special Issue:

Celebrating 60 Years of Excellence, 153 (July): 24–34. https://doi.org/10.1016/j.bcp.2018.01.007.

Romaniuk, Krzysztof, Anna Ciok, Przemyslaw Decewicz, Witold Uhrynowski, Karol Budzik, Marta Nieckarz, Julia Pawlowska, Marek K. Zdanowski, Dariusz Bartosik, and Lukasz Dziewit. 2018. 'Insight into Heavy Metal Resistome of Soil Psychrotolerant Bacteria Originating from King George Island (Antarctica)'. *Polar Biology* 41 (7): 1319–33. https://doi.org/10.1007/s00300-018-2287-4.

Romero, Diego, Matthew F. Traxler, Daniel López, and Roberto Kolter. 2011. 'Antibiotics as Signal Molecules'. *Chemical Reviews* 111 (9): 5492–5505. https://doi.org/10.1021/cr2000509.

Rondon, M. R., P. R. August, A. D. Bettermann, S. F. Brady, T. H. Grossman, M. R. Liles, K. A. Loiacono, et al. 2000. 'Cloning the Soil Metagenome: A Strategy for Accessing the Genetic and Functional Diversity of Uncultured Microorganisms'. *Applied and Environmental Microbiology* 66 (6): 2541–47. https://doi.org/10.1128/AEM.66.6.2541-2547.2000.

Röttig, Marc, Marnix H. Medema, Kai Blin, Tilmann Weber, Christian Rausch, and Oliver Kohlbacher. 2011. 'NRPSpredictor2—a Web Server for Predicting NRPS Adenylation Domain Specificity'. *Nucleic Acids Research* 39 (Web Server issue): W362–67. https://doi.org/10.1093/nar/gkr323.

Rubin, Benjamin E., Spencer Diamond, Brady F. Cress, Alexander Crits-Christoph, Yue Clare Lou, Adair L. Borges, Haridha Shivram, et al. 2022. 'Species- and Site-Specific Genome Editing in Complex Bacterial Communities'. *Nature Microbiology* 7 (1): 34–47. https://doi.org/10.1038/s41564-021-01014-7.

Rutledge, Peter J., and Gregory L. Challis. 2015. 'Discovery of Microbial Natural Products by Activation of Silent Biosynthetic Gene Clusters'. *Nature Reviews Microbiology* 13 (8): 509–23. https://doi.org/10.1038/nrmicro3496.

Sagova-Mareckova, Marketa, Ladislav Cermak, Jitka Novotna, Kamila Plhackova, Jana Forstova, and Jan Kopecky. 2008. 'Innovative Methods for Soil DNA Purification Tested in Soils with Widely Differing Characteristics'. *Applied and Environmental Microbiology*, May. https://doi.org/10.1128/AEM.02161-07.

Saleh, Orwah, Tobias Bonitz, Katrin Flinspach, Andreas Kulik, Nadja Burkard, Agnes Mühlenweg, Andreas Vente, et al. 2012. 'Activation of a Silent Phenazine Biosynthetic Gene Cluster Reveals a Novel Natural Product and a New Resistance Mechanism against Phenazines'. *MedChemComm* 3 (8): 1009–19. https://doi.org/10.1039/C2MD20045G.

Sarkisova, Svetlana A., Shalaka R. Lotlikar, Manita Guragain, Ryan Kubat, John Cloud, Michael J. Franklin, and Marianna A. Patrauchan. 2014. 'A Pseudomonas Aeruginosa EF-Hand Protein, EfhP (PA4107), Modulates Stress Responses and Virulence at High Calcium Concentration'. *PLOS ONE* 9 (6): e98985. https://doi.org/10.1371/journal.pone.0098985.

Schalk, Isabelle J., Mélissa Hannauer, and Armelle Braud. 2011. 'New Roles for Bacterial Siderophores in Metal Transport and Tolerance'. *Environmental Microbiology* 13 (11): 2844–54. https://doi.org/10.1111/j.1462-2920.2011.02556.x.

Schatz, Albert, Elizabeth Bugie, and Selman A. Waksman. 2005. 'Streptomycin, a Substance Exhibiting Antibiotic Activity against Gram-Positive and Gram-Negative Bacteria. 1944'. *Clinical Orthopaedics and Related Research*, no. 437 (August): 3–6. https://doi.org/10.1097/01.blo.0000175887.98112.fe.

Schirmer, Andreas, Rishali Gadkari, Christopher D. Reeves, Fadia Ibrahim, Edward F. DeLong, and C. Richard Hutchinson. 2005. 'Metagenomic Analysis Reveals Diverse Polyketide Synthase Gene Clusters in Microorganisms Associated with the Marine

Sponge Discodermia Dissoluta'. *Applied and Environmental Microbiology* 71 (8): 4840–49. https://doi.org/10.1128/AEM.71.8.4840-4849.2005.

Schlatter, Daniel C., Anita L. DavelosBaines, Kun Xiao, and Linda L. Kinkel. 2013. 'Resource Use of Soilborne Streptomyces Varies with Location, Phylogeny, and Nitrogen Amendment'. *Microbial Ecology* 66 (4): 961–71. https://doi.org/10.1007/s00248-013-0280-6.

Schmidt, T M, E F DeLong, and N R Pace. 1991. 'Analysis of a Marine Picoplankton Community by 16S RRNA Gene Cloning and Sequencing.' *Journal of Bacteriology* 173 (14): 4371–78.

Schrijver, Adinda De, and René DeYR 1999 Mot. n.d. 'A Subfamily of MalT-Related ATP-Dependent Regulators in the LuxR Family'. *Microbiology* 145 (6): 1287–88. https://doi.org/10.1099/13500872-145-6-1287.

Seemann, Torsten. (2013) 2022. 'Barrnap'. Perl. https://github.com/tseemann/barrnap.

Selbmann, Laura, Laura Zucconi, Serena Ruisi, Martin Grube, Massimiliano Cardinale, and Silvano Onofri. 2010. 'Culturable Bacteria Associated with Antarctic Lichens: Affiliation and Psychrotolerance'. *Polar Biology* 33 (1): 71–83. https://doi.org/10.1007/s00300-009-0686-2.

Sélem-Mojica, Nelly, César Aguilar, Karina Gutiérrez-García, Christian E. Martínez-Guerrero, and FanciscoYR 2019 Barona-Gómez. 2019. 'EvoMining Reveals the Origin and Fate of Natural Product Biosynthetic Enzymes'. *Microbial Genomics* 5 (12): e000260. https://doi.org/10.1099/mgen.0.000260.

Semrau, Jeremy D, Alan A DiSpirito, Parthiba Karthikeyan Obulisamy, and Christina S Kang-Yun. 2020. 'Methanobactin from Methanotrophs: Genetics, Structure, Function and Potential Applications'. *FEMS Microbiology Letters* 367 (5): fnaa045. https://doi.org/10.1093/femsle/fnaa045.

Seow, K. T., G. Meurer, M. Gerlitz, E. Wendt-Pienkowski, C. R. Hutchinson, and J. Davies. 1997. 'A Study of Iterative Type II Polyketide Synthases, Using Bacterial Genes Cloned from Soil DNA: A Means to Access and Use Genes from Uncultured Microorganisms'. *Journal of Bacteriology* 179 (23): 7360–68. https://doi.org/10.1128/jb.179.23.7360-7368.1997.

Sereika, Mantas, Rasmus Hansen Kirkegaard, Søren Michael Karst, Thomas Yssing Michaelsen, Emil Aarre Sørensen, Rasmus Dam Wollenberg, and Mads Albertsen. 2021. 'Oxford Nanopore R10.4 Long-Read Sequencing Enables near-Perfect Bacterial Genomes from Pure Cultures and Metagenomes without Short-Read or Reference Polishing'. bioRxiv. https://doi.org/10.1101/2021.10.27.466057.

Sevim, Volkan, Juna Lee, Robert Egan, Alicia Clum, Hope Hundley, Janey Lee, R. Craig Everroad, et al. 2019. 'Shotgun Metagenome Data of a Defined Mock Community Using Oxford Nanopore, PacBio and Illumina Technologies'. *Scientific Data* 6 (1): 285. https://doi.org/10.1038/s41597-019-0287-z.

Shao, Wen, Sonny Khin, and William C. Kopp. 2012. 'Characterization of Effect of Repeated Freeze and Thaw Cycles on Stability of Genomic DNA Using Pulsed Field Gel Electrophoresis'. *Biopreservation and Biobanking* 10 (1): 4–11. https://doi.org/10.1089/bio.2011.0016.

Shao, Zengyi, Guodong Rao, Chun Li, Zhanar Abil, Yunzi Luo, and Huimin Zhao. 2013. 'Refactoring the Silent Spectinabilin Gene Cluster Using a Plug-and-Play Scaffold'. *ACS Synthetic Biology* 2 (11): 662–69. https://doi.org/10.1021/sb400058n.

Sharrar, Allison M., Alexander Crits-Christoph, Raphaël Méheust, Spencer Diamond, Evan P. Starr, and Jillian F. Banfield. 2020. 'Bacterial Secondary Metabolite Biosynthetic Potential in Soil Varies with Phylum, Depth, and Vegetation Type'. *MBio* 11 (3). https://doi.org/10.1128/mBio.00416-20.

Shekh, Raeesh M., P. Singh, S. M. Singh, and Utpal Roy. 2011. 'Antifungal Activity of Arctic and Antarctic Bacteria Isolates'. *Polar Biology* 34 (1): 139–43. https://doi.org/10.1007/s00300-010-0854-4.

Sidda, John D., Lijiang Song, Vincent Poon, Mahmoud Al-Bassam, Orestis Lazos, Mark J. Buttner, Gregory L. Challis, and Christophe Corre. 2013. 'Discovery of a Family of γ-Aminobutyrate Ureas via Rational Derepression of a Silent Bacterial Gene Cluster'. *Chemical Science* 5 (1): 86–89. https://doi.org/10.1039/C3SC52536H.

Siebert, J., and P. Hirsch. 1988. 'Characterization of 15 Selected Coccal Bacteria Isolated from Antarctic Rock and Soil Samples from the McMurdo-Dry Valleys (South-Victoria Land)'. *Polar Biology* 9 (1): 37–44. https://doi.org/10.1007/BF00441762.

Sievers, Fabian, Andreas Wilm, David Dineen, Toby J Gibson, Kevin Karplus, Weizhong Li, Rodrigo Lopez, et al. 2011. 'Fast, Scalable Generation of High-Quality Protein Multiple Sequence Alignments Using Clustal Omega'. *Molecular Systems Biology* 7 (October): 539. https://doi.org/10.1038/msb.2011.75.

Silpe, Justin E., Joel W. H. Wong, Siân V. Owen, Michael Baym, and Emily P. Balskus. 2022. 'The Bacterial Toxin Colibactin Triggers Prophage Induction'. *Nature*, February, 1–6. https://doi.org/10.1038/s41586-022-04444-3.

Silva, Averlane Vieira da, Adeildo Junior de Oliveira, Ithallo Sathio Bessoni Tanabe, José Vieira Silva, Tiago Wallace da Silva Barros, Mayanne Karla da Silva, Paulo Henrique Barcellos França, et al. 2021. 'Antarctic Lichens as a Source of Phosphate-Solubilizing Bacteria'. *Extremophiles* 25 (2): 181–91. https://doi.org/10.1007/s00792-021-01220-5.

Silva, Tiago R., Alysson W. F. Duarte, Michel R. Z. Passarini, Ana Lucia T. G. Ruiz, Caio Haddad Franco, Carolina Borsoi Moraes, Itamar Soares de Melo, Rodney A. Rodrigues, Fabiana Fantinatti-Garboggini, and Valéria Maia Oliveira. 2018. 'Bacteria from Antarctic Environments: Diversity and Detection of Antimicrobial, Antiproliferative, and Antiparasitic Activities'. *Polar Biology* 41 (7): 1505–19. https://doi.org/10.1007/s00300-018-2300-y.

Singleton, Caitlin M., Francesca Petriglieri, Jannie M. Kristensen, Rasmus H. Kirkegaard, Thomas Y. Michaelsen, Martin H. Andersen, Zivile Kondrotaite, et al. 2020. 'Connecting Structure to Function with the Recovery of over 1000 High-Quality Activated Sludge Metagenome-Assembled Genomes Encoding Full-Length RRNA Genes Using Long-Read Sequencing'. *BioRxiv*, May, 2020.05.12.088096. https://doi.org/10.1101/2020.05.12.088096.

Skinnider, Michael A., Nishanth J. Merwin, Chad W. Johnston, and Nathan A. Magarvey. 2017. 'PRISM 3: Expanded Prediction of Natural Product Chemical Structures from Microbial Genomes'. *Nucleic Acids Research* 45 (W1): W49–54. https://doi.org/10.1093/nar/gkx320.

Smanski, Michael J., Ryan M. Peterson, Scott R. Rajski, and Ben Shen. 2009. 'Engineered Streptomyces Platensis Strains That Overproduce Antibiotics Platensimycin and Platencin'. *Antimicrobial Agents and Chemotherapy* 53 (4): 1299–1304. https://doi.org/10.1128/AAC.01358-08.

Smith, Jacques J., Lemese Ah Tow, William Stafford, Craig Cary, and Donald A. Cowan. 2006. 'Bacterial Diversity in Three Different Antarctic Cold Desert Mineral Soils'. *Microbial Ecology* 51 (4): 413–21. https://doi.org/10.1007/s00248-006-9022-3.

Staley, James T., and Allan Konopka. 1985. 'Measurement of in Situ Activities of Nonphotosynthetic Microorganisms in Aquatic and Terrestrial Habitats'. *Annual Review of Microbiology* 39 (1): 321–46. https://doi.org/10.1146/annurev.mi.39.100185.001541.

Stoilova-Disheva, Margarita, Evgenia Vasileva-Tonkova, and Iva Tomova. 2014. 'Characterization of Heavy Metals Resistant Heterotrophic Bacteria from Soils in the

Windmill Islands Region, Wilkes Land, East Antarctica'. *Polish Polar Research; 2014; No 4; 593-607*.

Sulheim, Snorre, Tjaša Kumelj, Dino van Dissel, Ali Salehzadeh-Yazdi, Chao Du, Gilles P. van Wezel, Kay Nieselt, Eivind Almaas, Alexander Wentzel, and Eduard J. Kerkhoven. 2020. 'Genome-Scale Model Constrained by Proteomics Reveals Metabolic Changes in Streptomyces Coelicolor M1152 Compared to M145'. bioRxiv. https://doi.org/10.1101/796722.

Sweerts, Jean-Pierre R. A., Dirk De Beer, Lars Peter Nielsen, Henk Verdouw, Johannes C. Van den Heuvel, Yehuda Cohen, and Thomas E. Cappenberg. 1990. 'Denitrification by Sulphur Oxidizing Beggiatoa Spp. Mats on Freshwater Sediments'. *Nature* 344 (6268): 762–63. https://doi.org/10.1038/344762a0.

Tamaki, Hideyuki, Satoshi Hanada, Yuji Sekiguchi, Yasuhiro Tanaka, and Yoichi Kamagata. 2009. 'Effect of Gelling Agent on Colony Formation in Solid Cultivation of Microbial Community in Lake Sediment'. *Environmental Microbiology* 11 (7): 1827–34. https://doi.org/10.1111/j.1462-2920.2009.01907.x.

Técher, Didier, Claudia Martinez-Chois, Marielle D'Innocenzo, Philippe Laval-Gilly, Amar Bennasroune, Laurent Foucaud, and Jairo Falla. 2010. 'Novel Perspectives to Purify Genomic DNA from High Humic Acid Content and Contaminated Soils'. *Separation and Purification Technology* 75 (1): 81–86. https://doi.org/10.1016/j.seppur.2010.07.014.

Tidjani, Abdoul-Razak, Jean-Noël Lorenzi, Maxime Toussaint, Erwin van Dijk, Delphine Naquin, Olivier Lespinet, Cyril Bontemps, and Pierre Leblond. 2019. 'Massive Gene Flux Drives Genome Diversity between Sympatric Streptomyces Conspecifics'. *MBio* 10 (5): e01533-19. https://doi.org/10.1128/mBio.01533-19.

Tietz, Jonathan I., Christopher J. Schwalen, Parth S. Patel, Tucker Maxson, Patricia M. Blair, Hua-Chia Tai, Uzma I. Zakai, and Douglas A. Mitchell. 2017. 'A New Genome-Mining Tool Redefines the Lasso Peptide Biosynthetic Landscape'. *Nature Chemical Biology* 13 (5): 470–78. https://doi.org/10.1038/nchembio.2319.

Ting, Chi P., Michael A. Funk, Steve L. Halaby, Zhengan Zhang, Tamir Gonen, and Wilfred A. van der Donk. 2019. 'Use of a Scaffold Peptide in the Biosynthesis of Amino Acid Derived Natural Products'. *Science (New York, N.Y.)* 365 (6450): 280–84. https://doi.org/10.1126/science.aau6232.

Tomm, Hailey A, Lorena Ucciferri, and Avena C Ross. 2019. 'Advances in Microbial Culturing Conditions to Activate Silent Biosynthetic Gene Clusters for Novel Metabolite Production'. *Journal of Industrial Microbiology and Biotechnology* 46 (9–10): 1381–1400. https://doi.org/10.1007/s10295-019-02198-y.

Tomova, Iva, Margarita Stoilova-Disheva, Irina Lazarkevich, and Evgenia Vasileva-Tonkova. 2015. 'Antimicrobial Activity and Resistance to Heavy Metals and Antibiotics of Heterotrophic Bacteria Isolated from Sediment and Soil Samples Collected from Two Antarctic Islands'. *Frontiers in Life Science* 8 (4): 348–57. https://doi.org/10.1080/21553769.2015.1044130.

Tong, Yaojun, Christopher M. Whitford, Helene L. Robertsen, Kai Blin, Tue S. Jørgensen, Andreas K. Klitgaard, Tetiana Gren, Xinglin Jiang, Tilmann Weber, and Sang Yup Lee. 2019. 'Highly Efficient DSB-Free Base Editing for Streptomycetes with CRISPR-BEST'. *Proceedings of the National Academy of Sciences* 116 (41): 20366–75. https://doi.org/10.1073/pnas.1913493116.

Trigodet, Florian, Karen Lolans, Emily Fogarty, Alon Shaiber, Hilary G. Morrison, Luis Barreiro, Bana Jabri, and A. Murat Eren. 2021. 'High Molecular Weight DNA Extraction Strategies for Long-Read Sequencing of Complex Metagenomes'. bioRxiv. https://doi.org/10.1101/2021.03.03.433801.

Trindade, Marla, Leonardo Joaquim van Zyl, José Navarro-Fernández, and Ahmed Abd Elrazak. 2015. 'Targeted Metagenomics as a Tool to Tap into Marine Natural Product Diversity for the Discovery and Production of Drug Candidates'. *Frontiers in Microbiology* 6. https://doi.org/10.3389/fmicb.2015.00890.

Turner, John, Hua Lu, Ian White, John C. King, Tony Phillips, J. Scott Hosking, Thomas J. Bracegirdle, Gareth J. Marshall, Robert Mulvaney, and Pranab Deb. 2016. 'Absence of 21st Century Warming on Antarctic Peninsula Consistent with Natural Variability'. *Nature* 535 (7612): 411–15. https://doi.org/10.1038/nature18645.

Tveit, Alexander T., Anne Grethe Hestnes, Serina L. Robinson, Arno Schintlmeister, Svetlana N. Dedysh, Nico Jehmlich, Martin von Bergen, et al. 2019. 'Widespread Soil Bacterium That Oxidizes Atmospheric Methane'. *Proceedings of the National Academy of Sciences* 116 (17): 8515–24. https://doi.org/10.1073/pnas.1817812116.

Tyc, Olaf, Chunxu Song, Jeroen S. Dickschat, Michiel Vos, and Paolina Garbeva. 2017. 'The Ecological Role of Volatile and Soluble Secondary Metabolites Produced by Soil Bacteria'. *Trends in Microbiology* 25 (4): 280–92. https://doi.org/10.1016/j.tim.2016.12.002.

Tyson, Gene W., Jarrod Chapman, Philip Hugenholtz, Eric E. Allen, Rachna J. Ram, Paul M. Richardson, Victor V. Solovyev, Edward M. Rubin, Daniel S. Rokhsar, and Jillian F. Banfield. 2004. 'Community Structure and Metabolism through Reconstruction of Microbial Genomes from the Environment'. *Nature* 428 (6978): 37–43. https://doi.org/10.1038/nature02340.

Tytgat, Bjorn, Elie Verleyen, Dagmar Obbels, Karolien Peeters, Aaike De Wever, Sofie D'hondt, Tim De Meyer, Wim Van Criekinge, Wim Vyverman, and Anne Willems. 2014. 'Bacterial Diversity Assessment in Antarctic Terrestrial and Aquatic Microbial Mats: A Comparison between Bidirectional Pyrosequencing and Cultivation'. *PLOS ONE* 9 (6): e97564. https://doi.org/10.1371/journal.pone.0097564.

Uritskiy, Gherman V., Jocelyne DiRuggiero, and James Taylor. 2018. 'MetaWRAP—a Flexible Pipeline for Genome-Resolved Metagenomic Data Analysis'. *Microbiome* 6 (1): 158. https://doi.org/10.1186/s40168-018-0541-1.

Van Goethem, Marc W., Andrew R. Osborn, Benjamin P. Bowen, Peter F. Andeer, Tami L. Swenson, Alicia Clum, Robert Riley, et al. 2021. 'Long-Read Metagenomics of Soil Communities Reveals Phylum-Specific Secondary Metabolite Dynamics'. *Communications Biology* 4 (1): 1–10. https://doi.org/10.1038/s42003-021-02809-4.

Vander Schaaf, Nicole A., Anna M. G. Cunningham, Brandon P. Cluff, CodyJo K. Kraemer, Chelsea L. Reeves, Carli J. Riester, Lauren K. Slater, Michael T. Madigan, and W. Matthew Sattley. 2015. 'Cold-Active, Heterotrophic Bacteria from the Highly Oligotrophic Waters of Lake Vanda, Antarctica'. *Microorganisms* 3 (3): 391–406. https://doi.org/10.3390/microorganisms3030391.

Vartoukian, Sonia R., Richard M. Palmer, and William G. Wade. 2010. 'Strategies for Culture of "Unculturable" Bacteria'. *FEMS Microbiology Letters* 309 (1): 1–7. https://doi.org/10.1111/j.1574-6968.2010.02000.x.

Vaser, Robert, Ivan Sović, Niranjan Nagarajan, and Mile Šikić. 2017. 'Fast and Accurate de Novo Genome Assembly from Long Uncorrected Reads'. *Genome Research* 27 (5): 737–46. https://doi.org/10.1101/gr.214270.116.

Ventura, Marco, Carlos Canchaya, Andreas Tauch, Govind Chandra, Gerald F. Fitzgerald, Keith F. Chater, and Douwe van Sinderen. 2007. 'Genomics of Actinobacteria: Tracing the Evolutionary History of an Ancient Phylum'. *Microbiology and Molecular Biology Reviews*, September. https://doi.org/10.1128/MMBR.00005-07.

Viaene, Tom, Sarah Langendries, Stien Beirinckx, Martine Maes, and Sofie Goormachtig. 2016. 'Streptomyces as a Plant's Best Friend?' *FEMS Microbiology Ecology* 92 (8): fiw119. https://doi.org/10.1093/femsec/fiw119.

Walker, Bruce J., Thomas Abeel, Terrance Shea, Margaret Priest, Amr Abouelliel, Sharadha Sakthikumar, Christina A. Cuomo, et al. 2014. 'Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement'. *PLOS ONE* 9 (11): e112963. https://doi.org/10.1371/journal.pone.0112963.

Wallenstein, Alexander, Nadine Rehm, Marina Brinkmann, Martina Selle, Nadège Bossuet-Greif, Daniel Sauer, Boyke Bunk, et al. 2020. 'ClbR Is the Key Transcriptional Activator of Colibactin Gene Expression in Escherichia Coli'. *MSphere*, July. https://doi.org/10.1128/mSphere.00591-20.

Wang, Gaoyan, Zhiying Zhao, Jing Ke, Yvonne Engel, Yi-Ming Shi, David Robinson, Kerem Bingol, et al. 2019. 'CRAGE Enables Rapid Activation of Biosynthetic Gene Clusters in Undomesticated Bacteria'. *Nature Microbiology* 4 (12): 2498–2510. https://doi.org/10.1038/s41564-019-0573-8.

Wang, Mingxun, Jeremy J. Carver, Vanessa V. Phelan, Laura M. Sanchez, Neha Garg, Yao Peng, Don Duy Nguyen, et al. 2016. 'Sharing and Community Curation of Mass Spectrometry Data with Global Natural Products Social Molecular Networking'. *Nature Biotechnology* 34 (8): 828–37. https://doi.org/10.1038/nbt.3597.

Wang, Weishan, Xiao Li, Juan Wang, Sihai Xiang, Xiaozhou Feng, and Keqian Yang. 2013. 'An Engineered Strong Promoter for Streptomycetes'. *Applied and Environmental Microbiology*, July. https://doi.org/10.1128/AEM.00985-13.

Wang, Xiuhong, and Xilin Zhao. 2009. 'Contribution of Oxidative Damage to Antimicrobial Lethality'. *Antimicrobial Agents and Chemotherapy*, April. https://doi.org/10.1128/AAC.01087-08.

Wang, Yunhao, Yue Zhao, Audrey Bollas, Yuru Wang, and Kin Fai Au. 2021. 'Nanopore Sequencing Technology, Bioinformatics and Applications'. *Nature Biotechnology* 39 (11): 1348–65. https://doi.org/10.1038/s41587-021-01108-x.

Watson, Mick. 2018. 'The Genomic and Proteomic Landscape of the Rumen Microbiome Revealed by Comprehensive Genome-Resolved Metagenomics'. University of Edinburgh. The Roslin Institute. Royal (Dick) School of Veterinary Studies. https://doi.org/10.7488/ds/2470.

Watve, Milind G., Rashmi Tickoo, Maithili M. Jog, and Bhalachandra D. Bhole. 2001. 'How Many Antibiotics Are Produced by the Genus Streptomyces?' *Archives of Microbiology* 176 (5): 386–90. https://doi.org/10.1007/s002030100345.

Weber, T., K. Welzel, S. Pelzer, A. Vente, and W. Wohlleben. 2003. 'Exploiting the Genetic Potential of Polyketide Producing Streptomycetes'. *Journal of Biotechnology*, Genome Research on Bacteria Relevant for Agriculture, Environment and Biotechnology, 106 (2): 221–32. https://doi.org/10.1016/j.jbiotec.2003.08.004.

Wheeler, Travis J., Jody Clements, and Robert D. Finn. 2014. 'Skylign: A Tool for Creating Informative, Interactive Logos Representing Sequence Alignments and Profile Hidden Markov Models'. *BMC Bioinformatics* 15 (1): 7. https://doi.org/10.1186/1471-2105-15-7.

Wichard, Thomas. 2016. 'Identification of Metallophores and Organic Ligands in the Chemosphere of the Marine Macroalga Ulva (Chlorophyta) and at Land-Sea Interfaces'. *Frontiers in Marine Science* 3. https://www.frontiersin.org/article/10.3389/fmars.2016.00131.

Wick, Ryan R., Louise M. Judd, Claire L. Gorrie, and Kathryn E. Holt. 2017. 'Unicycler: Resolving Bacterial Genome Assemblies from Short and Long Sequencing Reads'.

PLOS Computational Biology 13 (6): e1005595. https://doi.org/10.1371/journal.pcbi.1005595.

Widdick, David, Sylvain F. Royer, Hua Wang, Natalia M. Vior, Juan Pablo Gomez-Escribano, Benjamin G. Davis, and Mervyn J. Bibb. 2018. 'Analysis of the Tunicamycin Biosynthetic Gene Cluster of Streptomyces Chartreusis Reveals New Insights into Tunicamycin Production and Immunity'. *Antimicrobial Agents and Chemotherapy* 62 (8): e00130-18. https://doi.org/10.1128/AAC.00130-18.

Wnuk, Ewa, Adam Waśko, Anna Walkiewicz, Piotr Bartmiński, Romualda Bejger, Lilla Mielnik, and Andrzej Bieganowski. 2020. 'The Effects of Humic Substances on DNA Isolation from Soils'. *PeerJ* 8 (July): e9378. https://doi.org/10.7717/peerj.9378.

Wong, Sin Yin, James C. Charlesworth, Nicole Benaud, Brendan P. Burns, and Belinda C. Ferrari. 2019. 'Novel Quorum Sensing Activity in East Antarctic Soil Bacteria'. bioRxiv. https://doi.org/10.1101/749861.

Worobo, R. W., M. J. Van Belkum, M. Sailer, K. L. Roy, J. C. Vederas, and M. E. Stiles. 1995. 'A Signal Peptide Secretion-Dependent Bacteriocin from Carnobacterium Divergens'. *Journal of Bacteriology* 177 (11): 3143–49. https://doi.org/10.1128/jb.177.11.3143-3149.1995.

Wu, Yu-Wei, Yung-Hsu Tang, Susannah G. Tringe, Blake A. Simmons, and Steven W. Singer. 2014. 'MaxBin: An Automated Binning Method to Recover Individual Genomes from Metagenomes Using an Expectation-Maximization Algorithm'. *Microbiome* 2 (1): 26. https://doi.org/10.1186/2049-2618-2-26.

Xia, Haiyang, Xinqiao Zhan, Xu-Ming Mao, and Yong-Quan Li. 2020. 'The Regulatory Cascades of Antibiotic Production in Streptomyces'. *World Journal of Microbiology and Biotechnology* 36 (1): 13. https://doi.org/10.1007/s11274-019-2789-4.

Xu, Min, and Gerard D Wright. 2019. 'Heterologous Expression-Facilitated Natural Products' Discovery in Actinomycetes'. *Journal of Industrial Microbiology and Biotechnology* 46 (3–4): 415–31. https://doi.org/10.1007/s10295-018-2097-2.

Yan, Peiying, Shugui Hou, Tuo Chen, Xiaojun Ma, and Shuhong Zhang. 2012. 'Culturable Bacteria Isolated from Snow Cores along the 1300 Km Traverse from Zhongshan Station to Dome A, East Antarctica'. *Extremophiles* 16 (2): 345–54. https://doi.org/10.1007/s00792-012-0434-3.

Yan, Qiang, and Stephen S. Fong. 2017. 'Challenges and Advances for Genetic Engineering of Non-Model Bacteria and Uses in Consolidated Bioprocessing'. *Frontiers in Microbiology* 8. https://www.frontiersin.org/article/10.3389/fmicb.2017.02060.

Yergeau, Etienne, Kevin K. Newsham, David A. Pearce, and George A. Kowalchuk. 2007. 'Patterns of Bacterial Diversity across a Range of Antarctic Terrestrial Habitats'. *Environmental Microbiology* 9 (11): 2670–82. https://doi.org/10.1111/j.1462-2920.2007.01379.x.

Yi Pan, Shing, G. Y. Annie Tan, Peter Convey, David A. Pearce, and Irene K. P. Tan. 2013. 'Diversity and Bioactivity of Actinomycetes from Signy Island Terrestrial Soils, Maritime Antarctic'. *Advances in Polar Science* 24 (4): 208–12. https://doi.org/10.3724/SP.J.1085.2013.00208.

Zaffiri, Lorenzo, Jared Gardner, and Luis H. Toledo-Pereyra. 2012. 'History of Antibiotics. From Salvarsan to Cephalosporins'. *Journal of Investigative Surgery* 25 (2): 67–77. https://doi.org/10.3109/08941939.2012.664099.

Zhang, Hui, Yuji Sekiguchi, Satoshi Hanada, Philip Hugenholtz, Hongik Kim, Yoichi Kamagata, and KazunoriYR 2003 Nakamura. n.d. 'Gemmatimonas Aurantiaca Gen. Nov., Sp. Nov., a Gram-Negative, Aerobic, Polyphosphate-Accumulating Micro-Organism, the First Cultured Representative of the New Bacterial Phylum

Gemmatimonadetes Phyl. Nov.' *International Journal of Systematic and Evolutionary Microbiology* 53 (4): 1155–63. https://doi.org/10.1099/ijs.0.02520-0.

Zhang, Mingzi M., Yajie Wang, Ee Lui Ang, and Huimin Zhao. 2016. 'Engineering Microbial Hosts for Production of Bacterial Natural Products'. *Natural Product Reports* 33 (8): 963–87. https://doi.org/10.1039/c6np00017g.

Zhang, Xiafei, Hindra, and Marie A Elliot. 2019. 'Unlocking the Trove of Metabolic Treasures: Activating Silent Biosynthetic Gene Clusters in Bacteria and Fungi'. *Current Opinion in Microbiology*, Antimicrobials, 51 (October): 9–15. https://doi.org/10.1016/j.mib.2019.03.003.

Zhang, Zheng, Jianing Wang, Jinlan Wang, Jingjing Wang, and Yuezhong Li. 2020. 'Estimate of the Sequenced Proportion of the Global Prokaryotic Genome'. *Microbiome* 8 (1): 134. https://doi.org/10.1186/s40168-020-00903-z.

Zhu, Yuanjun, Lifei Wang, Yu Du, Songmei Wang, Tengfei Yu, and Bin Hong. 2011. 'Heterologous Expression of Human Interleukin-6 in Streptomyces Lividans TK24 Using Novel Secretory Expression Vectors'. *Biotechnology Letters* 33 (2): 253–61. https://doi.org/10.1007/s10529-010-0428-0.

Ziemert, Nadine, Anna Lechner, Matthias Wietz, Natalie Millán-Aguiñaga, Krystle L. Chavarria, and Paul Robert Jensen. 2014. 'Diversity and Evolution of Secondary Metabolism in the Marine Actinomycete Genus Salinispora'. *Proceedings of the National Academy of Sciences of the United States of America* 111 (12): E1130-1139. https://doi.org/10.1073/pnas.1324161111.

Ziller, Antoine, and Laurence Fraissinet-Tachet. 2018. 'Metallothionein Diversity and Distribution in the Tree of Life: A Multifunctional Protein'. *Metallomics* 10 (11): 1549–59. https://doi.org/10.1039/C8MT00165K.

Zopfi, Jakob, Thomas Kjær, Lars P. Nielsen, and Bo Barker Jørgensen. 2001. 'Ecology of Thioploca Spp.: Nitrate and Sulfur Storage in Relation to Chemical Microgradients and Influence OfThioploca Spp. on the Sedimentary Nitrogen Cycle'. *Applied and Environmental Microbiology* 67 (12): 5530–37. https://doi.org/10.1128/AEM.67.12.5530-5537.2001.

Zou, Zhengzhong, Deyao Du, Yanyan Zhang, Jihui Zhang, Guoqing Niu, and Huarong Tan. 2014. 'A γ-Butyrolactone-Sensing Activator/Repressor, JadR3, Controls a Regulatory Mini-Network for Jadomycin Biosynthesis'. *Molecular Microbiology* 94 (3): 490–505. https://doi.org/10.1111/mmi.12752.

7 Appendix A

Supplementary Table 1: Primers used for BGC amplification.

Primer	Sequence	Product length	PCR #	BGC	vector	promoter
2_215	GTGTTGTAAAGTCTGGTGTACCTA AGTTCGTCATCCTACGGACCGC	4355	22	contig_13212_region2 (terpene)	g2	sp44
2_216	TTTGGAGATTTTCAACGTAGGTTC TGGCTCACATTACGATCTGGG					
2_217	GTGTTGTAAAGTCTGGTGTACCTA AGGACCCACTTCAGTGGCGAG	21000	19	contig_2148 NRPS	g2	sp44
2_218	TTTGGAGATTTTCAACGTAGGTTC ATCAAGTTTCGCACCCGCTA					
2_219	GTGTTGTAAAGTCTGGTGTACCTA AGTGAGCATAAAGCGTACTCCGAA	13558	12	contig_11044 NRPS	g2	sp44
2_220	TTTGGAGATTTTCAACGTAGGTTC CTCGAGATTGCACGGGAGTA					
2_221	GTCCTAGTATGGTAGGATGAGCAA AGTCACCACAGGGAAGCGTT	9007	1	contig_11044 NRPS	g2	p21
2_222	CCAGATCTGCAACCTCTTAAGATT CGATTTCTTGCGCTGTGC					
2_223	GTGTTGTAAAGTCTGGTGTACCTA AGAGTTTGCAACTCGGCGCATC	12403	13	contig_24847 NRPS	g2	sp44
2_224	TTTGGAGATTTTCAACGTAGGTTC TGCCCTCTGCTTAGTTTCCG					
2_225	CTAGTATGGTAGGATGAGCAAGTT TAACGTGAAACGGGAGACAGAC	9229	2	contig_115 lassopeptide	g2	sp24/p21
2_226	CTAATGTAAAGTCGTGGCCAATTT TCTCGAATACTGTGCAGCCC					
2_227	CTAGTATGGTAGGATGAGCAAGTT TAGGATCGAGCTGACACGGG	7651	3	contig_13589 terpene	g1	sp24/p21
2_228	CTAATGTAAAGTCGTGGCCAATTT GGTGCTCCTTTGGCTGACC					
2_229	GTGTTGTAAAGTCTGGTGTACCTA AGGTACAACTAGCGATCAGGGGG	10594	14	contig_13212 terpene	g2	sp44
2_230	TTTGGAGATTTTCAACGTAGGTTC GAGGGGTATGACGATGTCGG					
2_231	GTCCTAGTATGGTAGGATGAGCAA AGTGGTACAGTATCCGTCCGC	7498	4	contig_4743 NRPS	g2	p21
2_232	CCAGATCTGCAACCTCTTAAGAGT CGACGACGTCATATTCGC					
2_233	GTGTTGTAAAGTCTGGTGTACCTA AGATCCGCCAAACAGACGAGAG	8490	5	contig_4743 NRPS	g2	sp44
2_234	TTTGGAGATTTTCAACGTAGGTTC GCGAATGGCAAAGTGCTCAA					
2_235	GTGTTGTAAAGTCTGGTGTACCTA AGATGCGGAAAGAAGGCCATAG	23158	20	contig_736 NRPS	g2	sp44
2_236	TTTGGAGATTTTCAACGTAGGTTC TCCTTCAACTTCATGGCACC					
2_237	CTAATGTAAAGTCGTGGCCAATTT CTCGCAATCGCAGGCACC	15354	15	contig_9172 NRPS	g1	sp24/p21
2_238	CTAGTATGGTAGGATGAGCAAGTT TACGGTAAGGTGGTGGAACGGA					
2_239	GTGTTGTAAAGTCTGGTGTACCTA AGGCATTGGGCGAATCTCTGC	7909	6	contig_4314 lassopeptide	g2	sp44
2_240	TTTGGAGATTTTCAACGTAGGTTC AAATCGTTGTCCCGGCGTA					
2_241	GTGTTGTAAAGTCTGGTGTACCTA AGCCTTTTGTTCGGACTTTCGACA T	20181	21	contig_36584 PKS	g2	sp44
2_242	TTTGGAGATTTTCAACGTAGGTTC AGCCGGACAGAGTAAGAGAC					

Supplementary Table 1 (continued)

2_243	GTCCTAGTATGGTAGGATGAGCAA AGAAGAAATTAGGCCGCTGACG	6877	7	contig_36584 PKS	g2	p21
2_244	CCAGATCTGCAACCTCTTAAGAGG GTTGACTGGAACACGCTG					
2_245	GTGTTGTAAAGTCTGGTGTACCTA AGTCACGATTTCATCACCCGAACA	9579	16	scaffold_45328 PKS	g2	sp44
2_246	TTTGGAGATTTTCAACGTAGGTTC GACTTTGGAAGCTCGCTCGTT					
2_247	CCAGATCTGCAACCTCTTAAGACG GTTATCTCTGCGGGAAGT	4613	8	scaffold_45328 PKS	g2	p21
2_248	GTCCTAGTATGGTAGGATGAGCAA ACTTCGTCGCGAAACGGTGAG					
2_249	GTGTTGTAAAGTCTGGTGTACCTA AGCATGCACGTTCCATGGCT	8390	9	contig_6994 PKS	g2	sp44
2_250	TTTGGAGATTTTCAACGTAGGTTC CCGCCGACTACGGTTAGAG					
2_251	CCAGATCTGCAACCTCTTAAGATA CTCCCTGTCGTTGGCTG	5198	10	contig_6994 PKS	g2	p21
2_252	GTCCTAGTATGGTAGGATGAGCAA AACCGCCAACACTTACAATCTTCA					
2_253	CTAGTATGGTAGGATGAGCAAGTT TACAAGACTTCGCGCCGTCAA	11886	17	contig_228 lassopeptide	g1	sp24/p21
2_254	CTAATGTAAAGTCGTGGCCAATTT CAGGGATCGAGTGTGGACAT					
2_255	GTGTTGTAAAGTCTGGTGTACCTA AGCCGTCACTATCGAGCCAATGT	7263	11	Contig_15892 lassopeptide	g2	sp44
2_256	TTTGGAGATTTTCAACGTAGGTTC TTAATTGTGCCGCGACTGGA					
2_257	CTAGTATGGTAGGATGAGCAAGTT TAGACGAGCGCGCTAGATGAA	13505	18	contig_25828 lassopeptide	g1	p21
2_258	CTAATGTAAAGTCGTGGCCAATTT ACGATAGATGGGTGGCAACG					
2_259	CTAATGTAAAGTCGTGGCCAATTT CCTTCGTCAAGTTCGGCCAG	13125	39	contig_5955: PKS	g1	sp24/p21
2_260	CTAGTATGGTAGGATGAGCAAGTT TACGATACGGAGGGAGCAACAT					
2_261	GATGGAGGCCGCGCTTGACGTTAT CCTGTCCAGCCGTG	2166	23	contig_5955: PKS	g1	sp24/p21
2_262	TCGGCGGCGTAGCCCTCGCTGAGA GCGCCCCTTG					
2_265	CTAATGTAAAGTCGTGGCCAATTT CCGAGGGGTGGACGAGAATG	9273	33	contig_10649: T3PKS	g1	sp24/p21
2_266	CTAGTATGGTAGGATGAGCAAGTT TATCTGCATTGTGGAATGTCCGTG					
2_267	GTGTTGTAAAGTCTGGTGTACCTA AGCATGGAAAATGGCTGGCACG	5231	34	contig_10669 : T3PKS	g2	sp44
2_268	TTTGGAGATTTTCAACGTAGGTTC TACGCGCATCCGGAAACTAA					
2_269	CTAGTATGGTAGGATGAGCAAGTT TATCGCCGCCCTATAATCGTTC	5372	35	contig_23551: T3pks	g1	sp24/p21
2_270	CTAATGTAAAGTCGTGGCCAATTT GTCTCCGTTGGCTTCTTTCG					
2_271	GTGTTGTAAAGTCTGGTGTACCTA AGCACTGATCAATCCGCCCCTG	13520	40	scaffold_13961: T1PKS	g2	sp44
2_272	TTTGGAGATTTTCAACGTAGGTTC GCGAATCCGAAGCTGGAAGT					
2_273	CCAGATCTGCAACCTCTTAAGACT CGAGCAACTGGGTAATTCG	2053	24	scaffold_13961: T1PKS	g2	p21
2_274	GTCCTAGTATGGTAGGATGAGCAA ATCTGCCATGGTAACAACCTGA					
2_275	CTAGTATGGTAGGATGAGCAAGTT TAGAGGCTCCGGGTGCTAACAT	18477	44	contig_10632: NRPS PKS hybrid	g1	sp24/p21
2_276	CTAATGTAAAGTCGTGGCCAATTT CGTCGTTACGCTGCTACTCT					

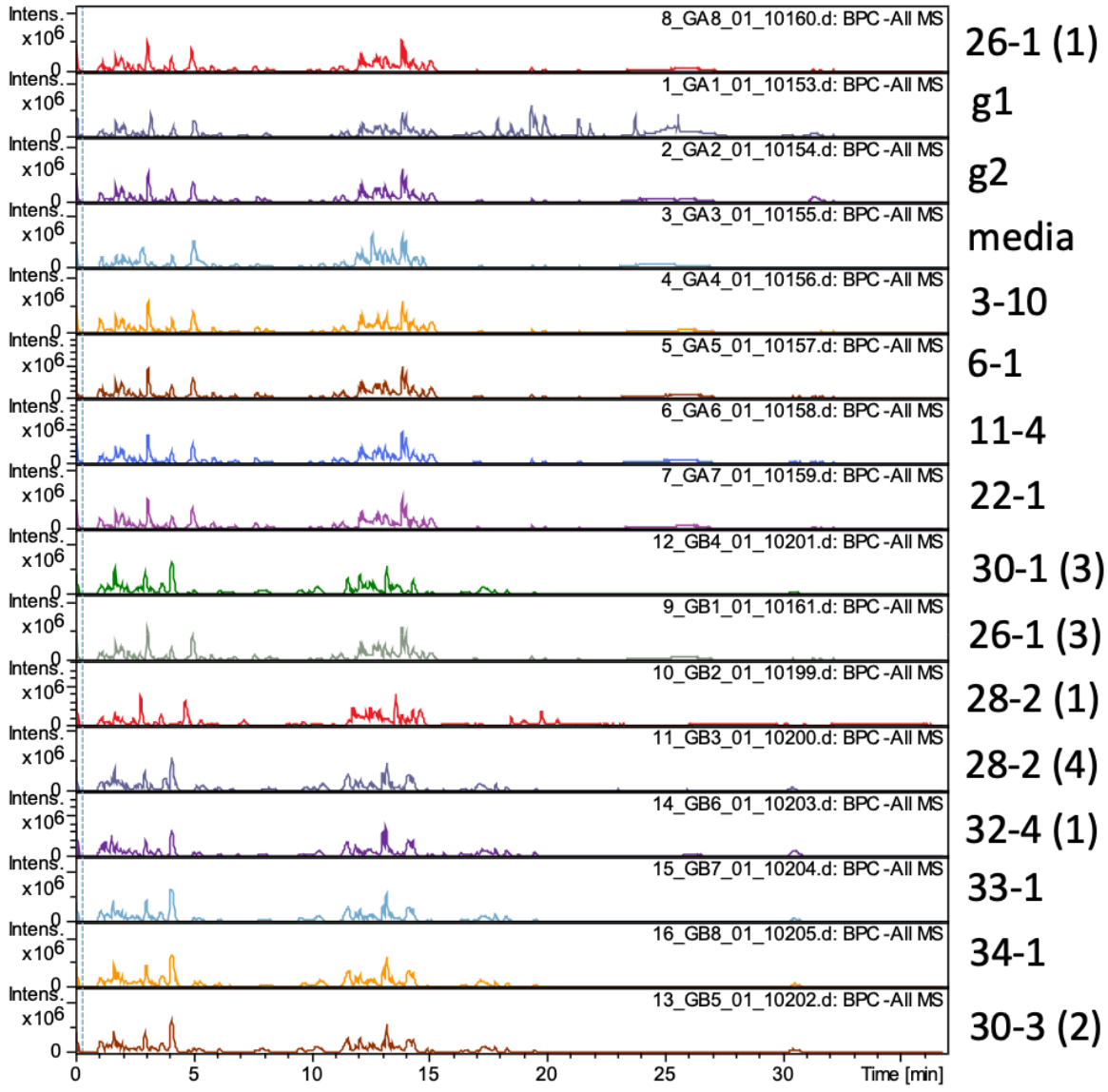
Supplementary Table 1 (continued)

2_277	GTGTTGTAAAGTCTGGTGTACCTA AGCCTCGCGCCGTTCCTGTC	17320	45	contig_7544: T1PKS	g2	sp44
2_278	TTTGGAGATTTTCAACGTAGGTTC ATGGATACCAGACAGCCTCCTGC					
2_279	CTAGTATGGTAGGATGAGCAAGTT TAATTACAGGCTTCATGCCACG	10189	36	scaffold_35893 lassopeptide	g1	sp24/p21
2_280	CTAATGTAAAGTCGTGGCCAATTT GAGTCTACGGAATAAGGCCGC					
2_281	TTTGGAGATTTTCAACGTAGGTTC TGCCCAACTTCCTGATCCGA	15013	41	contig_13679 lassopeptide	g1	sp44
2_282	GTGTTGTAAAGTCTGGTGTACCTA AGCCTGTATTGGAGCCGGGTT					
2_283	CTAGTATGGTAGGATGAGCAAGTT TATCGCAGACCACATTTGACACA	12791	42	contig_6313 lassopeptide	g1	p21
2_284	CTAATGTAAAGTCGTGGCCAATTT CGTCTTCACACATTTGCGCTC					
2_285	CTAGTATGGTAGGATGAGCAAGTT TAGAAGGGATCGCGCTGTAGG	6994	37	contig_1186 lanthipeptide	g1	sp24/p21
2_286	CTAATGTAAAGTCGTGGCCAATTT CTGCTGAGACCGCCACG					
2_287	CTAGTATGGTAGGATGAGCAAGTT TACACCAGGGTTGTTGCTTACAG	18248	46	contig_6994 BGC3 NRPSPKS	g1	p21
2_288	CTAATGTAAAGTCGTGGCCAATTT GACTCGCACTAAGACAGGCG					
2_289	GTGTTGTAAAGTCTGGTGTACCTA AGTTACCCTGTAGGACGGAACGAC	23750	47	scaffold_11847 NRPS	g2	sp44
2_290	TTTGGAGATTTTCAACGTAGGTTC AAGCTTTTGAAGGGCGTCGG					
2_291	GTGTTGTAAAGTCTGGTGTACCTA AGGCCGGCTTATCTATCCACCA	2202	25	contig_665: DUF692	g2	sp44
2_292	TTTGGAGATTTTCAACGTAGGTTC TTACCGATTCCGGCCCTCTA					
2_293	CCAGATCTGCAACCTCTTAAGAGC TTCAGTGGTTTAGTCGCC	1178	26	contig_665: DUF692	g2	p21
2_294	GTCCTAGTATGGTAGGATGAGCAA AGCGTCACGACGGTTTAAAAT					
2_295	GTGTTGTAAAGTCTGGTGTACCTA AGTAGCGGCCGGCTTATCTATC	2183	27	contig_414: DUF692	g2	sp44
2_296	TTTGGAGATTTTCAACGTAGGTTC TGGCGCATGACGTATTAGGA					
2_297	CCAGATCTGCAACCTCTTAAGAAG GCGACCCTCTCTAAAACG	3597	28	contig_414: DUF692	g2	p21
2_298	GTCCTAGTATGGTAGGATGAGCAA ATGTCGCTTCACGACGGTTTG					
2_299	GTGTTGTAAAGTCTGGTGTACCTA AGCTATCCACCATCAGGAGACCAT	2163	29	contig_291: DUF692	g2	sp44
2_300	TTTGGAGATTTTCAACGTAGGTTC CTGGCGCATGGCGTATTAGG					
2_301	CCAGATCTGCAACCTCTTAAGAGG CGGCTTCAATTCTGTGAA	3447	30	contig_291: DUF692	g2	p21
2_302	GTCCTAGTATGGTAGGATGAGCAA ACGCGTCACGACGGTTTG					
2_303	GTGTTGTAAAGTCTGGTGTACCTA AGGCCGGCTTATCTGTCCACC	2042	31	contig_14956: DUF692	g2	sp44
2_304	TTTGGAGATTTTCAACGTAGGTTC GGGCAAGGCAGTGAGTATCC					
2_305	CCAGATCTGCAACCTCTTAAGAGG GCCTCTCGTTTTGAGCAT	3267	32	contig_14956: DUF692	g2	p21
2_306	GTCCTAGTATGGTAGGATGAGCAA AGCAGCGGGGGAATAATAGACG					
2_307	GTGTTGTAAAGTCTGGTGTACCTA AGCATCAGTATCGTTAAACTGTTA CCC	16666	48	contig_2807: NRPS part 1	g2	sp44
2_308	TTTGGAGATTTTCAACGTAGGTTC AGCAATAGTTTCGGCGGTCA					

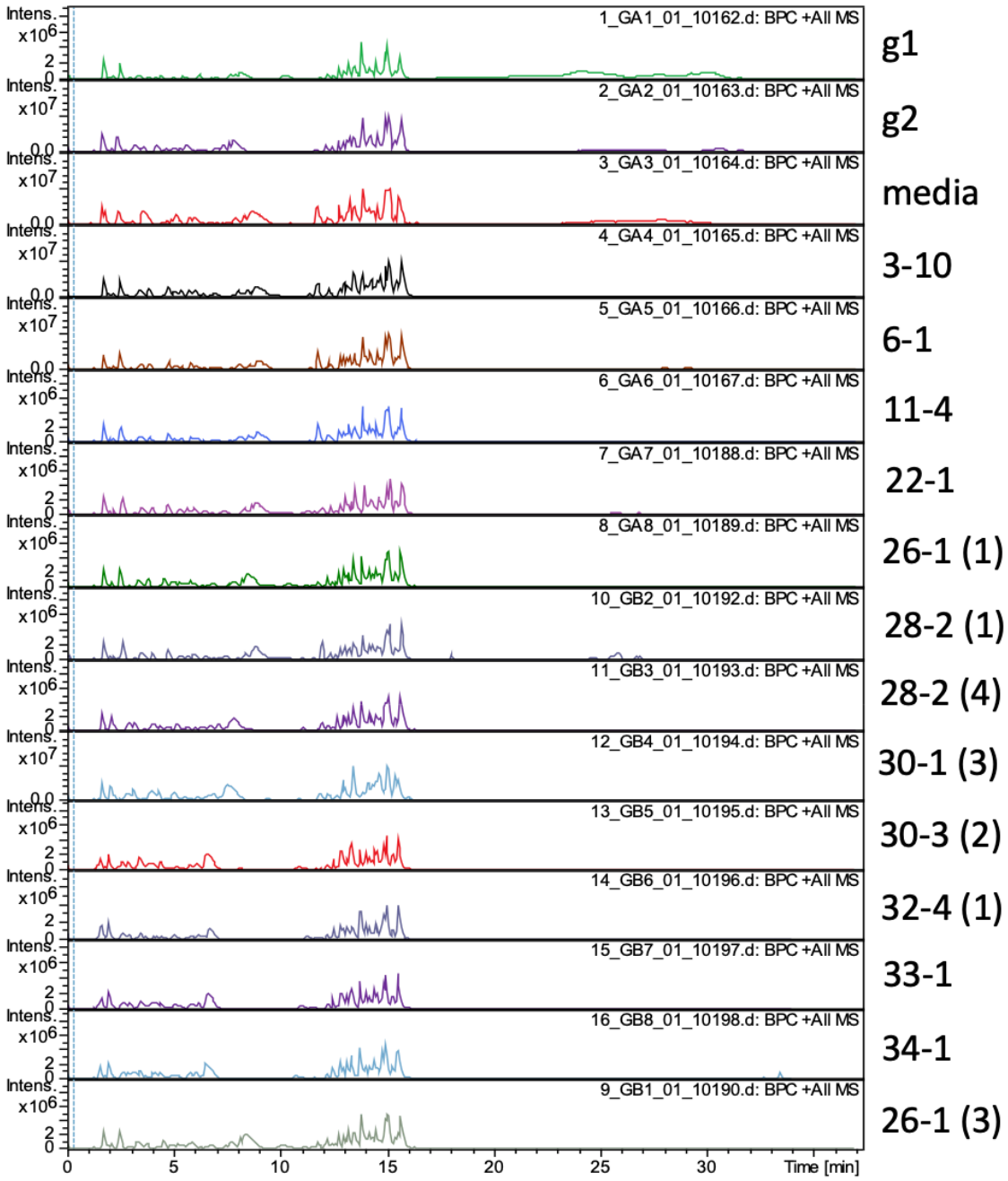
Supplementary Table 1 (continued)

2_310	GGCCTATGCCGTTGAACAAG	17420	49	contig_2807: NRPS part 2	g2	sp44
2_311	TTTGGAGATTTTCAACGTAGGTTC AGCTCGGTGAAAGACTCACA					
2_312	CTAGTATGGTAGGATGAGCAAGTT TACGATCATCTCCAGCCGCAA	6255	38	contig_3134 NRPS	g1	sp24/p21
2_313	CTAATGTAAAGTCGTGGCCAATTT ATGCCGAGATCATGCGCTAC					
2_315	CTAGTATGGTAGGATGAGCAAGTT TACTCACGCGACTGATGGATGAC	14863	43	contig_3134 NRPS	g1	sp24/p21
2_316	CTTGCGGCTGGAGATGATCGGACT CGTCCCCATGAAGCGG					
2_317	GTGTTGTAAAGTCTGGTGTACCTA AGGGACGCGAGACTAGCTCAT	25812	50	contig_11857 NRPS	g2	sp44
2_318	TTTGGAGATTTTCAACGTAGGTTC TACCCGCTCTGGTCGTTTCT					
2_319	GTGTTGTAAAGTCTGGTGTACCTA AGGAGCATCCACGCTGCCTAC	18946	51	contig_13147 NRPS PKS	g2	sp44
2_320	TTTGGAGATTTTCAACGTAGGTTC ATCTTCAACGGCAGCCTCC					
2_321	GGGAGTGTAAGCGTATGCGA	7227	58	contig_11044 NRPS fragment 1 (with 2_219)	g2	sp44
2_322	GGCAGACGGAATGAGTGTCA	7114	59	contig_11044 NRPS fragment 2 (with 2_220)	g2	sp44
2_323	GCCCCGGAATCCTGTATGAG	7492	60	contig_24847 NRPS fragment 1 (with 2_023)	g2	sp44
2_324	TGTTTCAACCATTGCGGCAG	6860	61	contig_24847 NRPS fragment 2 (with 2_224)	g2	sp44
2_325	GTGTTGTAAAGTCTGGTGTACCTA AGCCGATGAGGTCAGGTCCTT	11062	62	contig_13212 terpene	g2	sp44
2_326	TTTGGAGATTTTCAACGTAGGTTC AACGTCTATGCCGTGGTCTC					
2_327	CTGTCTGAACGTGAGGCGAA	c 8000	63	contig_9172 NRPS fragment 1 (with 2_237)	g1	sp24/p21
2_328	GGCAGCTCGTGACCGATAG	c 8000	64	contig_9172 NRPS fragment 2 (with 2_238)	g1	sp24/p21
2_329	CTACGCCATCACCTGGCATC	c 9000	65	contig_10632: NRPS PKS hybrid fragment 1 (with 2_275)	g1	sp24/p21
2_330	GGCAAGGTGAGGGGATTGG	c 9000	66	contig_10632: NRPS PKS hybrid fragment 2 (with 2_276)	g1	sp24/p21
2_331	GAGACTCAGACGACGAGGTG	c 9000	67	contig_6994 BGC3 NRPSPKS fragment 1 (with 2_287)	g1	sp24/p21
2_332	CTCAGCTCGGCCTCAATCTC	c 9000	68	contig_6994 BGC3 NRPSPKS fragment 2 (with 2_288)	g1	sp24/p21
2_333	GTGTTGTAAAGTCTGGTGTACCTA AGCACCTGACCGAAAGATAGCCAG	10425	55	contig_2313 terpene	g2	sp44
2_334	TTTGGAGATTTTCAACGTAGGTTC GTGGCTCGTTTGAGACTGAC					
2_335	TTTGGAGATTTTCAACGTAGGTTC ACTTTGGCGACTACGACCTC	8272	56	contig_795: carotenoid	g2	sp44
2_336	GTGTTGTAAAGTCTGGTGTACCTA AGGGAGCATACATGCGGCTAGA					
2_337	GTGTTGTAAAGTCTGGTGTACCTA AGGGCTCTCCGATCTGCTCTTC	5754	57	contig_4: carotenoid	g2	sp44
2_338	TTTGGAGATTTTCAACGTAGGTTC TGATCGTCCTCAACATCGGC					

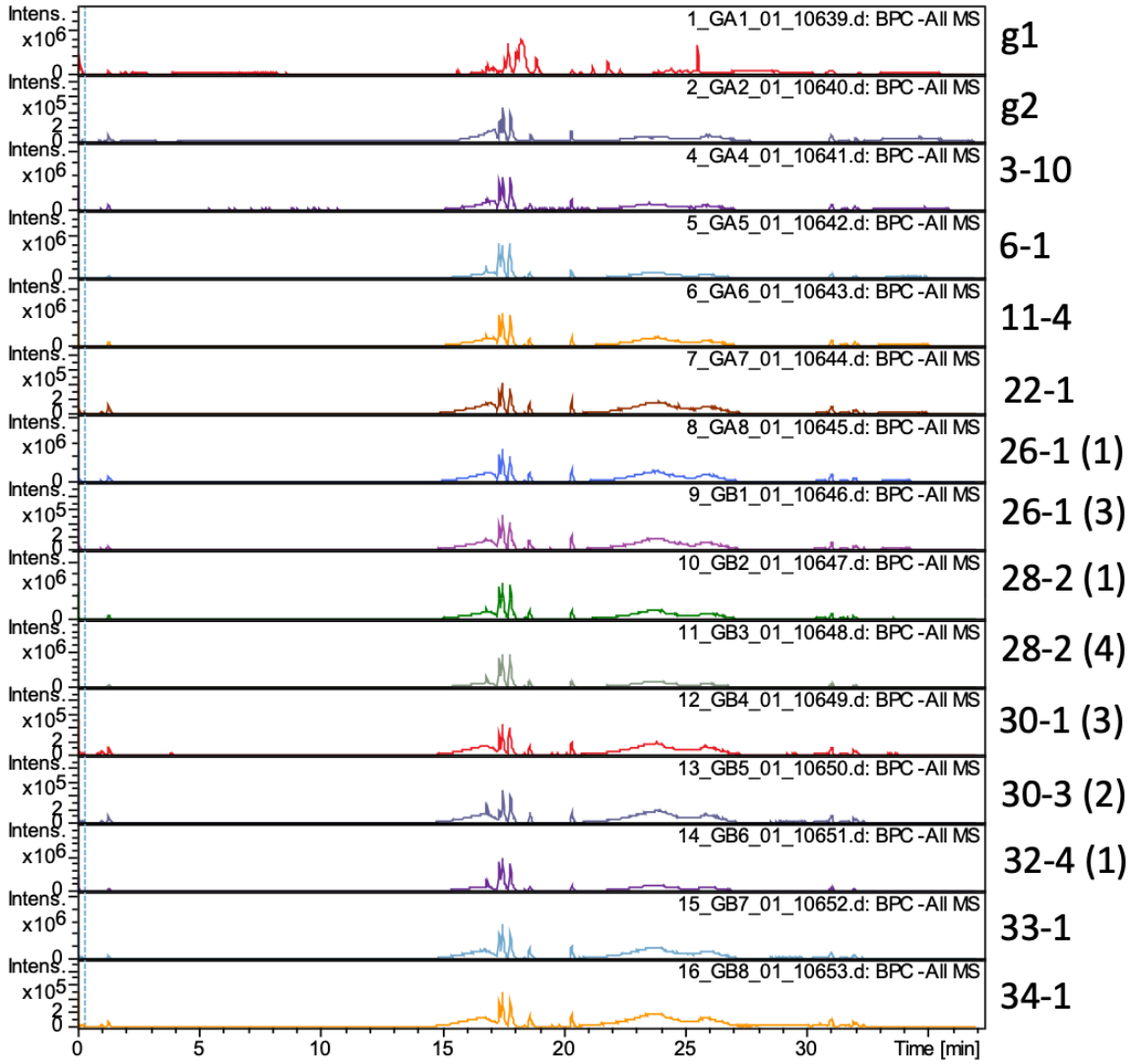
8 Appendix B



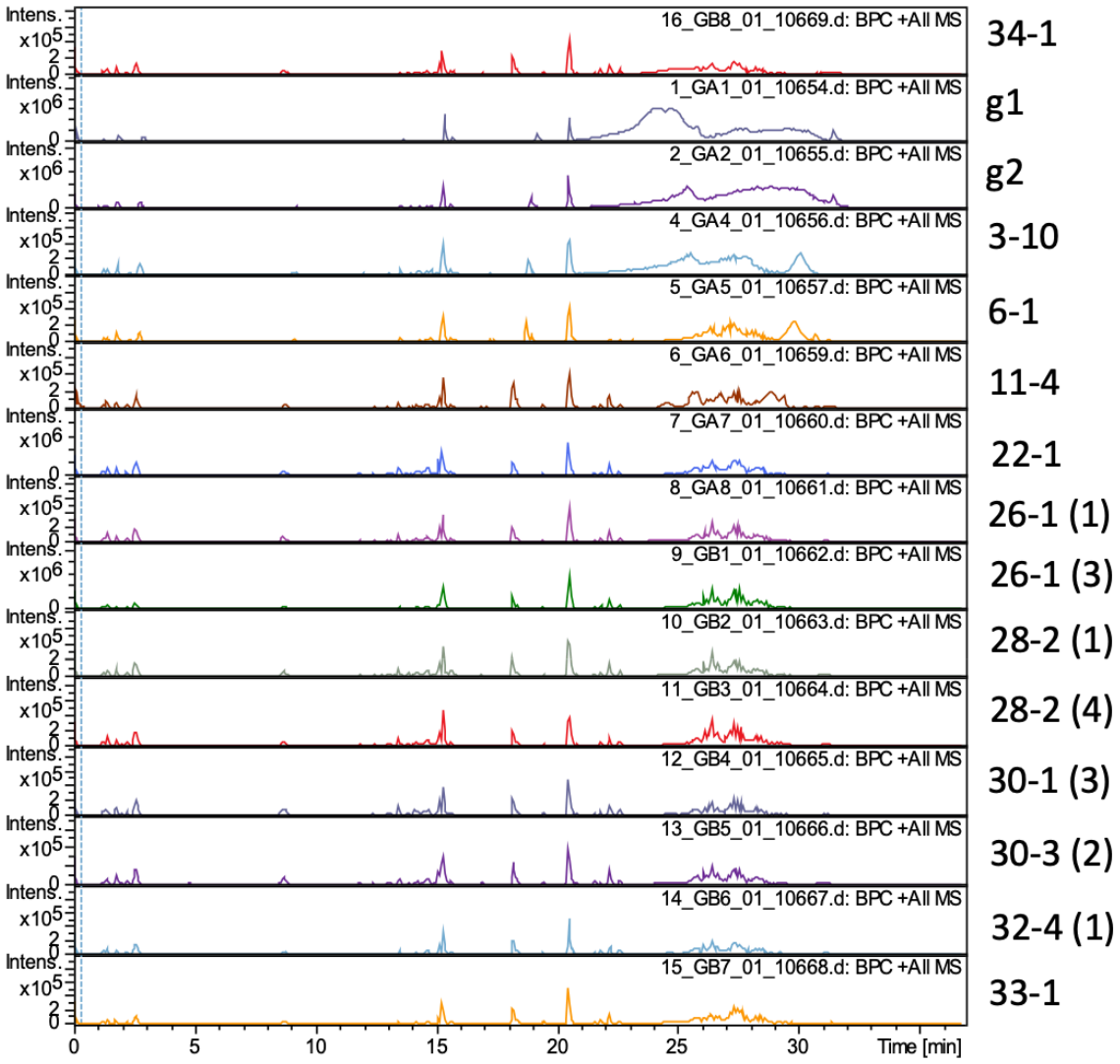
Supplementary Figure 1: *P. putida* transformants supernatant, base peak chromatograms of negative ion mode



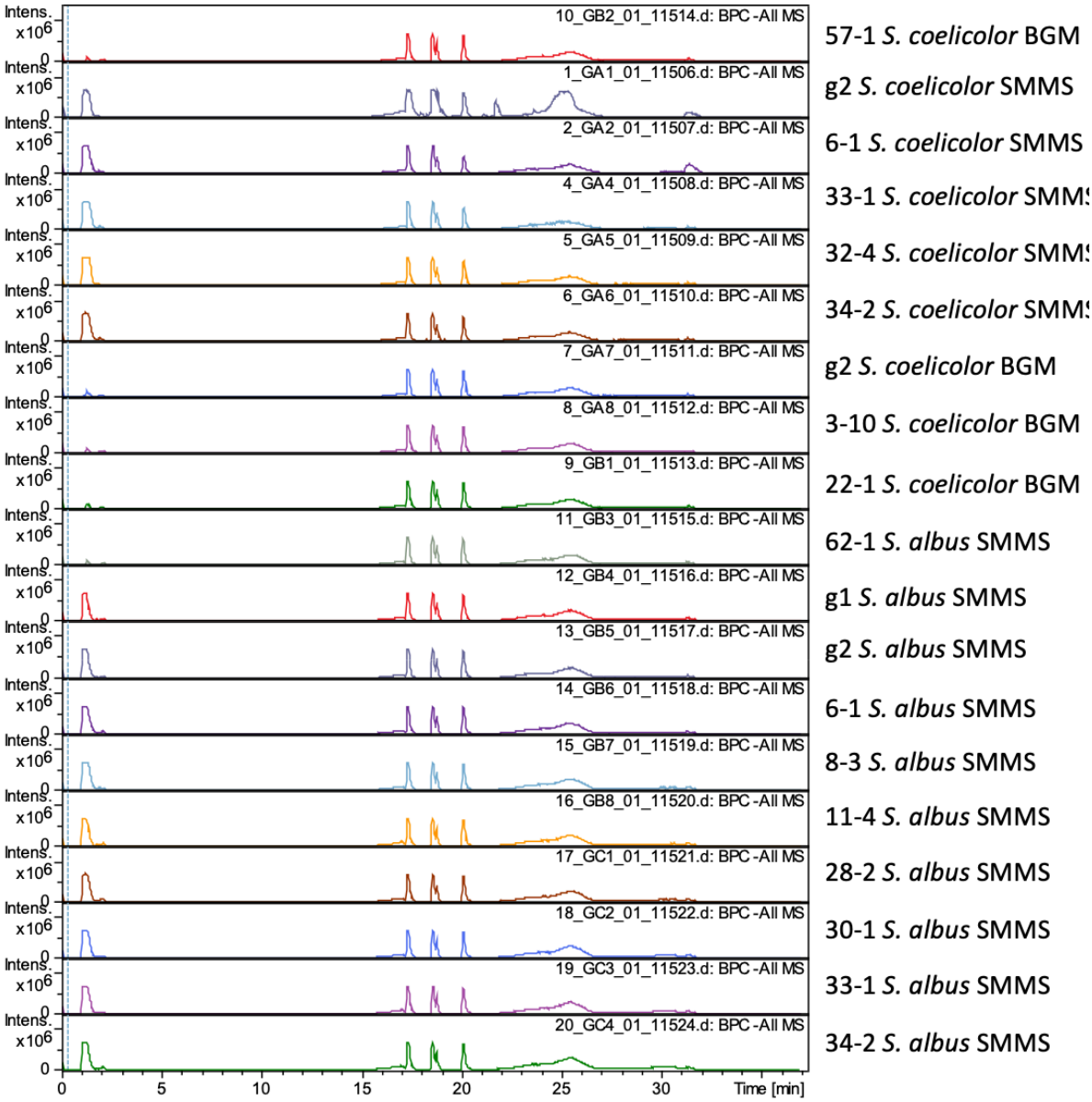
Supplementary Figure 2: *P. putida* transformants supernatant, base peak chromatograms of positive ion mode



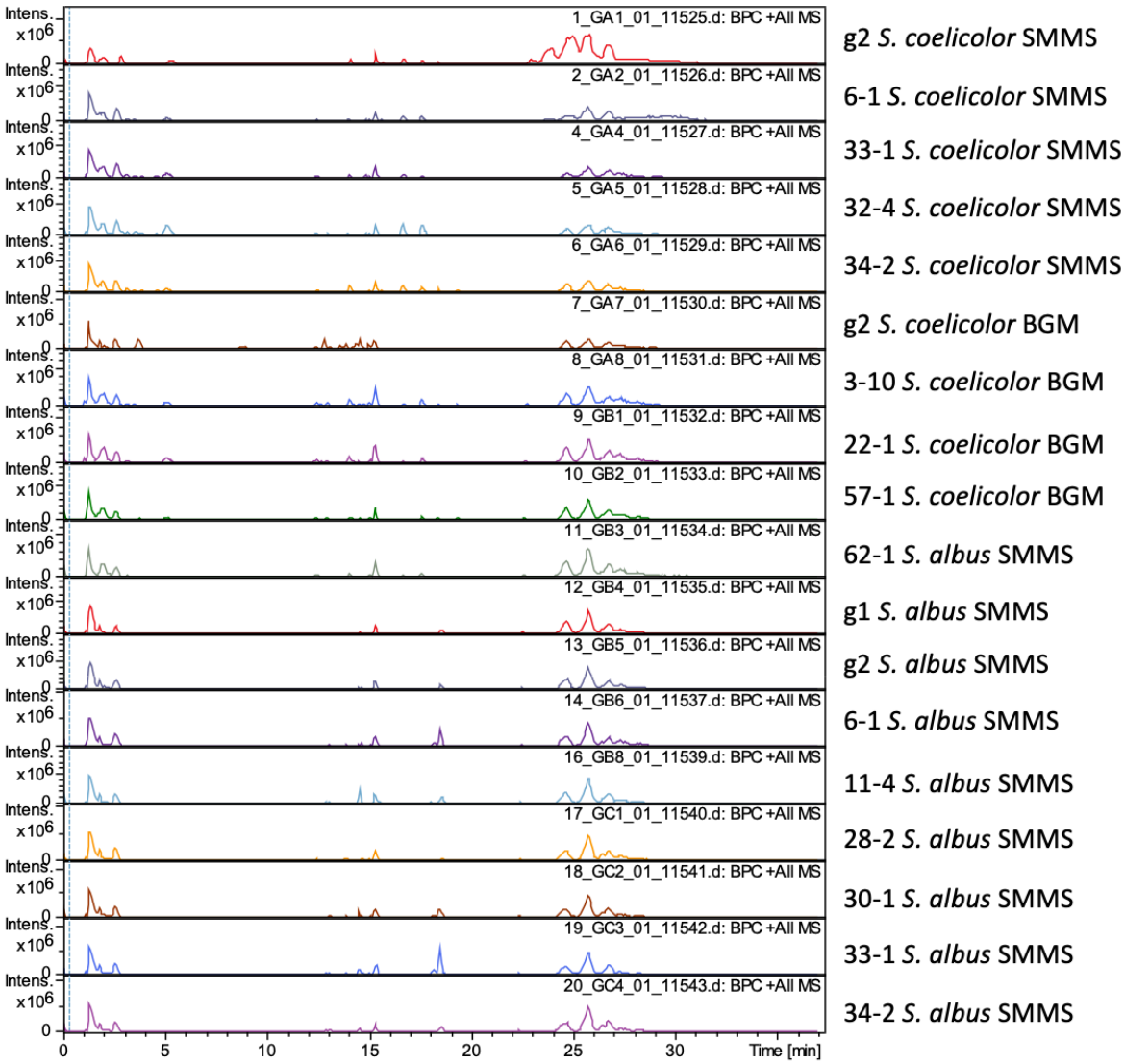
Supplementary Figure 3: *P. putida* transformants cell pellet MeOH extract, base peak chromatograms of negative ion mode



Supplementary Figure 4: *P. putida* transformants cell pellet MeOH extract, base peak chromatograms of positive ion mode

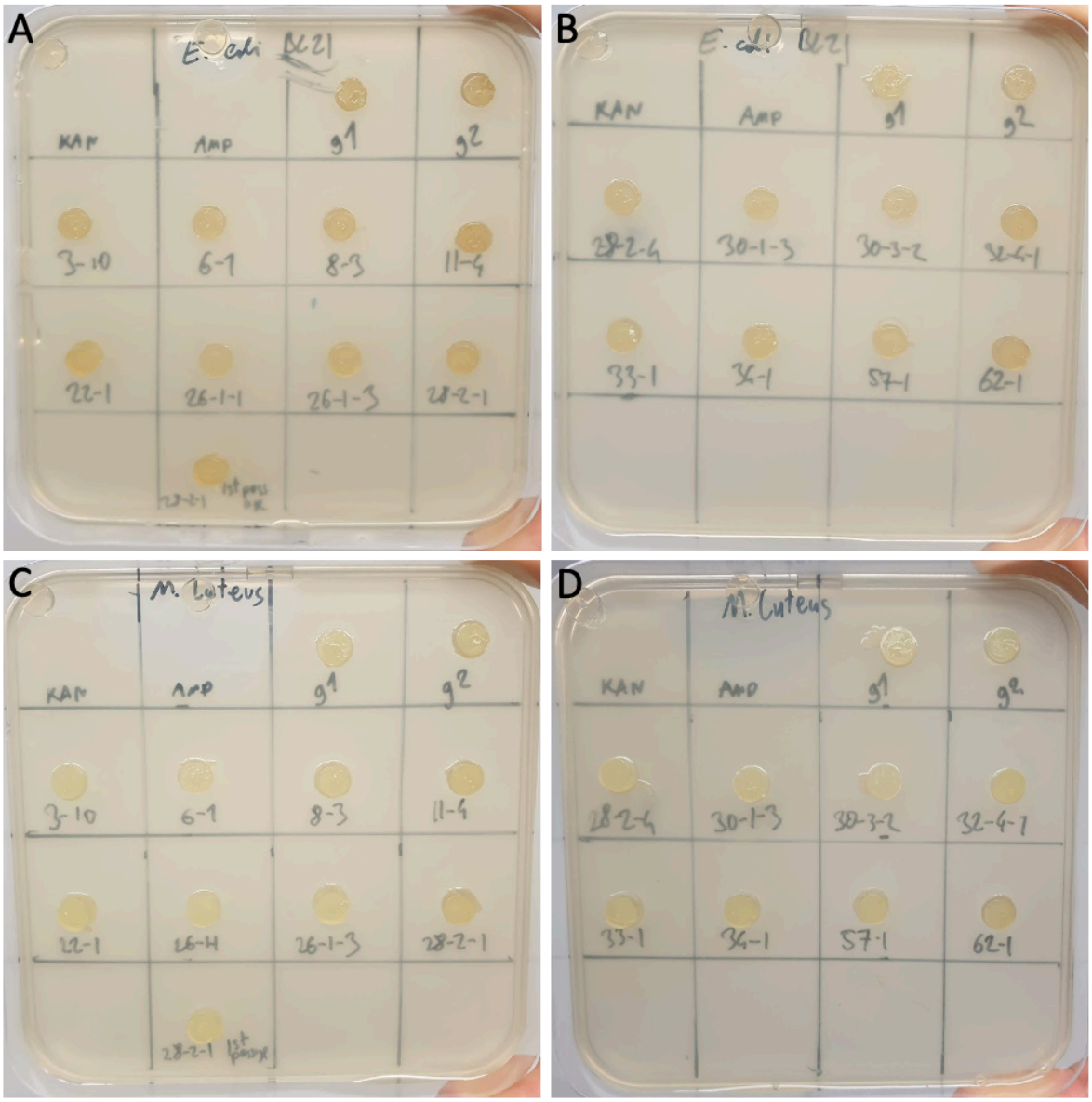


Supplementary Figure 5: *Streptomyces* exconjugants MeOH extract of agar plates, base peak chromatograms of negative ion mode

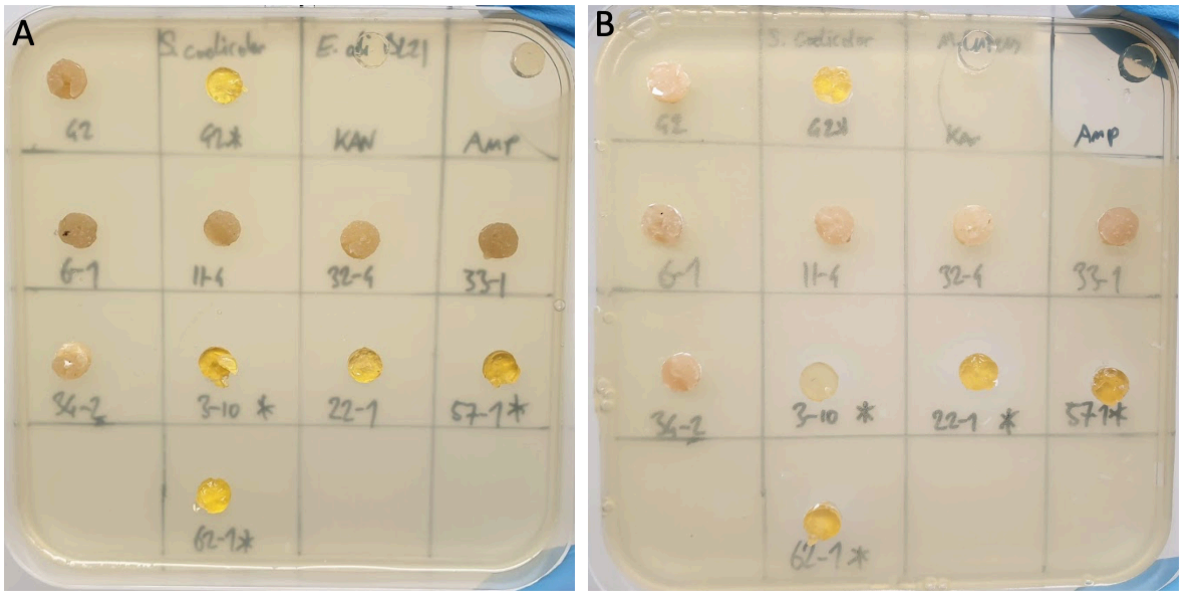


Supplementary Figure 6: *Streptomyces* exconjugants MeOH extract of agar plates, base peak chromatograms of positive ion mode

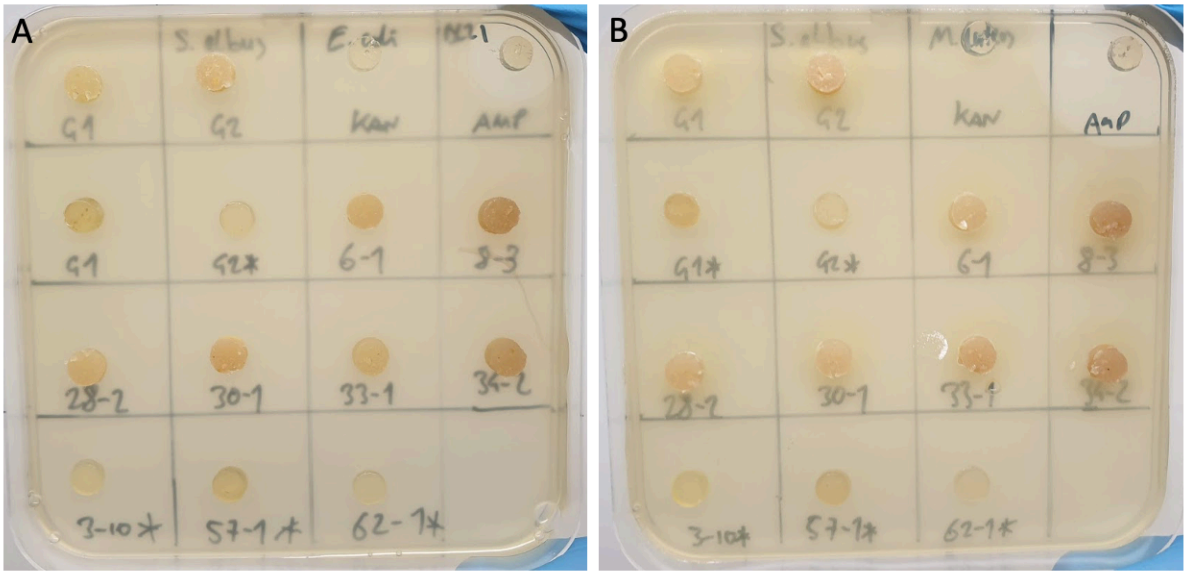
Antimicrobial assays of transformants and exconjugants.



Supplementary Figure 7: Results of *P. putida* transformants agar plug diffusion assay. No inhibition could be observed for (A, B) *E. coli* BL21 or (C, D) *M. luteus*.



Supplementary Figure 8: Results of *S. coelicolor* exconjugants agar plug diffusion assay. No inhibition could be observed (A) in *E. coli* BL21. (B) In *M. luteus*, a slight halo can be seen for all samples marked with an asterisk, including the pBCkBAC-g1 empty plasmid control. These samples were grown on BGM and under illumination to stimulate carotenoid production.



Supplementary Figure 9: Results of *S. albus* exconjugants agar plug diffusion assay. No inhibition could be observed (A) in *E. coli* BL21 or (B) in *M. luteus*.