UCL SCHOOL OF PHARMACY
BRUNSWICK SQUARE

Prospecting Novel Microbiomes for Antibiotic Compounds using Metagenomics and Genome Mining

Tim Ali Charles Walker

Thesis submitted in fulfilment of the requirements of the UCL degree of Doctor of Philosophy

2020

I, Tim Walker, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Signed:  ----------

Abstract

There has been a void in the discovery and development of new antibiotic classes over the past four decades due, in part, to the traditional bioprospecting pipeline becoming inefficient from high compound rediscovery rates and high costs. The need for new antibiotic classes is urgent as antimicrobial drug resistant infections are now a major public health concern. Strategies such as exploring novel environments, use of next-generation sequencing, and metagenomics may reduce rediscovery rates and costs which could help accelerate lead discovery and encourage greater participation in bioprospecting.

Whole genome-sequencing and analysis was used to characterise four bacterial strains (Y1-4) isolated from raw honey that were shown to have antibiotic activity. The isolates were identified as *Bacillus* and were closely related but distinctive strains with variations amongst their secondary metabolite profiles. All isolates contained a gene cluster homologous to AS-48, a circular bacteriocin produced by *Enterococcus faecalis*, which has broad-spectrum antibiotic activity. To date, no example of this bacteriocin has been reported in *Bacillus*. This work demonstrated the value of whole-microbial genome sequencing for dereplication.

A pipeline for the low-cost sequencing and assembly of bacterial genomes using Oxford Nanopore MinION was developed in order to produce contiguous and accurate genome assemblies for taxonomic and bioprospecting analysis. The pipeline developed used a combination of Nanopore draft assembly by Canu and polishing with RACON and Nanopolish, with final polishing with Illumina reads using Pilon. The Nanopore-only assembly of *Streptomyces coelicolor* A3(2) produced was contiguous and covered 98.9 % of reference. AntiSMASH analysis identified the full secondary metabolite profile of the genome through homology searches. However, indel rates were high (66.82 per 100 kbp) causing fragmented gene annotations which limited secondary metabolite structure prediction. Illumina read polishing reduced indels (2.03 per 100 kbp) and enabled accurate structure prediction from the identified biosynthetic pathways. This demonstrates that Nanopore sequencing can provide a viable dereplication strategy by detection of known biosynthetic pathways. Additionally, supplementation with Illumina sequencing can allow for structure prediction of biosynthetic pathways which could inform chemical extraction strategies for novel pathways.

Nanopore sequencing was further utilised to characterise an antibiotic producing isolate (KB16) active against methicillin-resistant *Staphylococcus aureus* and vancomycin-resistant *Enterococcus* from the hot spring of the Roman Baths, UK. Genomic analysis showed KB16

to be highly related to *Streptomyces canus* and to contain 26 putative secondary metabolite gene clusters - some of which were potentially novel. One of the gene clusters was identified as encoding the antibiotic albaflavenone. Attempts to chemically identify the antibiotic produced by KB16 showed that it may produce multiple antimicrobial compounds. These findings demonstrate the value in prospecting underexplored environments such as the Roman Baths for microbially-derived antimicrobial leads.

A PCR screen was used to amplify NRPS and PKS gene fragments from a human oral metagenome. Analysis of the fragments suggested that some are from uncharacterised gene clusters. Nanopore shotgun metagenomic sequencing was used to profile the water of the Roman Baths which revealed a diverse microbiome of species with reported metabolic characteristics that are in keeping with the known geochemistry of the waters and aligned with 16S rRNA analysis. Further analysis also identified putative heavy metal resistance genes which can be a co-marker for their metabolism and aligned with the chemical properties of the water. These findings demonstrate the potential value in these sites for bioprospecting whilst also giving insight that can inform bioprospecting strategies. The investigations also highlight the utility of Nanopore sequencing for taxonomic and functional gene profiling of environmental microbiomes.

In combination these findings have all contributed information on novel environments, potential isolate leads, and cost-efficient methodologies to accelerate the discovery of microbially-derived antibiotics.

Impact Statement

Discovery of new antibiotic classes has stalled over the past four decades. This has been caused in part by the 'traditional' model of bioprospecting becoming less economical from high rediscovery rates and resource costs. During this time, antimicrobial resistant infections have continued to proliferate and are a major global public health concern. Leading to warnings from public health authorities of the risk of a 'post-antibiotic age' where current antibiotics have lost all effectiveness. The potential costs of such a situation to the global economy, public health, and quality of life are estimated to be high. Therefore, the need for the development of new antibiotic classes with novel modes of action is urgent.

To fill the discovery void more novel antibiotic lead compounds need to be discovered. By utilising new technologies and approaches that may mitigate the issues that led to the decline in bioprospecting then more research groups and industry may begin searching again.

The work in this thesis has detailed attempts to discover novel bacterial isolates producing antibiotic compounds and develop strategies to improve the bioprospecting pathway using Oxford Nanopore long-read sequencing and genome-mining for dereplication.

To this end, five bacterial isolates (four from honey and one from the hot springs of the Roman Baths) with antibiotic activity were characterised. The analysis identified aspects of the isolates phylogenetics and secondary metabolism profiles. Which revealed that the isolates may have biosynthetic gene clusters for potentially novel compounds. The impact of this is that it forms the basis for further isolation of these compounds.

A pipeline for the whole-genome sequencing and annotation of isolates using Oxford Nanopore technology was developed and validated in the analysis of a novel isolate from the Roman Baths, UK. This pipeline can allow researchers to produce accurate genome assembles to dereplicate and prioritise lead isolates quickly and efficiently, and thus make bioprospecting an economical and viable strategy for more researchers to perform. The results of the analysis revealed the discovery of known antibiotic compounds and highlighted how taxonomy alone is not a marker of expected secondary metabolite potential. Genome-mining also revealed the presence of gene pathways for AS-48 bacteriocin in the *Bacillus* isolates from honey. This is the first report of this compound in this genus and further highlights the disconnect between taxonomy and biosynthetic potential. This is an important finding for bioprospectors to ensure that their dereplication strategies do not discount lead isolates wrongly.

In addition, the microbiome of the Roman Baths, UK was profiled using Oxford Nanopore sequencing. This is a UNESCO world heritage site of significant cultural and geological importance, and this was the first survey of its kind to be performed on this site. The findings revealed microbial taxa not previously reported to be present in the waters and gives a basis by which researchers can plan studies into the site from both a bioprospecting and ecological point of view. The Oxford Nanopore sequencing pipeline used to profile the microbiome also revealed evidence for the presence of metal resistance genes, which demonstrates the potential of the pipeline for functional gene profiling of the site.

Table of Contents

List of Figures

List of Tables

Abbreviations

%	Percentage
°C	Degrees Celsius
©	Copyright
®	Registered trademark
AMR	Antimicrobial resistance
ANI	Average nucleotide identity
ATCC	American Type Culture Collection
BAC	Bacterial Artificial Chromosome
BBSRC	Biotechnology and Biological Sciences Research Council
BGC	Biosynthetic gene cluster
BHI	Brain Heart Infusion
bp	Base pair
CFU	Colony forming units
CLIMB	Cloud Infrastructure for Big Data Microbial Bioinformatics
DDH	DNA-DNA hybridisation
Dha	Dehydroalanine
Dhb	Dehydrobutyrine
DNA	deoxyribonucleic acid
DSM	Deutsche Sammlung von Mikroorganismen (*German Collection of Microorganisms*)
eDNA	Environmental deoxyribomnucelic acid
EDTA	Ethylenediaminetetraacetic acid
g	Grams
Gbp	Gigabase pair
GC	guanine-cytosine
gDNA	Genomic deoxyribonucleic acid
GmbH	Gesellschaft mit beschränkter Haftung (*German company with limited liability*)
HMW	High molecular weight
HPLC	High pressure liquid chromatography
Inc.	Incorporated Company
IPTG	Isopropyl β- d-1-thiogalactopyranoside
kb	Kilobase pair
kDa	Kilodalton
L	Litre
LAPs	Linear azol(in)e-containing peptides
LB	Lysogeny broth
LCA	Lowest common ancestor
Ltd.	Limited Company
M	Mole
MAG	Metagenome assembled genome
Mbp	Megabase pair
MDA	Multiple displacement amplification

MDR	Multidrug resistant
mg	Milligrams
MIBiG	Minimum Information about a Biosynthetic Gene cluster
mL	Millilitres
mm	Millimetres
mM	Millimoles
mRNA	Messenger ribonucleic acid
MRSA	Methicillin resistant *Staphylococcus aureus*
MSSA	Methicillin sensitive *Staphylococcus aureus*
NB	Nutrient Broth
NCBI	National Center for Biotechnology Information
NCTC	National Culture and Tissue Collection
ng	Nanograms
nm	Nanometres
nM	Nanomoles
NMR	Nuclear magnetic resonance
NRP	Non-ribosomal protein
NRPS	Non-ribosomal protein synthase
O/N	Overnight
OLC	Overlap layout consensus
ONT	Oxford Nanopore Technology
ORF	Open reading frame
PABA	Paraaminobenzoate
PBP	Penicillin binding protein
PBS	Phosphate buffered solution
PCR	Polymerase chain reaction
PDR	Pan-drug resistant
pH	Potential of Hydrogen
PK	Polyketide
PKS	Polyketide synthase
PVDF	Polyvinylidene difluoride
QC	Quality control
Rf	Retention factor
RiPP	Ribosomal peptide product
RNA	ribonucleic acid
RP-SPE	Reverse phase solid-phase extraction
rpm	Revolutions per minute
rRNA	Ribosomal ribonucleic acid
Secs	Seconds
SFM	Soya flour media
SNP	Single nucleotide polymorphism
SPRI	Solid Phase Reversible Immobilization
TAE	Tris-acetate-EDTA

TBA	Tryptic soya broth
TE	Tris-EDTA
TLC	Thin-layer chromatography
TM	Trademark
tRNA	Transfer-ribonucleic acid
UK	United Kingdom
US$	United States Dollar
USA	United States of America
UV	Ultraviolet
V	Voltage
v/v	Volume/volume
VRE	Vancomycin resistant *Enterococcus*
WHO	World Health Organisation
XDR	Extremely drug resistant
λ	Lambda
µg	Micrograms
µL	Microliter
µM	Micromoles

Chapter 1.

Introduction

1.1. Antibiotics and their Modes of Action and Resistance Mechanisms

Antibiotics are chemical agents which either kill or prevent the replication of bacteria. The discovery and usage of antibiotics has revolutionised modern medicine and quality of life around the world. Effective treatments of serious bacterial infections have contributed to an increase in life expectancy and reduction of infant mortality. Additionally, the use of antibiotics to prevent infections after surgical procedures has improved patient outcomes and helped pave the way for advances in the field, such as organ transplantation. Antibiotics elicit their effect upon bacterial cells through targeting and disrupting an aspect of primary metabolism or an essential component of the cellular structure. The particular target and method by which it causes this disruption varies but, for current antibiotics in clinical usage, these can be broadly classified into the following mechanisms; membrane disruption, inhibition of nucleic acid synthesis, inhibition of protein synthesis, inhibition of cell wall synthesis, or inhibition of folic acid synthesis. Additionally, antibiotics can be subclassified into either bacteriostatic or bactericidal. Bacteriostatic antibiotics arrest bacterial cell replication but otherwise do not directly affect viability. The effect of bacteriostatic antibiotics agents tends to be reversible after removal or degradation. In contrast, bactericidal antibiotics illicit an effect that is lethal to the bacterial cell.

1.1.1. Inhibition of Cell Wall Synthesis

The core of the bacterial cell wall is a crystal lattice structure formed of linear chains of cross-linked peptidoglycan which provides protection and rigidity to the bacterial cell. Peptidoglycan is formed of alternating units of *N*-acetylglucosamine and *N*-acetylmuramic acid connected by a β-(1,4)-glycosidic bond. A peptide chain of between 4-5 amino acids is attached to the *N*-acetylmuramic acid units and the lattice structure is formed through the cross-linkages between these peptide chains (Vollmer, Bianot & De Pedro, 2008). Composition of the peptide chains can vary between species; for example it is L-alanine, D-glutamic acid, meso-diaminopimelic acid, and D-alanine in *E.coli*. And L-alanine, D-glutamine, L-lysine, and D-alanine in *Staphylococcus aureus*.

Growth and division of bacterial cells requires formation of new cross-links between the nascent peptidoglycan polymers. Some antibiotic classes inhibit the formation of these cross-links and thus inhibit cellular growth. β-lactam antibiotics, such as penicillin, bind to and inhibit the activity of the enzyme, penicillin binding protein (PBP), which catalyses the formation of cross-links between the peptidoglycan polymers (Yocum, Rasmussen & Strominger, 1980). Whilst glycopeptides, such as vancomycin, form hydrogen bonds with the

D-alanyl-D-alanine residues at the terminal of the *N*-acetylmuramic acids unit peptide chains, thus blocking the formation of cross-linkages between them (Watanakunakorn 1984).

1.1.2. Membrane Disruption

The membrane of bacterial cells is comprised of a fluid bilayer of phospholipid with the lipid 'tails' of the phospholipids arranged to form an internal hydrophobic layer sandwiched by the phosphate head units. The membrane serves to provide structure, protection, and regulation of osmotic pressure within the cell. The structural integrity of the membrane is therefore vital to cell survival. Peptide antibiotic molecules in clinical usage such as gramicidin and colistin, as well as many other small peptides with antibiotic activity, such as bacteriocins, disrupt this membrane by penetration into the lipid bilayer. Exposure to the hydrophobic environment of the internal lipid layers forces conformational changes in the peptides tertiary structure to form pore channels that span across the membrane. This causes a loss of osmotic pressure within the cytoplasm and cellular death (Hancock 1997).

1.1.3. Inhibition of Nucleic Acid Synthesis

Bacterial nucleic acid synthesis is inhibited by several classes of antibiotics which target different enzymes vital to the synthesis of RNA and DNA. Replication of DNA requires the activity of topoisomerase and gyrase enzymes which stabilise the template chromosome from tension as the supercoiled structure of the DNA is unwound to allow access of the polymerase machinery. The type I topoisomerases reduce helix negative supercoiling through cleavage of one strand and formation of a phosphoester bond between the exposed 5' end and a tyrosine residue within the enzyme. This stabilises the strand to allow for the uncut strand to then move between the single-strand break before being resealed. In contrast, gyrases increase negative supercoiling, by forming complexes with the DNA and catalysing the double strand cleavage and reannealing. Fluoroquinolones bind to gyrases when in complex with the DNA, inhibiting the enzymes ability to catalyse the reannealing of cleaved strands (Blondeau 2004). Rifamycin inhibits synthesis of RNA by binding with high affinity to RNA polymerase, creating steric clash between the polymerase active site and RNA substrates (Campbell et al., 2001).

1.1.4. Inhibition of Protein Synthesis

Proteins essential to primary metabolism are synthesised by the ribosome-mRNA complex. Many classes of antibiotics target sites of this ribosomal complex to inhibit translation of the mRNA template. For example, tetracyclines inhibit access of amino acid carrying tRNA to the A site by steric hindrance caused by binding to the 16S rRNA of the 30S ribosomal subunit (Yoneyama & Katsumata, 2006). Aminoglycosides also interact with the 16S rRNA region via hydrogen bonds leading to premature termination of translation (Kapoor, Saigal & Elongavan, 2017). Macrolides, lincosamides, streptogramins B, and chloramphenicol interact with the 50S subunit of the ribosome complex. This disrupts normal passage of the nascent peptide through the ribosome and early termination of translation.

1.1.5. Inhibition of Folic Acid Synthesis

Folic acid serves as a cofactor in nucleotide synthesis and thus is essential to continued growth of the cell. The synthesis of folic acid by the cell is through complex metabolic pathways which presents many targets for antibiotic activity. Sulphonamides inhibit the folic acid biosynthesis pathway at the point of conversion of paraaminobenzoate (PABA) to dihydropteroate by acting as a competitive inhibitor to the dihydropteroate synthase enzyme which catalyses this reaction (Bauman 2015).

1.1.6. Biochemical Mechanisms of Antibiotic Resistance

Some bacterial species may have resistance to the toxic effects of an antibiotic, and these resistance mechanisms generally fall within three groupings; 1) reduction of intracellular drug concentrations, 2) modification of antibiotic target, 3) inactivation of antibiotic. The resistance may be intrinsic due to innate physiological or biochemical aspects of the bacterial cell or acquired due to genetic plasticity (discussed more in Section 1.3). Resistance mutations may result in a ‘fitness cost’ to the bacterium which can manifest phenotypically as higher energy expenditure, lower growth rate or lower virulence. However, later secondary mutations that mitigate these costs can occur and is important for the establishment of resistant bacteria (Schulz et al., 2010). For example, mutations in *rpoA* and *rpoC* have been noted to reverse fitness costs associated with *rpoB* mutations in Rifampin-Resistant *Mycobacterium tuberculosis* (de Vos et al., 2013).

1.1.6.1. Reduction in Intracellular Concentrations

Membrane permeability is a major cause of the intrinsic resistance that Gram-negative bacteria exhibit against certain classes of antibiotics. The outer membrane of Gram-negatives contains saturated lipopolysaccharides that are tightly packed which reduces the permeability of the membrane (Cox & Wright, 2013). This is exemplified by the glycopeptides and β-lactams which inhibit growth by disruption of the formation of peptidoglycan cross-linking in Gram-positives. However, these drugs cannot access their targets at the cell wall in Gram-negatives due to the presence of the lipopolysaccharide outer membrane that encapsulates the peptidoglycan cell wall.

Alterations in cell wall thickness have been associated with increased resistance of *Staphylococcus aureus* to vancomycin (Cui et al., 2003). This is hypothesised to be caused by vancomycin molecules becoming clogged in the thickened peptidoglycan due to the increased d-alanyl-d-alanine residue content of the cell wall serving as a false target (Cui et al., 2006).

Protein channels, known as porins, which span Gram-negative membranes and facilitate diffusion of molecules between the cytoplasm and environment are a means of entry for some antibiotic drugs. Down regulation of porin channels results in reduced susceptibility to carbapenems and cephalosporins in clinical *Enterobacteriaceae* (Cornaglia et al., 1996). Additionally, mutations which alter the structure and selectivity of the porin channel can also reduce the uptake of drug molecules and has been seen to cause resistance to β-lactams and tetracycline in *Neisseria gonorrhoeae* (Gill et al., 1998).

Formation of biofilms by infecting pathogens such as by *Pseudomonas aeruginosa* in lung infections can also confer resistance or increased drug tolerance by reducing access of antibiotics to the cell. For example, Tseng and co-workers (2013) used florescent microscopy to demonstrate that tobramycin is sequestered at the edges of *Pseudomonas aeruginosa* biofilms.

Active expulsion of the antibiotic molecules from the cytoplasm via efflux pumps is another mechanism by which intracellular drug concentration is reduced. All bacteria have various varieties of efflux pump mechanisms which span the cellular membranes. Specificity of the pumps can be for one specific compound, or broader in range which can contribute to multidrug resistance (Webber & Piddock, 2003). The presence of innate efflux pumps encoded on the chromosome contributes to the intrinsic resistance to some antibiotic classes of Gram-negative bacteria. For example, the intrinsic resistance of *Pseudomonas aeruginosa*

to tetracycline, chloramphenicol, and norfloxacin is associated with its efflux systems (MexAB-OprM, MexCD-OprJ, MexEF-OprN and MexXY-OprM) (Poole, 2000). Efflux pump genes have also been found on mobile genetic elements which leads to the acquisition of efflux mediated resistance mechanisms in new species (Blair et al., 2015). Overexpression of efflux systems can also lead to acquired resistance such as in the case of *Pseudomonas aeruginosa* overexpressing the multidrug MexAB efflux system causing reduced susceptibility to a range of antibiotics; β-lactams, chloramphenicol and trimethoprim as well as to some common biocides such as triclosan (Chuanchuen et al, 2001).

1.1.6.2. Modification of Antibiotic Target

Mutations that cause a structural modification to an antibiotic target to reduce the affinity by which the antibiotic binds can confer resistance to that antibiotic. For example, mutations in DNA gyrase genes (*gyrA*, *gyrB*) confers resistance in clinical *E. coli* against fluoroquinolones, agents which work by disrupting DNA replication by binding to this protein (Redgrave et al., 2014). Genes on mobile genetic elements which encode for modified forms of the protein targets can also be acquired. Such as *mecA*, which encodes for Penicillin Binding Protein (PBP 2a) which has a lower affinity for β-lactams and confers high-level resistance to methicillin in *Staphylococcus aureus* (Wielders et al., 2002). Sulphonamide resistance has been observed in Gram-negative bacteria through the acquisition of a plasmid encoded structural variant of the drug target dihydrofolate reductase enzyme (Sköld, 2001).

1.1.6.3. Inactivation of Antibiotic

Production by bacteria of enzymes which degrade the antibiotic drug molecule confer resistance to the drug. The most prominent example of this are the β-lactamases, a large and diverse collection of enzymes which cleave the β-lactam ring of clavams, carbapenems, monobactams, and cephalosporin antibiotics. Hundreds of β-lactamases have been isolated which target different individual β-lactam drugs to differing extents (Bush & Jacoby, 2010). Extended-spectrum β-lactamases (ESBLs), which are plasmid-mediated and capable of hydrolysing most β-lactam drugs are currently of serious clinical concern (Rawat & Nair, 2010). Carbapenemases are another class of β-lactamases able to hydrolyse many of β-lactam drugs. Critically this class can also inactivate carbapenem which is resistant to ESBLs (Queenan & Bush, 2007). Carbapenem have a broad-spectrum activity against Gram-positive and Gram-negative infections and have been used as a "drug of last resort" in treatment of patients with serious MDR infections (Papp-Wallace et al., 2011). The development of β-lactamase inhibitor enzymes which inactivate β-lactamases, along with their co-administration

with β-lactam antibiotics has been used as a strategy to overcome β-lactamase mediated resistance (Tooke et al., 2019). But it has been found that some β-lactamases, such as TEM-1 or TEM-2 β-lactamases, can develop resistance to these compounds by mutation and can also develop into complex mutants with resistance to extended-spectrum cephalosporins (Cantón et al., 2008). Transfer and acquisition of β-lactamase genes encoded on mobile genetic elements between species has been a major contributing factor to emergence of multidrug resistant *Klebseilla pneumoniae*, *Escherichia coli*, *Pseudomonas aeruginosa* and *Acinetobacter baumannii* isolates which are difficult to treat effectively (Blair et al., 2015).

Inactivation of antibiotics by modification of the molecule through transfer of a functional group which reduces the target binding efficiency also confer resistance. Enzymatic modification of aminoglycosides exposed through transfer of acetyl moieties, phosphorus, or nucleotides to hydroxyl and amine groups has been shown to confer high level resistance (Ramirez & Tolmasky, 2010; Blair et al., 2015).

1.2 Microbially-Derived Antibiotics

Microorganisms are the oldest, most ubiquitous, and populous form of life on Earth. They populate environments as diverse as freshwater lakes, oceans, the subterranean surface, air, soil, animals, and plants (Whitman, Coleman & Wiebe, 1998). The term used to describe the microorganisms that populate a particular environment is the microbiome (soil microbiome, human microbiome, insect microbiome). To facilitate survival in complex ecosystems, microorganisms produce small molecules known as secondary metabolites or 'Natural Products' that are used, amongst other things, to suppress competing organisms (Challinor & Bode, 2015). The ability of microbes to secrete compounds that can kill pathogenic bacteria was first reported by Fleming (1929), who famously chanced upon lysed *Streptococcus* colonies surrounding a *Penicillium* contaminant on an agar plate. However, a systematic search for antibiotic compounds from microbes had begun two years previously by Renné Dubos, a soil microbiologist at the Rockefeller Institute (Van Epps, 2006). This led to the discovery of a bactericidal extract from *Bacillus brevis* named tyrothricin (Dubos, 1939). This extract was discovered to be a mixture of the two polypeptides, gramicidin and tyrocidine (Hotchkiss & Dubos, 1941). The toxicity of the peptides limited the usefulness of the tyrothricin mixture against systemic infections. However, the gramicidin component was successfully developed into a topical ointment for the treatment of infected wounds and is still an active ingredient in some topical antibiotics today. Gramicidin therefore became the first commercial antibiotic discovered from the systematic screening of microbes, and this spurred on the commercial development of Fleming's initial observations into penicillin in the 1940s,

(Van Epps, 2006). A period known as the 'Golden Age' of antibiotic discovery then followed, as scientists began systematically searching for, and discovering, novel classes of antimicrobial compounds through a process of bioassay-led screening of environmental microorganisms known as bioactivity-guided isolation (**Figure 1.1**). The principles of bioactivity-guided isolation are described in more depth in **Section 1.3.** To date, eighteen classes of antibiotics have been discovered from microorganisms that have subsequently been developed into commercial drugs (**Table 1.1**). Ubiquitous soil-dwelling microorganisms in the genera *Streptomyces* and *Bacillus* have been found to be prolific producers of antibiotic leads, and traditional bioassay-led screening programmes have tended to focus on the soil microbiome.

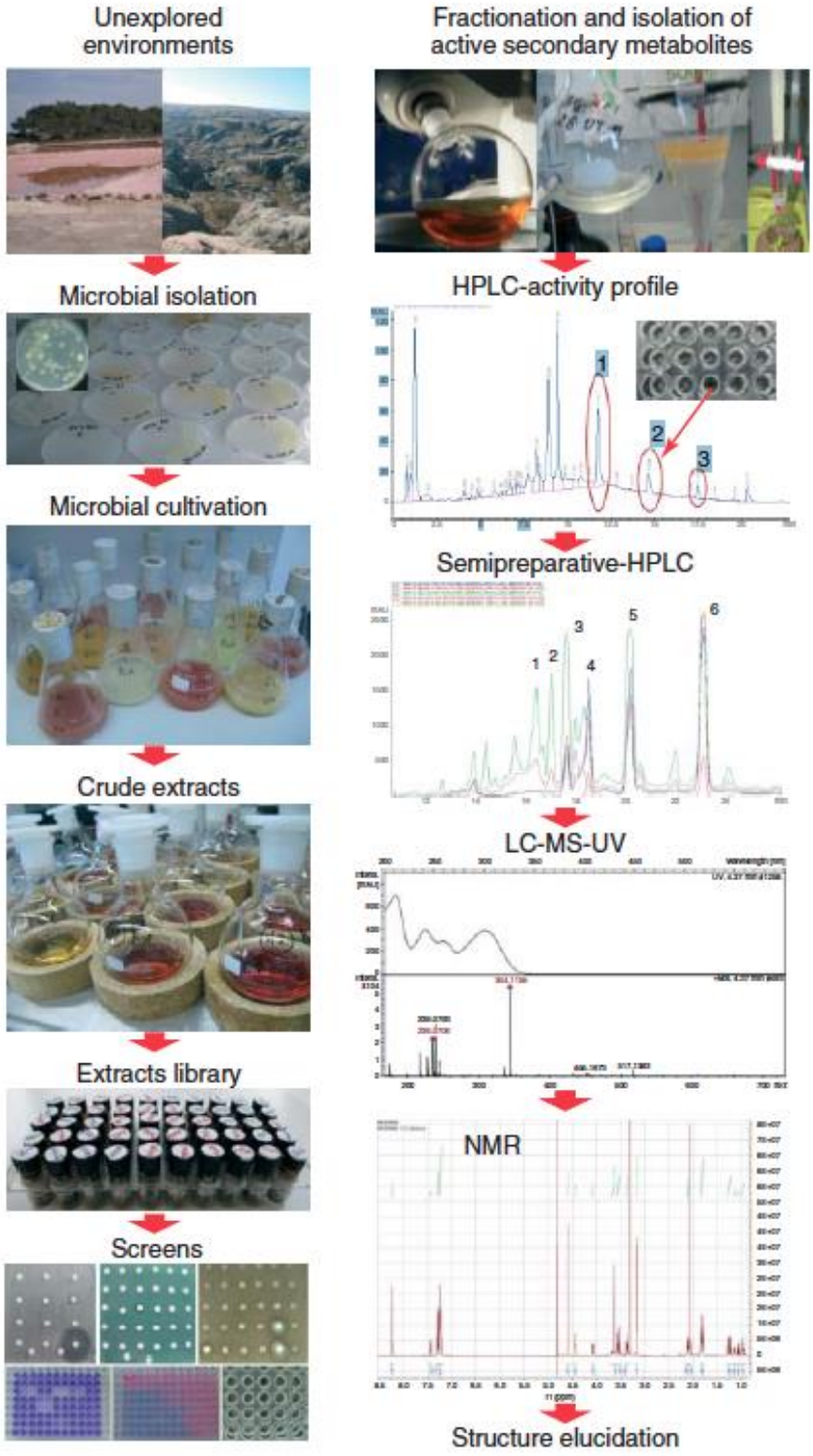


Figure 1.1. Schematic of the 'traditional' bioprospecting pipeline for microbially-derived antibiotics, known as bioactivity-guided isolation. Image taken from Gualerzi et al., (2013). Microorganisms are cultivated from an environmental source (such as soil) and fermented in culture media to create crude extracts. The extracts are screened for antibiotic activity using a bioassay. The active component of the extract is isolated, and its structure is determined.

Table 1.1. Commercial antibiotic classes derived from microbial sources, year the class was first reported, the first reported example and producing organism, the mechanism of action, and the biosynthetic class.

Antibiotic Class	First Reported	Example – Producing Organism	Mechanism of Action	Biosynthesis Class	References
β-Lactams	1929	Penicillin - *Penicillium chrysogenum* (originally identified as *P. rubrum*)	Inhibiting bacterial cell wall synthesis by inhibiting formation of peptidoglycan cross-links	Non-ribosomal protein	Fleming, 1929; Smith et al., 1990; Gualerzi et al., 2013
Polypeptides	1939	Gramicidin – *Bacillus brevis*	Increasing cell membrane permeability by penetrating membrane and forming channel pores	Non-ribosomal protein	Dubos 1939; Hotchkiss & Dubos 1941; Roskoski et al., 1970; Hancock 1997
	1945	Bacitracin – *Bacillus licheniformis*	Inhibiting bacterial cell wall synthesis by interaction with a peptidoglycan carrier molecule	Non-ribosomal protein	Johnson et al., 1945; Toscano & Storm, 1982
Aminoglycosides	1944	Streptomycin - *Streptomyces griseus*	Inhibiting protein synthesis by binding to 30S ribosome subunit	Other	Schatz et al., 1944; Gromadski & Rodnina 2004; Kudo & Tadashi, 2009

Amphenicols	1947	Chloramphenicol - *Streptomyces venezuelae*	Inhibiting protein synthesis by binding to 50S ribosome subunit.	Non-ribosomal protein	Ehrlic et al., 1947; Wolfe & Hanh, 1965; He et al., 2001
Tetracyclines	1948	Aureomycin (Chlortetracycline) - *Streptomyces aureofaciens*	Inhibiting protein synthesis by binding to the 16S RNA component of ribosome	Type II polyketide	Dugger, 1948; Nelson & Levy, 2011; Zu et al., 2013
Lipopeptides	1950	Colistin - *Paenibacillus polymyxa* (formerly *Bacillus polymyxa* var. *colistinus*)	Increasing cell membrane permeability by penetrating the membrane and forming channel pores	Non-ribosomal protein	Kawahara et al., 1997; Hancock 1997; Tambadou et al., 2015
Tuberactinomycins	1951	Viomycin - *Streptomyces vinaceus*	Inhibiting protein synthesis by binding to the 70S ribosome complex	Non-ribosomal protein	Finlay et al., 1951; Thomas et al., 2003; Stanley et al., 2010
Oxazolidinones	1952	Cycloserine – *Streptomyces* K-300	Inhibiting *Mycobacterium* cell wall synthesis by binding to D-alanine Ligase	Other	Kurosawa 1952; Kumagai et al., 2010; Prosser et al., 2013
Macrolides	1952	Erythromycin - *Saccharopolyspora erythraea (*formally *Streptomyces erythraeus)*	Inhibiting protein synthesis by binding to the 50S ribosomal subunit	Type I polyketide	McGuire et al., 1952; Labeda, 1987; Rawlings, 2001; Tenson 2003
Coumarins	1955	Novobiocin - *Streptomyces niveus*	Inhibiting DNA synthesis by binding to Gyrase B	Other	Hoeksema 1955; Steffensky et al., 200; Tse-Dinh 2007
Glycopeptides	1955	Vancomycin – *Amycolatopsis orientalis* (formally *Streptomyces orientalis*)	Inhibiting second stage of cell wall synthesis	Non-ribosomal protein	Griffith & Peck 1955; Watanakunakorn 1981; Xu et al., 2014
Streptogramins	1960	Virginiamycin - *Streptomyces virginiae*	Two component antibiotic that inhibits protein synthesis by binding to the 30S ribosomal subunit	Non-ribosomal protein polyketide hybrid	Barnhart et al., 1960; Cocito et al., 1985; Pulsawat et al., 2007
Ansamycins	1960	Rifamycin - *Amycolatopsis rifamycinica* (formally *Streptomyces mediterranei*)	Inhibiting RNA synthesis by binding to RNA Polymerase	Type I polyketide	Margalith & Beretta, 1960; August et al., 1998; Campbell et al., 2001; Bala et al., 2004

Fusidanes	1962	Fusidic Acid – *Fusidium coccineum*	Inhibiting protein synthesis by inhibiting elongation of nascent peptide on the ribosome.	Other	Godtfredsen et al., 1962; Godtfredsen et al., 1968; Gudkov, 2001.
Lincosamides	1964	Lincomycin - *Streptomyces lincolnensis* var*. lincolnensis*	Inhibiting protein synthesis by binding to 50S ribosomal subunit	Other	MacLeod et al., 1964; Tenson, 2003; Koberská, 2008
Phosphonic Acid	1969	Fosfomycin – *Streptomyces fradia, S. viridochromogenes & S. wedmorensis*	Inhibits cell wall synthesis by binding to phosphoenolpyruvate synthetase	Other	Hendlin et al., 1969; Woodyer et al., 2006; Argyris et al., 2011
Pseudomonic acids	1971	Mupirocin - *Pseudomonas fluorescens*	Inhibiting protein synthesis by binding to isoleucyl t-RNA synthetase.	Type I polyketide	Fuller et al., 1971; Hughes & Mellows, 1978; Gao et al., 2014
Lipopeptides	1987	Daptomycin – *Streptomyces roseosporus*	Disrupts functioning of the cell membrane and cell division. The compound aggregates in the cell membrane which distorts membrane structure and leads to redirection of the localisation of cell division proteins.	Non-ribosomal protein	Allen et al., 1987; Miao et al., 2005; Pogliano et al., 2012

1.3. Principles of Bioactivity Guided Isolation

Natural products with properties of commercial interest, such as antibiotics, have traditionally been discovered from environmental microorganisms through a process of bioactivity-guided isolation (summarised in **Figure 1.1**) (Gualerzi et al., 2013). After isolation and cultivation from the environment, the microorganisms are screened for the desired trait by means of a bioassay to identify leads for further investigation. A methodical process is then used to purify the natural product responsible for the activity from an often-crude microbial fermentation through a process of repeated fractionation (Sterner, 2012). The fractions are tested at each iteration for bioactivity to identify and purify the one containing the active compound (Heinrich et al., 2004).

Initially isolates are screened for antibiotic activity by means of a simple and cost-effective assay (Gualerzi et al., 2013). The cross-streak assay or spot-on-lawn assays are whole-cell plate-based bioassays whereby an indicator strain is directly challenged by the test isolate through co-incubation. Antibiotic activity by the test isolate against the indicator strain is then assessed by visible inhibition of growth in the region of the agar plate nearest to the test isolate. These assays are advantageous in that they are relatively quick, simple, and cost-effective to perform. The assays are also easy to reproduce and interpret. The assays can also be adapted by using differing indicator strains to screen for isolates with different antibiotic activity profiles. However, these assays are also relatively labour intensive and not well suited to high-throughput screens (Balouiri, Sadiki & Ibnsouda, 2016).

After lead isolates are identified, then a crude fermentation extract is made for bioactivity-guided isolation of the natural product. The generation of the crude microbial extract can be obtained by various means. The most common is to cultivate the organism in a liquid medium to allow the generation of biomass and the subsequent production of the natural product of interest which is secreted into the liquid (Gualerzi et al., 2013). This method is often favoured due to relative ease in which liquid fermentation production can be upscaled and the relative ease by which the microbial biomass can be filtered from the liquid prior to natural product extraction (Sterner 2012). Alternatively, microbes can be grown on solid agar and after suitable cultivation of the microbe, the compounds secreted into the agar can be extracted. This method was adopted by Qin and co-workers (2017) in the isolation and identification of the novel antifungal compound formicamycin from *Streptomyces formicae* F5, which is a novel species of bacteria that the researchers had previously isolated and identified as having antibiotic activity. Cultivating the bacteria on solid media can present a challenge in separating it from the agar prior to extraction. It is often desirable to do this to avoid an excessively

complex crude extract containing cellular material that can complicate subsequent isolation steps (Sterner, 2012).

Environmental isolates may often not be tractable in laboratory conditions, and generation of bioactive material of sufficient quantity for investigation can be a challenge. Often finding the optimal media and growth conditions can be a process of trial and error involving differing media formulations, conditions, and timepoints. Often these different methods are trialled on a small scale to find one that works before upscaling. This can be a time-consuming and resource intensive process which may yield no return (Heinrich et al., 2004, Gualerzi et al., 2013).

The initial extraction of compounds from the fermentation will usually by done by mixing the fermentation broth or agar with immiscible solvents of differing polarity. Compounds in the fermentation will then be separated based on polarities. This approach is favoured because of its simplicity and to minimise any potential degradation of compounds. In the case of microbial fermentations, a mid-polar ethyl acetate is often used as an extraction solvent. This is because the extraction is targeting compounds which have been secreted into the aqueous culture media and so are generally unlikely to be efficiently extracted using a highly non-polar solvent. However, sequential use of solvents of increasing polarity, starting with a highly non-polar solvent such as hexane, can also be performed. While this increases workload and the number of extracts required for subsequent bioassay testing, it does also serve as a simple first-stage separation and partitioning step to yield less complex extracts. This could give the benefit of simplifying further purification steps of the active extracts (Heinrich et al., 2004).

Extracts are tested for antibiotic activity to confirm that they contain the compound of interest. For the case of antibiotic compounds this is most commonly done by use of a disc diffusion assay against indicator strains. Fractions containing the antibiotic activity are identified by observation of a zone of inhibition around the disc treated with the extract. This assay is relatively cheap, straightforward to perform and reproduce, and the results can be interpreted easily by visual inspection (Balouiri, Sadiki & Ibnsouda, 2016).

Fractionation of active extracts can be performed using a variety of different chromatography techniques which can separate compounds based on different physicochemical characteristics such as size, ionic-charge, polarity, or solubility (Sterner, 2012). The method(s) researchers choose may be determined by what types of compounds they are seeking to isolate, the amount of input material available, and operational considerations such as resource availability and throughput (Heinrich et al., 2004, Gualerzi et al., 2013).

The purified compound can then be identified by structural elucidation often with analytical chemistry techniques such as mass-spectrometry or nuclear magnetic resonance (Reynolds, 2017). Determining the structure of the compound can allow researchers to learn if the compound is novel. If it is a novel compound, then studies can continue to try and convert this compound into a viable drug for administration to a patient. This involves studies into mechanism of action, pharmacokinetics, toxicity, manufacturing, clinical trials, and regulatory approvals (Heinrich et al., 2004).

Bioactivity-guided isolation has, in some form or another, been the standard 'traditional' approach to bioprospecting for microbially-derived antibiotics since its earliest days of Dubos and Waksman nearly a century ago (Gualerzi et al., 2013). And has been a very successful approach, with all classes of microbially-derived antibiotics in clinical use having been discovered through utilisation of this approach in some form. However, a bioactivity-guided isolation approach alone does contain drawbacks which are commonly regarded to have contributed to decline in bioprospecting activity by industry and academia from the 1980s onwards. Mainly, that the process can be highly resource, time, and capital intensive. Typically, it can take 12-15 years from crude screen to release of a clinical product, with an estimated lead attrition rate of 1 in 10,000 (Heinrich et al., 2004).

Since the 'Golden Age' of antibiotic discovery, the returns on these investments have been harder to obtain as bioprospecting screens started to return the same known compounds with increasing frequency. It is therefore important to determine if active extracts or fractions may contain already known compounds early in the process so that they can be excluded from further investigation – a process known as dereplication (Heinrich et al., 2004, Sterner, 2012, Gualerzi et al., 2013, Hubert, Nuzillard & Renault, 2017). Researchers in the past have often relied upon data obtained from analytical techniques such as retention times from LC-MS or GC-MS, as well as Mass-Spec and NMR peaks in an attempt to identify known compounds (Hubert, Nuzillard & Renault, 2017). However, these approaches rely upon generation of a sufficient amount of compound and may rely upon optimisation of multiple fractionation steps in order to obtain spectra of sufficient quality to perform such analysis. The method also relies upon searching the literature and multiple chemical databases to ensure that the compounds in question have been thoroughly dereplicated. This still remains a time and resource intensive process and is limited by the comprehensiveness of the database (Hubert, Nuzillard & Renault, 2017).

1.4. Biosynthetic Classes of Major Microbially-derived Antibiotics

1.4.1. Polyketide Synthases and Non-Ribosomal Synthases

As shown in **Table 1.1**, the majority of antibiotic classes that have derived from microbial sources are either non-ribosomal peptides (NRP) or polyketides (PK). These compounds are synthesised by large multi-enzymes called non-ribosomal protein synthases (NRPS) and polyketide synthases (PKS) respectively. The catalytic domains of NRPSs and type-1 PKSs are organised into functional units termed 'modules', with each module responsible for a specific round of chain extension. The sequence of the modules, and of the catalytic domains within them, correlates with the order in which they act, and so they have been dubbed "molecular assembly lines" (Weissman & Müller 2008). Type-II polyketide synthases are aggregates of monofunctional enzymes as opposed to a large multi-enzyme and are more common in fungi than in prokaryotes (Weissman & Müller 2008).

Type-I PKS enzymes catalyse the synthesis of polyketides through step-wise elongation of simple ketone-based subunits such as acetyl-CoA, propionyl-CoA or butyryl-CoA by decarboxylative condensation. This process is analogous to fatty acid biosynthesis, except that the PKS can utilise a wider variety of ketone-based starter and elongation units, and the building blocks are not fully reduced during elongation (Khosla et al., 1999). Further chemical complexity of the nascent polyketide chain is achieved by varying the modification of each elongation intermediate through a fixed sequence of ketoreduction, dehydration, or enoyl reduction catalytic domains (Khosla et al., 1999; Weissman & Müller 2008).

Figure 1.2 outlines the arrangement of a generic type-I PKS and the movement of a nascent polyketide chain through this system. The process begins with a ketone containing starter unit (R1), catalysed by the acyl transferase (AT) domain, being loaded onto the acyl carrier protein (ACP) domain of the loading module. This is then transferred downstream to the ketosynthase (KS) domain of the first elongation module, catalysed by that KS domain. The ketone containing elongation unit (R2) is loaded onto the ACP domain of EM1 by its AT domain. The KS-starter unit and ACP-elongation unit react at the KS-bound end by a Claisen type condensation, resulting in a free KS domain and an ACP-elongated intermediate (R1+R2). This intermediate is then transferred upstream to the KS domain of the second elongation module where it will react with this modules' ACP-elongation unit (R3). The processing domains of EM2 modify the β-ketone functional group of the ACP-bound intermediate (R1+R2+R3). The ketoreductase (KR) domain then reduces the β-ketone to a β-hydroxyl group; the dehydration (DH) domain then removes H_2O from the β-hydroxyl group to create

a α-β-double bond. The enoyl reduction (ER) domain then reduces the double bond to a single bond. The nascent polyketide chain continues to grow as this cycle is repeated for each elongation module until it is released from the PKS by a thioesterase (TE) domain, which hydrolyses it from the downstream ACP domain (Weissman & Müller 2008).

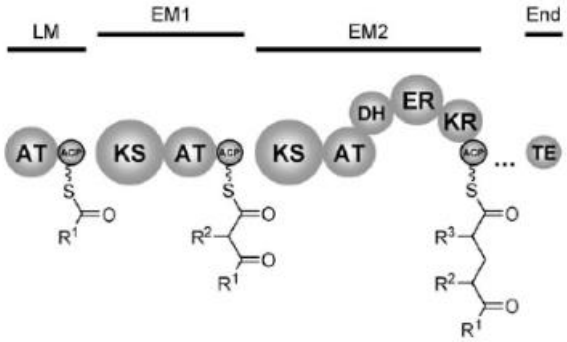


Figure 1.2. Generic outline of type I PKS system. LM = loading module, EM1 = first elongation module, EM2 = second elongation module with processing domains (DH, ER, KR). The type and number of processing domains present within a module can vary and will determine the structural modifications made on the subunit during each step in the synthesis of the nascent product. Adapted from Weissman & Müller (2008).

NRPSs synthesise long chains of amino acids in a similar step-wise fashion to polyketide synthases (**Figure 1.3**). Proteinogenic or non-proteinogenic amino acids can be utilised and the amino acids can be further modified by the presence of processing domains such as reduction (R), epimerase (E), N- and C-methyl- transferase (N/C-MT), and oxidase (Ox) (Schwarzer et al., 2003; Wiesmann & Muller, 2008).

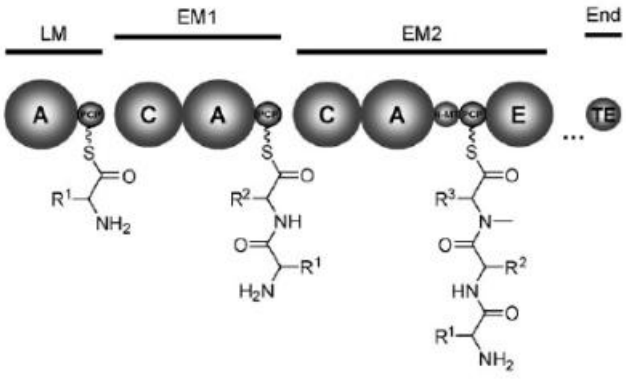


Figure 1.3. Generic outline of NRPS system. LM = loading module, EM1 = first elongation module, EM2 = second elongation module with processing domains. Adapted from Weissman & Müller (2008).

Figure 1.3 outlines the arrangement of a generic NRPS and the movement of a nascent peptide chain through this system. The process begins at the loading module (LM) where the adenylation (A) domain selects a specific amino acid (R1) and uses ATP to activate the amino acid into aminoacyl adenylate and load it onto the peptidyl carrier protein (PCP) domain. The peptide elongation unit (R2) is loaded onto the PCP domain of first elongation module (EM1) by its A domain. Amide bond formation between the starter amino acid and the elongation units is catalysed by the condensation (C) domain of the upstream module, creating a nascent peptide intermediate (R1+R2) in the PCP domain of EM1. The intermediate forms an amide bond with the elongation unit (R3) in the second elongation module (EM2) to create a peptide chain (R1+R2+R3) bound to EM2 PCP domain. EM2 contains some modification domains. The N-methylation (N-MT) catalyses the methylation of R3, and the epimerase (E) domain epimerises R3 into the D-configuration. The nascent peptide chain continues to grow as this cycle is repeated for each elongation module until it reaches the final PCP domain where it then undergoes macrocyclization and release by a thioesterase (TE) domain (Schwarzer et al., 2003; Wiessmann & Muller, 2008).

After the PK and NRP dissociate from the enzyme they are often modified by tailoring enzymes. These tailoring enzymes, along with associated regulation and transport genes, are all clustered together with the NRPS/PKS genes into biosynthetic gene clusters (BGC) (Rutledge & Challis, 2015; Medema & Fischbach; 2015).

1.4.2. Ribosomally Synthesized and Post-Translationally Modified Peptides

Peptide natural products (10 kDa or less) which are synthesised by ribosome machinery and undergo post-translational modifications are a broad class of natural products with diverse structures and characteristics. Due to their breadth, multiple terminology has evolved in the literature to attempt to categorise this class of natural products. Ribosomally synthesized and post-translationally modified peptides (RiPPs) is a term commonly used amongst prominent researchers in the field (Arnison et al., 2013). Multiple further subclassifications of these peptides can be made based either on shared structural features, producing strain, post-translational tailoring enzymes, or specific bioactivity. RiPPs have also been grouped according to their function. Bacteriocins is a term which predates ‘RiPP’ that is also frequently used in literature to describe small ribosomal peptides produced by bacteria that have antibiotic activity. Bacteriocins have also previously been grouped according to their producing organisms such as ‘Microcins’ for antimicrobial peptides from Gram-negative bacteria, or ‘Colicins’ to describe those specifically isolated from *Escherichia coli*. In this

thesis, the nomenclature convention detailed by Arnison et al., (2013) is followed where RiPPs are grouped into families based upon a shared core biosynthetic logic.

RiPPs undergo a common biosynthetic process (**Figure 1.4**). A precursor peptide is synthesised from a structural gene according to the central dogma of molecular biology. This precursor peptide can range in size from between ~20-110 residues and consists of a core peptide segment and upstream leader segment. The leader peptide is used for recognition by modifying enzymes (Arnison et al., 2013). It has been observed in bioprospecting studies that novel RiPPs will share the same leader peptide to analogues from related bacterial strains (Portmann et al., 2008). The use of a leader peptide in biosynthesis may therefore assist in evolution of new analogues as the modifying enzymes will recognise only the leader peptide irrespective to mutations in the core peptide. And Some families of RiPPs may also have recognition sequences downstream of the core peptide to assist in cyclisation or excision (Arnison et al., 2013). Residues on the core segment of the precursor are enzymatically modified, and non-core segments up-and-down stream of the modified core are lysed resulting in the mature RiPP. The type of modifications that occur to the peptide define the RiPP family.

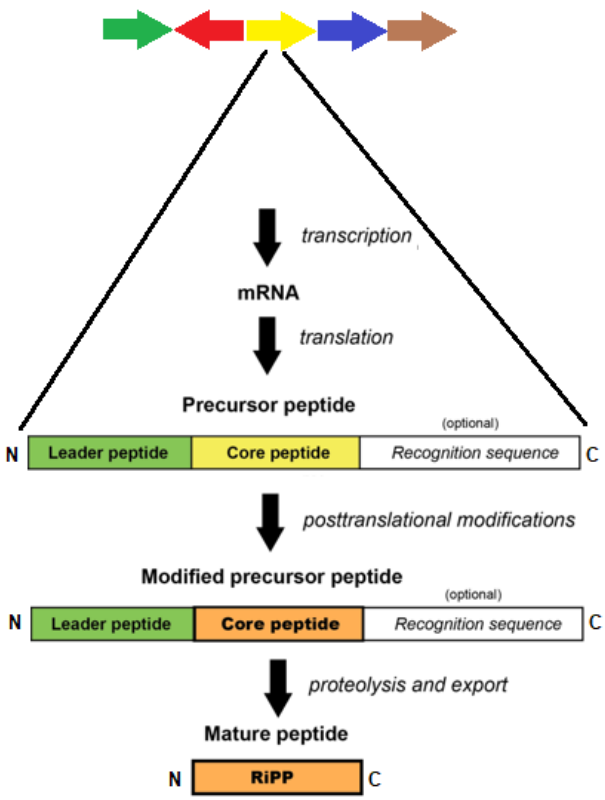


Figure 1.4. Graphical representation of the common RiPP biosynthesis process. The precursor peptide is translated from the structural gene. Modifying/tailoring enzymes alter the structure of the core section of the precursor peptide and remove the leader and recognition sequences to create a mature peptide. The nature of the posttranslational modifications that occur define a RiPP family.

1.4.2.1. Lantipeptides

Lantipeptides are polycyclic peptides that are defined by the presence of thioether-cross linked modified residues lanthionine and methyllanthionine. Lantipeptides may also contain other modified amino acids such as the dehydrated alanine residues dehydroalanine or dehydrobutyrine, cyclases with cystine or carbon-crosslinked labionin residues (Zhang et al., 2012). The family is split into four classes based upon the post-translational machinery used to create the cross-linked lanthionine and methyllanthionine residues (Goto et al., 2010) (**Figure 1.5**). In class I Lantipeptides two discrete enzymes carry out dehydration (LanB) of serine and therione residues and cyclisation (LanC) respectively (Zhang et al., 2012). For classes II-IV both dehydration and cyclisation are performed by a bifunctional enzyme but with differences in the enzymes structure and mode of activity. The dehydration domain of the class II lantipeptide synthase (LanM) has a unique sequence but the C-terminal cyclase domain is homologous with the domains of all other classes. The class III (LanKC) and class IV (LanL) lantipeptide synthases perform dehydration through the successive action of a kinase domain and a separate phosphoSer/phosphoThr lyase domain. LanKC is differentiated from LanL due to the fact it lacks zinc-binding sites present in the cyclisation domains of the other enzymes (Hegemann & Süssmuth, 2020). The linaridins are rare RiPPs which contain thioether cross-linkages similar to lanthipeptides, but their biosynthesic gene clusters appear to lack lanthipeptide-like dehydratases (Claesen & Bib, 2010). The status of the linaridins as either a subset of the lantipeptides or a family in their own right is debated by researchers (Arnison et al., 2013), and it is thought they are modified via a different biosynthetic route and so are not considered true lantipeptides (Mo et al., 2017, Ma & Zhang, 2020).

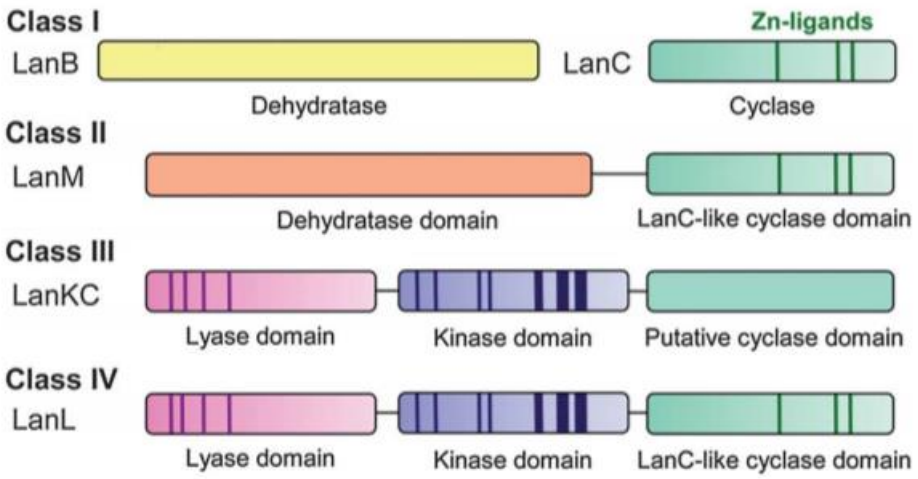


Figure 1.5. Schematic of the biosynthetic enzymes that form the thioether-cross linked modified residues lanthionine and methyllanthionine for the four different lantipeptide classes. Catalytic domains with homology share the same colour, and the dark regions are conserved regions with catalytic activity. From Arnison et al., (2013).

Lantipeptides are a well-studied family of RiPPs and there are many microbially-derived lantipeptides with antibiotic activity that have been identified by researchers. The most prominent example is Nisin, a class I lantipeptide produced by *Lactococcus lactis* that has broad-spectrum activity against Gram-positive bacteria and spores. Nisin has been used for 40 years as an additive to suppress food spoilage (Kitagawa et al., 2019). Duramycin is a class II lantipeptide that is now used as an antibiotic in veterinary medicine. There are also lantipeptides which have undergone clinical trials as antibiotic treatments in human medicine. A semisynthetic derivative of actagardine, produced by *Actinoplanes garbadinensis*, called NVB302 has been evaluated for treatment of *Clostridioides difficile* infection (Crowther et al., 2013). Due to their known biosynthetic logic and the development of genome-mining strategies for bioprospecting, these classes of natural products are receiving increasing focus as potential sources of novel antibiotics.

1.4.2.2. Thiopeptides

These peptides are defined by the presence of a highly modified six-membered nitrogenous macrocycle containing several thiazole rings with dehydrated residues (Arnison et al., 2013; Zhang & Liu, 2013; Just-Baringo, Albericio, & Álvarez, 2014; Zheng et al., 2015). Thiopeptides have been isolated from microorganisms that have been shown to have activity against Gram-positive pathogens (Morris et al., 2009). Whilst not shown to be active against Gram-negatives, there have been reports of activity against some eukaryotic parasitic pathogens (Bagley et al., 2005). The general mode of activity of thiopeptides is understood to be through inhibition of protein synthesis by different mechanisms of action. One known mechanism is through formation of a complex at the 50S subunit adjacent to the GTPase site. This blocks interaction of GTP-hydrolysing transcription factors at this site which are required to power movement of the ribosome along the mRNA template during translation (Bagley et al., 2005). Alternatively, some thiopeptides form direct complexes with elongation factors, such as elongation factor Tu (EF-Tu), and thus preventing their interaction with the ribosome to deliver aminoacylated tRNAs (Morris et al., 2009).

Prominent examples of antibiotic thiopeptides include micrococcin which was first isolated in 1948 from *Micrococcus* spp. (Su 1948) but also found in genus *Bacillus* (Heatley & Doery, 1951). This compound has been shown to have activity against parasite *Plasmodium falciparum* (Rogers, Cundliffe & McCutchan, 1998), and *Mycobacterium tuberculosis* (Degiacomi et al., 2016). Thiostreptons are another family of thiopeptides that have been isolated from several species of *Streptomyces* (Trejo et al., 1977) and also shown to have

activity against *Plasmodium falciparum* (Bagley et al., 2005) as well as broad-spectrum activity against Gram-positive (Anderson, Hodgkin & Viswamitra, 1970). Thiocillins are a family of related thiopeptides isolated from the genus *Bacillus* that have been shown to be active *in-vitro* against *Streptococcus pyogenes*, *Bacillus anthracis*, *Bacillus subtilis*, and *Staphylococcus aureus* (Shoji, et al., 1976). Microbially-derived thiopeptides could potentially be a source of new antibiotics for clinical usage. For example, a semi-synthetic derivative of the natural thiopeptide GE2270 A from *Planobispora rosea* (Selva et al., 1991) called LFF571 has been evaluated in clinical trials for treatment of *Clostridioides difficile* (LeMarche et al., 2012).

1.4.2.3. Lasso Peptides

Lasso peptides are defined by the presence of a specific knotted structure known as a lasso fold (Maksimov & Link, 2014). This unique structure makes lasso peptides highly stable and resistant to proteases and denaturing agents. Lasso peptides with antibiotic activity against Gram-positive and Gram-negative bacteria have been isolated from species of *Streptomyces* and *Rhodococcus* as well as from Proteobacteria in the genera *Escherichia* and *Burkholderia* (Arnison et al., 2013). Genome-mining study by (Maksimov, Pelczer & Link, 2012) has suggested that Lasso peptides may be ubiquitous amongst microorganisms, having predicted 76 lasso peptide biosynthetic gene clusters across nine bacterial phyla and an archaeal phylum. Their broad-spectrum activity and high stability make Lasso peptides an attractive prospect as novel antibiotic medicines.

The Lasso peptide family is subdivided into three classes based upon the presence of disulphide bonds (Tan, Moore & Nidwell, 2019). Class I contain two such bonds and a N-terminal cysteine. Class II lack disulphide bonds and have a N-terminal glycine, and Class III contain a single disulphide bond and a glycine at the N-terminal.

A common mechanism of action identified amongst lasso peptides with antimicrobial activity is the inhibition of RNA polymerase. Examples include acinetodin from human-associated *Acinetobacter* which has been shown to inhibit RNA polymerase of *Escherichia coli*. Microcin J25 produced by *Escherichia coli* which inhibits the growth of Gram-negative pathogens in the genera *Salmonella*, *Shigella* and *Escherichia* through RNA polymerase interaction (Wilson et al, 2003; Naimi et al., 2018). Capistruin from *Burkholderia thailandensis* E264 has been shown to inhibit species in the genera *Burkholderia* and *Pseudomonas* (Kuznedelov et al., 2011). Additionally, citrocin from the *Citrobacter pasteurii*

and *Citrobacter braakii* has activity against some against *Escherichia coli* and *Citrobacter* spp. (Cheung-Lee et al., 2019).

1.4.2.4. Bottromycins

Bottromycins are defined by the presence of a macrocyclic amidine and a decarboxylated C-terminal thiazole as well as C-methylated amino acids (Shimanmura, 2009). The eponymous bottromycin was first isolated from *Streptomyces bottropensis* (Waisvisz et al., 1957) and the derivatives have been shown to have activity against drug-resistant Gram-positives such as methicillin-resistant *Staphylococcus aureus* (MRSA) and vancomycin-resistant enterococci (VRE) (Kobayashi et al., 2010). The general mechanism of action of this family is believed to be through inhibition of protein synthesis by binding to the ribosome A site to prevent interaction with aminoacyl-tRNAs (Otaka & Kaji, 1976).

1.4.2.5. Cyanobactins

This is a broad family of small cyclic peptides produced by the phylum Cyanobacteria (Sivonen et al., 2010). Their structures may also contain oxazolines or thiazolines. Over 200 cyanobactins have been isolated from cyanobacteria and share a common biosynthetic process (Arnison et al., 2013) whereby the leader peptide is cleaved from the core and cyclised by serine proteases. Further tailoring enzymes may also introduce further modifications such as oxazolines, thiazolines, or their oxidized derivatives oxazoles and thiazoles, methylations, or Isoprenoids (Sivonen et al, 2010). Antibiotic cyanobactins include Kawaguchipeptin B isolated from *Microcystis aeruginosa* and shown to have activity against *Staphylococcus aureus* (Ishida et al., 1997).

1.4.2.6. Glycocins

This family is defined by the presence of glycosylated cystine residues (Arnison et al., 2013). Examples include sublancin isolated from *Bacillus subtilis* 168 which has inhibitory activity against some Gram-positive species including *Staphylococcus aureus* ATCC 12600, *Streptococcus pyogenes* ATCC 49399, and *Bacillus subtilis* ATCC 6633 but no effect against Gram-negatives (Paik, Chakicheria & Jansen, 1998). Glycocin F produced from *Lactobacillus plantarum* KW30 has a narrow spectrum of inhibitory activity against other strains within the species *Lactobacillus plantarum* (Kelly, Asmundson & Huang, 1996).

1.4.2.7. Bacterial Head-to-Tail Cyclised Peptides

This family is defined by its relatively large size of between 35-70 residues in addition to the presence of a peptide bond between the N- and C-terminal residues which results in a macrocylised structure. The large macrocylised structure confers resistance to high temperatures up to 100 ºC, pH fluctuations, and protease lysis, and they are also highly hydrophobic (Arnison et al., 2013). Studies into the tertiary structures of this family have shown them to have very similar structures regardless of differences in primary sequence homology, which may contribute to consistent physicochemical characteristics in the family (Martin-Visscher et al., 2009). There are three classes within the family; class I and class II are of similar sizes (60-70 residues) and structures are differentiated by their ionic states at neutral pH with class I being cationic and class II being anionic or neutral (Huang et al., 2009). Class III are smaller in size (~20-40 residues) and contain additional sactipeptide crosslinks between its residues (Arnison et al., 2013).

The high stability of these peptides makes them a potential source for novel antibiotic medicines. Several examples of this family with antimicrobial activity have been isolated from Gram-positive bacteria and it is believed that the general mechanism of action of these peptides is due to the hydrophobic nature of the peptides which leads to the creation of pores in target cell membranes resulting in a loss of osmotic pressure and cell death (Grande Burgos et al., 2014). The most prominent example in this family is AS-48, a class I which has been isolated from species in the *Enterococcus* genus and has *in-vitro* broad-spectrum activity against both Gram-positives and Gram-negatives such as *Listeria monocytogenes* (Ananou et al., 2004), *Staphylococcus aureus* (Perales-Adán et al., 2018), *Mycobacterium tuberculosis* (Aguilar-Pérez et al., 2018), and *Escherichia coli* (Gálvez et al., 1989). Preclinical evaluation of toxicity and safety of this peptide against human cells *in-vitro* and in *in-vivo* mouse models suggests that AS-48 is a promising lead for further therapeutic development with a lack of haemolytic of pro-inflammatory effects (Cebrián et al., 2019). AS-48 is discussed again in (**Chapter 3**). Other examples include uberolysin isolated from *Streptococcus uberis* strain 42 and shown to have activity against multiple strains of *Staphylococcus*, *Streptococcus*, *Lactococcus*, and *Enterococcus* (Wirawan et al., 2007), and circularin A which was isolated from *Clostridium beijerinckii* ATCC 25752 and is active against *Clostridium tyrobutyricum* and has been suggested to have utilisation in the food industry (Kawai et al., 2004).

1.4.2.8. Linear azole(in)e-containing peptides (LAPs)

This family are defined by a linear structure and the presence of thiazole and oxazole or their reduced forms which are formed by heterocyclisation of cysteine, serine, and threonine residues (Molohon et al., 2011). Microcin B17 is an example of a LAP with antimicrobial activity. Microcin B17 is a plasmid-encoded peptide present in some strains of *Escherichia coli* species. It's mechanism of action is to inhibit DNA replication by binding to DNA-Gyrase complex and stabilising it in a confirmation to prevent rejoining of the DNA strands (Heddle, et al., 2001; Parks et al., 2007; Collin & Maxwell, 2019). Another example is plantazolicin, isolated from strains of *Bacillus methylotrophicus* and *Bacillus pumilus*, which has a narrow spectrum of activity against *Bacillus anthracis*. The proposed mechanism of action of plantozolicin is to localize to and exacerbate structurally compromised regions of the bacterial membrane to cause cell lysis (Molohon et al., 2016).

1.4.2.9. Proteusins

Proteusins are a rare family of RiPPs. They are defined by a complex 48-mer structure which contains a N-acyl unit, and a high proportion of nonproteinogenic residues. Polytheonamides are an example of this class which was isolated from a marine sponge (Hamada et al., 1994), but believed to be produced by a member of its microbiome (Freeman et al., 2012). These compounds exhibit a high cytotoxicity through the formation of channels in the cytoplasmic membrane. They achieve this in part through the formation of a β-helical secondary structure which inserts into the membrane. This is guided by the lipophilic N-acyl unit to ensure that insertion is longitudinal and the large size of the molecule ensures the channel spans across the membrane (Hamada et al., 2010). Genome mining study by Freeman and co-workers (2012) suggest that the post-translational modification pathway is complex, with six enzymes potentially catalysing up to 48 modifications. It remains a possibility that bacteria from other microbiomes may be potential sources for these novel peptide compounds.

1.4.3. Terpenes

Terpenes are a diverse class of natural product that are synthesized through polymerisations reactions and modifications between two precursor C5 subunits dimethylallyl pyrophosphate (DMAPP) and isopentenyl pyrophosphate (IPP). The polymerisation of the precursor subunits is catalysed by terpene synthases. Modifications of these core structures, which can include cyclisation, hydroxylation, and glycosylation, are catalysed by cytochrome P450s (Kuzuyama 2017). DMAPP and IPP precursors are products produced via the mevalonate pathway which

is a common part of the primary metabolism of eukaryotic organisms. While examples of the mevalonate pathway in bacterial cells has been found, in bacteria and plants these subunits are more commonly derived from an alternative pathway, non-mevalonate pathway, known as the MEP pathway which utilises pyruvate (**Figure 1.6**).

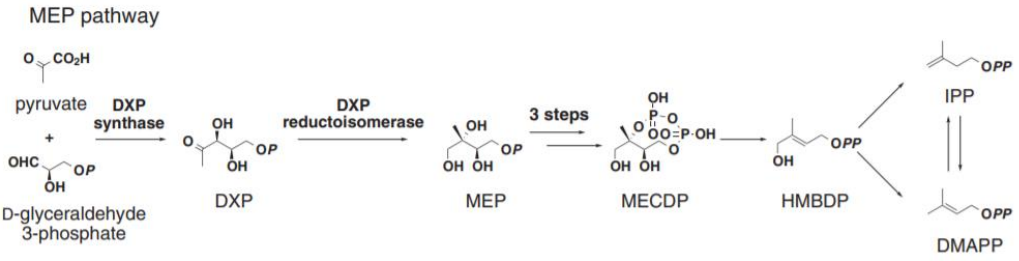


Figure 1.6. Outline of the MEP pathway. The metabolic pathway which produce the terpene precursor compounds dimethylallyl pyrophosphate (DMAPP) and isopentenyl pyrophosphate (IPP) amongst most prokaryotic life. DXP synthase catalyses the reaction of pyruvate and D-glyceraldehyde 3-phosphate to form -deoxy-D-xylulose 5-phosphate (DXP). This is converted to 2-C-methylerythritol 4-phosphate by the action of DXP reductoisomerase. MEP is converted to 2-C-methyl-D-erythritol 2,4-cyclodiphosphate (MECDP) through a three-step reaction catalysed by 4-diphosphocytidyl-2-C-methyl-D-erythritol synthase, 4-diphosphocytidyl-2-C-methyl-D-erythritol kinase, and 2-C-methyl-D-erythritol 2,4-cyclodiphosphate synthase. HMBPP synthase then catalyses the conversion to 1-hydroxy-2-methyl-2- (E)-butenyl-4-diphosphate (HMBDP). This is reduced to IPP or DMAPP by HMBPP reductase. And IPP isomerase catalyses the reversable isomerisation between these two molecules. Adapted from Kuzuyama, (2017).

Terpenes are classified into families according to the number of C5 units in their structure. Hemiterpenes (C5) are the simplest and are created through modifications to either a DMAPP or IPP subunit. Monoterpenes (C10) are synthesised through the condensation of IPP and DMAPP to produce a 10 carbon scaffold known as geranyl pyrophosphate (GPP) before subsequent modifications. Polymerisation with additional IPP/DMAPP subunits leads to creation of more complex terpenes, with most notable terpenes with biological activity being monoterpene, sesquiterpenes (C15), diterpenes (C20), triterpenes (C30), or tetraterpenes (C40) (Heinrich et al., 2004). The production of terpenes is ubiquitous to all domains of life, and over 5000 have been isolated from natural sources.

Terpenes can serve many functions in the cellular metabolism of an organism, including as signalling molecules, hormones, pigments, and antibiotics. The biggest commercial applications of terpene natural products have included ingredients in perfumes, flavourings, and food dyes (Cane & Watt, 2003; Yamada et al., 2015). Fusidic Acid is an example of a terpene antibiotic in clinical usage and was first isolated from the fungus *Fusidium coccineum* (Godtfredsen et al., 1962; Godtfredsen et al., 1968; Gudkov, 2001).

Overall, the vast majority of isolated terpenes have come from plants and fungi, with relatively few isolated from bacteria (Cane & Watt, 2003; Yamada et al., 2015). However, bacteria do produce terpene compounds. Prominent examples include 2-methylisoborneol and geosmin (**Figures 1.7a, 1.7b**) – a monoterpene and sesquiterpene respectively both produced by cyanobacteria and actinobacteria. These volatile compounds are attributed to the "earthy" aroma and taste often associated with these species (Jüttner & Watson, 2007; Suurnäkki et al., 2015).

Some terpenes isolated from bacteria have been shown to have antibiotic activity. For example, pentalenolactone (**Figure 1.7c**) is a sesquiterpene produced by several species of *Streptomyces* and has been demonstrated to have antifungal, antitumour, and antiviral activity *in-vitro* (Cane et al., 1990), and albaflavenone (**Figure 1.7d**), also a sesquiterpene with antibiotic activity *in-vitro*. Albaflavenone has been shown to be produced by several species of *Streptomyces* that also produce other antibiotic natural products (Gurtler et al., 1994; Zhao et al., 2007; Moody et al., 2011) (discussed further in **Chapter 5**).

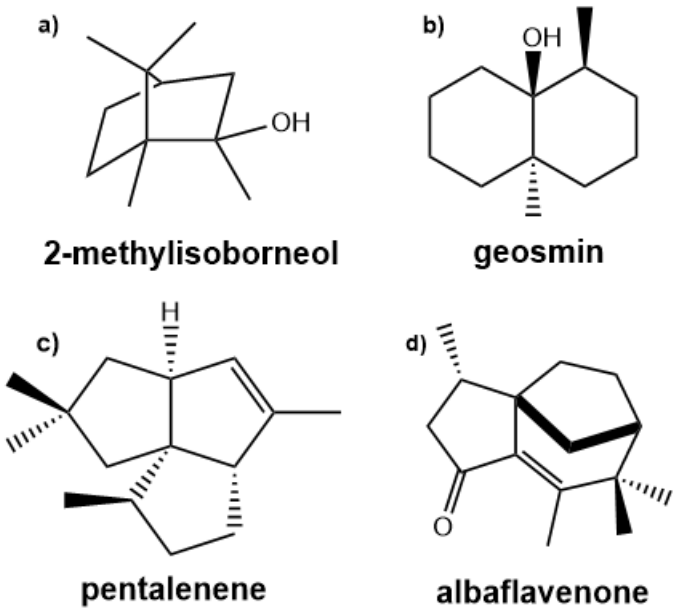


Figure 1.7. Structures of a) 2-methylisoborneol b) geosmin c) pentalenolactone d) albaflavenone. Drawn in ChemDraw Professional version 16

Bacterial terpene synthases differ from those of plant and fungal species in that they exhibit low levels of mutual sequence homology with other bacterial synthases or between themselves (Cane & Watt, 2003; Yamada et al., 2015). This has hampered the ability to apply a biosynthetic logic that can be utilised for PCR-based screens or bioinformatic led bioprospecting investigations into bacterial terpenes. Along with the relative rarity of bacterial

terpenes in comparison to plants had led some to assume that terpene natural products were not highly abundant amongst this domain. For example, as of 2003 only approximately a dozen terpene natural products were associated with the *Streptomyces* genus despite the high proliferation of commercially and medicinally important natural products associated with this genus (Cane & Watt, 2003). However, the development of profiles via Hidden Markov Models through multisequence alignments of 140 characterized terpene synthase sequences by Yamada and co-workers (2015) has led to the suggestion that terpene biosynthesis pathways may be far more abundant and widely distributed amongst bacteria than previously assumed. The authors of this study applied this model to 8,759,463 predicted bacterial proteins from public genome databases and identified what they claim to be 262 presumptive mono-, sesqui-, and diterpene synthases from bacterial genomes in orders *Actinomycetales*, *Myxococcales*, *Oscillatoriales*, *Nostocales*, *Burkholderiales*, *Herpetosiphonales*, *Rhizobiales*, *Chlamydiales*, *Flavobacteriales*, *Chromatiales*, *Ktedonobacterales*, *Sphingobacteriales*, and *Pseudomonadales* – but did not detect any in the phylum Firmicutes. This suggested that terpene natural products may be a potentially rich source of new bacterial natural product leads including antibiotic compounds.

1.5. The Decline in Antibiotic Discovery and the Need for New Antibiotics

The use of antibiotics creates a selection pressure that drives the evolution of resistant bacterial strains, which reduces the effectiveness of the antibiotic. When a population of bacteria is subjected to a toxic compound, any subpopulations which harbour mutations which confer resistance to the biochemical mode of activity of the antibiotic are able to proliferate at the expense of the sensitive wild-type and become the dominant strain. This is a natural process and inevitable outcome of antibiotic chemotherapy predicted by its pioneers such as Fleming (Rosenblatt-Farrell, 2009). For example, streptococcal pathogens that had acquired resistance to penicillin were first reported soon after its introduction (Barber, 1947).

While the onset of bacterial resistance is a natural process, several social factors also accelerate its development. For example, environmental contamination with antibiotics and sub-optimal treatment dosages increase exposure of bacteria to antibiotics and thus drive this selection pressure. This occurs through poor antibiotic stewardship, such as the overuse of antibiotics in livestock farming, usages of antibiotics for patient treatment when it is not appropriate, and patients not completing treatment courses or disposing of surplus medicine appropriately (World Health Organisation, 2018). In addition to the development of single drug resistance through genetic plasticity, the genetic promiscuity of bacteria allows for drug resistant genes to be shared amongst bacteria across different genera readily through

horizontal gene transfer. The mechanisms that drive this are mobile genetic elements such as plasmids, introns, and transposons, which have the ability to move across genomes. The mechanisms by which these gene elements are exchanged include as conjugation (exchange of plasmids between bacteria via direct contact between cells through formation of pilus "bridges"), transduction (via bacteriophage infection), and transformation (direct uptake of environmental DNA) (**Figure 1.8**).

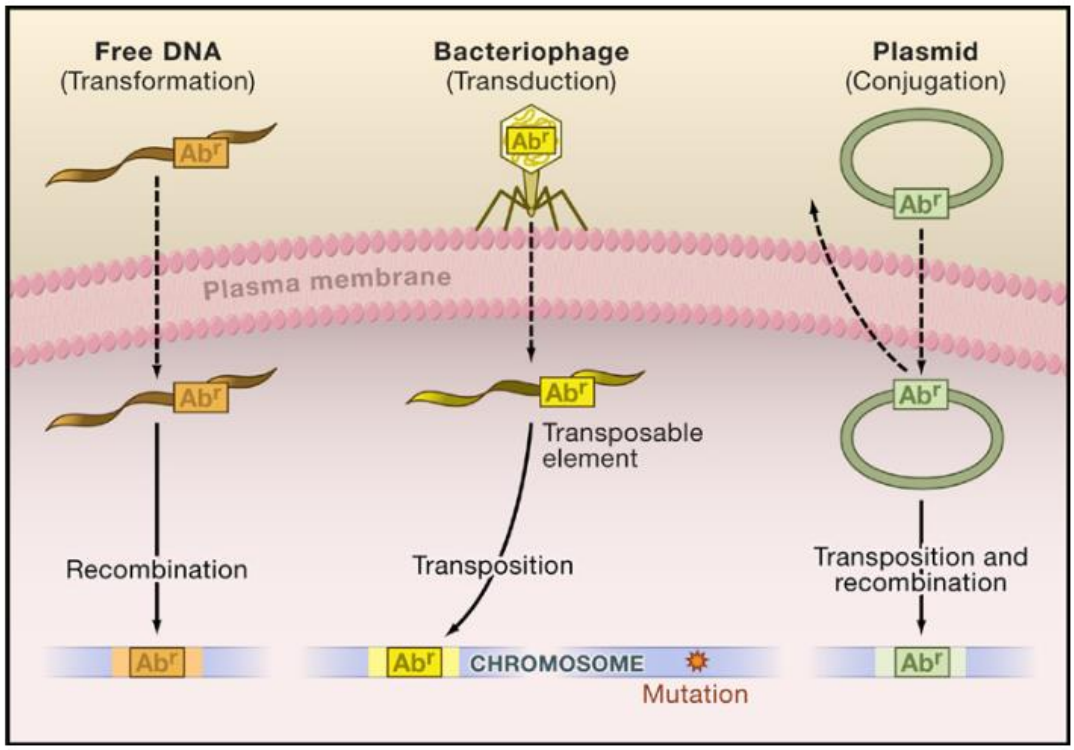


Figure 1.8. Outline of the mechanisms by which antibiotic resistant genes (Abr) can be acquired by bacteria through horizontal gene transfer. Resistance genes located on mobile genetic elements (transposons, bacteriophage genetic elements, plasmids, and introns) are acquired through conjugation, transduction, or transformation and integrated into the chromosome. Adapted from Alekshun & Levy (2007).

A combination of these effects means that over time bacteria can acquire resistance genes to multiple antibiotics. Multiple drug resistant (MDR) and pan/extensive drug resistant (PDR or XDR) pathogens have become an established threat to public health. Of the most concern include nosocomial infections caused by *Enterococcus faecium*, *Staphylococcus aureus*, *Klebsiella pneumoniae*, *Acinetobacter baumannii*, *Pseudomonas aeruginosa*, and *Enterobacter* spp. – known as the "ESKAPE" group. MDR infections from this group are difficult to treat and are associated with poorer clinical outcomes in patients and have been identified by the World Health Organisation (WHO) as amongst pathogens for which new antibiotic treatments are urgently needed (Tacconelli et al., 2018). A report by the Centers for Disease Control and Prevention (CDC, 2019) has estimated that over 2.8 million antibiotic-

resistant infections occurr in the U.S.A each year and that over 35,000 die from these infections. Treatment of patients with MDR infections is often very costly to the healthcare system. A recent report by Nelson et al., (2021) estimated that healthcare costs for the treatment of MDR infections in hospitalised patients was $4.6 billion in 2017.

Some MDR pathogens of concern are known as opportunistic and can cause infection in vulnerable patients with compromised immune systems or are undergoing invasive medical treatments. An example of this is *Staphylococcus aureus* which is a common commensal bacterium but can cause invasive infections if people with weakened immune systems or in skin wounds (Krismer et al., 2017). Methicillin-resistant *Staphylococcus aureus* (MRSA) are resistant to multiple antibiotic drugs and can cause catheter-related bacteraemia (Cuervo et. al., 2015) or infection of surgical wounds (Kurinchi et al., 2013). It has been estimated that there were 323,700 hospital-associated MRSA infections in the U.S.A. in 2017 and 10,600 fatalities (CDC, 2019).

Escherichia coli is another example of a human commensal which is common cause of healthcare-associated opportunistic antibiotic-resistant infections (Kourtis et al., 2020). MDR *E. coli* has been shown to cause infections in medical implant sites (Pfang et al., 2019), bacteraemia (Alhashash, Weston, Diggle & McNally, 2013) and urinary tract infections (Mutters et al., 2018). Healthcare-associated MDR *E. coli* infections are increasing globally (de Kraker et al., 2012). Annual cases of MDR *E. coli* infections in hospitalised patients have been rising in the U.S.A. year-on-year from an estimate of 131,900 in 2012 to 196,400 in 2017. The estimated number of deaths caused by MDR *E. coli* in U.S.A. in 2017 is 9,100 (CDC, 2019). A study by Ibrahim, Bilal & Hamid (2012) isolated 232 *E. coli* strains from hospital patients in Khartoum state, Sudan. The authors found the majority of these (65.1 %) to be from urine and 92.2% of the isolates were multidrug resistant.

Multi-drug-resistant tuberculosis (TB), which is resistant to isoniazid and rifampicin, is a major public health concern in the developing world, causing an estimated 170,000 deaths globally in 2012 (Seung, Keshavjee & Rich, 2015). MDR *Neisseria gonorrhoeae* causes the sexually transmitted infection (STI) gonorrhoea and is also an increasingly significant global public health concern (Unemo & Shafer, 2014). *Neisseria gonorrhoeae* infections have been treated successfully with antimicrobials for nearly 80 years, but the emergence of strains resistant to most antimicrobials has led to an increase in cases of gonorrhoea infections which cannot be effectively treated (Blomquist, Miari, Buddulph & Charalambous, 2014). The CDC has estimated that there has been a 124% increase in drug-resistant gonorrhoea infections in the U.S.A between 2012-2017 and have classed this as an “urgent threat” to public health

(CDC, 2019). The CDC also consider drug resistant *Clostridioides difficile* as an urgent threat, with an estimated 223,900 cases amongst hospitalised patients and 12,800 deaths in the U.S.A in 2017 (CDC, 2019). A recent systematic review and meta-analysis of *C. difficile* resistance studies concluded that instances of resistance to vancomycin have increased since 2012 alongside increased usage of this drug to treat the infection, and that the highest rates of resistance were detected in the Americas and Asia (Saha et al., 2019).

There is therefore an urgent need for the development of new classes of antibiotics that can by-pass the established resistance mechanisms (Gualerzi et al., 2013), and during the 20th century the discovery and development of new antibiotic classes with new biochemical modes of action mitigated the proliferation of resistance. However, despite the constant need for new antibiotic classes, discovery rates via the traditional method of bioactivity-guided isolation (detailed in **Section 1.3**) rapidly declined from the 1960s onwards and pharmaceutical companies subsequently turned their attention away from screening microbial sources towards synthetic chemical libraries and target-based drug development. This move was in part due to an increasing tendency to rediscover already known compounds from screens of traditional microbial sources (Clardy, Fischbach & Walsh, 2006; Rutledge & Challis, 2015) and also due to the large costs and long time periods required to develop a clinical product (Gualerzi et al., 2013).

While antibiotic development has declined, antibiotic resistance amongst pathogens has continued to proliferate. Multi-drug and pan-drug resistant pathogens are now becoming more frequent, which could lead to a 'post-antibiotic age' where current antibiotics have lost all effectiveness (Ventola, 2015; Wright & Brown, 2016). A UK Government review (O' Neill, 2016) estimates that if the current pace of antibiotic resistance proliferation is not slowed, then global fatality rates due to drug-resistant infections could rise from the current level of 700,000 to 10 million a year by 2050. The review also estimates a loss of over US$100 trillion to the global economy, and states that without effective antibiotic prophylactics, many modern medical procedures such as abdominal surgery, joint replacements, and chemotherapy would be too dangerous to perform.

1.6. New Approaches to Bioprospecting

There is now an urgent need for the discovery and development of new antibiotic drugs, which has led to a renewed focus on bioprospecting of microbiomes for antibiotic leads. And the development of new experimental, genomic, and computational approaches, that may address some of the limitations of the 'traditional' bioactivity-guided isolation approach are being

developed and utilised to support the discovery of novel antibiotic leads.

1.6.1. Culturing the "Unculturables"

It is estimated that less than 1% of environmental microorganisms are culturable in standard laboratory conditions (Pham & Kim, 2012). This means that nearly all microbial diversity, and the resulting chemical diversity, is missed when screening a microbiome using traditional culture techniques (**Figure 1.1**). New culturing techniques are expanding the range of microorganisms that can be isolated and this may in turn lead to an increase in the discovery of novel natural product antibiotic leads. A recent invention called the iChip (isolation chip) is a semi-permeable chamber containing membranes that capture and cultivate environmental microorganisms in their natural environment. Once the colonies have grown to a sufficient density, they can be removed from the iChip and maintained in the laboratory (Nicols et al., 2010). Screening of soil microorganisms isolated using the iChip has led to the recent discovery of a new antibiotic class from a new species of bacteria. Teixobactin is a non-ribosomal protein produced by a β-proteobacteria called *Eleftheria terrae* that is active against *Staphylococcus aureus* and *Mycobacterium tuberculosis*. It inhibits cell wall synthesis by binding to lipid II and lipid III, which are precursor molecules to peptidoglycan and teichoic acid respectively (Ling et al., 2015). This new technology offers the prospect that other antibiotic classes could be discovered. However, while a significant advance, the technology does not isolate all microorganisms. It is estimated that approximately 50% of soil bacteria can be captured on the iChip (Nicols et al., 2010). A possible concern about the iChip is the process of isolation and screening the bacteria from an iChip is costly and labour intensive as ten thousand isolates were screened to yield the discovery of teixcobactin (Ling et al., 2015).

1.6.2. Metagenomic Techniques

Another approach used to access the biosynthetic potential of unculturable microorganisms is to extract and study the genetic material directly from the environmental sample. This is known as metagenomics and is broadly divided into functional and sequence-led approaches. Functional metagenomics involves the cloning of environmental DNA (eDNA) into heterogeneous expression hosts and then screening the recombinant clones for the desired phenotype in a bioassay. The recombinant biosynthetic gene clusters and their products from lead clones can then be characterised. Brady & Clardy (2000) first reported on the discovery of an antibiotic compound using functional metagenomics. The compound that was discovered is a long-chain N-acyl amino acid and was identified by screening soil eDNA cloned into *E. coli* expression hosts for antimicrobial activity against *Bacillus subtilis*. This

study was also noteworthy as the first identification of a long-chain N-acyl amino acid biosynthesis gene. The researchers also discovered an isocyanide with anti-*Bacillus* activity from soil eDNA (Brady & Clardy, 2005), and expanded the approach beyond soil to discover a long-chain N-acyl palmitoylputrescine that inhibits *Bacillus subtilis* from Costa Rican bromeliad tank water eDNA (Brady & Clardy, 2004). A different phenotypic screening approach has led to the discovery of two triaryl cations called turbomycins A and B from soil eDNA cloned into *E. coli* expression hosts. The recombinant colonies produced a brown pigment, which was shown to have broad spectrum activity against Gram-positive bacteria, Gram-negative bacteria, and yeast (Gillespie et al., 2002). Two other pigmented compounds, indirubin and indigo, have been found in a forest soil eDNA library and were shown to have antimicrobial activity against *Bacillus subtilis* (Kim et al., 2005).

The sensitivity of the functional metagenomics approach is limited by the fact that not all recombinant DNA can be readily expressed in a heterogeneous host (Brady, 2014). Sequence-led metagenomic approaches utilise the conserved nature of certain biosynthesis genes to target and identity novel biosynthetic gene clusters. After the gene clusters have been isolated, they are heterogeneously expressed in a suitable host and characterised. Two unique polyketides, fasamycin A and fasamycin B with activity against methicillin-resistant *Staphylococcus aureus* and vancomycin-resistant *Enterococcus faecalis* were discovered in a soil eDNA *E. coli* cosmid library by using a PCR screen with polyketide synthase-specific degenerate primers. Recombinant vectors selected by the PCR screen where then cloned into *Streptomyces albus* for phenotype screening and metabolite characterisation (Feng et al., 2011). The antibiotic activity against Gram-positives was found to be due to inhibition of type II fatty acid biosynthesis (Feng et al., 2012). Bioinformatics tools have also been used in combination with a 'sequence-tagging' approach to discover several potentially medically relevant compounds from a soil eDNA cosmid library (Owen et al., 2013). The researchers used barcoded degenerate primers for NRPS and PKS genes to create a library of amplicons which were sequenced and analysed for homology with NRPS and PKS genes of known medical value. Lead amplicons were tracked back to the cosmid vector template (with the use of the barcodes) and the recombinant eDNA fragment was sequenced and annotated. This study identified putative BGCs for analogues of the antibiotics teicoplanin and friulimicin.

The sensitivity of the sequence-led approach is limited to metabolites that are synthesised by gene clusters with high homology and through a known core biosynthetic logic. Some antibiotic compound classes do not have high genetic similarity. For example, functional screening of soil eDNA libraries discovered 11 long-chain N-acyltyrosines with antimicrobial activity. However, analysis of the recombinant vectors found that the N-acyltyrosine synthase

genes did not share any similarity to enable the development of a PCR-based screen (Brady, Chao & Clardy, 2004).

1.6.3. Bioinformatics Techniques

It has been discovered that many microorganisms contain secondary metabolite gene clusters that are not readily expressed in laboratory conditions, these are termed "silent or cryptic biosynthetic gene clusters". The earliest and most striking example of silent BGCs comes from the model *Streptomyces* organism *S. coelicolor*. When its complete genome was published in 2002, the annotation revealed that the genome encoded 18 putative BGCs for common secondary metabolite classes that had not been discovered through the previous 40 years of traditional bioassay-led approaches (Bentley et al., 2002). The development of cost-effective and fast next-generation sequencing (NGS) platforms such as Illumnia, Pac-Bio, and Oxford Nanopore, in addition to powerful bioinformatics tools, have led researchers to adopt a "genome-mining"-led approach to microbial antibiotic discovery. Through analysis of the annotation of these sequenced genomes, researchers can make predictions of putative BGCs that the genome maybe encoding through an understanding of the biosynthetic logic of that BGC.

AntiSMASH (antibiotics & Secondary Metabolite Analysis Shell) has become a major bioinformatics tool for the detection of putative BGCs encoded in microbial genomes. The software searches genomes for profile Hidden Markov Models (pHMMs) that are generated through multisequence alignments of experimentally characterised signature proteins that are specific for certain biosynthetic classes (e.g. NRPS, PKS, lantibiotics, bacteriocins, etc.) (Medema et al., 2011). While this pipeline is used for the detection of known secondary metabolite classes, the latest version v3.0 (Weber et al., 2015) also includes the 'ClusterFinder' algorithm for the detection of putative biosynthetic gene clusters for unknown metabolite classes. The algorithm assumes that the BGCs of unknown metabolite classes would still utilise broadly similar catalysis families (such as oxidoreductases, methyltransferases, etc.) and that these enzymes would be clustered together as with other BGC types. The algorithm searches the genome against the Pfam dataset to identify clusters of putative biosynthetic enzyme genes and highlight them as possible BGCs. The BGC for the antibiotic tarromycin A, a derivative of daptomycin, was discovered using AntiSMASH to analyse the draft genome sequence of the marine actinomycete *Saccharomonospora* sp. CNQ-490. Once identified, the putative gene cluster was transformed into a heterogeneous expression host and characterised (Yamanaka et al., 2014).

Genome-mining can therefore identify microorganisms as potential leads for antibiotic production. For example, McClerren and co-workers (2006) used a genome-mining approach to discover a two-component lantibiotic called haloduracin. The researchers searched published bacterial genomes for lantibiotic analogous and identified a putative lantibiotic gene cluster in the genome of *Bacillus halodurans* C-125, which had not been reported previously. The researchers obtained the strain and purified the lantibiotic and demonstrated its activity against *Lactococcus lactis*.

1.6.3.1. Genome-mining to Inform Bioactivity-Guided Isolation Strategies

Genome-mining approaches may offer hope to facilitate the dereplication and lead prioritising process in microbial bioprospecting screens by offering insight into the identity of the isolates and what biosynthetic compounds they may be producing. Genome-mining can also help researchers predict the physicochemical characteristics of putative compounds which can help inform chromatography strategies used.

Effectively integrating genome-mining data with bioactivity-guided isolation pipelines can therefore help streamline the process of dereplication and lead prioritisation whilst improving efficiency of compound purification. Possible examples being if genome-mining analysis reveals a lead isolate to encode several large peptide products of interest to the researcher, then methods such as size-exclusion chromatography or tandem mass-spec could be employed to target the isolation and identification of these peptides over other compounds. Alternatively, if a lead isolate appears to encode for a known compound with published spectra data but also encodes potentially novel compounds of interest then the knowledge of the known compounds spectra and properties can be used to quickly discount it or confirm it as the bioactive compound.

Some researchers have developed algorithms to automate the matching of chemical spectra data of natural product compounds to the putative biosynthetic gene clusters for high-throughput studies. An early study by Kersten and co-workers (2011) generated de novo tandem mass-spec data of crude peptide extracts of several model Streptomyces species and compared these data with predicted structures from annotated BGCs to match the chemotypes with their biosynthetic gene clusters. The authors claimed that the combined information from both datasets allowed for the identification and characterisation of a range of modified peptide products (lantipeptides, lasso peptides, linardins, formylated peptides and lipopeptides) which had not been identified previously. Software packages have been developed for researchers to

use to integrate such datasets such as Pep2Path (Medma et al., 2014), RIPPquest (Mohimani et al., 2014a), NRPquest (Mohimani et al., 2014b) and iSNAP (Ibrahim et al., 2012). However these approaches are all limited to peptide natural products only due to the known biosynthetic logic of these compounds. The potential for genome-mining data to inform bioactivity-guided isolation strategies has been demonstrated in several studies which have taken different approaches to integrate the two. For example, Kersten and co-workers (2013) isolated a novel DNA-interfering polyketide lomaiviticin from a strain of the marine actinobacteria *Salinispora tropica*. The authors identified a putative PKS biosynthetic gene cluster from the annotated genome data of this strain. Using PCR-based mutagenesis they created a knockout mutant strain and compared LC-MS spectra data of extracts from both wild-type and mutant to identify the peak associated with this BGC. This knowledge then allowed the authors to optimise their purification strategy to isolate and characterise this compound.

1.6.4. Antibiotics from Novel Microbiomes

The soil microbiome has been a major focus for the search for antimicrobial producing microorganisms since the earliest bioprospecting programmes of the 1930s and '40s. The rationale for this came from the conclusion of 19th century microbiologists that the absence of certain human pathogens in soil could be due to the presence of soil microorganisms that supress these pathogens (Waksman & Woodruff, 1940). With the renewed focus on searching for antimicrobial compounds, bioprospectors have turned their attention to other microbial ecosystems in the belief that novel environmental niches will contain novel biology which will in turn lead to the discovery of novel chemical compounds.

1.6.4.1. Bioprospecting Plant and Insect Microbiomes

Microbes from plants and animals have been shown to be potential sources of antibiotic compounds. For example, genome-mining of the plant-associated *Bacillus amyloliquefaciens* FZB42 found it encoded BGCs for polyketide antibiotics bacillaene and difficidin (Chen et al., 2007). Computational analysis of the genomes of several *Pseudonocardia* strains isolated from Panamanian leafcutter ants found them to encode several novel polyketides which had similarity with the antifungal nystatin. It is hypothesised that *Pseudonocardia* acts as a probiotic to protect the ants' fungal food source from pathogens (Holmes et al., 2016).

Olofsson and co-workers (2014) hypothesised that honey may contain antimicrobial compounds secreted by the microbiome of the Honeybee and successfully isolated *Lactobacillus* from both the bee gut and its fresh honey which had broad-spectrum activity.

The group demonstrated that the isolates had antimicrobial activity against methicillin-resistant *Staphylococcus aureus* (MRSA), *Pseudomonas aeruginosa* and vancomycin-resistant *Enterococcus* (VRE). Many of the strains were also isolated in fresh honey (1-3 days old) but died once the honey had matured after several weeks due to a lack of water. However, the compounds secreted by the isolates remain in the mature honey and may contribute to its antimicrobial properties.

1.6.4.2. Bioprospecting the Human Microbiome

Some researchers have begun focusing on isolating antimicrobial compounds from the human microbiome under the rationale that our commensal bacteria may suppress invading species (Yang et al., 2011; Dobson et al., 2012). Several examples of small ribosomal peptides have been isolated from human commensals that are active against a narrow range of related bacterial strains (Dobin & Fischbach, 2015). These molecules are termed 'bacteriocins' and have not yet been successfully developed into clinical drugs (Mills et al., 2011). A large-scale genome mining effort of the genome data from the US National Institute of Health (NIH) human microbiome project was performed by Donia and co-workers (2014). They discovered that BGCs for many classes of secondary metabolites, including PKS and NRPS, were abundant and widely distributed throughout all sites of the human microbiome (oral, skin, gut, vaginal), demonstrating it as a potentially rich source of novel chemistry. The researchers also identified putative BGCs for thiopeptides, a class of antibiotic currently in clinical trials. From this analysis, the researchers were able to purify a thiopeptide antibiotic, lactocillin, from the vaginal commensal *Lactobacillus gasseri* and demonstrated its activity against the vaginal pathogen *Corynebacterium aurimucosum*. This was a significant finding as it is the first example of a drug-like antibiotic being isolated from the human microbiome. Zipperer and co-workers (2016) screened nasal cavity commensals for activity against *Staphylococcus aureus* which is an opportunistic pathogen found in ~30 % of people. This led to the discovery of lugdunbin, a non-ribosomal cyclic peptide from *Staphylococcus lugdunensis*. This is the first example of a non-ribosomal peptide antibiotic isolated from the human microbiome. This is significant because many classes of commercial antibiotics are of the non-ribosomal class, so this discovery further supports the findings of Donia and co-workers (2014) of the potential of the human microbiome as a source of antibiotics with commercial potential.

1.6.4.3. Bioprospecting Aquatic Environments

Aquatic environments are an example of areas where physical conditions are very different to that of terrestrial environments. Over 70 % of the Earth's surface is covered in water and physical conditions such as temperature, nutrient composition, pressure, gas content, pH, and flow rates vary enormously between different locations. This diversity makes aquatic environments an attractive source of novel microbes that can produce novel chemistry. There are numerous examples of novel antibiotic compounds being discovered from microbes from aquatic environments. An early example is a strain of *Streptomyces griseus* isolated from shallow sea sediment in Japan. *S. griseus* is commonly associated with terrestrial environments and is a known producer of the antibiotic streptomycin. However, this strain required seawater for growth and produced an antibiotic compound with a novel chemical structure that was different to examples that had been isolated from terrestrial strains in that it had a boron-containing polyether ionophore structure. The compound has antibiotic activity against Gram-positives and *Plasmodium* and so was named aplasmomycin (Sato et al., 1978). More recently, a Streptomyces isolated on saline based media from marine sediments from Santa Barbara, USA was found to produce a novel antibiotic compound named anthracimycin. The compound showed broad spectrum activity against both Gram-positives and Gram-negatives, including activity against *Bacillus anthracis* (Jang et al., 2013).

Some aquatic environments have physical conditions that are considered extreme and microorganisms adapted to survive these conditions are known as 'extremophiles'. There is currently interest in extremophiles not just for their potentially novel secondary metabolites but also for potential biocatalysis purposes. This is because the extreme conditions (such as temperature and pH) their metabolic enzymes are optimised to work in, as well as the uncommon substrates they may utilise, give these enzymes potential applications in industrial processes (Coker, 2016).

Thermal springs, which are sites were geothermally heated groundwater rises through the Earth's crust to the surface are an example of such an extreme aquatic environment. The geochemistry of every thermal spring around the world is unique because of a combination of different factors such as the temperature, water quantity and flow rates, the rock types that the water travels through, and also the nature of the habitat around the spring site. It is therefore expected that the microbiome of each thermal spring to also be unique, making each site a potentially rich source of novel microbiology and natural product chemistry. Examples of microbes with novel biology isolated from such environments include *Thermus aquaticus* – first isolated from the Mushroom Spring in Yellowstone National Park, USA (Brock and

Freeze, 1969). It can tolerate temperatures in the range 50-80 °C and has had a huge biotechnological impact through the now indispensable use of its thermostable DNA polymerase (Taq polymerase) in PCR (Saiki et al., 1988). Another example is *Thermoanaerobacter mathranii*, which was isolated from thermal spring sediments in Iceland and has been shown to be capable of metabolising plant biomass to produce ethanol, and is therefore a focus of interest in the development of sustainable biofuels (Larsen, Nielsen and Ahring, 1997).

There is also a focus on finding novel antibiotic compounds from thermal spring microorganisms. For example, Esikova and co-workers (2002) reported on the discovery of two thermophilic *Bacillus* strains isolated from the thermal springs in the Kamchatka Peninsula, Russia which produced fermentative extracts that inhibited growth of several Gram-positive bacteria. Saikia and co-workers (2010) reported the isolation of a *Brevibacillus* species from a thermal spring in Assam, India that exhibited anti-fungal activity. A bacteriocin named Putadicin T01 was reported to be produced by a strain of *Pseudomonas* isolated from a thermal spring in Tunisia. The bacteriocin was shown to have inhibitory activity against other *Pseudomonas* strains and *E. coli* (Ghrairi, Braiek and Hani, 2015). Puri and co-workers (2010) isolated a strain of the fungal species *Elaphocordyceps ophioglossoides* from thermal spring water in Japan. Analysis of extracts of the strain showed it to produce the tetramic acid antibiotic equisetin and a new analogue that the authors named ophiosetin – which showed weak antibiotic activity against *Enterococcus faecalis*. A strain of *Geobacillus* isolated from a thermal spring in Jordan has been reported as producing three potentially novel antimicrobial peptides with activity against *Bacillus subtilis* and *Salmonella* Typhimurium (Alkhalili et al., 2016).

Studies to determine the microbiome of thermal spring sites around the world have revealed a greater than expected abundancy and diversity of microbial life, which varies between different spring sites and seems to be defined by the geochemistry of each particular site. Investigations into several thermal spring sites at Yellowstone National Park showed a variation in the microbiomes between springs but the dominant phyla were Aquificae and Thermodeulfobacteria. This led the researchers to suspect that aerobic hydrogen oxidation was an important metabolic reaction in these microbial ecosystems, and they subsequently confirmed the presence of this reaction in the water using analytical techniques (Meyer-Dombard, Shock and Amend, 2005). Researchers studying the thermal groundwater of the Uzon Caldera, Russia (Mardanov et al., 2011) found it to contain a high sulphate content and that the most abundant bacterial phyla were Acidithiobacillus and Verrucomicrobia which are known to oxidise sulphates. This study also revealed that over 70 % of the microbiome was comprised archaea, most of which affiliated with “uncultured” lineages. This suggested an

abundance of novel microorganisms yet to be discovered reside in this site. An investigation on the costal thermal springs of the Reykjanes Peninsula, Iceland revealed the influence that temperature has on microbiome composition (Hobel et al., 2005). These springs are unique for the large temperature differences between them, ranging between 45-95 °C. Some of the springs are also covered by sea water during high tide, meaning that temperatures can fluctuate by up to 100 °C daily. The researchers found that springs that had the highest temperatures had greater proportions of thermophilic bacterial whilst moderate temperature springs had a higher abundance of mesophilic marine microbes.

1.7. Aims & Objectives

The work detailed in this thesis aimed to contribute towards addressing the antibiotic discovery void by bioprospecting underexplored environmental sites for microbially-derived antibiotic compounds. The sites explored in this thesis include a human oral microbiome, raw honey, and the King's Bath thermal spring of the Roman Baths, UK. The hypothesis for choosing these sites is that novel environmental niches would contain novel ecosystems and microorganisms adapted to survive in these conditions which may therefore produce secondary metabolites of novel structure and function.

In addition to techniques in microbiology and natural product chemistry, metagenomic and genome-mining techniques were also utilised in order to attempt to address some of the limitations of the bioprospecting pipeline such as high rediscovery rates and untargeted chemical extraction strategies. The rationale for this approach being that it would limit resource spent on unviable leads by introducing early dereplication and also provide insights into the nature of the secondary metabolite potential of both individual isolates and microbiomes so that effective bioactivity-guided isolation strategies could be utilised. Additionally, taxonomic surveys of prospective microbiomes could help inform effective culturing strategies.

A final aim of the work in this thesis was to utilise and develop pipelines for the use of Oxford Nanopore long-read sequencing for genome mining and metagenomic sequencing that could be used in dereplication and microbiome profiling. Effective genome-mining requires access to high quality genomes which can be uneconomical for small-scale studies when using other sequencing platforms. Additionally, common short-read sequencing may not lead to effective assembly of long and repetitive biosynthetic pathways which can reduce confidence in their annotations. Likewise, recreation of biosynthetic pathways from short-read metagenomic datasets is extremely challenging and resource intensive. To address these limitations, a pipeline for the genome sequencing, assembly and annotation of microbial isolates using Oxford Nanopore long-read sequencing was developed for use in dereplication and predictions of secondary metabolite pathways. Additionally, a pipeline was developed to use long-read shotgun sequencing to profile the taxonomy of the microbiomes and to annotate long biosynthetic pathways without the need for high-power computational processing.

The objective of the work detailed in **Chapter 2** was to identify putative NRPS and PKS genes in a human oral microbiome. A PCR screen was used to amplify NRPS and PKS gene

fragments from a human oral metagenome, which were then examined for homology to other NRPS/PKS genes in the NCBI database to determine their novelty.

The objective of the work detailed in **Chapter 3** was to use whole genome sequencing and genome-mining to help characterise and dereplicate for four antibiotic producing bacteria. These bacteria were isolated from raw honey were screened for antibiotic activity using a cross-streak bioassay in order to identify leads with antibiotic potential. The next step was to determine if the isolates were distinctive which was done through a combination of molecular typing and genomic digestion. Finally, genome-mining was used as a dereplication strategy to prioritise leads. The genomes of the isolates were sequenced and assembled using Illumina short-read sequencing technology and bioinformatics tools. The assembled draft genomes were analysed using bioinformatics tools to identify putative natural product biosynthetic gene clusters and predict the product structures.

The objective of the work detailed in **Chapter 4** was to develop a pipeline for the sequencing and assembly of bacterial genomes using Oxford Nanopore MinION for the purposes of bioprospecting using *Streptomyces coelicolor* A3(2). The accuracy of this data for dereplication of known BGCs and prediction of potentially novel BGCs was assessed by comparison to reference genome sequence. This pipeline was then implemented to analyse the genome of a novel antibiotic-producing microorganism in **Chapter 5**.

The objective of the work detailed in **Chapter 5** was to characterise the taxonomy of an antibiotic producing isolate from a thermal spring and to attempt to isolate and identify the antibiotic compounds it may be producing using a combination of genome-mining and bioactivity-guide isolation. A bacterial isolate with antibiotic activity that was isolated from the hot-spring of the Roman Baths, UK was characterised using the genome-sequencing pipeline developed in **Chapter 4.** Information on taxonomy and biosynthetic potential of the isolate was determined and used to inform a bioactivity-guided strategy to isolate the pure compound of interest.

The objective of the work detailed in **Chapter 6** was to develop a pipeline to profile taxonomic and functional gene diversity of the microbiome of the King's Bath in the Roman Baths by means of 16S rRNA gene and shotgun metagenomic sequencing using Oxford Nanopore technology. The site was investigated due to the discovery of the antibiotic isolate detailed in **Chapter 5**. A metagenome of water was profiled by 16S rRNA sequencing to get an understanding of the microbiome diversity and to serve as a comparison against shotgun sequencing for taxonomic profiling. Two approaches to preparing the metagenome for

shotgun sequencing (multiple displacement amplification and PCR) were compared for performance and taxonomic accuracy. Attempts were then made to analyse the data for the detection of long biosynthetic genes and also singular genes of metabolic interest such as metal resistance.

Chapter 2.

Sequence-Led Investigation to Identify Putative NRPS and PKS Genes in the Human Oral Metagenome

2.1. Introduction

The oral cavity includes the tongue, teeth, gingival sulcus, checks, palates and tonsils, all of which are distinctive habitats colonised by microorganisms. The Human Oral Microbiome Database (Chen et al., 2010) contains 16S rRNA sequences from the human oral microbiome mapped to 619 taxa across 13 phyla; Actinobacteria, Bacteroidetes, Chlamydiae, Chloroflexi, Euryarchaeota, Firmicutes, Fusobacteria, Proteobacteria, Spirochaetes, SR1, Synergistetes, Tenericutes, and TM7. Additionally, the oral cavity is a major site of interaction and exchange of material between the human body and the environment, and there have been studies reported which suggest a correlation between the composition of the oral microbiome and human disease (Seymour et al, 2007). The human oral microbiome is therefore a complex and diverse ecosystem which may have an important role in the maintenance of health and prevention of disease. From evidence of natural products isolated from other human microbiome sites, it is possible that part of this contribution could be through natural products produced by some members of this ecosystem. Analysis by Floyd and co-workers (2010) into the human oral microbiome by 16S rRNA gene profiling determined that there may be up to 1,179 taxa present and 68% are uncultivated phylotypes. These putative novel species of bacteria present in the oral microbiome may produce natural products with novel chemistry.

The work detailed in this chapter used sequence-led metagenomics to investigate the natural product biosynthetic potential of uncultivated microorganisms of the oral microbiome from healthy human volunteers. A PCR screen was used to amplify NRPS and PKS gene fragments from a human oral metagenome, which were then examined for homology to other NRPS/PKS genes in NCBI database to determine their novelty. This work could then inform on the potential of the oral microbiome as a source of natural product leads.

2.2. Material and Methods

2.2.1. Extraction of Genomic DNA from Human Saliva

Nine healthy human volunteers each donated approximately 2-3 mL of saliva, giving an approximate total of 22 mL. Collection of human saliva for research was performed under UCL ethics committee project ID: 5017/001.

The saliva from nine donors was first pooled together then then split into two 50 mL Falcon tubes (approximately 11 mL in each) and centrifuged for 10 mins at 5,000 rpm / 4 °C. The supernatant was removed, and the two pellets were each resuspended in 4 mL of 'cell suspension solution' from the Gentra® Puregene® Yeast/Bact. Kit (Qiagen, Germany). Four 1 mL aliquots were made from each resuspension into 1.5 mL microcentrifuge tubes to produce a total of eight aliquots from the original pooled saliva samples. The eight aliquots were each centrifuged for 5 mins at 4,000 rpm / 25 °C and the supernatants were removed, and the pellets were resuspended in 600 µL 'cell suspension solution'. Genomic DNA was extracted from each of these eight suspensions using the Gentra® Puregene® Yeast/Bact. Kit (Qiagen) protocol for Gram-positive bacteria. Two modifications were made to the manufacturer's protocol, which were to increase the standard volume of lytic enzyme and RNAase A used in each reaction to 7.5 µL and 5 µL respectively. The eight genomic DNA (gDNA) extractions were dissolved in TE buffer (10 mM Tris pH 8.0, 1 mM EDTA) and analysed on a Nanodrop.

2.2.2. Gel Analysis of Genomic DNA Extractions

Approximately 100 ng of each extraction was run on a 0.8 % agarose gel containing 0.5 % GelRed™ (Biotium, Inc, USA) in TAE buffer (40mM Tris, 20mM acetic acid, and 1mM EDTA) for 90 mins at 80 V. Prior to loading, each sample was mixed with Purple Gel Loading Dye (New England BioLabs Inc., USA) according to the manufacturer's protocol. As a molecular size marker, λ DNA-*Hin*dIII Digest (New England BioLabs Inc., USA) was used according to the manufacturer's protocol. The gel bands were visualised on an UV transilluminator.

2.2.3. Storage of Oral Metagenome

After gel analysis, the gDNA was pooled together and diluted to ~50 ng/µL in TE buffer. An aliquot (100 µL) was stored at -20 °C to be used as a working stock for PCR reactions. The remaining metagenome was then aliquoted (50 µL) across sixty-one 0.5 mL PCR tubes and stored at -20 °C with the tops of the tubes sealed with parafilm.

2.2.4. PCR Amplification of Adenlyation and Ketosynethase Domains from Oral Metagenome

Table 2.1. details the primers used in this study. The reaction mixture (50 µL) contained: 25 µL 2X MyTaq™ Red Mix (Bioline Reagents Ltd., UK), 4 µL oral metagenome working stock (50 ng/µL), 5 µL of each primer (10 µM), 11 µL H_2O. Reaction conditions for amplification using the A3F/A7R primer pair were 94 °C for 4 min, 35 cycles [94 °C for 30 secs, 67.5 °C for 30 secs, 72 °C for 60 secs], and 72 °C for 5 mins. For the degKS2F/degKS2R primer pair, conditions were 94 °C for 4 min, 35 cycles [95 °C for 40 secs, 56.3 °C for 40 secs, 72 °C for 75 secs], and 72 °C for 5 mins. *Streptomyces niveus* gDNA was used as a positive control.

Table 2.1. Details of the degenerate primers designed to amplify sections of NRPS and PKS genes

Primer pair (5'-3')	Intended product	Product length (bp)	Source
A3F – GCSTACSYSATSTACACSTCSGG **A7R – SASGTCVCCSGTSCGGTA**	**Conserved A3 and A7 regions of the adenylation domain of NRPS genes**	**~700 bp**	**Ayuso-Sacido & Genilloud (2005)**
degKS2F – GCIATGGAYCCICARCARMGIVT **degKS2R – GTICCIGTICCRTGISCYTCIAC**	**Conserved regions of the ketosynthase domain of PKS genes**	**~700 bp**	**Schirmer et al. (2005)**

2.2.5. Gel Analysis of PCR Products

PCR reaction products were run on a 1 % agarose gel containing 0.5 % GelRed™ (Biotium Inc., USA) in TAE buffer for 45 mins at 100 V. Prior to loading, each sample was mixed with Purple Gel Loading Dye (New England BioLabs Inc., USA) according to the manufacturer's protocol. As a molecular size marker, 100 bp ladder (New England BioLabs Inc., USA) was used according to the manufacturer's protocol.

2.2.6. Gel Extraction of PCR Products of Interest

Bands of interest were cut out from the 1 % agarose gel and extracted using the Monarch® DNA Gel Extraction Kit (New England BioLabs Inc., USA) according to the manufacturer's protocol.

2.2.7. Cloning of PCR Products

The gel extracts were ligated into pGEM®-T Easy Vector (Promega Corp., USA) and transformed into *E. coli* cells (α-Select Silver Competent Cells; Bioline Reagents Ltd., UK) according to the manufacturers' protocols. The transformant cultures were spread onto LB/ampicillin/IPTG/X-Gal plates (LB agar, 0.2 % ampicillin (50 mg/mL), 0.1 % IPTG (0.5 M), 0.4 % X-Gal (20 mg/mL)) and incubated at 37 °C overnight.

2.2.8. Selection of Transformant Colonies and Plasmid Extraction

Transformant colonies were selected using blue/white screening. White colonies were subcultured onto fresh LB/ampicillin/IPTG/X-Gal plates and incubated overnight at 37 °C. Subcolonies which still appeared white where then inoculated into 10 mL LB broth containing 200 µL/mL ampicillin and incubated overnight at 37 °C with shaking (200 rpm). The recombinant vectors were extracted using Monarch® Plasmid Miniprep Kit (New England BioLabs Inc., USA) according to the manufacturer's protocol.

2.2.9. Selection and Sequencing of Recombinant Plasmid Inserts

Aliquots (1 µL) of each plasmid extraction were digested with *Eco*RI (New England BioLabs Inc., USA) according to the manufacturer's protocol to release the insert. The inserts were analysed on 1 % agarose gel as previous described (Section 2.5). Plasmids that contained inserts that were of the intended size were sequenced using T7 5′-d(TAATACGACTCACTATAGGG)-3′ and SP6 5′-d(TATTTAGGTGACACTATAG)-3′ primers on an ABI 3730xl DNA Analyzer by Genewiz Inc., with the samples prepared to the supplier's specification.

2.2.10. Analysis of Insert Sequences

The chromatograms of obtained nucleotide sequences were manually assessed for quality using SnapGene Viewer 3.0.1. (GSL Biotech LLC, USA). Sequences deemed of sufficient quality were trimmed by aligning the insert sequences against the flanking regions of the vector using Bioedit 7.2.5. (Ibis Biosciences, USA). The trimmed forward and reverse sequences were then aligned to check for discrepancies using the CLUSTAL W package in Bioedit. Any discrepancies were investigated by manually reviewing the chromatograms in SnapGene Viewer. Once a consensus sequence was obtained it was saved in FASTA file format in Bioedit. The FASTA files were converted to a translated sequence and compared for homology using BLASTx (http://blast.ncbi.nlm.nih.gov, National Center for Biotechnology Information, USA).

2.3. Results and Discussion

2.3.1. Quality Assessment of Oral Genomic DNA

Table 2.2 shows the nanodrop analysis of the eight human saliva genomic DNA extractions. Each extraction was initially suspended into 100 µL of TE buffer. However, the DNA pellets in extractions 3, 5, and 7 did not fully dissolve so the volume was increased to 500 µL. Estimation of the concentration of each sample shows that a total of ~277.8 µg of DNA had been extracted from the 22 mL of human saliva.

Table 2.2. Nanodrop analysis of each genomic DNA extraction of pooled saliva samples. Results show that the extracts produced a high yield of DNA but some contamination remained in the samples.

Sample (volume in µL)	Concentration (ng/µL)	Absorbance ratios	
		260/280	260/230
1 (100)	311.4	1.64	0.86
2 (100)	162.6	1.74	0.98
3 (500)	97.60	1.62	0.62
4 (100)	325.2	1.63	0.85
5 (500)	106.1	1.58	0.58
6 (100)	339.6	1.62	0.82
7 (500)	90.40	1.61	0.63
8 (100)	168.5	1.80	1.23

The absorbance ratio at 260/280 nm of each gDNA extraction was taken to assess their purity. Nucleic acid absorbs strongly at 260 nm while other common biological molecules, such as amino acids, absorb strongly at 280 nm. It is generally accepted that a 260/280 nm ratio of 1.7-2.0 is suitable for most downstream molecular cloning purposes (Sambrook & Russell, 2001). The mean 260/280 nm ratio of the gDNA extractions is slightly below this threshold at 1.65. This could be due to a high carry over of amino acids from the extraction process. A secondary ratio measurement at 260/230 nm was also taken. Typically expected 260/230 ratio values for a "pure" DNA sample is in the range 2.0-2.2. The mean 260/230 ratio recorded for the eight extractions was 0.82. This could be due to the presence of carbohydrates that absorb strongly at 230 nm being carried over from the extraction process. The standard Gentra® Puregene® protocol is designed for use with a 0.5 mL sample of overnight (~1 x 10^9 per mL) bacterial culture in a simple medium (such as LB broth). The protocol was adapted for use with a higher volume of bacterial cells in human saliva by modifying the protocol to introduce two cell washing steps, increase the volumes of lysis solution and RNAase A used, split the extraction across eight replicates. However, it is possible that due to the complex nature of human saliva (Neyraud et al., 2012) and the higher volume of bacterial cells present in it, there was a higher than normal carry over of peptides and carbohydrates into the final DNA

suspensions. These impurities could interfere with downstream cloning processes, so further purification of the DNA using techniques such as alcohol precipitation (Sambrook & Russell, 2001) could be done to remove these impurities before cloning.

The eight DNA extractions were each analysed on an agarose gel to assess the size of gDNA fragments that were obtained (**Figure 2.1**). The Gentra® Puregene® Yeast/Bact. Kit DNA extraction kit was used because it is designed to extract high molecular weight DNA in the range 100–200 kbp. Most biosynthetic gene clusters of NRP and PK antibiotics are relatively large. For example, the gene clusters for penicillin, chloramphenicol, and erythromycin are 59.5, 22.0, and 54.6 kbp respectively (Genomic Standards Consortium, 2017). Therefore, it is important to obtain high molecular weight DNA in order to maximise the chance of capturing and cloning NRP and PK biosynthetic gene clusters in their entirety. If a partial gene cluster is isolated from a cloning vector, then further downstream steps would have to be undertaken to isolate the remaining gene cluster and reassemble it (Kim et al., 2010).

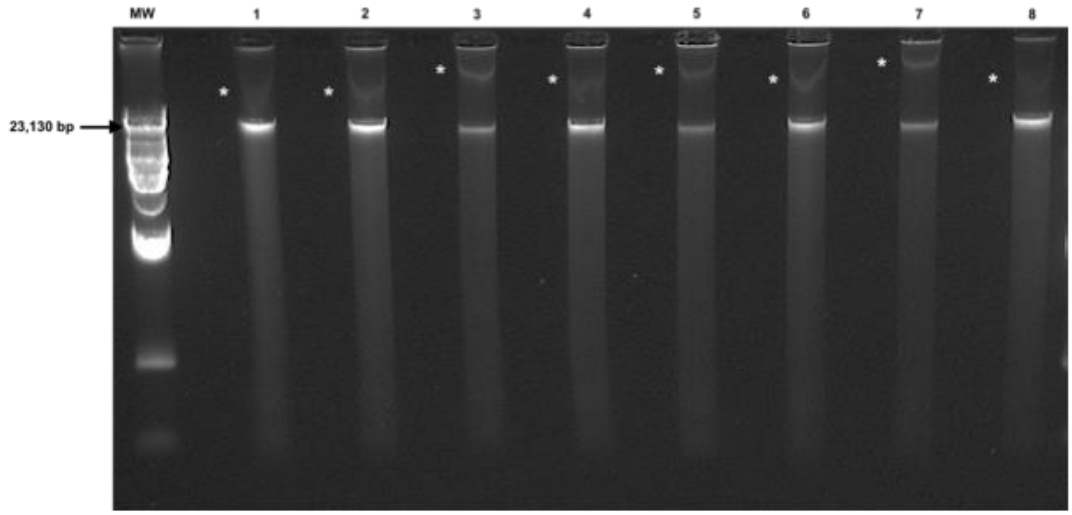


Figure 2.1. Agarose gel (0.8 %) of 1 µL (~ 100 ng) of each of the eight genomic DNA extractions made from the pooled saliva from nine human volunteers. MW = λ DNA-*Hin*dIII digest molecular size marker, 1 - 8 = saliva DNA extraction.

All eight extractions show a distinct band that has migrated along with the 23 kbp band of the molecular marker, and faint bands upstream from this which are marked with an '*'. There is also smearing beneath the distinct bands that continue down to the 2 kbp marker. The presence of a distinct band at the 23 kbp marker was suggestive of high molecular weight DNA. If the DNA had been totally degraded during the extraction, either by nucleases or physical shearing, then there would be no distinct band but only a smeared band that would migrate below the 23 kbp band. Standard agarose gels cannot distinguish the size of DNA above ~15 kbp, as DNA at this size and above will migrate together. However, the presence of faint and

misshapen bands upstream from the 23 kbp marker suggests the presence of DNA fragments of a length that could contain an entire NRP or PK biosynthetic gene cluster. The smearing seen could be due to degradation of the DNA, however it could also be an artefact caused by the presence of contaminants in the DNA samples that absorb UV, or because of an overloading of the sample into the agarose gel. Pulse-field electrophoresis is a technique designed to separate large DNA fragments (up to 2 Mbp) by applying alternating voltage fields across different axis. This technique could be used to assess the true size range of the gDNA extraction. If it is found that there is low molecular weight DNA present in the extractions (below 50 kbp), then the larger fragments could be purified by gel extraction and dialysis before cloning.

2.3.2. PCR Reaction and Cloning of PCR Products

Figure 2.2 shows PCR products using both the degenerate degKS2F/degKS2R and A3F/A7R primer sets. These two primers sets were chosen to screen for NRPS and PKS genes because they have been reported in previous studies to be successful (Ayuso-Sacido & Genilloud, 2005; Schirmer et al., 2005; Owen et al., 2013). Both primer sets also amplify an expected PCR product of ~700 bp, this length is suitable for TA-cloning and can be reliably sequenced using Sanger sequencing. The gel analysis of the PCR reactions shows that there are bands present between the 500-1000 bp markers for both the degKS2F/degKS2R or A3F/A7R reactions of the oral metagenome (**Figure 2.2: lanes 1 and 4**), which is in agreement with the intended product size of ~700 bp. The reactions with *S. niveus* gDNA also showed bands in this region (**Figure 2.2: lanes 2 and 5**). *S. niveus* gDNA was used as a positive control to compare the efficacy of the PCR reaction. The *S. niveus* genome has been sequenced and the presence of NRPS and PKS genes have been identified (Flinspach et al., 2014). The negative control reactions using water in place of gDNA was performed to assess if the PCR reaction mix contained any nucleic acid contaminants that may result in false positive results. The absence of bands in the negative control reactions (**Figure 2.2: lanes 3 and 6**) indicate that the reaction mix was free from contaminants. The bands within the 500-100 bp range were all excised from the gel and ligated into cloning vectors and transformed into competent *E. coli* cells.

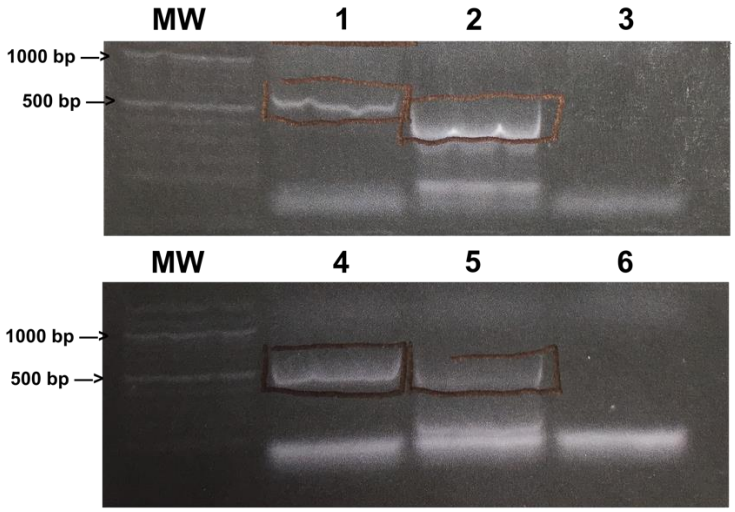


Figure 2.2. Agarose gel (1 %) of the PCR reactions (50 µL) of oral metagenome and *Streptomyces niveus* gDNA with degKS2F/degKS2R or A3F/A7R primer sets. MW = molecular size marker, 1 = oral metagenome with degKS2F/degKS2R, 2 = *S. niveus* gDNA with degKS2F/degKS2R, 3 = degKS2F/degKS2R negative control, 4 = oral metagenome with A3F/A7R, 5 = *S. niveus* gDNA with A3F/A7R, 6 = A3F/A7R negative control.

Blue/white screening was used to distinguish between recombinant and non-recombinant transformants. This is because, during the cloning protocol, some plasmid vectors will self-ligate and be transformed into the competent cells. So, using only the antibiotic resistance gene contained in the vector to select for transformants can give false positive results. The competent *E. coli* cells contain an inactive mutant β-galactosidase (*LacZ*) gene which has a section deleted. The vector contains the deleted section of this gene (*lacZα*). When the vector is transformed into the competent cells, the two genes complement to form a functional β-galactosidase enzyme which is able to metabolise X-gal present in the agar to form a blue by-product. Therefore, non-recombinant transformants appear blue. The *lacZα* gene is positioned within the multiple cloning site of the vector. If a fragment of DNA is successfully ligated into this region it disrupts the *lacZα* gene, preventing the production of β-galactosidase. Therefore, transformants with a recombinant vector remain white (Sambrook & Russell, 2001).

After cloning of the PCR amplicons, 50 white transformant colonies each from the A3F/A7R and degKS2F/degKS2R oral metagenome reactions, and 10 each from the *S. niveus* reactions were subcloned for further analysis. The transformants were subcloned in order to obtain a pure colony of each and check for the stability of the recombinant insert. **Table 2.3.** shows that many of the initially white transformant colonies become blue after subcloning. The low efficiency of subcloning could be due to the DNA inserts not being stable in the vector. Assessment of the recombinant insert sequences (data not shown) show the inserts are GC-rich and contain multiple nucleotide repeating sequences. Repetitive and GC-rich DNA can form secondary structures that places super-helical stress on circular plasmids, leading to

deletion of the insert. Repetitive DNA, if translated by the host cell machinery, can be deleterious to the host and create a selection pressure for deletion of the insert (Godiska et al., 2010).

Table 2.3. Percentage of transformants that remained white after subcloning

Transformation reaction	Percentage of tranformants that remain white after subcloning
A3F/A7R oral metagenome	22 % (11/50)
degKS2F/degKS2R oral metagenome	14 % (7/50)
A3F/A7R *S. niveus* gDNA	20 % (2/10)
degKS2F/degKS2R *S. niveus* gDNA	10 % (1/10)

2.3.3. Bioinformatics Analysis of PCR products

The recombinant inserts from positive subclones were excised from their vectors by *Eco*RI digestion and gel analysed to check that the inserts were of the expected size range (data not shown). The inserts were then sequenced and analysed by BLASTx. **Tables 2.4.** and **2.5.** show the top results from this analysis. Five out of ten (50 %) A3F/A7R oral metagenome derived inserts (AD_4, AD_5, AD_6, AD_8, and AD_12) showed similarity with NRPS genes. Four out of seven (57 %) degKS2F/degKS2R oral metagenome derived inserts (KS_4, KS_6, KS_8, and KS_9) showed similarity with PKS genes. For the *S. niveus* inserts, one out of two (50 %) A3F/A7R, and the only degKS2F/degKS2R (100 %) derived insert showed a 99 % identity match with the NRPS and PKS genes in the *S. nivues* NCBI reference genome sequence (WP_031228505.1) respectively. All of the amplicons with matches to NRPS/PKS genes were within ≤ 28 bp of the intended size range of 700 bp. These results demonstrated that both PCR reactions successfully amplified segments of NRPS and PKS genes but that the reactions were not efficient as some unintended genes were also amplified.

Table 2.4. BLASTx results of sequenced amplicons from PCR using A3F/A7R. All amplicons are from the oral metagenome except AD_1_Strep_niveus & AD_2_Strep_niveus, which are from *S. niveus* gDNA. Inserts of interest are highlighted grey. All inserts are the consensus sequence of forward and reverse sequencing except for AD_9 and AD_14.

Query (size in bp) (GC %)		Description	Max score	Total score	Query cover	E value	Ident	Accession
AD_2 (691 bp) (55.7 %)		peptide-methionine (S)-S-oxide reductase [*Lachnoanaerobaculum* sp. ICM7]	162	310	85%	3e-75	80%	WP_009663155.1
AD_3 (941 bp) (54.7 %)		peptide ABC transporter substrate-binding protein [*Cryptobacterium curtum*]	358	358	57%	1e-117	96%	WP_012802862.1
AD_4 (713) (68.6 %)		type I polyketide synthase [*Lautropia mirabilis*]	479	479	99%	9e-153	99%	WP_005673253.1
AD_5 (714) (64.0 %)		amino acid adenylation domain-containing protein [*Variovorax* sp. NFACC29]	317	317	99%	1e-105	63%	SEG99475.1
AD_6 (722 bp) (66.3 %)		Amino acid adenylation domain protein [Delftia sp. RIT313]	265	329	99%	8e-78	59%	EZP44834.1
AD_8 (713 bp) (67.1 %)		non-ribosomal peptide synthetase [*Actinomyces* sp. oral taxon 170]	464	671	99%	1e-147	95%	WP_034514267.1
AD_9	**Forward (392 bp) (57.9 %)**	N-acetyl-1-D-myo-inosityl-2-amino-2-deoxy-alpha-D-glucopyranoside deacetylase MshB [*Rothia aeria*]	233	233	90%	7e-73	95%	BAV88691.1
	Reverse (372 bp) (56.2 %)	Possible GTP-binding translation elongation factor [*Mycobacterium tuberculosis*]	202	202	79%	2e-62	95%	CNI88481.1
AD_12 (713 bp) (64.2 %)		amino acid adenylation domain-containing protein [*Variovorax* sp. NFACC29]	320	320	99%	9e-107	64%	SEG99475.1
AD_13 (950 bp) (45.9 %)		ABC transporter substrate-binding protein [*Sphaerochaeta pleomorpha*]	289	289	82%	9e-92	50%	WP_014271833.1
AD_14	**Forward (392 bp) (57.9 %)**	N-acetyl-1-D-myo-inosityl-2-amino-2-deoxy-alpha-D-glucopyranoside deacetylase MshB [*Rothia aeria*]	233	233	90%	7e-73	95%	BAV88691.1
	Reverse (372 bp) (56.2 %)	Possible GTP-binding translation elongation factor [*Mycobacterium tuberculosis*]	202	202	79%	2e-62	95%	CNI88481.1
AD_1_Strep_niveus (709 bp) (68.3 %)		non-ribosomal peptide synthetase [*Streptomyces niveus*]	488	692	99%	8e-156	99%	WP_031228505.1
AD_2_Strep_niveus (792 bp) (67.6 %)		short-chain dehydrogenase/reductase [*Streptomyces niveus*]	350	350	73%	2e-119	99%	WP_023540658.1

Table 2.5. BLASTx results of sequenced amplicons from PCR using degKS2F/degKS2R. All amplicons are from oral metagenome except KS_1_Strep_niveus, which is from *S. niveus* gDNA. Inserts of interest are highlighted grey.

Query (size in bp) (GC %)	Description	Max score	Total score	Query cover	E value	Ident	Accession
KS_3 (870 bp) (57.1 %)	7-cyano-7-deazaguanine synthase [*Porphyromonas somerae*]	381	381	66%	1e-131	95%	WP_060935291.1
KS_4 (676 bp) (60.5 %)	type I ketosynthase [uncultured bacterium]	271	271	99%	2e-89	60%	ADD65217.1
KS_5 (724 bp) (66.7 %)	peptidylprolyl isomerase [*Actinomyces* sp. HMSC035G02]	277	277	59%	7e-90	97%	WP_070835585.1
KS_6 (662 bp) (60.4 %)	type I ketosynthase [uncultured bacterium]	259	259	99%	2e-84	59%	ADD65217.1
KS_7 (672 bp) (55.2 %)	hypothetical protein [*Daphnia magna*]	107	107	38%	5e-26	63%	JAN83343.1
KS_8 (676 bp) (60.2 %)	type I ketosynthase [uncultured bacterium]	268	268	99%	4e-88	59%	ADD65217.1
KS_9 (683 bp) (63.1 %)	polyketide synthase [*Corynebacterium matruchotii*]	457	457	99%	2e-147	99%	WP_005525091.1
KS_1_Strep_niveus (682 bp) (72.1 %)	type I polyketide synthase [*Streptomyces niveus*]	417	1263	99%	6e-131	99%	WP_023543239.1

Insert AD_8 showed 95 % similarity with an NRPS gene from the genome of an Actinomycetes associated with the human oral microbiome. Insert KS_9 showed 99 % similarity with a PKS gene in the *Corynebacterium matruchotii* genome, which is also an Actinomycetes associated with the human oral microbiome. Bacteria in the order Actinomycetales are a prolific source of diverse natural products. It is estimated that over 5000 antibiotic compounds have been isolated from Actinomycetales (Challinor & Bode, 2015).

Despite being derived from primers designed to isolate NRPS genes, the top hit for insert AD_2 showed a 99 % similarity with a PKS gene in *Lautropia mirabilis* (NCBI reference genome sequence WP_005673253.1). *L. mirabilis* is a Gram-negative in the order Burkholderiales, which is associated with the oral microbiome (Gerner-Smidt et al., 1994). Inspection of the genome annotation record (data not shown) showed that it is a NRPS-PKS hybrid gene. PKS-NRPS hybrids catalyse the synthesis of molecules containing both amino acid and ketone starter units. Virginiamycin is an example of an NRPS-PKS hybrid antibiotic (Pulsawat et al., 2007). Bacteria from the order Burkholderiales have recently been reported as a potentially rich source of diverse natural products (Challinor & Bode, 2015). *Burkholderia ambifaria*, which is an opportunistic pathogen of cystic fibrosis sufferers, was found to produce an NRPS-PKS hybrid called enacyloxin IIa which has antibiotic activity against MDR Gram-negative pathogens (Mahenthiralingam et al., 2011).

Inserts AD_5 and AD_12 showed 63 % and 64 % similarity with NRPS genes from the genome sequences of a species of *Variovorax* bacteria and insert AD_6 showed 59 % similarity with the NRPS gene in a species of *Delftia*. *Variovorax* and *Delftia* are also in the order Burkholderiales, and *Variovorax* species has been isolated from the human oral cavity (Anesti et al., 2005) while a *Delftia* species has been isolated from aortic aneurysms alongside well characterized oral bacteria (da Silva et al., 2006). A search of the literature did not return any examples of antibiotic compounds being isolated from *Variovorax* or *Delftia* species. The similarly scores of 59-64 % for these three inserts suggested that they could be NRPS gene fragments from an uncharacterized bacteria.

Inserts KS_4, KS_6 and KS_8 showed similarity of 60 %, 59 %, and 59 % respectively to a PKS gene isolated from a soil metagenomics study (Xhao et al., 2011). It is therefore also possible that these three inserts are of PKS genes from uncharacterized bacteria. These findings demonstrate the potential for this PCR screening approach to isolate novel NRPS and PKS gene clusters from the oral microbiome.

Percentage identity matrices of the lead amplicons (**Tables 2.6 & 2.7.**) showed that the similarity between amplicons is low. These findings showed that the PCR screen had amplified a diverse range of potentially novel NRPS and PKS genes. All but one of the amplicons sequenced were GC-rich (**Tables 2.4. & 2.5.**) and this may be the reason for the low transformation efficiency of the PCR products. The reason why all the amplicons were GC-rich could be due to the primer sets used having a greater efficiency for GC-rich templates. The NRPS primer set (A3F/A7R) was created by alignment of GC-rich NRPS templates from Actinomycetes (Ayuso-Sacido & Genilloud, 2004) and were used because they have been reported to successfully used in prospecting of many environmental metagenomes by other researchers (Reddy et al., 2012; Amos et al., 2015; Zachary et al., 2016; Lemetre et al., 2017; Hover et al., 2018). Repeating the PCR reaction using a primer set based on AT-rich templates could lead to the amplification of AT-rich inserts and so increase the diversity of NRPS and PKS genes identified.

Table 2.6. Percentage identity matrix of sequenced amplicons from PCR using A3F/A7R that had a BLASTx match to a NRPS gene.

	AD_4	AD_5	AD_6	AD_8	AD_12	AD_1_strep_niveus
AD_4	-	27.4 %	38.5 %	29.8 %	43.1 %	25.6 %
AD_5	27.4 %	-	27.1 %	25.2 %	28.1 %	34.7 %
AD_6	38.5 %	27.1 %	-	29.6 %	42.6 %	25.9 %
AD_8	29.8 %	25.2 %	29.6 %	-	27.6 %	27.7 %
AD_12	43.1 %	28.1 %	42.6 %	27.6 %	-	28.6 %
AD_1_strep_niveus	25.6 %	34.7 %	25.9 %	27.7 %	28.6 %	-

Table 2.7. Percentage identity matrix of sequenced amplicons from PCR using degKS2F/degKS2R that had a BLASTx match to a PKS gene.

	KS_4	KS_6	KS_8	KS_9	KS_1_Strep_niveus
KS_4	-	23.9 %	98.6 %	24.5 %	34.1 %
KS_6	23.9 %	-	23.9 %	26.0 %	26.8 %
KS_8	98.6 %	23.9 %	-	25.0 %	34.0 %
KS_9	24.5 %	26.0 %	25.0 %	-	24.1 %
KS_1_Strep_niveus	34.1 %	26.8 %	34.0 %	24.1 %	-

Each insert was sequenced twice (forward: 5’ -> 3’ & reverse: 5’ <- 3’) and the two sequences were aligned to ensure accuracy. This was not possible for AD_9 and AD_14 as both sequence reactions terminated abruptly (**Figure 2.3**). This could be due to hairpin structures in the template DNA, formed because of the repetitive and GC-rich nature of the DNA, retarding the sequence reaction (Kieleczawa, 2006). Addition of DMSO to the sequencing reaction can

help to relax the hairpin structures of the template by competing for hydrogen bonds between the DNA bases.

Figure 2.3. Screenshots of the reverse and forward sequence chromatograms of inserts AD_9 and AD_14. The arrow highlights the area were the sequence quality suddenly drops.

2.3.4. Discussion and Future Directions

The study focused on PKS and NRPS genes because these pathways are a prominent class of natural products for which many antibiotic drugs in clinical use are derived from. The modular nature of these enzymes and conserved biosynthetic logic have allowed for the design of degenerate primers which have been reliably used in metagenomic investigations by multiple research groups. However, the primers only target relatively small and conserved regions of the large NRPS/PKS genes. Studying these amplicons in isolation precludes the identification of the full BGC pathway. Therefore, from the data generated from this study it is not possible to predict the potential structure of the scaffold to determine if the putative compound is novel.

Additionally, the use of a PCR screen with degenerate primers means that analysis is limited to BGCs with a highly conserved biosynthetic logic that is conducive to the design of such primers. Whilst NRPS and PKS compounds have been strong sources of antibiotic in the past, and so it is logical to focus attention on these, when wishing to find new classes of antibiotics then different biosynthetic classes should also be considered.

These findings add further support to the hypothesis that microbiomes from novel and extreme environments may play host to novel natural products. However, this conclusion assumes a strong association between phylogenetic diversity and natural product diversity. Studies have revealed horizontal gene transfer of PKS and NRPS clusters between different genera of bacteria (Nongkhlaw et al., 2016), fungi (Theobald, Vesth & Andersen, 2019), and even between the kingdoms of bacteria and fungi (Lawrence et al., 2011). It has also been noted by

Antony-Babu and co-workers (2017) that there is a weak association between 16S rRNA sequences and natural product profiles within the genera *Streptomyces*. These factors cannot be accounted for fully without isolation and analysis of full BGCs.

BLASTx was used to search the translated sequences as opposed to the primary nucleotide sequence in order to achieve a higher resolution on the potential novelty of the amplicons. This is based on the premise that missense mutations within conserved enzyme domain regions are selected at a lower frequency than silent mutations because of the affect the amino acid substitution may have on the function of the protein. However, missense mutations can also have minimal or no effect on protein structure and function if the substituted amino acids have similar physiochemical properties. Therefore, the novelty of the small amplicon sequence may not have a direct correlation on the novelty of the enzyme or its product. Also, as with all sequence-led metagenomic studies, no indication as to the potential function of the PKS/NRPS products. Therefore, no determination can be made if they may have antibiotic or other functional properties of commercial or medicinal interest.

It was noted previously that the amplicons isolated had a high GC content and this may have been caused through the use of primers optimised for high GC sequences. It is also possible that the diversity of amplicons obtained may have been influenced by the DNA extraction method chosen giving bias towards more efficient extraction for some species. The DNA extraction method used in this study relies upon a chemical lysis method with purification relying upon salt and alcohol precipitation steps. It was determined that this method would offer the best yield of HMW DNA as use of a mechanical lysis method may have led to shearing of DNA. This may not influence the efficiency of isolation of small amplicons by PCR, but full PKS/NRPS biosynthetic gene clusters are typically in the range of 10-100s kbp in size. Therefore, fragmentation of the oral metagenome could preclude the potential to target and study the full BGCs that the amplicons are isolated from at a later date. However, the composition and thickness of bacterial cell walls varies widely between species and so a "gentle" chemical-based lysis may not effectively lyse all species with more complex cell walls. A review of the efficiency of chemical and mechanical gDNA extraction methods against vegetative cells and spores of a range of bacterial species found larger variations in efficiency of between methods for each species (de Bruin & Birnboim, 2016). A potential amendment to the experimental design could be to use different DNA extraction protocols against the subsets of the sample and pools successful extractions that return HMW to attempt to increase the breadth of species lysed.

Further investigation into this metagenome was suspended due to a move to a different laboratory. It was not feasible to continue the work in this new laboratory due to space and budgetary constraints and due to the administrative issues. However, there are many future directions to expand upon this study and address some of the limitations discussed, and the data presented merits further investigation to determine the potential novelty of the PKS and NRPS genes identified, as well as the bioactivity of the products they produce.

Shotgun sequencing and analysis of the metagenome could be used as a strategy for annotating entire BGCs *in silico*. This strategy could also allow for analysis of other BGCs besides NRPS and PKS for which is it is less viable to design degenerate PCR primers. Shotgun sequencing can also provide a taxonomic profile of the microbiome which would allow researchers to understand the nature of the microorganism in the environment. The strategy of shotgun sequencing of metagenomes for surveying biosynthetic potential is explored in the work detailed in **Chapter 6** regarding the microbiome of the Roman Baths, UK.

The study could be directed to isolate the complete biosynthetic gene clusters for the novel NRPS and PKS gene fragments that have been amplified. This would allow for reliable annotation of the full PKS and NRPS enzymes and a prediction of the core structure of their products to determine their novelty. To achieve this, high molecular weight DNA from the metagenome would be ligated into BAC vectors and cloned into competent *E. coli* to create an oral metagenome library. The isolated NRPS and PKS amplicons would be used as probes to isolate transformants containing the NRPS and PKS gene clusters by DNA hybridisation. The isolated vectors could then be extracted and sequenced by primer walking, and the architecture of the gene cluster and structure of the product it encodes could be predicted using bioinformatics tools. A similar approach was taken by Owen and co-workers (2013) in a study into the PKS and NRPS diversity of a soil metagenome. In that study the authors cloned eDNA into cosmid vectors and performed PCR screens on transformant colonies using barcoded degenerate primers. Amplicons identified to be PKS or NRPS genes were tracked back to the template transformant using barcodes and the full BGC was sequenced. The authors claimed to have identified putative BGCs for analogues of the antibiotics teicoplanin and friulimicin using this approach. This approach could also serve the additional benefit that the isolated BGCs could be expressed and purified for assessment of its bioactivity and structural elucidation.

Additional to sequence-led screens, a BAC library could be used for functional screens to isolate other putative natural products with desired bioactivity. However, functional screens performed by other researchers to isolate antibiotic compounds have focused predominately

on small peptide products rather than large and complex BGCs such as PKS and NRPS. This is in part due to the extra complexity of cloning larger BGCs such as PKS/NRPS. Long-length DNA requires very careful handling and may shear easily, resulting in partial BGC being cloned. Also, efficient expression of large BGC in heterozygous hosts is a complex task as the required substrates, which are often precursors produced by the native strain metabolism, would need to be present. The expression of host translational machinery can also bias the efficient expression of recombinant sequences. It has been estimated that the commonly used *Escherichia coli* expression hosts may only successfully translate up to 40 % of genes from a cloned metagenome library and this may be due to differences in relative GC content of the recombinant sequence differs significantly to the host (Vrancken, Van Mellaert & Anné, 2010).

As the amplicons isolated in this study had a high GC bias, it may be necessary to use an expression host with a similarly high native GC content. *Streptomyces* strains have been used for such purpose but expression and isolation of the recombinant natural products can be complicated by the presence of the expression hosts native natural product pathways competing for substrates and being co-isolated from extracts. Fazal and co-workers (2020) and have attempted to develop efficient expression host systems in high GC *Streptomyces* for natural product BGC expression. This work has involved removing other natural product pathways from the host genome which may interfere with the recombinant pathway expression and produce noise in functional screens and chromatograms. Hover and co-workers (2018) reported on the discovery and characterisation of a novel NRP called malacidin which had antibiotic activity against Methicillin-resistant *Staphylococcus aureus* by first taking a screening-led approach similar to that used by Owen and co-workers (2013) and then transferring the template BGC from the initial recombinant host to a modified *Streptomyces* host for heterogenous expression.

The method to isolate PKS/NRPS amplicons used in this study involved cloning the product into vectors and for transformation into recombinant host cells. Each amplicon in the mixture of PCR products would ideally be ligated into an individual vector and transformed successfully into a host cell whereby it is propagated for extraction and sequencing. This approach was chosen because it utilises well established gene cloning protocols and allows for propagation of the individual amplicons within the recombinant hosts to produce enough mass and purity for efficient Sanger sequencing. However, this approach is low throughput in its design and may introduce bias in which amplicon sequences are recovered. Efficiency of ligation and transformation of each amplicon may vary depending upon its sequence. It was noted in **Table 2.3** that transformation efficiency appeared low and this could have been due

to the high GC content of the amplicons. Amplicon sequences may also be toxic to recombinant hosts or unstable in the vector, leading or loss of the amplicon or death of the host. Use of high throughput Next-Generation sequencing could be utilised to address some of these potential limitations.

The study could also be expanded to include an analysis of the phylogenetic diversity of the oral metagenome. This may give a stronger indication of the potential novelty and diversity of PKS and NRPS genes present in the sample. Borsetto and co-workers (2019) recently combined high throughput 16 rRNA gene profiling with PKS and NRPS amplicon sequencing of 13 soil metagenomes from around the world. The authors contrasted the abundance and distribution of PKS/NRPS genes isolated with their phylogenetic profiles to estimate which phlya may be most prolific producers of such products. The authors found an association between PKS/NRPS gene abundance and the phyla Actinobacteria, Proteobacteria, Firmicutes, and Cyanobacteria in most soil types. They also noted that in soils from the extreme environment of the Antarctic there was an association between NRPS and PKS genes and the less characterised phyla Bacteroidetes and Verrucomicrobia.

If the work had been continued, then the next step taken would have most likely been to create a BAC library of the large fragments from the oral metagenome. The PKS and NRPS amplicons identified in this study could then have been used as probes to identify the template fragments within the BAC library by Southern blotting to recover and characterise the full BGC. This approach would have had the advantage of being able to potentially identify the BGCs without a requirement for the compounds to be expressed by the host. This approach would also offer versatility in the sense that the BAC library could also be used as the basis for functional-led screens. These functional screens could be used to identify novel BGCs which are not detected by the sequence-led screening methods because they have a currently unknown biosynthetic logic which precludes the design of molecular probes to target them.

2.3.5. Conclusions

In summary, the data presented here shows that a diverse range of NRPS and PKS genes fragments have been amplified from a human oral microbiome. Identity matches of the inserts with the NCBI database suggest that these inserts are from as yet uncharacterised gene clusters. These findings suggest that the human oral microbiome may be a source of novel natural products. The data obtained in this study has given an indication into the diversity of PKS and NRPS genes present in a human oral microbiome. Work did not continue due to practical limitation of a move of laboratories. However, the data presented merits further investigation to determine their potential novelty and the bioactivity of the products they produce

Chapter 3.

Genome-Mining Investigation to Identify Putative Natural Products by Antibiotic Producing Bacteria Isolated from Honey

3.1. Introduction

Honey has been used as an antimicrobial agent in many cultures throughout history and was often applied topically to clear infected wounds (Olaitan, Adeleke, & Ola, 2007). Modern *in vitro* investigations by various researchers have reported pure honeys to have a broad-spectrum inhibitory effect on the growth of many different Gram-positive, Gram-negative, and fungal species (Mullai & Menon, 2007; Sherlock et al., 2010). The antimicrobial activity of the substance is mainly attributed to the physical and chemical aspects of the substance such as low pH, hight osmolarity, and high peroxide content which makes conditions inhospitable to microorganisms. Some microorganisms do survive in honey, most often spore-forming species which are able to withstand the harsh conditions, such as *Bacillus* and *Streptomyces* (Gilliam & Prest, 1987). Whilst raw honey does not contain a microbiome *per se*, it has been reported to act as a reservoir of environmental microorganisms due to its frequent contamination with the microbiomes of the Honeybee, plants, pollen, as well as contaminants from the wind. Honey could therefore harbour antimicrobial-producing environmental bacteria that could be a source for bioprospecting. As detailed in **Section 1.5.4.1**, other researchers have reported on the isolation of antimicrobial-producing bacteria isolated from both the Honeybee microbiome and fresh honey from its hive (Olofsson et al., 2014).

Recent advances in next-generation sequencing technology has made whole genome sequencing of bacteria more attainable to researchers around the world. Combined with advances in genome-mining approaches, these advances have allowed bioprospectors to predict the biosynthetic potential of bacteria from these microbiomes. The advantages of this approach is that it allows researchers to make decisions on which lead isolates with bioactivity should be prioritised for further investigation based on the potential novelty of the biosynthetic genes predicted to be encoded in the genomes. It also allows researchers to identify isolates that are encoding known natural products so that they can be excluded from further investigation earlier in the discovery pipeline – a practice known as dereplication.

Based on the hypothesis that honey may contain microorganisms which produce antimicrobial compounds, a swab of honey was collected, and the bacteria was cultivated from it. The isolates were investigated for antimicrobial activity and genome-mining was utilised as a dereplication strategy to identify which leads were of potential interest for further investigation.

3.2. Materials and Methods

3.2.1. Cultivation of Bacterial Isolates from Honey

A swab of raw honey (supplied by Dr Jorge Gutierrezs from the University of Surrey) was inoculated onto 4% Brain Heart Infusion (BHI) (Oxoid Ltd., UK) agar at 37 ˚C for 16 hours. Each of the four cultivated isolates from this swab were inoculated in addition to isolates obtained from other sources and studies (these other isolates do not form part of the work in this thesis) into a well of a microtiter plate containing 100 µL BHI broth and incubated at 37 ˚C with shaking (200 rpm) for 16 hours. The plates were maintained at -80 ˚C after the addition of 20% (v/v) glycerol.

3.2.2. Indicator Strains Used in Antimicrobial Activity Assay

Escherichia coli NCTC 10418 and *Staphylococcus aureus* NCTC 12981 were maintained at -80 ˚C with 20% (v/v) glycerol and on Nutrient Agar slopes (Oxoid Ltd., UK). Strains were propagated directly from stocks by streaking to purity onto Nutrient Agar (Oxoid Ltd., UK) and incubating at 37 ˚C for 24 hours.

3.2.3. Screening of Honey Isolates for Antimicrobial Activity

The isolates were screened for antibiotic activity using the cross-streak assay. The test isolate was inoculated across a third of a Nutrient Agar (Oxoid Ltd., UK) plate and incubated at 37 ˚C for 24 hours. After initial incubation, a sterile loop was used to inoculate the indicator strains (*Escherichia coli* NCTC 10418 and *Staphylococcus aureus* NCTC 1298) in a straight line across the clear area of the plate, perpendicular to the test isolate. The indicators were streaked to within ~1 mm of the edge of the test isolate with care taken to not touch the test isolate with the loop. The plates were then incubated at 37 ˚C overnight, after which they were examined for antibiotic activity.

3.2.4. Extraction of Genomic DNA from Honey Isolates

Isolates were incubated in 5 mL BHI broth at 37 ˚C with shaking (200 rpm) for 24 hours, and gDNA was extracted from the cultures using Gentra® Puregene® Yeast/Bact. Kit (Qiagen, Germany) to the manufacturer's protocol.

3.2.5. Molecular Typing of Honey Isolates

The 16S rRNA and gyrase B genes of each isolate was amplified using the below primers (**Table 3.1**).

Table 3.1. Details of the degenerate primers used for molecular typing of honey isolates

Primer pair (5'-3')	Intended cDNA product	cDNA product length	Source
27F - AGAGTTTGATCCTGGCTCAG 1492R - GGTTACCTTGTTACGACTT	16S rRNA gene	~1500 bp	Lane et al. 1991
UP1 – GAAGTCATCATGACCGTTCTGCA YGCNGGNGGNAARTTYGA UP2R – AGCAGGGTACGGATGTGCGAGC CRTCNACRTCNGCRTCNGTCAT	Gyrase B gene	~1200 bp	Yamamoto & Harayama, 1995

The reaction mixture (50 µL) contained: 25 µL 2X MyTaq™ Red Mix (Bioline), 200 ng of gDNA, 20 uM of each primer, H_2O up to 50 µL. Reaction conditions for amplification using the 27F/1492R primer pair were 95 °C for 1 min, 30 cycles [95 °C for 15 secs, 54 °C for 15 secs, 72 °C for 90 secs], and 72 °C for 5 mins. For the UP1/ UP2R primer pair, conditions were 95 °C for 1 min, 30 cycles [95 °C for 15 secs, 60 °C for 15 secs, 72 °C for 72 secs], and 72 °C for 5 mins.

PCR products were purified using QIAquick PCR Purification Kit (Qiagen, Germany) according to the manufacturer's protocol and sequenced on an ABI 3730xl DNA Analyser by Genewiz Inc., with the samples prepared to the supplier's specification. The 27F/1492R amplicons were sequencing using the same primer pair whilst the UP1/ UP2R amplicons were sequenced using modified primer pair (UP1S - GAAGTCATCATGACCGTTCTGCA / UP2SR - AGCAGGGTACGGATGTGCGAGCC) (Yamamoto & Harayama, 1995).

3.2.6. Restriction Enzyme Digestion of Honey Isolate Genomic DNA

Genomic DNA (400 ng) was digested using *Hin*cII (New England BioLabs Inc., USA) according to the manufacturer's protocol. The digest was run on a 1% agarose gel containing 0.5 % GelRed™ (Biotium, Inc, USA) in TAE buffer for 3 hours at 50 V. Prior to loading, each sample was mixed with Purple Gel Loading Dye (New England BioLabs Inc., USA) according to manufacturer's protocol. As a molecular size marker, 1 kbp ladder (New England BioLabs Inc., USA) was used according to the manufacturer's protocol

3.2.7. Whole Genome Sequencing and Draft Genome Assembly of Honey Isolates

A single colony of each lead bacterial isolate was mixed in 100 µL sterile phosphate buffer solution (PBS) and streaked to purity onto 1.5% BHI agar and incubated at 37 °C for 16 hours. All of the bacterial culture was removed from the agar plate using a sterile 10 µL loop and mixed into a 2 mL Microbank™ beaded cryovial (Pro-Lab Diagnostics). After mixing the tube by inversion, excess cryopreservative solution was removed using a sterile pasteur pipette until it was level with the top of the beads. The tubes were then sealed and barcoded and transported to MicrobesNG by courier.

Whole genome sequencing and draft genome assembly was provided by MicrobesNG (BBSRC & The University of Birmingham). Sequencing was performed on the Illumina MiSeq and HiSeq 2500 platforms. Paired-end reads (2 x 250 bp) were analysed using Kraken (Wood & Salzberg, 2014) to determine the closest available reference genome. The reads were then mapped to the reference using BWA mem (http://bio-bwa.sourceforge.net) for quality analysis. *De novo* assembly of the reads was performed using SPAdes (Nurk, Bankevich et al., 2013), and the reads were mapped to the resultant contigs using BWA-mem for quality analysis. Automatic annotation of the contigs was performed using Prokka (Seemann, 2014).

3.2.8. Computational Analysis of Biosynthetic Potential of Honey Isolate Draft Genomes

GenBank files (.gbk) for each draft genome were analysed using AntiSMASH v3.0.5 (Weber et al., 2015). Settings: ClusterFinder algorithm - disabled, BLAST comparisons to other gene clusters – enabled, smCOG analysis for functional prediction and phylogenetic analysis of genes – enabled, Active site finder – enabled. The returned results were downloaded in GenBank format and used to generate a database which was used to compare the individual BGCs for homology using MutliGeneBlast (Medema, Takanko & Breitling, 2013). Predicted

NRPS/PKS catalytic domains, substrate specificities, and product structures predicted in AntiSMASH were manually curated using the NCBI conserved domain database (http://blast.ncbi.nlm.nih.gov, National Center for Biotechnology Information, USA). Putative bacteriocin BGCs were further analysed by searching for homologs in the bactibase database (Hammami et al., 2010). Putative NRPS and PKS multi-enzyme complexes were searched for homogeny in the MIBIG database (Medema et al., 2015).

3.3. Results and Discussion

3.3.1. Antibiotic Activity of Honey Isolates

Figures 3.1 shows images from the cross-streak assay used to assess the antibiotic potential of the four isolates from honey. The rationale of the assay is that once the test isolate has been established on the agar plate, any antibiotic compounds it produces will diffuse into the agar. These compounds will then inhibit the indicator strains that are streaked perpendicular to the test isolate, and this effect will be seen in the absence of growth of the indicator strains in the area closest to the test isolate. *Escherichia coli* NCTC 10418 and *Staphylococcus aureus* NCTC 12981 were used as indicators because these are characterised antibiotic-sensitive strains. Sensitive strains were used instead of resistant ones because the sensitive strains will be more susceptible to lower concentrations of antibiotic compounds, making it easier to observe if the test isolates are producing any antibiotics that may be present in low concentration.

Control assays using *Escherichia coli* NCTC 10418 were performed as it is known it does not produce compounds that would inhibit the growth of the indicators and so allow for comparison of the indicator streaks between the test isolates. The control assay (**Figure 3.1e**) showed no inhibition of the indicator strains as streaks can be seen to have grown to the within ~ 1 mm of the edge of the test isolate. In Y1 (**Figure 3.1a**), *E. coli* growth was absent at ~ 5 mm from the test isolate edge whilst *S. aureus* was absent at ~ 3 mm. In Y2 (**Figure 3.1b**), growth of both indicators was absent at ~ 5 mm from the test isolate edge. Isolate Y3 (**Figure 3.1c**) showed the strongest activity of all the isolates, with growth of both indicators absent at ~ 10 mm from the edge of the indicator. Isolate Y4 (**Figure 3.1d**) showed no inhibition of *E. coli* as growth can be seen within ~ 1 mm of the edge of the test isolate, as with the control assay. However, the growth of *S. aureus* was inhibited at ~ 5 mm of the test isolate edge. The results of the assay therefore showed that isolates Y1, Y2, and Y3 had broad-spectrum antibiotic activity against both *Escherichia coli* and *Staphylococcus aureus* but at varying levels, whilst Y4 had antibiotic activity against only *S. aureus*. The difference in activity profiles could be due to different antibiotic compounds produced by each isolate, or the ability of the antibiotics to diffuse through the agar. However, this is a qualitative assessment and variations in activity could also be down to differences in the density of the test isolate inocula leading to different natural product production levels.

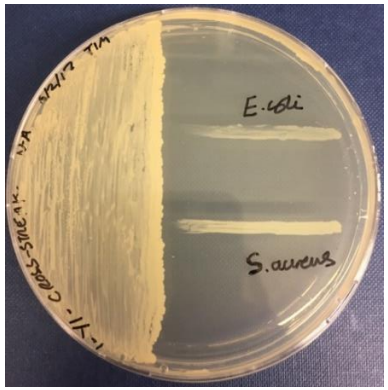


Figure 3.1a. Isolate Y1

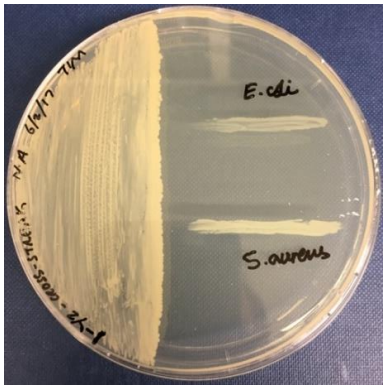


Figure 3.1b. Isolate Y2

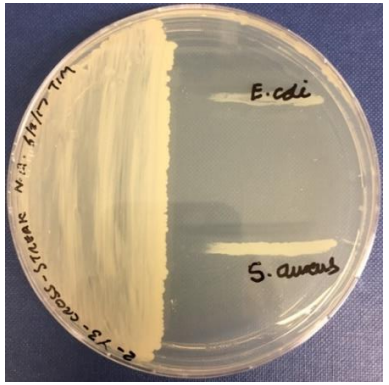


Figure 3.1c. Isolate Y3

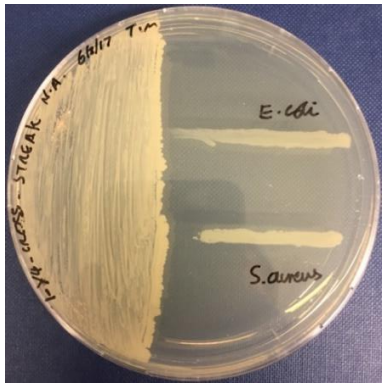


Figure 3.1d. Isolate Y4

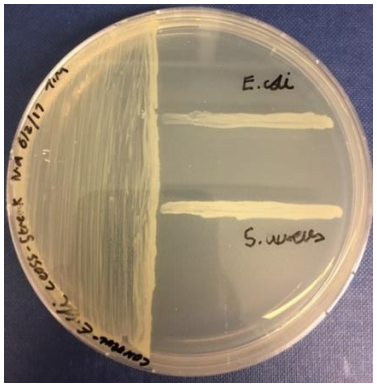


Figure 3.1e. *Escherichia coli* NCTC 10418

Figure 3.1. Representative images of the cross streak assay for each isolate (Y1-Y4) against *Escherichia coli* NCTC 10418 (top) and *Staphylococcus aureus* NCTC 12981 (bottom). *Escherichia coli* NCTC 10418 was also used as a control test isolate (n=3).

3.3.2. Phylogenetic Analysis of the Isolates from Honey

Phylogenetic analysis of the partial 16S rRNA gene sequences of the four honey isolates showed them to have the closest similarity to *Bacillus stratosphericus*, *Bacillus altitudinis*, and *Bacillus aerius* (**Figure 3.2**). These three species were first reported by Shivaji and co-workers in 2006 as distinct but related species of *Bacillus pumilis*. However, the status of these three species is currently under review (Branquinho et. al., 2015; Dunlap, 2015), and so these isolates can be considered as being part of the *Bacillus pumilis* phylogenetic group (Liu et. al., 2013). A search of the literature returned only one report of an antimicrobial secondary metabolite from *Bacillus pumilis* – Pumilicin 4, reported as a bacteriocin with a molecular mass of 1994.62 Da and broad-spectrum antibiotic activity (Aunpad & Na-Bangchang, 2007). However, no genetic or structural data relating to this compound has yet been reported. A manual search also returned no records of annotated secondary metabolite BGCs associated with verified genomes of this species. The absence of reports of antibiotic compounds from *B. pumilis* suggest that this group may be an underexplored source of novel antibiotic compounds that is worthy of further investigation.

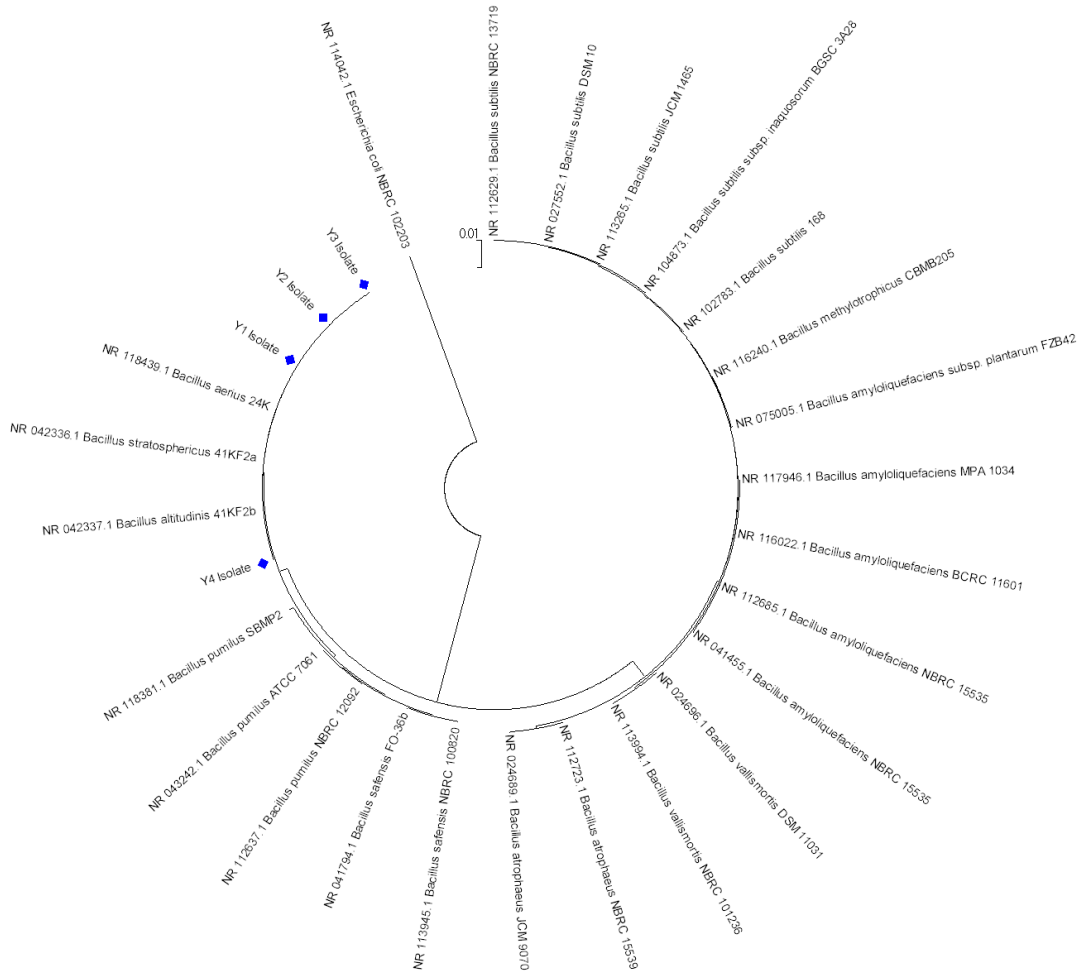


Figure 3.2. Neighbour-joining phylogenetic tree of the honey isolates (Y1, Y2, Y3, and Y4) partial 16S rRNA gene sequences and the top 23 BLASTn similarity matches from the NCBI 16S reference dataset. *E. coli* NBRC 102203 is used as an outgroup. The honey isolates are marked with a blue diamond (◆). Scale bar indicates the number of base substitutions per site.

Percentage identity comparison of the 16S sequences of the four honey isolates showed the sequences for Y1, Y2, and Y3 to be identical (**Table 3.2**). Based on this analysis alone it could be concluded that isolates Y1, Y2, and Y3 are clones and it is unnecessary to sequence the genomes all three. However, different *Bacillus* species are known to have high 16S rRNA gene sequence similarity making differentiation by 16S rRNA gene sequencing alone insufficient (Liu et. al., 2013). To obtain a higher discrimination resolution, the partial gyrase B (*gyrB*) housekeeping gene of the four isolates was also analysed.

Table 3.2. Percentage identify matrix of the 16S rRNA partial gene sequence (1222 bp) of the four honey isolates. Y1-3 all show complete alignment match.

	Y1_16s	Y2_16s	Y3_16s	Y4_16s
Y1_16s	-	100 %	100 %	99.9 %
Y2_16s	100 %	-	100 %	99.9 %
Y3_16s	100 %	100 %	-	99.9 %
Y4_16s	99.9 %	99.9 %	99.9 %	-

Table 3.3. Percentage identify matrix of the gyrase B partial gene sequence (1022 bp) of the four honey isolates. All isolates have different gyrase B sequences.

Like the 16S rRNA gene, *gyrB* is an essential gene. However, *gyrB* has a faster rate of molecular evolution and has been used as an alternative phylogenetic marker to distinguish closely related *Bacillus* species (Wang et. al., 2007; Liu et. al., 2013). Analysis of the *gyrB* sequences showed all four to be different (**Table 3.3**).

	Y1_GyrB	Y2_GyrB	Y3_GyrB	Y4_GyrB
Y1_GyrB	-	98.8 %	98.9 %	98.0 %
Y2_GyrB	98.8 %	-	99.1 %	98.8 %
Y3_GyrB	98.9 %	99.1 %	-	98.7 %
Y4_GyrB	98.0 %	98.8 %	98.7 %	-

This suggests that the four isolates are not clones but distinctive strains. This justifies sequencing the genomes of all four isolates to determine if there are different biosynthetic gene clusters in each of them. In addition to *gyrB* tying, the genomic DNA of the four isolates was digested using *Hinc*II and analysed on an agarose gel. *Hinc*II is an endonuclease that cleaves phosphodiester bonds of DNA at specific recognition sites (**Figure 3.3**). After digestion of the genomic DNA by the enzyme, the resultant fragments form a unique pattern when run on an agarose gel. Different genomes will contain recognition sites at different areas of the genome, resulting in different patterns of DNA fragments (Owen, 1989).

Figure 3.3. The recognition and cleaving site of *Hinc*II. The cleavage site is highlighted by arrows (▼ ▲). Y = pyrimidine bases (C/T), R = purine bases (A/G).

The banding patterns of the four isolates were not identical (**Figure 3.4**), giving further evidence that they are indeed distinctive strains and so justifying the genome sequencing of them all. However, the conclusion that the isolates are distinctive from the *Hinc*II digestion is not conclusive on its own as the results between the four samples can also be influenced by minor differences in the extent of enzyme digestion and in gel running. The use of a restriction enzyme with less common cleavage site than *HincII* may have been more suitable in showing a difference in banding patterns between the four isolates.

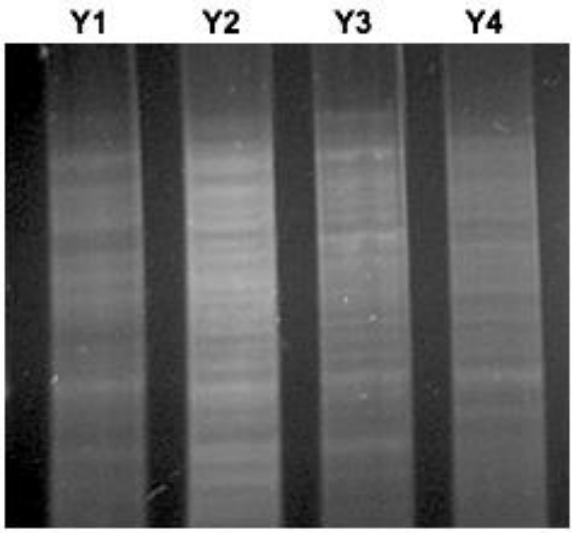


Figure 3.4. Close-up image of *Hin*cII digestion of genomic DNA of each isolate. Each lane shows a different banding patterns, suggesting the genomic DNA in each well is distinctive.

3.3.3. Bioinformatics Analysis of Biosynthetic Potential of the Isolates from Honey

The genomes of the four bacterial isolates were assembled and annotated (**Table 3.4**). All four isolates returned genomes of similar sizes within 3.7-3.8 kbp. The number of putative protein annotations (CDS) to genome size (Mbp) were all approximately 1:1000, which is consistent with expected gene numbers per genome size for bacterial genomes (Lynch & Conery, 2003) and gives confidence on the annotation. GC content is very similar in four strains and consistent with the GC content of other bacteria in the *Bacillus* genus. These QC statistics are within the expected range which suggests that the pipeline used to generate these assemblies is of good quality and confidence can be taken from analysis of the annotations.

Table 3.4. QC Analysis of the genome assemblies of the four isolates

Isolate	Contigs	Total size (Mb)	CDS	GC content (%)	L50	N50 (bp)
Y1	26	3.7	3766	41.19	3	618085
Y2	55	3.8	3832	41.15	6	182496
Y3	38	3.8	3869	41.25	5	251109
Y4	55	3.7	3765	41.27	5	268397

The four genome assembles are however fragmented and the number of contigs between each assembly varies from 26-55. Fragmented genome assemblies often occur due to the short (250 bp) sequences reads being unable to fully resolve across repetitive genome regions causing the assembler to break the contig at this point. Variations in the number and positions of such regions within each genome can cause the variance in contig number. This can also contribute to the variation in L50 (number of contig to cover half of genome length) and N50 (average

contig length) statistics seen. Fragmented genomes are therefore indicative that there are areas of the genome which are not accurately assembled. This can potentially compromise genome-mining as long repetitive BGCs (such as NRPS and PKS) can be misassembled or fragmented across contigs (Klassen & Currie, 2012; Rutledge & Challis, 2015; Miller, Chevrette & Kwan, 2017; Goldstein, et al., 2019).

Table 3.5 summarises the putative biosynthetic gene clusters predicted to be encoded within the genomes of the four isolates. The results show that each isolate potentially encodes secondary metabolites of various different classes. The results also suggest that there is variation in the number and classes of BGCs encoded within the genomes of the four isolates. The Y1 genome is predicted to encode 3 putative bacteriocins whilst the other isolates encode 2 each. Y2 is predicted to encode a lantipeptide BGC that has not been identified in the other isolates, and Y3 is predicted to encode a PKS-NRPS hybrid not identified in the other isolates. No pure PKS gene clusters were identified in any of the isolate genomes. This is perhaps not surprising as it has been noted by other researchers that many *Bacillus* associated polyketide compounds are found to be produced hybrid PKS-NRPS BGCs (Hertweck, 2009). Furthermore, a manual inspection of the MIBiG database of *Bacillus* associated BGCs returned 68 BGCs for proteins, 10 hybrid NRPS-PKS BGCs, and 5 PKS BGCs. Genome mining data also suggests that the isolates produce other natural products which cannot be classified and/or predicted based on genetic homology. Some of these low homology gene clusters may have antibiotic activity. An example of this is the long-chain N-acyltyrosines class, whose biosynthetic genes do not share homology (Brady, Chao & Clardy, 2004).

Table 3.5. Results of AntiSMASH analysis of the draft genomes of the four isolates obtained from Honey. *'Others' are highlighted as putative gene clusters due to a clustering of biosynthesis-like genes, but the products could not be predicted and/or could not be identified as belonging to a particular biosynthetic class. All isolates have a different secondary metabolite profile.

Isolate	Putative Secondary Metabolite Biosynthetic Gene Clusters						
	Ribosomal Peptides			NRPS	PKS-NRPS	Other*	Total
	Bacteriocin	Microcin	Lantipeptide				
Y1	3	1	-	2	-	5	11
Y2	2	1	1	2	-	3	9
Y3	2	1	-	2	1	4	10
Y4	2	1	-	2	-	4	9

3.3.3.1. Ribosomal Peptide Analysis

AntiSMASH analysis detected a putative microcin encoded in all four isolates that had similarity with the micJ25 profile hidden Markov model (pHMM), which is a profile based on alignment of microcin peptides. Microcins are small (<10 kDa) ribosomal antibacterial peptides that are usually associated with Enterobacteria (Severinov & Nair, 2012). MultiGeneBlast analysis showed that the microcin genes identified in all four isolates were identical. However, no homologues were found in the Bactibase database. Of the three putative bacteriocins identified in isolate Y1; two were also identified in the other isolates with ≥99% homology using MultiGeneBlast. However, there was a difference in which bacteriocins from Y1 were identified in Y2 compared to Y3 & Y4 (**Table 3.6**).

Table 3.6. Summary of MutliGeneBlast homology comparison of the putative bacteriocin BGCs found in isolate Y1 against the BGCs identified in the other three isolates. A '+' indicates that a homology was detected with a BGC at ≥99 %.

	Y1-Bacteriocin 1	Y1-Bacteriocin 2	Y1-Bacteriocin 3
Isolate Y1	+	+	+
Isolate Y2		+	+
Isolate Y3	+		+
Isolate Y4	+		+

Two of the bacteriocin BGCs that were found in Y1 (provisionally named Y1-Bacteriocin 1 and Y1-Bacteriocin 2) were identified as having homology with the TIGR03651 pHMM, which is a profile based on alignment of circular bacteriocin protein sequences. However, no matches to this gene were returned in Bactibase database. Also, no accessory genes expected to be encoded on the cluster, such as modification, transport or immunity genes, could be identified. The absence of any related accessory genes suggests that this finding may be an artefact. The third bacteriocin (Y1-Bacteriocin 3) is present in all isolates and was matched to

the Bacteriocin_IId pHHM. Further analysis using Bactibase showed the gene to have homology with the precursor gene for the circular bacteriocin AS-48. This is a circular cationic bacteriocin that has been discovered only in *Enterococcus* species. AS-48 is reported to have broad-spectrum antibiotic activity and high stability over wide pH and temperature ranges (Sanchez-Hidalgo et al., 2011). Homology was also found between genes surrounding the predicted bacteriocin gene and the modification and transport genes of the AS-48 gene cluster (**Figure 3.5**).

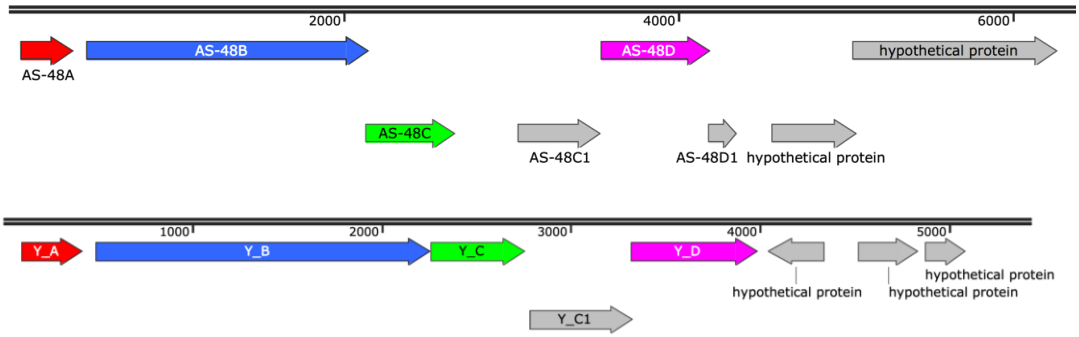


Figure 3.5. Diagram of the biosynthetic gene cluster of AS-48 (above) and the putative bacteriocin identified in the four isolates (below). Gene products with significant sequence similarity (homogeny) according to the bactibase database are shown by matching colour coding (red, blue, green, pink). No homologues could be found for genes colured grey. AS-48A is the bacteriocin precursor gene, AS-48B modifies the precursor into a mature peptide, AS-48C is an accessory factor for protein secretion, AS-48C1 and AS-48D are provisionally identified as ABC transporter genes, AS-48D1 is provisionally identified as being involved in conferring immunity to the AS-48. Names have been given to the genes in the honey isolate clusters based on their homologous gene in the AS-48 cluster. Gene Y_C1 did not show homogeny with AS-48C1, however the gene product does show homogeny with other ABC transporters in the NCBI database. No immunity gene homologues could be found in the Y isolate cluster.

These data suggest that these four isolates encode a circular bacteriocin homologous to AS-48. The inability to find an immunity gene homologue could be due immunity to the bacteriocin being conferred by a gene that is present elsewhere in the genome. Alternatively, the hypothetical proteins upstream of Y_D could have a role in immunity but are not homologous to other bacteriocin immunity genes. If the four isolates do produce an AS-48 analogue, this may be the cause of the broad-spectrum antibiotic activity seen in the cross-streak bioassays (**Figure 3.1**). However, this activity was not seen in Y4, which is also predicted to encode this compound. Possible reasons for this could be that the peptide was not expressed by this isolate due to different regulation mechanism, differences in the environmental conditions of the test (such as agar composition). It could also suggest that the antibiotic activity seen in the assay is caused by a different natural product or a combination of products. AS-48 gene cluster was found on a transferable plasmid (Gálvez, Valdivia &

Maqueda, 1990). This suggests the possibility that the isolates acquired the AS-48 genes through plasmid mediated horizontal transfer.

The membrane disrupting potential of AS-48 is proposed to be determined by the sequence of the alpha helical domain (Sánchez-Hidalgo et al., 2011). Clustal Omega alignment of this 25-residue sequence with the corresponding sequence from the Y isolates (Table 3.7) shows the two to be highly conserved. Out of the 25 residues, 18 were identical and 6 had strongly similar properties. This finding would suggest that the AS-48-like gene product in the Y isolates will likely have highly similar antibiotic properties to that of AS-48. Soon after submission of this thesis a study was published by Ross et al., (2020) whereby homology-based searches of the active helical domain of AS-48 were used to identify three previously uncharacterized AS-48-like bacteriocin genes in *Clostridium sordellii*, *Paenibacillus larvae*, and *Bacillus xiamenensis*. The authors used these identified sequences as scaffolds to design a synthetic minimal peptide library of 384 peptides and assessed their activity against Gram-negative and Gram-positive bacteria. The alignment matches reported by the authors between AS-48 and the genes in *C. sordellii*, *P. larvae*, and *B. xiamenensis* have a lower degree of conservation than that shown between AS-48 and the Y isolate gene. This further supports the conclusion that the AS-48 like gene in the Y isolates will have antibiotic properties highly similar to those of AS-48. The findings of Ross et al., (2020) also demonstrate that AS-48 like genes are widely distributed amongst different genera, and so further supports the conclusion made from the findings in this thesis that taxonomy is not a reliable means of dereplication due to the possibility of shared antibiotic BGCs across taxa.

Table 3.7 Results of Clustal Omega alignment of the alpha helical sequence of AS-48 (AS-48_alpha) and the corresponding sequence of the putative AS-48 analogue from the Y isolates (y_a). The symbols under the alignment show the extent of conservation (* asterisk) indicates positions which have a single, fully conserved residue, (: colon) indicates conservation between groups of strongly similar properties, (. period) indicates conservation between groups of weakly similar properties.

File	Parent Sequence
AS-48_alpha y_a	AGRESIKAYLKKEIKKKGKRAVIAW AGKETIRQYLKNEIKKKGRKAVIAW **:*:*: ***:******::*****

Isolate Y2 is predicted to encode a lantipeptide, which is a class of ribosomal peptides which is extensively post-translationally modified. The leader sequence of the immature peptide is cleaved by a modifying enzyme, and the hydroxyl amino acids of the remaining core sequence are dehydrated to dehydroalanine or dehydrobutyrine. Thioester bonds between cysteine and the dehydrated residues are catalysed to circularise the peptide. Some lantipeptides have broad-spectrum antibiotic activity such as, nisin, subtilin, and epidermin (Begley et al., 2009). **Figure 3.6** shows the predicted primary structure of the mature lantipeptide predicted to be encoded in Y2. Searching for this sequence in the Bactibase and NCBI databases returned no identical matches, suggesting that this predicative structure is a potentially novel lantipeptide.

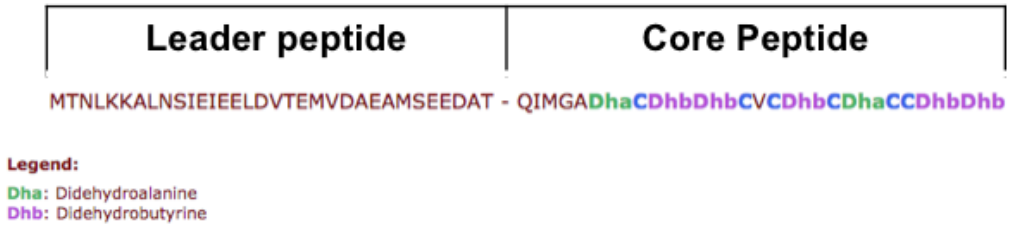


Figure 3.6. Predicted primary structure of the mature lantipeptide encoded in isolate Y2. The predicted leader peptide sequence is cleaved from the core peptide which is then dehydrated to contain dehydroalanine (Dha) and dehydrobutyrine (Dhb) residues which form cross-links with the cysteine residues.

3.3.3.2. Non-ribosomal Peptide and Polyketide Analysis

AntiSMASH analysis identified each isolate to encode two NRPS gene clusters (**Table 3.5**). Comparison of these clusters using MultiGeneBlast showed that the two NRPS clusters in each isolate were identical. The two clusters were named NRPS_1 and NRPS_2 respectively.

NRPS_1 multi-enzyme complex contains three catalytic modules (adenylation and condensation domain). Analysis of the conserved regions of the adenylation domains in these modules predicts their specificity to be for dihydroxybenzoate (Dhb), glycine (Gly), and threonine (Thr) respectively. This sequence is identical to that of bacillibactin, a trimeric ester siderophore produced by *Bacillus* bacteria (May et al., 2001). Comparison of the NRPS_1 complex with other BGCs in the MIBIG database confirmed that NRPS_1 is homologous with the bacillibactin NRPS complex. This suggests that NRPS_1 encodes an analogue of bacillibactin.

The NRPS_2 complex contains eleven catalytic modules (adenylation and condensation domain) across five open reading frames (OFR). **Figure 3.7** summaries the architecture of the complex, the amino acid specifics of each module, and a prediction of the core structure produced by the complex. Upstream of module 5, is predicted to be a type II thioesterase domain. Type I thioesterases are embedded in the C-terminus of final module and are responsible for release of the full-length peptide chain from the enzyme complex by hydrolysis. In contrast, type II thioesterases are discrete enzymes that function to remove incorrect amino acids that are obstructing catalysis, substrate selection, and product release. Type II thioesterases have been found associated with other *Bacillus* NRPS complexes, such as for tyrocidine and surfactin (Kotowska & Pawlik, 2014). The NRPS_2 multienzyme complex was searched for homogeny against BGCs in the MIBIG database but no homogenous complexes could be found. This suggests that NRPS_2 is a novel biosynthetic gene cluster which could potentially encode an uncharacterised product. However, three of the amino acid substrate specificities of three of the adenylation domains could not be accurately predicted. This reduces the detail and confidence of the predicted structure of the BGCs core NRP product synthesised by this BGC. In addition, it also reduces the confidence in any hypothesis of the novelty of this BGC and its product which can be made. Mispredictions such as this have been reported by other researchers and can be caused by short-read genome assemblers failing to adequately resolve across highly repetitive sequence regions such as NRPS BGCs (Miller, Chevrette & Kwan, 2017).

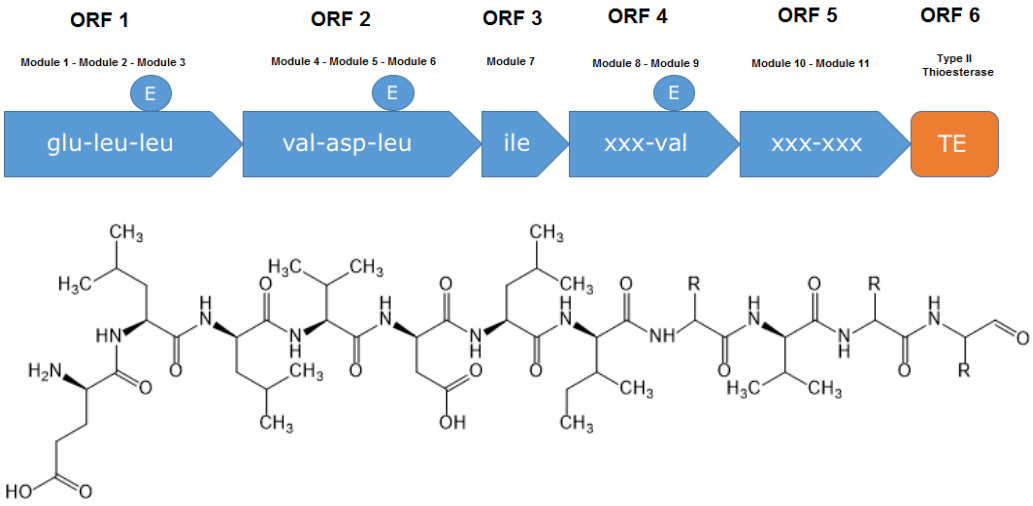


Figure 3.7. Diagram of the modules in the NRPS_2 multienzyme complex with prediction of the core structure it produces below. The core structure prediction assumes co-linearity of the biosynthetic modules and does not take into account tailoring reactions which may occur downstream. There are modules encoded across five ORFs. ORF 1 and ORF 2 contain three modules each, ORF 3 contains one module, and ORF 4 and 5 contain two modules each. The specificity of the adenylation domain within each module is listed: Glutamic acid (glu), leucine (leu), valine (val), aspartic acid (asp), isoleucine (ile). Specificity for three of the adenylation domains could not be predicted and are marked as 'xxx'. Downstream of ORF 5 is a type II thioesterase domain (TE). There are epimerase (E) modification domains encoded in modules 3, 6, and 9, which are predicted to epimerise the corresponding amino acids (3-leucine, 6-leucine, and 9-valine) into the D-configuration. Structure drawn in ChemDraw Professional version 16.

Isolate Y3 was predicted to encode a PKS-NRPS hybrid complex. PKS-NRPS hybrids catalyse the synthesis of molecules containing both amino acid and ketone starter units, an example of a NRPs-PKs hybrid product is the antibiotic virginiamycin (Pulsawat et al., 2007). The complex is comprised of six ORFs with three NRPS and four PKS modules encoded within them. A type II thioesterase domain is also encoded in an ORF between the first and second modules. **Figure 3.8** summarises the architecture of the complex, the amino acid specific of each module, and a prediction of the core structure produced by this complex. No homologs for this BGC could be found in the MIBIG database, which might suggest that this is a potentially novel biosynthetic gene cluster encoding a novel product. Unlike the BGC featured in **Figure 3.7**, the substrate specificities of each module are predicted which allows for a more detailed prediction of the core-structure of this BGC. This PKS-NRPS complex, which is predicted to only be present in isolate Y3, could be contributing to the stronger broad-spectrum antibiotic activity against both *S. aureus* and *E. coli* seen in the cross-streak assays (**Figure 3.1c**).

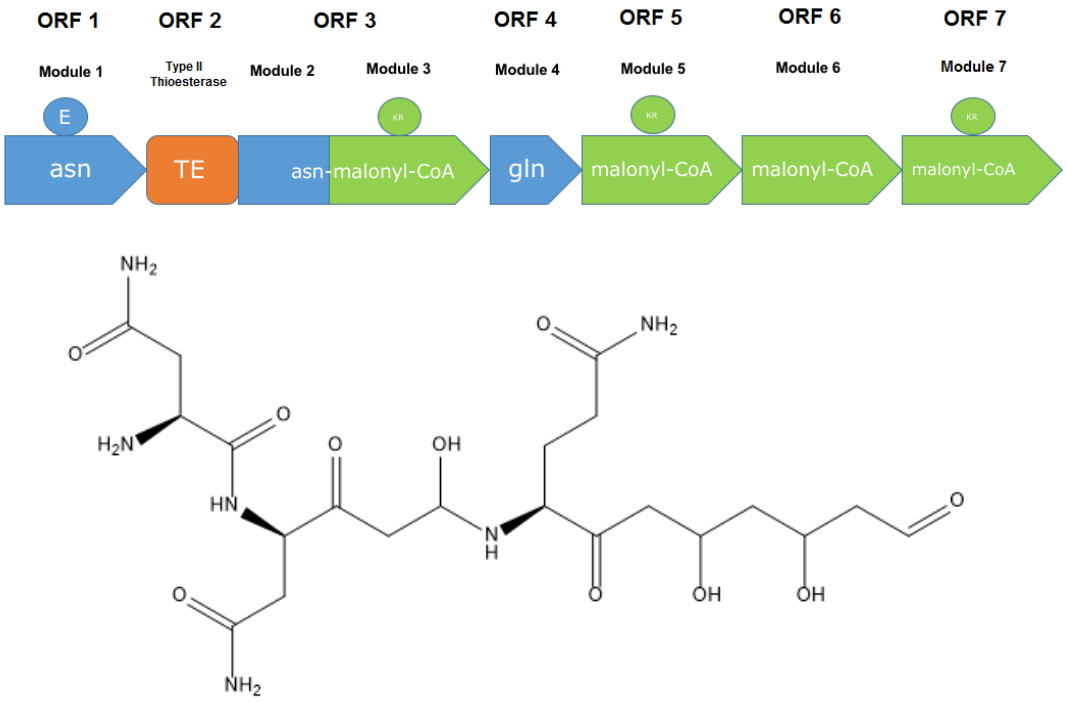


Figure 3.8. Diagram of the PKS-NRPS multienzyme complex with prediction of the core structure it produces below. The core structure prediction assumes co-linearity of the catalytic domains and does not take into account tailoring reactions which may occur downstream. There are seven biosynthesis modules in total and a type II thioesterase. ORF 1 contains an NRPS module (blue shading) specific for asparagine (asn). ORF 2 contains type II thioesterase (TE). ORF 3 contains two modules, one is an NRPS module specific for aspartic acid and the other is a PKS module (green shading) specific for malonyl-CoA. ORF 4 contains an NRPS module specific for glutamic acid (glu). ORF 5, 6, and 7 each contain a PKS module each that is specific for malonyl-CoA. The module in ORF 1 also contains an epimerase (E) modification domain which is predicted to epimerise the corresponding amino acid (1-apsartic acid) into the D-configuration. Modules 3, 5, and 7 each contain ketoreducatase (KR) modification domains which are predicted to reduce the β-ketone of the nascent substrate to a β-hydroxyl group. Structure drawn in ChemDraw Professional version 16.

3.3.4. Conclusions

Bacterial colonies were isolated from raw honey to investigate if the bacterial isolates have antibiotic potential. Four bacterial strains were isolated which have antibiotic activity. Three of the isolates (Y1, Y2, Y3) showed broad-spectrum inhibition of both *S. aureus* and *E. coli*, whilst Y4 showed activity against *S. aureus*. The morphologies of all four isolates were identical and 16S rRNA gene typing identified them to be related to related to the *Bacillus pumilis* phylogenetic group. With the 16S sequence of isolates Y1-3 being identical to each other and that of Y4 sharing 99.9 % alignment match with Y1-3. However, further molecular typing by RFLP analysis and *gyrB* typing revealed all four isolates are distinct. Whilst analysis of the genomes of the four isolates revealed that they encoded similar but varying biosynthetic profiles.

All four isolates are predicted to encode the same two NRPS complexes. NRPS_1 is a homolog of bacillibactin NRPS, a common *Bacillus* siderophore, whilst NRPS_2 is predicted to encode for a potentially novel product. Y3 is predicted to encode a NRPS-PKS hybrid complex which potentially produces a novel product. In addition, Y2 is also predicted to encode a potentially novel lantipeptide.

The findings from this study demonstrate that examining the 16S rRNA gene in isolation does not provide sufficient resolution to reliability distinguish related bacteria, which is consistent with the findings of Liu and co-workers (2013). The findings of the genome-mining analysis also demonstrate that species within a genus with the same 16S rRNA gene can encode different natural product profiles. This finding is consistent with the findings of other researchers (Antony-Babu et al., 2017). Therefore, it can be concluded that for purposes of dereplication, reliance on 16S rRNA gene phylogenetic typing alone risks missing out on potential novel natural products. In a future investigation of an antibiotic-producing isolate from the hot-spring waters of the Roman Baths (**Chapter 5**), a genome wide approach was therefore taken during phylogenetic analysis (**Chapter 5**).

All isolates are predicted to encode several ribosomal peptides with potential antibiotic activity based on pHMM hits. One gene cluster has homology with the gene cluster for AS-48, a Head-to-Tail Cyclised RiPP produced by *Enterococcus* which has broad-spectrum antibiotic activity and is under clinical trials (as discussed in **Section 1.4.2.7**) and had not previously been reported as present in *Bacillus* species. The genome mining data also suggests that the isolates produce other natural products which cannot be classified and/or predicted based on sequence homology.

These results suggest that these four isolates are potentially a rich reservoir of novel natural products. However, due to the presence of a well annotated analogue for AS-48 in all four isolates is was decided not to continue further investigation into these isolates to discover novel antibiotic compounds. This validates the benefit of using genome-mining in bioprospecting studies for allowing early dereplication. Additionally, these findings also demonstrate that analogous natural product gene clusters can be encoded in distinct genera. Therefore, it can be concluded from this result that taxonomy alone is not a reliable strategy for dereplication of antimicrobial producing bacteria.

The isolate genomes were fragmented because short-reads failing to resolve repetitive regions. This undermines the BGC annotations produced due to their long and repetitive nature and may have contributed to the low resolution annotation of NRPS_2 (**Figure 3.7**). Long-read sequencing technologies offer a potential solution to this by 'bridging' these ambiguous regions. An alternative for genome-mining studies is to utilise long-read technology to improve the contiguity of bacterial genome assemblies for natural product annotation. For this reason, a pipeline for long-read genome assembly and annotation is developed in **Chapter 4** which is then utilised in the genome-mining of isolate KB16 described in **Chapter 5**. However, the presence of a well annotated analogue for AS-48 in all four isolates which is a well characterised antibiotic molecule precluded these isolates from further investigation to discover novel antibiotic compounds in favour of another isolate described in **Chapter 5**.

Chapter 4.

Development of a Genome Mining Pipeline for Bioprospecting using Oxford Nanopore Long-Read Whole-Genome Sequencing

4.1. Introduction

4.1.1. The limits of Short-Read Genome Mining

Whole genome sequencing and analysis has potential in aiding the screening of environmental microorganisms for novel natural products with antibiotic (or other commercially useful) properties. The results of the genomic analysis of the *Bacillus* species isolated from honey described in **Chapter 3** highlighted this potential in identifying the pathway for a well characterised antibiotic natural product (AS-48) which had not been reported in *Bacillus* previously. However, the work detailed in **Chapter 3** also highlighted limitations with the genome-mining approach. These limitations were the production of fragmented genome assemblies (**Table 3.4**) and the failure to fully predict the substrate specificity of some of the adenylation domains of a NRPS BGC (**Figure 3.7**), which is due to the sole reliance on Illumina short read data to assemble the genomes.

Illumina based sequencing technology produces reads with high accuracy (>95.5 %) (Goodwin, McPherson & McCombie, 2016) However, the short length of these reads, which can range from 36 bp to 300 bp, is dependent on the particular machine platform used (Goodwin, McPherson & McCombie, 2016). Short reads create issues for assembly algorithms across homopolymer or repetitive stretches of the genomes that are larger than the length of the reads, resulting in either a break in the assembly or mis-assembly (Koren et al., 2013). Which can therefore impact the assembly and annotation of BGC with long repetitive regions such as NPRS and PKS enzymes in natural product biosynthetic gene clusters (Klassen & Currie, 2012; Rutledge & Challis, 2015; Miller, Chevrette & Kwan, 2017; Goldstein, et al., 2019).

Additionally, Illumina technology, being based on a “sequencing by synthesis” approach whereby input DNA is first amplified by a polymerase, also means that sequencing basis against regions with high GC content can be observed (Aird et al., 2011; Chen et al., 2013; Goodwin, McPherson & McCombie, 2016). Many bacteria with high natural product potential are from the phyla Actinobacteria which typically have a higher GC content. Another potential limitation of Illumina technology is around practical aspects of its use in smaller-scale screening projects as it typically has a high financial, resource, and computational cost (Goodwin, McPherson & McCombie, 2016).

4.1.2. The Potential of Long-Read Genome Mining

Long-read sequencing technologies such as the sequencing platforms from Pacific Biosciences (Pac-Bio) have been demonstrated to resolve some of these issues due to the production of long-reads that can span repetitive chromosome regions and therefore produce closed contigs (Koren et al., 2013). However, much like Illumina, Pac-Bio has a high capital cost (Goldstein et al., 2019) which may make it impractical for small scale projects.

A potential solution to these limitations is the use of long-read sequencing technologies such as the Oxford Nanopore (ONT) platform. Rather than amplify the input genetic material, the ONT approach involves direct sequencing of the DNA strands through protein pores embedded into a membrane within the flowcell of the device. A current is applied across the membrane, the voltage of which is disrupted by the movement of nucleotides through the pores. The change in voltages is characteristic to each nucleotide base and therefore the sequence of a DNA fragment can be inferred by measuring the voltage changes (**Figure 4.1**).

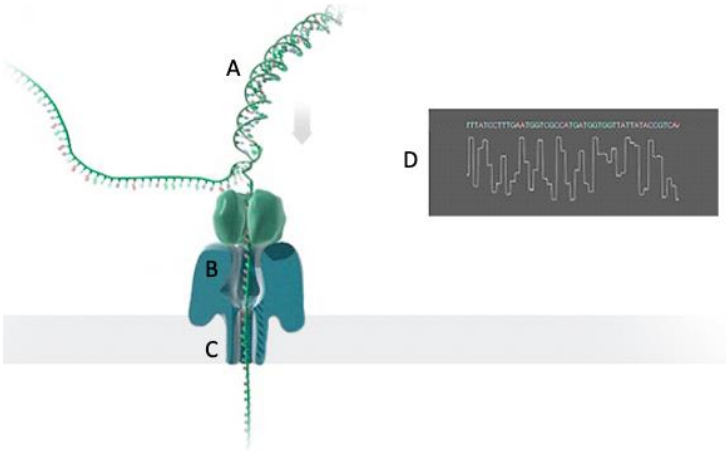


Figure 4.1. Schematic outlining the principle of Oxford Nanopore 'direct sequencing'. Nucleic acid (A) is threaded through protein channel pores (B) that are embedded into the membrane (C) within a flowcell. Passage of the nucleic acid across membrane disrupts the voltage applied across it in a manner characteristic to each base. These voltages change signatures are interpreted as the nucleic acid sequence (D). Image adapted from Oxford Nanopore Technologies Ltd. website: (https://www.nanoporetech.com/how-it-works)

This approach theoretically places no limit onto the length of DNA fragments that can be sequenced and therefore allows for much longer average read lengths than in Illumina sequencing. The longer reads have the potential to offer sufficient coverage across repetitive genome regions and so enable contiguous genome assemblies. Also, the technology is potentially more accessible to more researchers due to its greater simplicity and relatively lower financial and computational costs. Oxford Nanopore sequencing of microbial genomes has been shown to have potential applications in microbiology research such as in disease

outbreak monitoring (Quick et al., 2016), and antimicrobial resistance surveillance (Judge et al., 2016). However, a search of the literature shows few assessments of the potential of this technology towards bioprospecting. However, the technology is still in beta phase and is generally regarded to have lower accuracy than Illumina (Goodwin, McPherson & McCombie, 2016) which may limit its utility.

Strategies to process this data are also still being optimised by an active research community. Generally, comparisons of Nanopore microbial genome assemblies with Illumina have reported higher rates of complete contigs being assembled but with lower sequence accuracies. Several approaches have been taken by other researchers to produce closed contig assemblies of microbial genomes using Oxford Nanopore data. Loman, Quick & Simpson (2015) pioneered the development of a pipeline to assemble microbial genomes using only Nanopore reads. Their pipeline involved a three-stage process that first attempted to correct raw reads by detection and consensus calling of overlapping regions between reads. These corrected reads were then assembled into a raw scaffold and this scaffold was "polished" by alignment of raw reads back to the scaffold to create a consensus sequence. The authors reported recreation of the *Escherichia coli* K-12 MG1655 genome with 99.5 % accuracy. They reported notable errors in the Nanopore only assembly compared to the reference being in the under representation of 4-mer and 5-mer homopolymer regions. Additionally, other researchers have reported that protein annotations of Nanopore genome assemblies show high rates of truncated or fragmented CDS and pseudogenes. This has been attributed to a high instance of erroneous stop codons in annotated sequences caused by indel errors in Nanopore reads (Stewart et al., 2019; Watson & Warr, 2019).

Contiguous genome assemblies are desirable for natural product genome-mining as it may reduce misassembles of long BGCs, but this could be undermined by inaccurate protein annotations. Therefore, it is important to have a genome-assembly strategy to reduce these if using Nanopore data for such a purpose.

4.1.3. Genome Assembly using Long-Reads

Many open-source assemblers designed to specialise in the assembly of Nanopore long-reads have been developed since the release of the technology. A notable example is Canu (Koren et al., 2017), a fork of the Celera Assembler used in the pipeline by Loman, Quick and Simpson (2015). This assembler attempts to first correct noisy raw reads by finding consensus sequences between overlaps (an approach taken in the Loman pipeline). Overlaps of the corrected reads are then trimmed of low census sections. These processed reads are then

overlapped to form consensus contigs. The pipeline has been shown to produce contigs of higher accuracy than the raw reads thanks to its pre-correction steps. But is computationally expensive and time consuming to run.

Alternative assemblers have also been developed that take a simpler overlap layout consensus (OLC) based approach, whereby consensus contigs are formed by overlapping of the reads without any prior correction steps. Examples of such assemblers include Minasm (Heng et al. 2016) and SMARTdenovo (https://github.com/ruanjue/smartdenovo). These assemblers are significantly faster than Canu – often capable of producing a draft assembly within minutes compared to several hours or days. However, the accuracy of the contigs is the same as the raw reads. Another prominent example of a long-read assembler is Flye (Kolmogorov et al., 2019). Unlike most long-read assemblers that use an OLC based approach, Flye utilised de Bruijn graphs modelling typically used for short high accuracy reads. But Vaser and co-workers (2017) demonstrated a tool called RACON that can improve the accuracy of these contigs by realigning the long-reads to the contigs to create a new consensus – a process known as "polishing". They also demonstrated that further improvements to sequence accuracy is achievable through multiple iterations of RACON polishing. While RACON creates consensus through the alignment and overlap of basepairs, another polisher - called Nanopolish - polishes by using the raw voltage signal data created by the passage of each nucleotide through the MinION membrane pores to correct assembled sequence errors (https://github.com/jts/nanopolish). Using such tools, researchers have developed pipelines to assemble whole microbial genomes from Nanopore long-read data and improve their accuracy with multiple polishing steps.

4.1.4. Hybrid Genome Assembly Approaches

Some researchers have combined both long reads from either PacBio or Nanopore and higher accuracy Illumina short-reads to obtain the closed contigs that long-reads can provide along with the higher sequence accuracy offered by the shorter Illumina reads. Two broad approaches taken by researchers are to either perform a "hybrid" assembly of both read types where the longer reads serve to bridge gaps between the fragmented short read contigs. An example of a such a hybrid assembler is hybridSPAdes (Antipov et al., 2015). The assembly creates initial De Bruijn assembly graphs of overlapped Illumina reads and then maps the long reads to these graphs to bridge the gaps and create larger contigs. The authors report that the pipeline was able to assemble several *E. coli* reference genomes into single contigs. However, no assessment of the nucleotide or protein annotation accuracies was reported. Wick and co-workers (2017a) expanded upon this pipeline with Unicycler. This pipeline also creates de

Bruijn assemblies of Illumina reads using the SPAdes assembly algorithm. It then utilises RACON to bridge gaps and polish the assembly using long reads. Wick et al., (2017b) compared the Unicycler hybrid pipeline against Illumina-only and Nanopore-only methods to assemble 12 *Klebsiella pneumoniae* genomes. The authors reported that the hybrid method produced the most complete and accurate assemblies. All genomes were assembled into closed circular contigs, whereas only four out of twelve of the Nanopore-only assemblies and none for the Illumina-only assemblies were complete. Sequence inaccuracy rates (indels, SNPS, misassembled regions) of the hybrid assemblies compared to manually completed references matched the accuracy levels of Illumina-only assemblies at <0.000%, while the Nanopore-only assembly inaccuracy rates ranged between 0.3-1.0%.

An alternative hybrid method is to use Illumina reads to "polish" a Nanopore-only assembly. Tools such as Pilon (Walker, 2014) will align Illumina reads to a template to create consensus sequences and has been demonstrated to improve the accuracy of Nanopore based assemblies (George et al. 2017, Leim et al., 2018, Miller et al., 2018, De Maio et al., 2019, Shin et al., 2019). However, Kolmogorov and co-workers (2019) recently reported that the method is not perfect and can struggle to resolve regions with low short-read mapability.

While hybrid assembly approaches using both short and long read technologies may increase assembly sequence accuracy, it also increases cost and complexity of the procedure. This may make it unviable for some small-scale bioprospecting projects which are looking to leverage NGS data for simplistic dereplication purposes. Another consideration is that, while highest possible sequence accuracy is always something to strive for, it may not be essential to sufficiently answer the research question at hand. For example, Judge et al., (2016) compared Illumina-only and Nanopore-only assemblies of clinical bacteria and found that while the general sequence accuracy of the Nanopore-only assemblies was lower (~85 %), it was still of sufficient quality to accurately identify antimicrobial resistance genes. Given the greater speed and lower costs to sequence the isolates using Nanopore technology, they suggest that Nanopore is a viable option for AMR surveillance. For the purposes of bioprospecting, it is also unclear if Nanopore only assemblies provide sufficient levels of accuracy for dereplication and novel BGC prediction, or if additional Illumina reads would be necessary.

4.1.5. Aims and Objectives

The aims of the work described in this chapter are to optimise a pipeline for the sequencing and assembly of bacterial genomes using Oxford Nanopore MinION for the purposes of small-scale bioprospecting. This was done in order to utilise the potential benefits of cost-efficient long-read sequencing to produce a highly accurate contiguous genome assembly in order to address the limitations that short-read sequencing can have regarding accurate assembly and annotation of BGCs.

Streptomyces coelicolor A3(2) was used as a reference strain for the development of this pipeline. This bacterium was chosen because it is a well-established model organism for *Streptomyces* research and has been used to study the genetics of the species for nearly 60 years (Hopwood, 1999). The organism produces multiple antibiotic compounds and its complete genome was published in 2002 (Bentley et al., 2002), and the BGCs pathways are well annotated. Additionally, because many microbial antibiotic bioprospecting projects will likely isolate *Streptomycetes* or related high-GC Actinomycetes it was chosen as an appropriate representative organism for this study.

The most suitable pipeline for dereplication of known BGCs and prediction of potentially novel BGCs was assessed by comparing the annotated BGCs produced by different assembly methods to the annotated BGCs in the *Streptomyces coelicolor* A3(2) reference genome.

The pipeline optimised in this chapter was subsequently utilised to support the characterisation of an antibiotic-producing bacterium isolated from the hot-spring water of the Roman Baths, UK. (**Chapter 5**). Additionally, the Oxford Nanopore MinION was later used to utilise cost-efficient long-read sequencing to profile the microbiome of the source water and attempt to detect putative BGCs to help inform future bioprospecting strategies for this site (**Chapter 6**).

4.2. Materials and Methods

4.2.1. Microorganisms

A dried spore stock of *Streptomyces coelicolor* A3(2) was provided by Dr David Widdick of the John Innes Centre, Norwick, UK. The spores had been dried onto filter paper discs and were revived by resuspension in 100 µl of sterile water. This spore suspension was spread over Soya Flour Media (SFM) agars plates (2% soya flour, 2% mannitol, 2% agar) (Keiser et al., 2000) and incubated for 14 days at room temperature until sporulation of fresh pure colonies had occurred. These spores where cultivated from the plates and stored in Microbank™ tubes (Pro-Lab Diagnostics, Inc., Canada) at -80 °C.

4.2.2. DNA Extraction

Streptomyces coelicolor biomass was generated by incubation of a loop of spores in 100 mL tryptic soy broth (Sigma-Aldrich, UK) for 8 days at 30 °C / 240 rpm. Twenty mL of the culture was centrifuged at 4000 rpm / 5 mins to pellet biomass. The supernatant was removed, and the pellet was transferred to a 2 mL microcentrifuge tube and centrifuged for 5 mins a 13,000 rpm to remove remaining supernatant. The dry pellet was divided into three and genomic DNA extracted from each using DNAeasy ultraclean kit (QIAGEN GmbH, Germany) according to the manufacturers' protocol but with the modification that the cell suspensions were incubated in powerbead solution at 70 °C for 10 mins prior to addition of CB1 at step 3. The three genomic DNA (gDNA) extractions were combined in elution buffer (10 mM Tris pH 8.0) and analysed on a Nanodrop

4.2.3. Gel Analysis of DNA Extracts

One µL of DNA extraction was run on a 1 % agarose gel containing 0.5 % GelRed™ (Biotium, Inc, USA) in TAE buffer (40 mM Tris, 20 mM acetic acid, and 1 mM EDTA) for 120 mins at 65 V. Prior to loading, the sample was mixed with Purple Gel Loading Dye (New England BioLabs Inc., USA) according to the manufacturer's protocol. As a molecular size marker, λ DNA-*Hin*dIII Digest (New England BioLabs Inc., USA) was used according to the manufacturer's protocol. The gel bands were visualised on an UV transilluminator.

4.2.4. Preparation of Nanopore Sequencing Libraries

High molecular weight DNA was prepared for sequencing using the Nanopore Rapid Barcoding Sequencing Kit (SQK-RBK001) (Oxford Nanopore Technologies Ltd, UK) according to the manufacturer's protocol.

4.2.5. Sequencing of Nanopore Libraries

Sample libraries and a blank negative control sample library were loaded into a MinION™ Mk1B Sequencer (SKU-MIN101B) containing a 'SpotON' flow cell with R9 chemistry (SKU-FLOMIN106) (Oxford Nanopore Technologies Ltd, UK) according to the manufacturer's instructions, and the samples were sequenced for 22 hours without live basecalling.

4.2.6. Illumina Sequencing

Libraries of high molecular weight DNA were prepared by Pathogens Genomes Unit of UCL, using NEBNext II Ultra DNA kit (New England Biolabs, Inc., USA). Libraries were sequenced on an Illumina MiSeq (Illumina Inc., USA) using 500 v2 kit (2 x 250bp insert size).

4.2.7. Bioinformatics Analysis

4.2.7.1. Basecalling of Nanopore Raw Data

After Nanopore sequencing, raw .fast5 read files were based called using Albacore v2.1.3 (Oxford Nanopore Technologies Ltd, UK) with a q-score threshold setting of 7. The resultant .fastq files were filtered to a q-score of 10 using Nanofilt (De Coster et al., 2018) and demultiplexed and trimmed using Porechop (Wick et al., 2017). After Illumina sequencing, returned .fastq files were trimmed of their adapters using Trimmomatic (Bolger, Lohse & Usadel, 2014).

4.2.7.2. Nanopore-Only Draft Genome Assembly and Polishing

Draft genome assemblies of trimmed reads were performed using Canu v1.6 (Koren et al., 2017) with the following settings (genomeSize=9m \ -nanopore-raw \ corOutCoverage=1000 \ corMaxEvidenceErate=0.15). The raw nanopore reads were aligned back to the draft assembly scaffold files to create a corrected consensus sequence using RACON (Vaser et al.,

2017) for five iteration to create six sequences (raw draft assembly, and five corrected consensus sequences). Illumina reads were aligned against the Nanopore assembly using BWA-MEM and visualised in Tablet (Milne et al., 2013) to manually assess coverage. An Illumina polished consensus sequence was then made using PILON (Walker, 2014).

4.2.7.3. Illumina-Only and Illumina+Nanopore Hybrid Draft Genome Assembly

Two draft assemblies using just Illumina reads and Illumina and Nanopore reads were made using SPAdes and hybridSPAdes respectively (Antipov et al., 2015) (Parameters k: 21, 33, 55, 77, 99, 127).

4.2.7.4. QC of Genome Assemblies

Assemblies (Nanopore only, Illumina only, and Illumina+Nanopore Hybrid) were initially compared by contig numbers and lengths against the published genome reference files (GenBank: AL645882.2). Each assembly was analysed for completeness by ortholog analysis using BUSCO (Simão et al., 2015) and compared to the reference sequences using QUAST (Gurevich et al., 2013).

4.2.7.5. Comparison of House-keeping Genes

Genome assemblies and reference .fasta files were annotated using PROKKA v1.3 (Seemann, 2014) and selected housekeeping genes (16S rRNA, *gyrB*, *rpoB*, *rspL*, *atpD*, *recA*, *trpB*) from each were aligned with reference genes using the Clustal Omega algorithm and visualised in Jalview (Clamp et al., 2004).

4.2.7.6. Analysis of Secondary Metabolite Gene Clusters

Annotated .gbk files of Nanopore assemblies were analysed for secondary metabolite gene clusters using AntiSMASH v4.0.2 (Medema et al., 2011) and compared to the reference genome. MultiGeneBlast (Medema, Takano & Breitling, 2013) was used to visualise structural differences between annotated BGCs.

4.3. Results and Discussion

4.3.1. Quality Assessment of Genomic DNA Extraction

The genomic DNA (gDNA) extraction from *Streptomyces coelicolor* A3(2) was analysed on an agarose gel (**Figure 4.2**) to determine the integrity of the DNA. The results showed a tight band concentrated at the region of the 23 kbp band of ladder marker, which is suggestive of high molecular weight (HMW) DNA. In order to obtain long reads of good quality from Nanopore sequencing, high molecular weight DNA is required. Therefore, the extraction was deemed sufficient for sequencing. The extraction was also analysed on a Nanodrop (**Table 4.1**) to determine its concentration and purity.

Figure 4.2. Agarose gel (1 %) of 1 µL (~ 100 ng) of *Streptomyces coelicolor* A3(2) genomic DNA extraction. MW = λ DNA-*Hin*dIII digest molecular size marker, 1 = genomic DNA extraction. The extraction shows a band in ~23 kbp region which is indicative of HMW DNA.

Table 4.1. Nanodrop analysis of genomic DNA extraction from *Streptomyces coelicolor* A3(2). Concentration is adequate for genome sequencing, and absorbance scores suggest acceptable DNA purity.

Sample	Concentration (ng/µL)	Absorbance ratios	
		260/280	260/230
S. coelicolor A3(2)	98.7	1.92	2.12

As previously described in **Chapter 2**, 260/280 and 260/230 ratios of 1.7-2.0 and 2.0-2.2 respectively are deemed to represent "pure" DNA because indicates a high concentration of nucleic acid which has absorbance at 260 nm in comparison to common contaminating substances with high absorbances at 280 nm or 230 nm. DNA samples within these ranges is the recommended input for Nanopore library prep to minimise the presence of any salts or

biomolecules in the samples that may interfere with the efficient performance of the library prep chemistry. The extraction is within these purity ranges and so was deemed sufficient to progress with library prep.

4.3.2. Assessment of Genome Assemblies

The sequencing run produced 339,212 reads with a mean length of 4,476.6 bp and mean quality score of 8.4. The total throughput was 1,573,510,759 bp which equates to approximately 175X coverage of the *S. coelicolor* A3(2) genome. An initial genome assembly using only Nanopore long reads (≥1000 bp and ≥q10) was produced using Canu. This Nanopore-only assembly used 113,216 reads totalling 579,576,777 bp, which equates to approximately 64.39X coverage of the *S. coelicolor* A3(2) genome. The reference genome is comprised of a linear chromosome of 8,667,507 bp and two plasmids; linear plasmid SCP1 356,023 bp and circular plasmid SCP2 31,317 bp (Bentley et al., 2002). The draft assembly produced included 3 contigs of sizes that were similar to those of the reference genome (**Table 4.2**). In contrast, the Illumina read only and hybrid Nanopore+Illumina read assemblies produced using SPAdes and hybridSPAdes respectively produced fragmented assemblies of 177 and 17 contigs respectively. This shows that Nanopore only assemblies are capable of producing highly continuous contigs. However, the variations in the sizes between the draft contigs and the references varied, with the smaller contigs showing a greater size variation compared to the large contig.

Table 4.2. Comparison of contigs sizes of Nanopore-only assembly and reference genome. Coverage of plasmid contigs is higher but size difference against reference is greater

File	Contig	Size (bp)	Size difference (%)	Coverage
S. coelicolor A3(2) **Reference genome**	Chromosome	8667507	-	-
	SCP1	356023	-	-
	SCP2	31317	-	-
Nanopore-only assembly	Contig 1	8640937	0.31 %	60x
	Contig 2	295971	16.88 %	146x
	Contig 3	24062	23.17 %	91.4x

As contiguous genome assemblies are most desirable for high confidence genome-mining of long repetitive BGCs (Miller, Cevrette & Kwan, 2017), the Nanopore-only assembly was advanced to test approaches to further improve the accuracy of the assembly by polishing. The approach taken was to polish the assembled contigs aligning the raw Nanopore long reads using the polishers RACON, followed by Nanopolish. Other researchers have reported that taking an iterative approach and creating a series of multiple consensus sequences using the previous consensus as a template can show greater improvements in accuracy of consensus long read assemblies compared to reference genomes (Vaser et al., 2017), so therefore a series of five sequential consensus sequences was produced using RACON and assessed for quality.

Using the BUSCO tool the presence or absence of orthologous housekeeping genes was measured to determine the completeness of the genome assemblies. Because other researchers have reported that long-read bacterial assemblies, while contiguous, can have many indel errors which lead to fragmented protein annotations and a high proportion of pseudogenes. To test for this the PROKKA tool was used to annotate each assembly and compare the number of predicted CDS compared to the reference genome (**Table 4.3**). The BUSCO analysis of the multiple iterations of RACON polishing showed that each polishing step progressively improved the reported genome completeness up to a peak at the fourth iteration and that the extent of improvement reduced between each of these iterations.

PROKKA annotations of each assembly showed that there was a large increase in CDS compared to the reference genome of the initial Nanopore-only assembly. This could likely be due to the presence of indels in the assembly sequences leading to truncated and fragmented annotations. RACON polishing did show improvements in CDS numbers and so taken in conjunction with the improved detection of orthologous genes demonstrates that polishing Nanopore assemblies does improve protein annotations. This is in agreement with examples detailed by other researchers such as Warr & Wick (2019). The potential presence and influence of indels upon protein annotation was inspected further in later analysis. At the RACON fifth iteration there was a slight reduction in completeness scores and CDS differences. Therefore, the fourth iteration was selected for further polishing using Nanopolish which further improved the completeness and differences in annotation. Polishing this sequence using Illumina short-reads and Pilon showed a dramatic further improvement in completeness and CDS scores with the consensus file showing a completeness score slightly higher (0.3%) than the reference and 100 fewer CDS genes. From this result it is apparent that polishing Nanopore assemblies with Illumina short reads can improve sequences quality. To gain a better insight into the effect upon the assembly quality that Illumina polishing has,

both the polished sequence and its precursor were aligned to the reference genome sequence using QUAST (**Table 4.4**).

Table 4.3. Results of BUSCO and PROKKA analysis of Nanopore-only assembly files. Correlation of genome completeness scores and CDS annotations of assembly with reference genome improves with multiple polishing steps.

File	BUSCO Genome Completeness (%)			PROKKA Gene Annotation	
	Complete	Fragmented	Missing	CDS	+/- Difference compared to Reference
***S. coelicolor* A3(2) Reference genome**	99.4	0.3	0.3	8128	-
Nanopore-only assembly	23.0	48.9	28.1	14798	+ 6670
Nanopore+Racon 1	43.2	41.8	15.0	12519	+ 4391
Nanopore+Racon 2	48.9	36.1	15.0	12332	+ 4204
Nanopore+Racon 3	50.3	35.5	14.2	12192	+ 4064
Nanopore+Racon 4	51.1	34.1	14.8	12133	+ 4005
Nanopore+Racon 5	50.6	34.7	14.7	12172	+ 4044
Nanopore+Racon 4+Nanopolish	78.1	15.1	6.8	9303	+ 1175
Nanopore+Racon 4+Nanopolish+Pilon 1	99.7	0.0	0.3	8028	- 100

Table 4.4. Summary of QUAST alignments of Nanopore only and Nanopore+Illumina assemblies against *S. Coelicolor* A3(2) reference genome. Polishing with Illumina reads reduces indels.

File	Local Misassemblies	Genome fraction (%)	# mismatches per 100 kbp	# indels per 100 kbp
Nanopore+Racon 4+Nanopolish	5	98.9	6.2	66.82
Nanopore+Racon 4+Nanopolish+Pilon 1	5	98.9	2.19	2.03

Misassembled regions are defined by QUAST as regions where assembled contig flanking sequences align to opposing strands (inversion), or to different chromosomes (translocation), or if there is an inconsistency in size between the alignments of the left and right flanking sequences on the reference (relocation) (Gurevich et al., 2013). While polishing with Pilon did not improve upon these discrepancies there was a notable reduction in mismatched bases (from 6.2 to 2.19 per 100 kbp) and in the frequency of indels (66.82 per 100 kbp to 2.03 per 100 kbp). These results are suggestive that polishing of Nanopore assembled contigs with short Illumina reads of high accuracy does create high accuracy consensus sequences as described by other researchers previously (George et al. 2017, Leim et al., 2018, Miller et al., 2018, De Maio et al., 2019, Shin et al., 2019). It would seem that this improvement is through the correction of point errors in the sequences rather than in structural misassemblies. This is

to be expected as the shorter reads will be unable to overlap large misassembled regions to such an extent to correct them. This is also in keeping with findings by Kolmogorov and co-workers (2019) who have reported on the limitations of short reads to polish larger assembly errors. It is also worth noting that these misassembles reported in QUAST (relocation, inversion, translocation) could be true discrepancies which have occurred naturally and are now present in the genome of the strain sequenced for this study since its original genome report in 2002. These discrepancies could also be due to errors in the original reference genome assembly.

As a final check of the quality of the assemblies, a selection of housekeeping genes that had been annotated in PROKKA (*gyrB, rpoB, rspL, atpD, recA, trpB*) from the Nanopore+Racon 4+Nanopolish and the Nanopore+Racon 4+Nanopolish+Pilon 1 were aligned against the genes from the reference genome using Clustal. The results are summarised below in **Table 4.5**. The results revealed that for the Nanopore+Racon 4+Nanopolish assembly, there was partial or multiple fragments of three of the housekeeping genes which did not fully align with the reference and a SNP present in the alignment of *atpD* against the reference. However, for the Pilon polished genome, all of the above housekeeping genes were complete and aligned fully the reference. These results show that while the polishing steps using only Nanopore reads may have reduced the number of indels, those that remained did cause fragmented and truncated gene annotations. Whereas, polishing with Illumina based reads dramatically reduced the occurrence of indels and improved gene annotation.

Table 4.5. Summary of the results for alignment of annotated housekeeping genes between Nanopore-only assembly and Nanopore+Illumina polished assembly. Polishing with Illumina reads using Pilon improves alignments.

Housekeeping Gene	Nanopore+Racon 4+Nanopolish	Nanopore+Racon 4+Nanopolish+Pilon 1
gyrB	Partial fragment, partial alignment	Complete, full alignment
rpoB	Three fragments, partial alignment	Complete, full alignment
rspL	Complete, full alignment	Complete, full alignment
atpD	SNP at 299, full alignment	Complete, full alignment
recA	Complete, full alignment	Complete, full alignment
trpB	Partial fragment, partial alignment	Complete, full alignment

Accurate genome annotations are important in genome-mining bioprospecting studies to ensure confidence in any BGC predictions made. However, the use of Illumina data in conjunction with Nanopore increases costs and complexity which may act as a barrier to its use in small-scale bioprospecting studies. Therefore, both the genome assembly using only Nanopore reads and the Illumina polished assembly were analysed using AntiSMASH and compared against the reference genome to determine if a Nanopore only assembly is sufficient for accurately annotating BGC for the purpose of dereplication.

4.3.3. Biosynthetic Gene Cluster Annotations of Assemblies

AntiSMASH analysis identified 29 BGCs across 2 contigs (27 in chromosome and 2 in plasmid SCP1) within the reference genome file. The same numbers and types of BGCs were also identified by AntiSMASH in the Nanopore only assembly files. A summary of BGCs detected in the chromosome is given in **Table 4.6**. However, the order of BGCs was inverted which shows that the Nanopore assembly contigs were assembled in the opposite orientation of the reference contigs. The BGCs was classified as a particular compound type based on sequence motifs identified and homogeny with known BGCs on the MiBiG database. Each of the 29 BGCs were identified as the same types in both the reference and Nanopore-only assembly. This shows that AntiSMASH analysis of de novo Nanopore-only genome assemblies can accurately determine the number of and types of BGCs encoded within the genome. AntiSMASH reported mostly the same matches between the reference and Nanopore-only assembly, with only 4 differences between the reference and the Nanopore-only assembly.

The AntiSMASH analysis also lists a percentage identity match of each predicted BGCs against known BGCs in the MiBiG database. The percentage matches reported for the reference and Nanopore-Only assembly were broadly similar. However, some of the BGCs were matched to different known BGCs or matched to the same known BGC but with different levels of homogeny. Cluster 1 of the reference (cluster 27 of Nanopore-only assembly) which is classified as a T1pks-Otherpks in both files was given a 3 % match to Leinamycin in the reference and a 2% match to A54145 in the Nanopore-only assembly. Cluster 3 of the reference (cluster 25 in Nanopore-only assembly) was identified as a Lantipeptide BGC and given a 4% match to Sanglifehrin in the reference file while no match was identified in the Nanopore-only assembly. Cluster 20 in the reference (cluster 8 in the Nanopore-only assembly) was identified as a NRPS and given a 40% match to Nogalamycin in the reference file and 9% match to Borreliodin in the Nanopore-only assembly. Cluster 25 in the reference (cluster 3 in the Nanopore-only assembly) which is classified within the “Other” BGC type by AntiSMASH had no identity match in the reference but a 9 % match to Carbapenem_MM_4550 in the Nanopore-only assembly. However, all four of these clusters were given low matches in the reference file but still classified as the same BGC type in both reference and Nanopore-only assembly. This suggests that these discrepancies are also in part due to an absence of high confident matches in the MiBiG database.

Other clusters had variations in the percentage identified match but were still matched against the same known BGC. For example, cluster 9 in the reference genomes (cluster 19 in Nanopore-only assembly) was classified as a 100% match to Desferrioxamine B in the reference and 83% in the Nanopore-assembly. Cluster 11 (cluster 17) was classified as 100 % homologous to Actinorhodin in the reference and 90 % in the Nanopore-only assembly. Cluster 15 (cluster 13 in Nanopore-only assembly) was classified as a 100% match to Undecylprodigiosin in the reference and 90% in the Nanopore-only assembly. Cluster 21 (cluster 6 in Nanopore-only assembly) is a 100% match to Hopene in the reference and 92% match in the Nanopore-only assembly. While cluster 23 in the reference (cluster 5 in Nanopore-only assembly) has a 95% homology match to Arsenopolyketides while it is matched 100% in the Nanopore-only assembly. All of these matches were still of a high homology and so maybe due to variations in the sequences of the reference and Nanopore-only assembly such as an increased incidence of indels.

Despite the discrepancies reported, all of the BGCs identified in the reference genome were also identified in the Nanopore-only assembly as the same BGC type and those with significant homology matches to known BGCs in the MiBiG database were classified accordingly. Therefore, these results suggest that de novo Nanopore-only genome assemblies may be suitable for reporting the content, type, and similarly to known BGCs putatively encoded in prospected environmental microbial genomes. The AntiSMASH results were then further analysed to determine if the results could potentially be used to predict the structure of unknown predicted BGCs. This was performed by looking at the predicted substrates for certain NRPS and PKS BGCs that had been annotated by AntiSMASH in both the reference and Nanopore-only assembly.

Table 4.6. Summary of manual inspections of biosynthetic gene cluster annotations made by AntiSMASH of the Nanopore-only *S. coelicolor* A3(2) assembly in comparison to the reference genome file. Key: Y = Yes, N= No, - = Not observed/absent. The summary shows that AntiSMASH was able to identify BGCs by genetic homology in the Nanopore-only assembly with high accuracy. However, the gene clusters were fragmented in many cases, and in cases where substrate domain specificity or product structure prediction were made there was many errors.

Cluster (reference/Nanopore-only assembly)	Type	% homology with known BGC (Reference / Nanopore-only assembly)	Structure prediction match? (Y/N)	Substrate specificity match? (Y/N)	Gene fragmentation observed? (Y/N)
CHROMOSOME					
1/27	T1pks-Otherks	Leinamycin (2 % / 3 %)	N	N	Y
2/26	Terpene	Isorenieratene (100 % / 100%)	-	-	Y
3/25	Lantipeptide	Sanglifehrin_A (4 % / -)	Y	-	Y
4/24	Nrps	Coelichelin (100 % / 100 %)	Y	Y	Y
5/23	Bacteriocin	Informatipeptin (42 % / 42 %)	-	-	N
6/22	T3pks	Herboxidiene (8 % / 8 %)	-	-	Y
7/21	Ectoine	Ectoine (100 % / 100 %)	-	-	N
8/20	Melanin	Lactonamycin (3 % / 3 %)	-	-	Y
9/19	Siderophore	Desferrioxamine_B (100 % / 83 %)	-	-	Y
10/18	Nrps	Calcium-dependent_antibiotic (90 % / 90 %)	Y	Y	Y
11/17	T2pks	Actinorhodin (100 % / 90 %)	-	Y	Y
12/16	Terpene	Albaflavenone (100 % / 100 %)	-	-	N

13/15	T2pks	Spore_pigment (66 % / 66 %)	-	-	Y
14/14	Siderophore	-	-	-	Y
15/13	T1pks	Undecylprodigiosin (100 % / 90 %)	N	N	Y
16/12	Bacteriocin	-	-	-	N
17/11	Terpene	-	-	-	Y
18/10	Siderophore	Enduracidin (8 % / 8 %)	-	-	N
19/9	T1pks-Butyrolactone	Coelimycin (100 % / 100 %)	N	N	Y
20/8	Nrps	Nogalamycin (40 % / 9 %)	Y	N	Y
21/7	Lantipeptide	SAL-2242 (100 % / 100 %)	-	-	N
22/6	Terpene	Hopene (100 % / 92 %)	-	-	Y
23/5	T1pks-Otherks	Arsenopolyketides (95 % / 100 %)	-	N	Y
24/4	Lantpeptide	-	Y	-	Y
25/3	Other	Carbapenem_MM_4550 (- / 6 %)	-	-	-
26/2	Indole	Ravidomycin (5 % / 5 %)	-	-	Y
27/1	T3pks-Terpene-Nrps	Coelibactin (100 % / 100 %)	N	N	Y
SCP1					
1/1	Terpene	-	-	-	N
2/2	Butyrolactone-furan	Methylenomycin (61 % / 61 %)	-	-	N

Manual inspection of the BGCs of the reference and Nanopore-only assembly revealed a trend where these larger multi-domain enzymes (NPRS, PKS) are fragmented in the Nanopore-only assembly (**Table 4.7**). In some examples the substrate prediction of these domains was inaccurate in the Nanopore-only assembly, even if the BGC had been correctly matched to the same homologous gene cluster. Notable examples are cluster 4 (ref)/ 24 (assembly) which is of the Coelichelin BGC, which is a non-ribosomal peptide siderophore (Challis, 2000). The gene cluster in both the reference and Nanopore-only assembly files were correctly matched to the Coelichelin BGC cluster with 100 % homology and all the substrate specificity of all the adenylation domains of the NRPS were correctly predicted by AntiSMASH in both files, along with the structure chemical scaffold that these domains would produce. However, the NRPS and downstream efflux pump genes in the Nanopore-only assembly file appeared fragmented. Manual inspection of both sequences in Tablet confirmed this to be due to the presence of erroneous stop codons in the Nanopore-only assembly.

In another example, cluster 3 (ref) / 25 (assembly) which was called as a Lantipeptide by AntiSMASH show the same predicted peptide structure, however comparison of both gene clusters show fragmentation in the assembly in comparison to the reference.

Cluster 19 (ref) / 9 (assembly) was identified as a 100% match in AntiSMASH as the biosynthetic pathway for Coelimycin in both reference and assembly. This is a cryptic biosynthesis gene cluster, the biosynthesis of which has been determined through overexpression mutagenesis studies (Gomez-Escribano et al., 2012). Unlike with the Coelichelin BGC, the substrate specificity predictions of the PKS domains and product structural predictions provided by AntiSMASH differed with the reference file with five domains missing in the assembly annotation. Fragmentation and truncation of PKS genes was observed in the assembly annotation.

Cluster 27 (reference) / 1 (assembly) of Coelibactin is the putative product predicted to be encoded by the cryptic NRPS BGC (Zhao et al., 2012). Comparison of the two showed that, as with other clusters, the large NRPS genes were fragmented in the assembly file and there were several discrepancies in the substrate specificities of these genes' adenylation domains. This in turn lead to different scaffold structure predictions being reported for both the reference and assembly file.

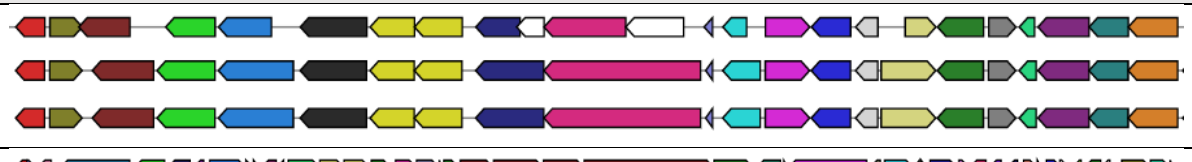
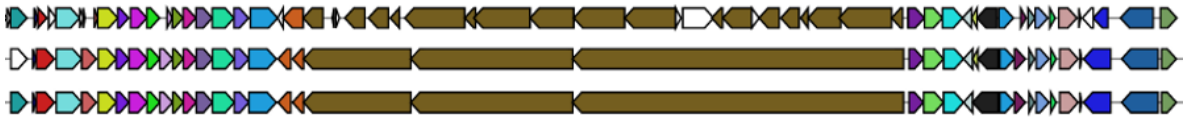
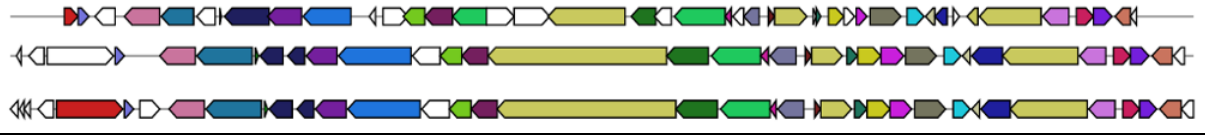
Cluster 10 (reference) / 18 (assembly) was matched to calcium-dependant antibiotic BGC (Hojati et al., 2002) with 90% gene homology in both the reference and assembly files. The gene cluster consists of three large NRPS genes each with multiple adenylation substrate

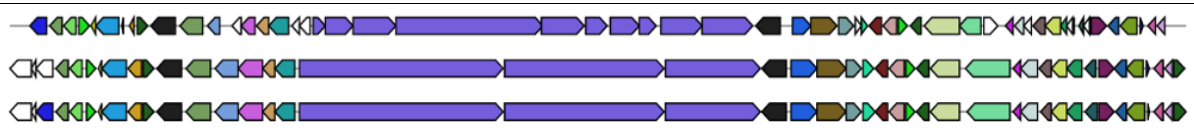
domains. In the Nanopore-only assembly these genes are fragmented and there is variation in two of the substrate domains are missing in comparison to the reference annotation. These lead to a difference in the core structure scaffold predicted for the assembly BGC compared to the reference file.

Cluster 15 (reference) / 13 (assembly) also showed fragmentation of the PKS genes in the Nanopore-only assembly in comparison to the reference. While substrate specifies and a scaffold structure prediction were reported by AntiSMASH in the reference file, neither predictions were made with the Nanopore-only assembly file. However, both annotations showed a strong gene homology match to the undecylprodiosin BGC of 100% and 90% for the reference and assembly annotations respectively.

Cluster 11 (reference) / 7 (assembly) was matched to Actinorhodin with 100% homology in the reference and 90% in the assembly. AntiSMASH did not provide a chemical structure prediction for the scaffold of this BGC, most likely because it is a Type 2 PKS and the algorithm that predicts scaffold predictions relies upon the specific substrate domains to be along the same coding regions limiting predictions to Type 1 PKS only. However, fragmentation and truncation of genes in this cluster was observed in the assembly annotation in comparison to the reference.

Table 4.7. Selected multigeneblast comparisons of AntiSMASH biosynthetic gene cluster annotations of the Nanopore-only and Nanopore+Illumina polished assemblies of *S. coelicolor* A3(2) genome against the reference genome file. The assemblies are aligned above the reference and genes that share the same colour indicate homology. The multigeneblast comparisons show that in the Nanopore only assemblies, there is examples of fragmented genes, especially in the long multidomain NRPS or PKS genes which are rectified after polishing with Illumina reads.

	Multigene Blast Results	Assembly	Cluster (reference/ assembly)
a)		Nanopore Only Nanopore+Illumina Polish Reference	3/25
b)		Nanopore Only Nanopore+Illumina Polish Reference	4/24
c)		Nanopore Only Nanopore+Illumina Polish Reference	10/18
d)		Nanopore Only Nanopore+Illumina Polish Reference	11/17
e)		Nanopore Only Nanopore+Illumina Polish Reference	15/13

f)		Nanopore Only Nanopore+Illumina Polish Reference	19/9
g)		Nanopore Only Nanopore+Illumina Polish Reference	27/1

Manual inspection of the above-mentioned sequences in Tablet confirm this to be due to the presence of erroneous stop codons in the Nanopore-only assembly. These findings suggest that AntiSMASH analysis of Nanopore-only microbial genome assemblies has the potential to provide enough accuracy to give a high confidence match to known gene clusters which would make this a simplistic and cost-efficient dereplication strategy in bioprospecting studies. However, the presence of indels and erroneous stop codons in these assemblies effects protein prediction of these genes, and this in turn effects the reliability of prediction of novel BGCs and their potential products.

The next step was to analyse the short read polished assembly in AntiSMASH to determine if it reduced the discrepancies seen between the Nanopore-only assembly and the reference. The results of these analyses showed that the short read polished genome showed a near perfect match to the reference genome annotations. Also, all examples of differences between the Nanopore-only assembly and reference in terms of product structural predictions, gene fragmentations, or specific subunit domain predictions were corrected in the short-read polished genome with a perfect match seen in all BGCs with the reference file. Manual inspection of some of these regions in Tablet showed that erroneous stop codons that were present in the Nanopore-only assembly had been rectified.

Therefore, it can be concluded that polishing with the Illumina short-reads had created a higher accuracy consensus sequence which removed indel errors and their subsequent downstream effects on protein annotation. This is in keeping with observations of other researchers such as Stewart and co-workers (2019) who demonstrated a correlation between reduced protein truncation and improved sequence accuracy in Nanopore genome assemblies after polishing. They achieved this by aligning protein annotations from raw and polished assemblies with their matches in the SwissProt and comparing discrepancies. The authors found that with increasing rounds of polishing with both Nanopore reads and Illumina reads reduced the size discrepancies between the predicted proteins and their database matches.

These results suggest that this pipeline of Nanopore long-read genome assembly followed by polishing with Illumina short-reads creates a genome assembly of high enough quality for prediction of natural product structures from annotated BGCs. This could be useful when attempting to isolate compounds from novel BGCs because having an estimation of the chemical scaffold structure could help inform of the potential chemical characteristics of the compound. These predictions could then be used to select the most appropriate chemical extractions methods.

A recently published study by Goldstein and co-workers (2019) also reports improvements in BGC annotations of Nanopore-assemblies after polishing with Illumina reads. The authors sequenced several *Pseudonocardia* strains using both Illumina and Nanopore technologies and compared differing assembly strategies across various metrics (contiguous, sequence accuracy, and BGC prediction). They noted that assemblies produced by Canu plus a polishing step (in their case either Nanopolish or Pilon alone, not in combination) had greatly improved representation and quality of BGC annotations compared to Illumina or Nanopore-only assemblies. However, they also noted that all assemblies had flaws. Unlike in this study their methodology did not employ multiple polishing steps on the same assembly, and nor did they use a benchmarking tool such as BUSCO to determine if each polish step had improved the assembly. Using a method such as this may have led to further improvements in BGC annotations.

4.3.4. Conclusions

Nanopore sequencing and genome assembly produced a contiguous genome assembly in comparison to the fragmented assemblies produced using both Illumina-only reads or a hybrid assembly with both data types. Contiguous assemblies are advantageous in bioprospecting in ensuring adequate coverage and annotation of long and repetitive BGCs. However high accuracy in gene annotations is also required to ensure reliable predictions of a genomes biosynthetic potential can be made.

Polishing the genome assemblies with the Nanopore-reads did improve the quality of genome annotations but errors still existed which was exemplified in the alignments of housekeeping genes to the reference (**Table 4.6**). A further polishing step using Illumina-reads further improved the gene annotations. This would suggest that both Nanopore and Illumina sequencing data is required to produce the highest quality genome assemblies. However, the use of both technologies adds cost and complexity to the pipeline which may be a bottleneck in some bioprospecting studies.

Comparison of the Nanopore-only assembly to the reference genome reveals that the Nanopore-only assembly gave a reliable overview of the number and types of biosynthetic gene clusters encoded within the genome. Where a reliable match exists, the analysis was also able to reliably match these BGCs to known BGCs based on genetic homology. However, deeper analysis showed that, even for BGCs that had been correctly matched by gene homology, chemical structure prediction was not as reliable. Many of the large multicatalytic genes were fragmented, some genes were misarranged, and predictions of the substrate

specificities of some of these domains were inaccurate. These results suggest that while the Nanopore-only assembly provided a useful guide for identifying known BGCs, it was not reliable for estimating the structures of unknown BGCs that maybe annotated in bacterial genomes. In contrast, the Illumina polished assembly corrected all discrepancies between the Nanopore-only assembly and reference in terms of product structural predictions, gene fragmentations, or specific subunit domain predictions corrected.

This demonstrates that Nanopore sequencing has utility in bioprospecting studies. The Nanopore-only assembly was able to detect known and well characterised BGCs with confidence. Therefore, it can have use for dereplication in bioprospecting studies. However, the high indel rate which causes fragmented gene annotations limits the utility of the Nanopore-only assembly in informing bioactivity-guided isolation strategies for any novel BGCs. This means that in order to fully integrate genome-mining with bioactivity-guided isolation the Nanopore+Illumina polishing pipeline would be the best approach. In addition to this, the higher accuracy annotations of housekeeping genes seen in the Illumina polished genome assembly would mean high accuracy in genome-wide phylogenetic analysis.

The Nanopore+Illumina polishing pipeline was therefore carried forward in the investigation to characterise an antibiotic-producing isolate (KB16) from the hot-spring water of the Roman Baths. The analysis of the genome assembly was used to inform the bioactivity-guided isolation strategy used to attempt to identify the antibiotic compound being produced (**Chapter 5**). Additionally, a parallel investigation was performed to determine the utility of the Oxford Nanopore MinION for profiling of the microbe of the source water. The aim was to utilise long-reads to attempt to identify putative BGCs directly from the source water to inform further bioprospecting of the site (**Chapter 6**).

The study showed that the pipeline used was able to assemble high GC *Streptomyces* reads well. However, many factors can influence the quality of a genome assembly. The nature of the genome itself plays a significant role, for example its structure (circular or linear), GC content and homopolymer sections. When performing *de novo* assembly on unknown organisms, these factors cannot be controlled and the influence they may have on genome sequencing and assembly is unknown. Testing this pipeline on a wider range of microorganisms with differing genome characteristics could help inform on how robust the pipeline is to such variances.

Other factors that can also influence genome assembly can include the length of reads generated, their quality, and the depth of sequence coverage generated. During bioinformatic

processing, a minimum quality threshold was applied of q10 (90 % accuracy) to the Nanopore reads alongside a minimum length threshold of 1000 bp. During assembly of these reads using Canu, a theoretical coverage limit of 1000X was applied rather than the default 60x in order to ensure the maximal number of reads were used during the assembly to ensure the best consensus sequence was generated. However, due to the nature of the library prep method used in this study, which fragments the input DNA randomly by use of a transposase, there is no definitive upper limit to read length and the relative distributions of read lengths sequenced will be random. Likewise, the upper limits of read coverage were not capped and for every sequencing project this output may be variable depending upon the size of the genome being sequenced and the number of quality reads obtained from the sequencing run. These variations can all potentially impact the quality of any genome assembly and, in the case of sequencing unknown microorganisms, would be difficult to control. Tests which look to assemble genomes using only subsets of read of different lengths distributions and coverages could be performed to test the impact that these factors may have on the pipeline developed in this study.

Bioinformatic tools used in this study were chosen due to their prominence within the field having been established in research publications. There is no agreed "Gold Standard" methodology for the assembly of genomes or in their assessment of quality and analysis, and there are dozens of open source packages available that can be utilised. This means that it is not possible to evaluate every potential tool through direct experimentation. For example, De Maio et al., (2019) performed a review of hybrid assembly approaches using a collection of 20 Enterobacteriaceae species. They reported that a hybrid assembly approach using the Unicycler pipeline produced higher sequence accuracies and greater completeness than a long-read assembly followed by short-read polishing. However, in the comparison study of Enterobacteriaeae plasmid assembly methods by George et al. (2017), authors reported Canu+Pilon pipeline superior to other hybrid and short-read only methods tested. However, these studies are not standardised between each other either in terms of the sample types used or pipelines tested, or the criteria or methods in which each pipeline was judged. A recent attempt to benchmark the performance of six common long-read assemblers (Canu, Flye, Miniasm/Minipolish, Raven, Redbean and Shasta) for computational efficiency, consistency, and accuracy across 620 data sets by Wick and Holt (2019) concluded that no single assembler performed well on all metrics. Therefore, the choices of what pipeline(s) to use should depend on limiting factors such as the nature of the dataset, the research question to hand, familiarity with the operation of the package, user skill level, and costs in terms of financial, time, and computational power.

Ultimately the most suitable pipeline is one that is robust to provide confident answers to the research question consistently across different datasets, while falling within the other limiting factors that the researcher works within. Comparisons of assembled data to reference files can help gauge the suitability of a particular pipeline. But full confidence is not obtainable for de novo assembly of unknown species due to the variety of unknown aspects of its genome that can affect its assembly. To this aim, the pipeline developed in this chapter offers a good starting point as a model for reliably identifying the taxa and the natural product biosynthetic potential of novel microorganisms in smaller scale bioprospecting studies when used in combination with other methods. This can aid in dereplication and so avoid wasted resource investigating producers of known compounds and was utilised in **Chapter 5**. But further expansion by testing the pipeline against other microbial reference species could add greater confidence to the pipeline and give insights into what its limitations may be.

Chapter 5.

Investigation to Characterise KB16, An Antibiotic Producing Isolate from the Roman Baths, Using Genome-Mining and Bioactivity-Guided Isolation

5.1. Introduction

As detailed in **Section 1.6.4.3**; aquatic environments have been a source for microbial bioprospecting which has yielded enzymes of commercial value as well as antibiotic compounds. The different and variable physical and chemical conditions in some these aquatic environments, in comparison to temperate terrestrial soil environments, appeal to bioprospectors attempting to isolate novel microbes capable of producing metabolites with novel chemical structures and activities. Especially, thermal springs which are characterised by higher temperatures and mineral contents.

In 2015, team members (SG and PS) isolated a bacterial strain (KB16) from swabs of the King's Bath thermal spring (The Roman Bath, UK) that exhibited antibiotic activity against Gram-positive bacteria, including methicillin-resistant *Staphylococcus aureus*. The swab was taken from inner stonewall surface of the King's Bath, approximately 6 cm below the waterline. The swab was stored for 24 hours at room temperature before being plated onto Nutrient Agar and incubated at room temperature for 14 days. To our knowledge there are currently no reported examples of antibiotic-producing bacteria having been isolated from this site before. KB16, and the putative antibiotic compounds it produces, had not yet been characterised.

Bioactivity-guided isolation (as detailed in **Section 1.3**) has been the standard process by which microbially-derived antibiotics have been isolated and characterised. However, the process is highly resource, time, and capital intensive, which can serve as a bottleneck for small-scale bioprospecting studies. Additionally, a robust dereplication strategy is necessary in order limit the chance of rediscovering known compounds.

The utility of genome-mining for dereplication had been demonstrated in the work detailed in **Chapter 3** which identified the AS-48 BGC in the four *Bacillus* species isolated from honey. However, the fragmented assemblies and poor resolutions of some BGC annotations reduced the confidence of predictions made. A high confidence annotation could yield high confidence predictions of the BGCs products which could also inform bioactivity-guided isolation approaches. To this end, a pipeline utilising Nanopore long-read sequencing was optimised for microbial characterisation and bioprospecting (**Chapter 4**).

The objective of the work detailed in this chapter is to characterise the taxonomy of KB16 and to attempt to isolate and identify the antibiotic compounds it may be producing using a bioactivity-guided isolation strategy. The Nanopore long-read sequencing pipeline (**Chapter**

4) optimised with *Streptomyces coelicolor* A3(2) was utilised in order to validate its use on unknown environmental isolates by producing a high-resolution phylogenetic profile of KB16, dereplicate known BGCs, and to inform decisions on the bioactivity-guided isolation strategies to take.

5.2. Material and Methods

5.2.1. Microorganisms

5.2.1.1. KB16 Cultivation and Maintenance

KB16 was streaked to purity and cultivated on nutrient agar (NA) (Oxoid Ltd, UK) and Soya Flour Media (SFM) (2 % soya flour, 2 % mannitol, 2 % agar) (Keiser et al., 2000) plates for up to 14 days at room temperature. Spores where cultivated from the plates and stored in Microbank™ tubes (Pro-Lab Diagnostics, Inc., Canada) at -80 °C. Cultivated plates of KB16 were maintained at -4 °C for up to 7 days with up to two subcultures made before returning to frozen stocks.

5.2.1.2. Indicator Strains Used in Antimicrobial Activity Assay

Escherichia coli NCTC 10418, Methicillin-Sensitive *Staphylococcus aureus* NCTC 12981 (MSSA), Methicillin-Resistant *Staphylococcus aureus* NCTC 13373 (MRSA), Vancomycin-Resistant *Enterococcus faecalis* NCTC 13379, and a multidrug-resistant clinical strain of *Klebsiella pneumoniae* 342 (Source: Paul Stapleton, UCL) were maintained at -80 ˚C with 20 % (v/v) glycerol and on Nutrient Agar slopes (Oxoid Ltd., UK). Strains were propagated directly from stocks by streaking to purity onto Nutrient Agar (Oxoid Ltd., UK) and incubating at 37 ˚C for 24-48 hours.

5.2.2. Screening of KB16 for Antimicrobial Activity

5.2.2.1. Antibiotic Activity Assays

5.2.2.1.1. Cross-Streak Assay

The test isolate KB16 was screened for antibiotic activity on solid media using the cross-streak assay against the above mentioned indicator strains (**Section 5.2.1.2.**). The test isolate was inoculated across a third of a Nutrient Agar (Oxoid Ltd., UK) plate and incubated at 37 ˚C for 24 hours. After initial incubation, a sterile loop was used to inoculate the indicator strains in a straight line across the clear area of the plate, perpendicular to the test isolate. The indicators were streaked to within ~1 mm of the edge of the test isolate with care taken to not touch the test isolate with the loop. The plates were then incubated at 37 ˚C overnight, after which they were examined for antibiotic activity.

5.2.2.1.2. Fermentation Broth Activity Assay

The test isolate KB16 was screened for production of antibiotic activity on liquid media. The test isolate was inoculated into 3 x 3 mL of Nutrient Broth (NB) (Oxoid Ltd, UK) and incubated for 4 days at room temperature with shaking (100 rpm). After establishment of growth, 250 µL aliquots were taken on daily time points for ten days. These aliquots were centrifuged at 13,000 rpm for 5 mins to pellet KB16 and 100 µL of supernatant was removed and inoculated into 96-well microtiter plates with a 100 µL PBS suspension of methicillin-sensitive *Staphylococcus aureus* NCTC 12981 at a concentration of 1 x 10^3 CFU/mL. The plates were incubated at 37° C O/N and the presence or absence of indicator strain growth was observed by eye. A mix of 100 µL of fermentation broth supernatant and 100 µL fresh Nutrient Broth and a mix of 100 µL indicator strain suspension and 100 µL fresh Nutrient Broth were used a control for the sterility of the supernatant and for growth of in the indicator strain respectively.

5.2.2.1.3. Disk Diffusion Assay

The activity of liquid fractions considered to contain antibiotic compounds produced by KB16 was tested by means of a disk diffusion assay. 10^8 CFU/mL suspensions of indicator strains were made by picking fresh overnight colonies and suspending into 1 mL of PBS and adjusting to OD_{600} 0.1. These suspensions were swabbed across the surface of Nutrient Agar plates and allowed to dry. While the inoculated plates were drying, 6 mm diameter sterile paper discs (Sigma-Aldrich, UK) were saturated with a liquid fraction and allowed to dry. Once dry, the discs were applied to the surface of the dry inoculated plates and incubated overnight at 37 °C. A suitable antibiotic disc (Sigma-Aldrich, UK) for each indicator strain was used a control.

5.2.2.1.4. Bioautography Assay

Thin-layer chromatograms of complex extractions were assayed using bioautography (Dewanjee et al., 2015) in order to determine which fraction had antibiotic activity. Freshly developed analytical TLC plates were dried of any residual solvents under a lamina flow hood for 10 mins and placed onto 15 cm^2 sterile petri dishes. Fifty mL of molten Nutrient Agar that was cooled to approximately 50 ºC and inoculated with a 50 µL PBS suspension of appropriate indicator strain at OD_{600} 0.1 and mixed by gentle agitation. The agar was then gently poured into the petri dish to ensure full and even coverage of the TLC plate, and left to set. Once the agar had solidified the plate was incubated overnight at 37 ºC. Zones of inhibition were

observed by covering the surface of the agar with a suspension of tetrazolium salt (MTT) in methanol and observing colour changes.

5.2.3. Isolation of Putative Antibiotic Compound(s) Produced by KB16

5.2.3.1. Solvent Extraction from Solid Media

Attempts to isolate antibiotic compound(s) secreted by KB16 into solid media were made. Spore suspensions (100 µL) of the test strain were spread over the surface of 100 Nutrient Agar plates and incubated at room temperature for 14 days until a dense confluent lawn across the entirety of the plate surface had been cultivated. The biomass was scrapped off the plates and the agar diced into approximately 5 mm^2 pieces. Approximately 400 g of diced agar was submerged into 500 mL of ethyl acetate in 2 litre flasks across three batches and sonicated for 90 mins, with manual inversion of the flasks every 15 mins. The solvent extraction was passed through Whatman filter paper (Camlab Ltd., UK) under vacuum and dried using a rotary evaporator. The mass of the dried extract was recorded and resuspended in ethyl acetate. Extractions of pure Nutrient Agar using the same methodology were also made as a control. An attempt was later made to upscale fermentation of KB16 onto ~600 Nutrient Agar plates.

5.2.3.2. Reverse Phase Solid Phase Extraction (RP-SPE)

Dried extracts from solid media (Section 2.2.2.1.) were resuspended in water using a STRATA® C18-E (50 µm / 70 A) 70 g/150 mL giga tube (Phenomenex Inc., U.S.A.) according to the manufacturer's protocol using methanol and water as eluting solvents. Eleven 200 mL fractions were collected, dried by both rotary evaporation and freeze drying and the mass was recorded.

5.2.3.3. Analytical Thin-Layer Chromatography (TLC)

Extractions were separated into fractions by analytical thin-layer chromatography (TLC). Dried extract was dissolved in ethyl acetate and 1 cm long bands of the extract was spotted onto the baseline of a 5 x 13 cm foil backed silica gel plates (Sigma-Aldrich, UK). Plates were developed in a mobile phase mixture containing 95 % ethyl acetate and 5 % acetonitrile that had been acidified with acetic acid. Developed plates were visualised under UV light at 254 nm and 366 nm, and also with 1 % vanillin in sulphuric acid.

5.2.3.4. Preparative Thin Layer Chromatography (Prep-TLC)

A subset of active fractions in ethyl acetate were deposited across a 10 cm band onto the baseline of a Plates were developed in a mobile phase mixture containing 95 % ethyl acetate and 5 % acetonitrile that had been acidified with acetic acid.

Spots corresponding to Rf values of interest were scrapped off developed plates into 0.7 mL deuterated methanol (Sigma-Aldrich, UK) and filtered through tissue paper to remove excess silica into 5 mm NMR tubes.

5.2.3.5. Nuclear Magnetic Resonance (NMR) Analysis of Prep-TLC samples

One dimensional (H) NMR analysis was performed upon prep-TLC samples using a Bruker Advance 400 spectrometer operating at 400 MHz and the spectra were visualised using TopSpin v2.1 (Bruker Corporation, U.S.A.)

5.2.3.6. Liquid Extraction Surface Analysis Mass Spectrometry (LESA-MS)

Replicate TLC plates of active fractions of interest were made as per the method described in **Section 5.2.3.3** and sent to the Centre for Mass Spectrometry Imaging at Sheffield Hallam University, UK for analysis by LESA-MS.

5.2.4. Sequencing of KB16

5.2.4.1. DNA extraction

KB16 biomass was generated by incubation of a loop of spores in 200 mL brain heart infusion broth (Sigma-Aldrich, UK) for 6 days at 37 °C / 240 rpm. Extraction of DNA from this biomass was performed using the method detailed in **Section 4.2.2**.

5.2.4.2. Gel Analysis of DNA Extracts

The DNA extract was analysed for integrity by gel electrophoresis using the method detailed in **Section 4.2.3**.

5.2.4.3. Phylogenetic typing of KB16

The 16S rRNA gene of the KB16 isolate was amplified using the primers and method detailed in **Section 3.2.5**. PCR products were purified using Monarch PCR & DNA Cleanup Kit (New England Biolabs Inc., USA) according to the manufacturer's protocol and sequenced on an ABI 3730xl DNA Analyser by (GATC GmbH, Germany), with the samples prepared to the supplier's specification using the 27F/1492R primer pair.

5.2.3.4. Preparation and Sequencing of Nanopore Libraries

The high molecular weight DNA was prepared as per the method detailed in **Section 4.2.4.** and sequenced on the Oxford Nanopore platform as per the method detailed in **Section 4.2.5**.

5.2.3.5. Illumina Sequencing

Libraries of high molecular weight DNA were prepared by the Pathogens Genomes Unit of UCL as previously described in **Section 4.2.6**.

5.2.5. Bioinformatic Analysis

5.2.5.1. Processing and QC of Raw Sequencing Reads

After Nanopore sequencing, raw .fast5 read files were based called using Albacore v2.1.3 (Oxford Nanopore Technologies Ltd, UK) with a q-score threshold setting of 7. The resultant .fastq files were analysed using QC purposes using NanoPlot (De Coster et al., 2018), filtered to a q-score of 10 using Nanofilt (De Coster et al., 2018), and demultiplexed and trimmed using Porechop (Wick et al., 2017). After Illumina sequencing, returned .fastq files were trimmed of their adapters using Trimmomatic (Bolger, Lohse & Usadel, 2014).

5.2.5.2. Genome Assembly of KB16

The genome of KB16 was assembled using both Nanopore long reads and Illumina short reads following the pipeline outlined in **Chapter 4**. The scaffold was assembled from Nanopore data using Canu v1.6 (Koren et al., 2017) with the following settings (genomeSize=9m \ -nanopore-raw \ corOutCoverage=1000 \ corMaxEvidenceErate=0.15). The raw Nanopore reads were aligned back to the scaffolds to create a corrected consensus sequence using RACON (Vaser et al., 2017) for five iterations. Each itineration was checked for completeness by ortholog analysis using BUSCO (Simão et al., 2015). The consensus sequence with the highness completeness score was polished with Illumina reads using PILON (Walker, 2014) and this polished consensus sequence was checked for completeness using BUSCO.

5.2.5.3. Annotation and Phylogenetic Analysis of KB16 Genome

The complete genome assembly was annotated using PROKKA v1.3 (Seemann, 2014). PhyloPhlAn (Segata et al., 2013) was used to align 400 orthologous genes against 102 *Streptomyces* genomes downloaded from the RefSeq database. The results were visualised as a cladogram using MEGA v6 (Tamura et al., 2013).

5.2.5.4. Identification of Putative Biosynthetic Gene Clusters in KB16

GenBank file (.gbk) of complete assembly was analysed using AntiSMASH v3.0.5 (Weber et al., 2015). Settings: ClusterFinder algorithm - disabled, BLAST comparisons to other gene clusters – enabled, smCOG analysis for functional prediction and phylogenetic analysis of genes – enabled, Active site finder – enabled. The returned results were downloaded in GenBank format and used to generate a database which was used to compare the individual BGCs for homology using MutliGeneBlast (Medema, Takanko & Breitling, 2013). Predicted NRPS/PKS catalytic domains, substrate specificities, and product structures predicted in AntiSMASH were manually curated using the NCBI conserved domain database (http://blast.ncbi.nlm.nih.gov, National Center for Biotechnology Information, USA). Putative bacteriocin BGCs were further analysed by searching for homologs in the bactibase database (Hammami et al., 2010). Putative NRPS and PKS multi-enzyme complexes were searched for homogeny in the MIBiG database (Medema et al., 2015).

5.3. Results and Discussion

5.3.1. Assessment of Antibiotic Activity of KB16

The antibiotic activity of KB16 was assessed against several indicator strains by means of a cross streak assay as seen in **Figure 5.1**. As previously discussed in **Chapter 3**, the rationale of the assay is that once the test isolate (in this case KB16) has formed an established confluent lawn upon the agar plate then any natural product compounds it is producing with antibiotic effect will diffuse through the agar from the leading edges of the lawn. This will then be shown by an inhibition of growth of any of the indicator strains that are streaked across the clear area of the plate perpendicular to KB16. The images in **Figures 5.1a-e** show that there is a consistent inhibition across all repeats of the three Gram-positive indicator strains; Methicillin-Sensitive *Staphylococcus aureus* NCTC 12981 (MSSA), Methicillin-Resistant *Staphylococcus aureus* NCTC 13373 (MRSA), and Vancomycin-Resistant *Enterococcus faecalis* NCTC 13379. While there is no observed inhibition of growth of the two Gram-negative indicator strains *Escherichia coli* NCTC 10418 and multidrug-resistant clinical *Klebsiella pneumoniae.* The control assay shown in **Figure 5.1f** showed that there was no inhibition of growth of any indicator strain as all streaks had visible growth up to ~1 mm of the leading edge of the control strain. Therefore, the results of this assay suggest that KB16 produces at least one antibiotic compound that can inhibit the growth of multiple Gram-positive bacteria but does not show any effect against the Gram-negative bacteria tested in this assay. Gram-negative bacteria are distinguished from Gram-positive bacteria in having a second lipopolysaccharide outer cell membrane encasing the peptidoglycan cell-wall, which can often act as an additional barrier to entry of toxins and is known to contribute to innate resistance to several clinical antibiotics such as β-lactam class (Miller, 2016). Therefore, it is possible that the compounds(s) secreted by KB16 which are showing inhibition of the Gram-positive bacteria tested in this assay are ineffectual against Gram-negative bacteria due to the presence of this additional outer-membrane.

The assay included both sensitive indicator strains such as *Staphylococcus aureus* NCTC 12981 and *Escherichia coli* NCTC 10418, as well as multidrug-resistant strains: Methicillin-Resistant *Staphylococcus aureus* NCTC 13373 (MRSA), Vancomycin-Resistant *Enterococcus faecalis* NCTC 13379 and a clinical *Klebsiella pneumoniae*. The reason for using highly drug sensitive strains is to ensure that any antibiotic activity was easy to observe by eye in a qualitative assay where expression levels, diffusion efficiency, and the potency of any antibiotic compound is unknown. The reason for also including multidrug resistant compounds is twofold. Firstly, it gives some indication as to whether the antibiotic compound

might be of clinical importance by having activity against important MDR pathogens. Secondly, it gives a level of confidence that there is a possibility that the compounds mode of action is novel to those of existing antibiotic medicines in clinical use. This is important as a dereplication mechanism to try to minimise the possibility of "rediscovering" known compounds. However, it is worth noting that this is not a failsafe approach, as drug resistance is measured in terms of dose response and the dosage of the compound(s) observed in this assay is unknown because the expression/production levels of the compounds by KB16 is unknown. Therefore, it could be the case that the compound(s) are not novel in their mechanism of action but just present at a high dosage. Another metric that this assay cannot measure is if the compound(s) structures are of a type that has a higher probability of being developed into a stable and safe medicine (i.e. is it "druggable"?). However, that cannot be assessed without further investigations and so therefore, from the basis of these results, it was determined that the antibiotic activity exhibited by KB16 was worthy of further investigation.

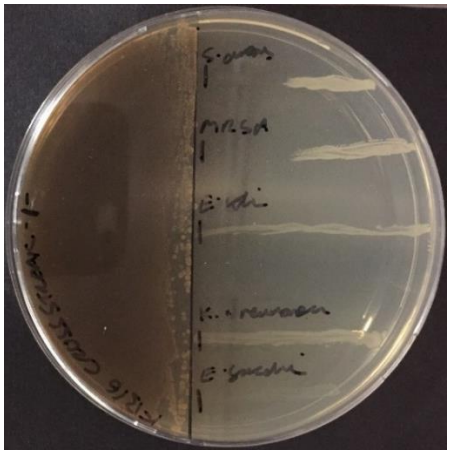


Figure 5.1a) KB16 replicate **1**

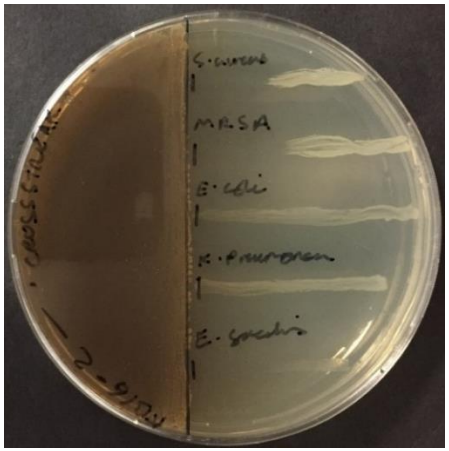


Figure 5.1b) KB16 replicate **2**

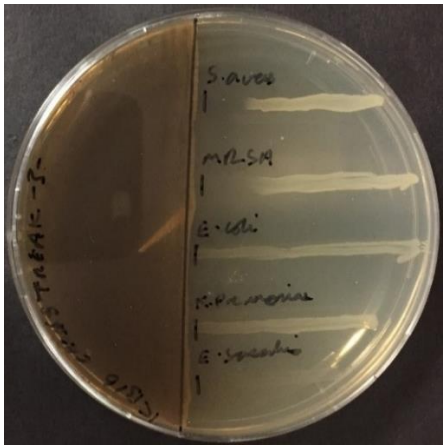


Figure 5.1c) KB16 replicate **3**

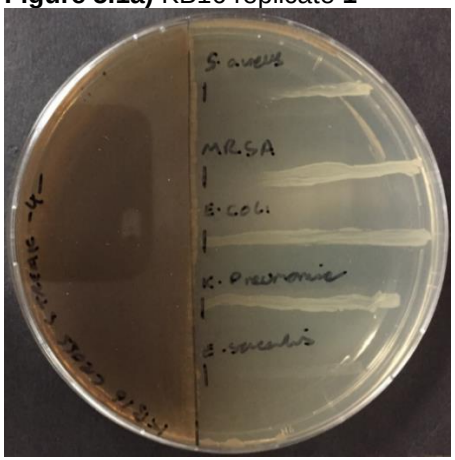


Figure 5.1d) KB16 replicate **4**

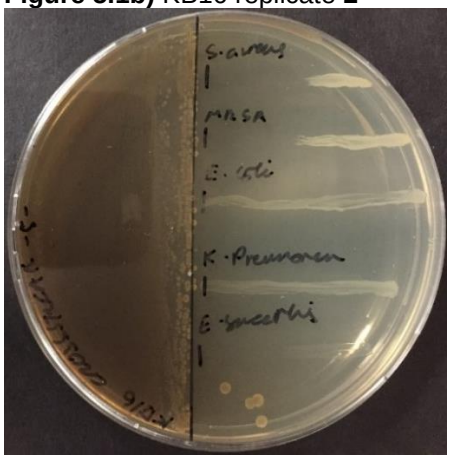


Figure 5.1e) KB16 replicate **5**

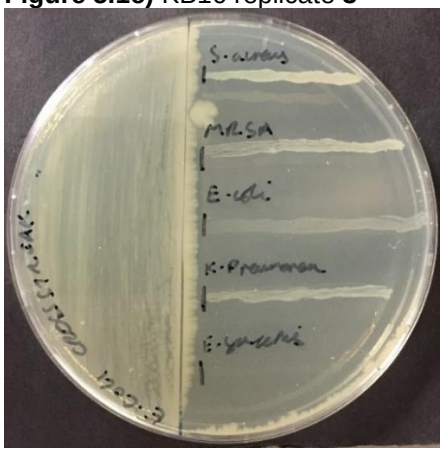


Figure 5.1f) *E. coli* control assay

Figure 5.1. KB16 Cross streak of KB16 incubated at 30 °C for 10 days against MSSA, MRSA, *E. coli*, *K. pneumoniae*, VRE. Figures 1a-e are replicates and Figure 1f is a control using *E. coli* as the test strain. All Gram-positive indicator strains are inhibited by KB16, but Gram-negatives are not inhibited

5.3.2. Taxonomy of KB16

5.3.2.1. Morphology of KB16

The first step after the assessment of the antibiotic activity of KB16 was to identify its taxonomy. This information is important because understanding the taxonomy of the isolate provides a reference point by which the literature can be researched to understand optimal methods and past precedent on the optimal handling of related isolates. Initial investigations focused on assessing the cellular and colony morphology of the isolate because appearances of microbes in different culture conditions can provide evidence to its taxonomy. The cross-streak assay (**Figure 5.1**) showed that on Nutrient Agar KB16 formed small circular colonies, with a flat elevation and linear margins. The colonies were a uniform beige colour, but a dark brown pigment was secreted into the agar which omitted a deep earthy odour. Such an odour is a characteristic of the genera *Streptomyces*, a notable producer of antibiotic natural products (Kieser et al., 2000).

Streptomyces are Gram-positive bacteria which have a lifecycle analogous to filamentous fungi. *Streptomyces* are filamentous bacteria that grow through the production of long chained branched hyphae which form a substrate mycelium for the absorption of nutrients from their environment. Once established, this mycelium will produce aerial hyphae supporting spores. These spores are distributed to new locations by the environment and germinate to form new mycelia colonies (Kieser et al., 2000).

The morphology of KB16 in different growth conditions was observed to determine if it was in accord with the characteristic morphology of *Streptomyces*. Firstly, the isolate was incubated in 3 mL of nutrient broth over the course of 14 days at room temperature with agitation. **Figure 5.2b** shows the culture after this incubation. The image shows that the culture has formed small discrete clusters of biomass. This morphology in liquid culture is in accord with that of *Streptomyces* and other filamentous microorganisms and is caused by the growing hyphae clustering into itself to form little “balls” of mycelia. A small aliquot of this liquid culture was Gram stained and observed under a light microscope (**Figure 5.2a**). The image reveals a pure culture that is Gram-positive and filamentous. The filaments also contained small dark spots along their lengths which is characteristic of spores. Characteristic appearance of sporulation had not been seen in colonies grown on Nutrient Agar. Therefore, KB16 was incubated on Soya Flour Media (SFM) agar for 14 days at room temperature. This agar recipe has been found by many researchers to reliably induce sporulation in many *Streptomyces* species – although the reason for this is not known (Kieser et al., 2000). **Figure 5.2c** shows the appearance of KB16 on SFM after 14 days growth. The appearance of the

colonies was very different to its appearance on Nutrient Agar. Colonies appear circular and with a raised elevation and curled margins. The colonies have a “fuzzy”-like texture rather than the bald appearance on Nutrient Agar, with some of the larger colonies forming a dark grey pigment. The earthy brown secretion was also still present. This appearance on SFM is characteristic of *Streptomyces* grown on this media and so taken together this data gives a strong indication that KB16 is of that genus. However, the evidence is not conclusive as other filamentous microbes (such as fungi) could behave in similar ways. Therefore, these initial observations were followed up with molecular 16S rRNA gene typing.

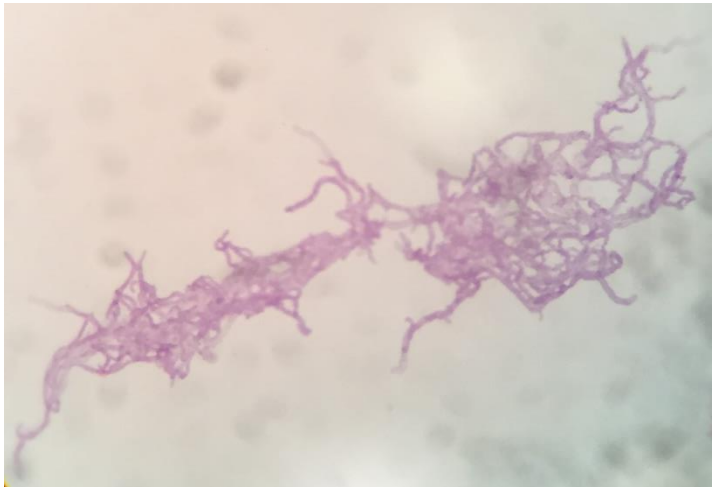


Figure 5.2a) Image of Gram-Stain of KB16 showing hyphal morphology.

Figure 5.2b) Image of KB16 in TBS after growth for 10 days at Room Temperature showing clusters of mycelia

Figure 5.2c) Image of KB16 incubated on SFM agar for 14 Days at Room Temperature showing evidence of sporulation and secretion of brown pigment

Figure 5.2. Images of morphology of KB16 in different conditions. The images all show that KB16 adopts a morphology characteristic of *Streptomyces* in each condition.

5.3.2.2 16S rRNA Gene Typing of KB16

Figure 5.3 shows the results of the 16S rRNA gene typing of KB16. A BLASTn search of its partial 16S rRNA gene sequence amplified by PCR against the NCBI 16S reference database confirms that the isolate is of the genus *Streptomyces*. The 16S rRNA sequence of KB16 has the strongest matches (99.9%) with sequences from species *Streptomyces canus*, *Streptomyces ciscaucasicus*, and *Streptomyces clavifer*.

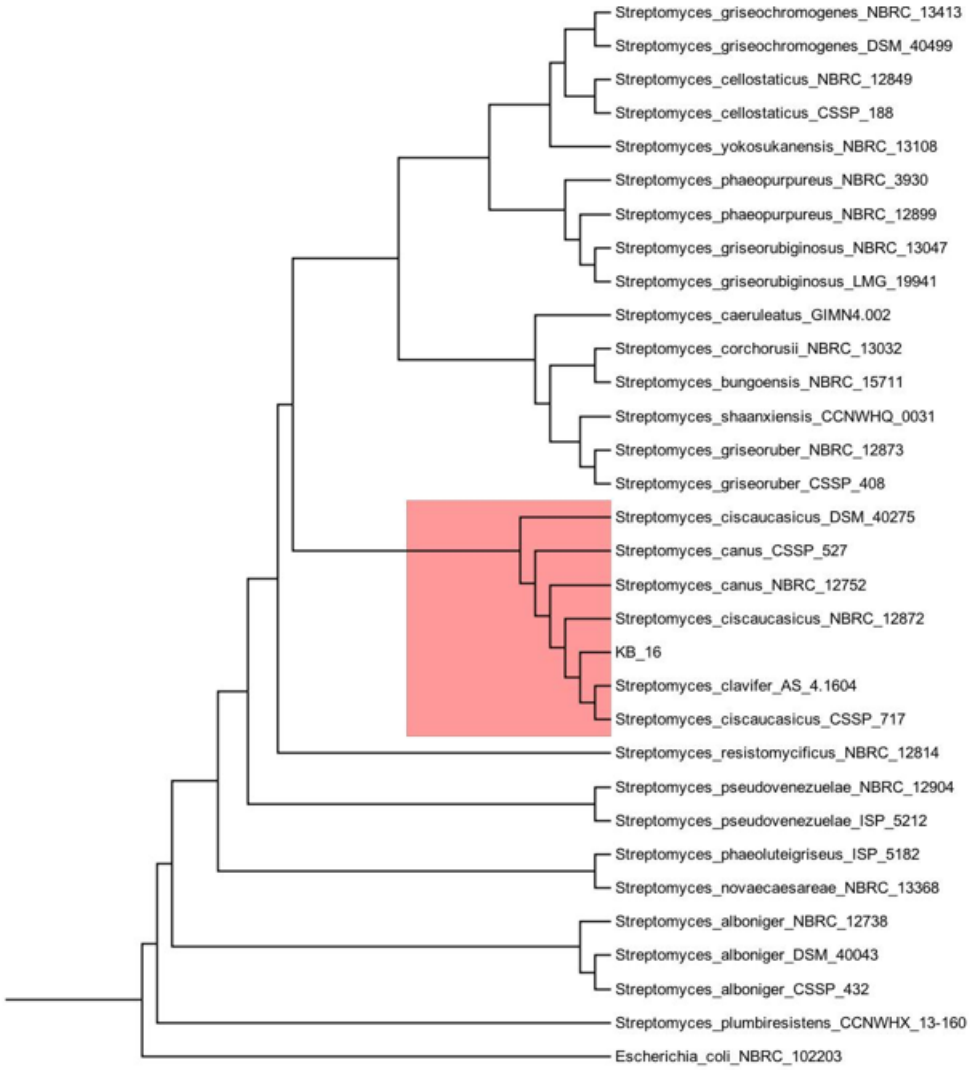


Figure 5.3. Neighbour-joining phylogenetic tree of KB16 partial 16S rRNA gene sequence and the top 30 BLASTn similarity matches from the NCBI 16S reference dataset. *E. coli* NBRC 102203 is used as an outgroup. The clade that KB16 sits within is highlighted in red.

A search of the literature reveals that Streptomyces *ciscaucasicus* is now considered a synomium of *Streptomyces canus* (Kämpfer et al., 1983). Additionally, *Streptomyces clavifer*, which was first published by Millard and Burr in 1926, has few records in the literature and is not presently a valid taxonomic name under the rules of the International Code of Nomenclature of Bacteria (Skerman, McGowan & Sneath (ed.), 1989; Tindall, Kämpfer & Euzéby, 2006). Therefore, the results of this assay suggest that KB16 is related to *Streptomyces canus* and is a member of this phylogenetic clade.

Antibiotic compounds that have been reported to be isolated from strains of *Streptomyces canus* are Telomycin (Fu et al., 2015; Liu, Li & Magarvey, 2016) and Amphomycin (Yang et al., 2004). The telomycins are a family of cyclic depsipeptides with antibiotic active against MRSA. Six analogues of this molecule have been isolated from various strains of *Streptomyces canus*. The structure of the family is defined by the presence of 11 amino acids including five non-proteinogenic ones, and a nonapeptide lactone ring. The telomycin biosynthetic gene cluster is comprised of three NRPS genes which encode 11 amino acid substrate modules. The amphomycins are a family of modified NRPS-PKS hybrid lipopetides that also contain two analogues named aspartocin D and aspartocin E. Amphomycins have been demonstrated to inhibit Gram-positive bacteria such as *S. aureus* and *B. subtilis* but appear to be ineffective against Gram-negative bacteria.

A high confidence identification of the taxonomy of KB16 at a species level cannot be determined using 16S rRNA data alone. This is because, as was previously discussed in **Chapter 3**, the relative level of molecular conservation of the 16S rRNA limits its resolution below genus level. This limitation is further compounded by the fact that most prokaryotes will have multiple copies of the 16S rRNA gene within their chromosomes and a PCR-type screen method will produce a consensus. Additionally, these genes can be transferred between species by horizontal gene transfer which can distort taxonomic classification (Sentausa & Fournier, 2013).

Other researchers have demonstrated the effectiveness of using multiple housekeeping genes to provide robust species level resolution of *Streptomyces* species. For example, Guo and co-workers (2008) performed PCR sequencing of six partial housekeeping genes (*atpD*, *gyrB*, *recA*, *rpoB*, *trpB* and 16S rRNA) across 45 species and subspecies within the *Streptomyces griseus* genera. The authors created phylogenetic trees of the concatenated gene sequences and concluded that the data provided for a robust method of species level resolution within the *Streptomyces griseus* genera. In addition to this, it has been shown by other researchers that there can be a wide variation in secondary metabolite structures produced within

Streptomyces species clades, and therefore single locus phylogenetic typing is an unreliable method of dereplication. For example, Seipke (2015) used genome-mining and genome-wide phytogenic approaches to investigate the diversity of secondary metabolite profiles within the *Streptomyces albus* species clade. The author analysed six genomes of *Streptomyces albus* available at the time on the NCBI database. The authors confirmed the genomes to belong to the *Streptomyces albus* clade and determining their phylogenetic relationships through multilocus typing of the *aptD, gyrB, recA, rpoB,* and *trpB* gene fragments similar to Guo and co-workers (2018). After genome-mining analysis the author identified 48 BGCs. Amongst these 48, 18 were present in all strains and were determined to be "core" to the species. A further 14 were encoded in one or more strain and were labelled "auxiliary". The remaining 16 BGC were all specific to a single strain and many of these BGCs are as yet undefined in laboratory settings. Analysis of the unique BGCs showed that *Streptomyces albus* strains with more distant phylogenetics appeared to harbour more unique BGCs. If 16S rRNA gene typing was used alone for dereplication then such rare and novel BGCs would be missed.

To determine a species level resolution of the taxonomy of KB16, multilocus typing was performed. Now that the ability to sequence and assemble a high-quality and continuous bacterial genome using Oxford Nanopore MinION technology was an option **Chapter 4**, it was decided that whole-genome phylogenetic analysis should be performed rather than the PCR-based screen as performed by Guo and co-workers and also in **Chapter 2** with the *Bacillus* isolates from honey. This would give the advantage of allowing for higher taxonomic resolution through the analysis of potentially hundreds of orthologous genes, as well as other taxonomic metrics to further confirm KB16 taxonomy such as its average nuclei identity (ANI). It would also have the benefit of providing the dataset required for prediction of natural product BCG encoded within the genome for more detailed dereplication as demonstrated in **Chapter 4**.

5.3.3. KB16 DNA Extraction

The pipeline for genome sequencing and assembly using both Oxford Nanopore and Illumina data developed in **Chapter 4** was utilised for the study of KB16. The genomic DNA (gDNA) extraction from KB16 was first analysed prior to genomic sequencing to ensure that it was a suitable quality and purity. The results of the gel electrophoresis **Figure 5.4** confirmed that the DNA extract was of high molecular weight (HMW) due to the presence of a tight band concentrated at the region of the 23 kbp band of ladder marker. Nanodrop analysis **Table 5.1** confirmed that the concentration of the extract and its purity was within the recommended ranges for both Illumina and Oxford Nanopore library preparation and sequencing.

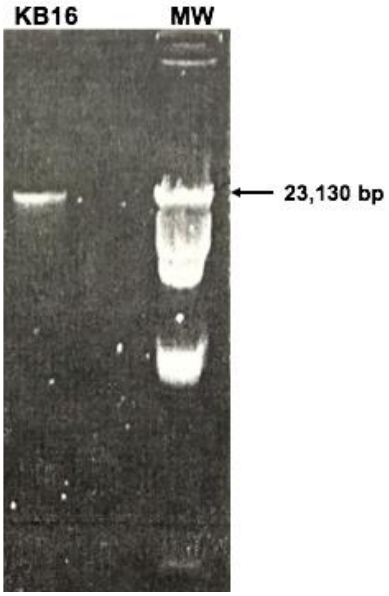


Figure 5.4. Agarose gel (1 %) of 1 µL of 1:10 dilution of KB16 genomic DNA extract (~ 40 ng). MW = λ DNA-*Hin*dIII digest molecular size marker. The DNA extract is seen forming a band ~23 kbp with no smearing beneath, suggestive of HMW DNA suitable for Nanopore sequencing.

Table 5.1. Nanodrop analysis of KB16 genomic DNA extract. Results show that both the concentration and absorbance ratios are within values sufficient for Nanopore sequencing.

Concentration (ng/µL)	Absorbance ratios	
	260/280	260/230
425.4	1.90	2.10

5.3.4. Whole-Genome Sequencing of KB16

The genome of KB16 was sequenced on Oxford Nanopore MinION and data generation was monitored during the sequencing run to determine by which point sufficient data has been generated (**Figure 5.5**). The intention was to generate enough reads of sufficient length (≥ 1000 bp) and quality (≥ q10) to assemble the KB16 genome with at least a 30× coverage depth as this is advised as the minimum input for assembly using Canu (Koren et al., 2017).

The molecular typing using the 16S rRNA gene had given an indication that KB16 may be related to *Streptomyces canus*, and a search of the literature had shown that genomes of this species averaged in the ~10 Mbp region. Therefore, this figure was taken as an assumed genome size for KB16 to use when benchmarking data generation. Having made an initial typing of the isolate with the partial 16S rRNA gene rather than sequence the entire genome immediately therefore offers benefits in allowing researchers to ensure sufficient coverage depth is obtained. After 22 hours, the sequencing run had generated 1,289,806 reads with an average length of 4,168.3 bp, and total bp yield was 1.7 Gbp - amounting to an assumed coverage depth of ~170x. This is deemed sufficient to produce a quality genome assembly of at least 30× and so the run was terminated. The reads were filtered to a minimum size of 1000 bp and quality score of q10 resulting in 141,464 reads with an average length of 3,943.0 bp and totalling 5.6 Mbp – an estimated coverage depth of ~56x. This demonstrates that there is a level of redundancy in the data generated by the MinION when it is filtered for size and quality. These filtered reads were used as input for genome assembly.

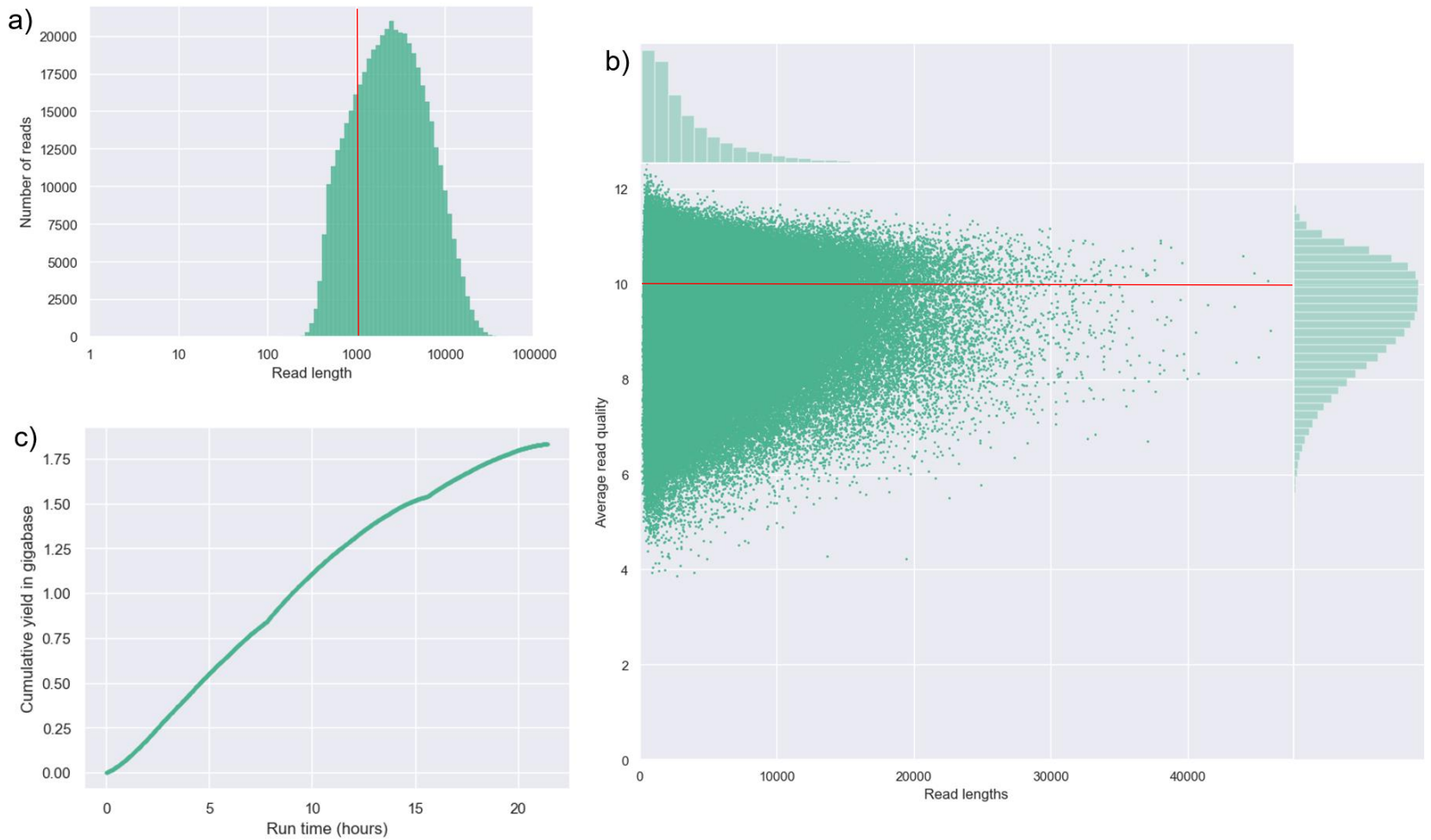


Figure 5.5. NanoPlot graphs summarising the data generation of the MinION sequencing of KB16. Figure 4a) histogram of reads by length. The filter point of (≥ 1000 bp) is indicated by red line. Figure 4b) Comparison of average read quality and read length. The filter point of (≥ q10) is indicated by red line. Figure 4c) cumulative generation of data in Gbp during the course of the run.

Table 5.2 shows the QC metrics that were used during assembly of the KB16 genome as per the pipeline developed in **Chapter 4**. As previously discussed, BUSCO was used to identify expected orthologues in each draft assembly file after each iteration of polishing. The presence of orthologues expected to be found within Actinobacteria is a useful tool to assess the "completeness" of the genome assembly and hence infer the likely accuracy of the assembly and reliability for analysis based upon it. As **Table 5.2** shows, each iteration of polishing with Nanopore reads using RACON improved the identification of complete orthologues for the first four iterations before showing a reduction after the fifth. The extent of improvement between each RACON polishing step appears to be greatest in the first iteration before progressively reducing as expected.

A subsequent polishing step using Nanopolish followed by polishing with Illumina reads using Pilon further improved orthologue identification with the final value being 99.7 %. These findings are in accord with those of observed using this pipeline for the assembly of *Streptomyces coelicolor* (**Chapter 4**) and give confidence in this pipeline producing a high-quality assembly of the KB16 genome for high resolution phylogenetic and biosynthetic gene cluster analysis. The finding also validated the pipeline developed in **Chapter 4** to produce high-quality contiguous genomes of environmental isolates for bioprospecting.

Table 5.2. Results of BUSCO and PROKKA Analysis of Nanopore-only Assembly Files. Completeness scores are improved with multiple polishing steps.

File	BUSCO Genome Completeness (%)		
	Complete	Fragmented	Missing
Nanopore-only assembly	22.7	42.3	35.0
Nanopore+Racon 1	35.5	40.6	23.9
Nanopore+Racon 2	39.8	38.6	21.6
Nanopore+Racon 3	40.6	37.8	21.6
Nanopore+Racon 4	42.6	36.4	21.0
Nanopore+Racon 5	39.8	41.2	19.0
Nanopore+Racon 4+Nanopolish	80.7	13.4	5.9
Nanopore+Racon 4+Nanopolish+Pilon 1	99.7	0.0	0.3

The final complete genome assembly of KB16 (**Table 5.3**) is comprised of a single linear chromosome of 10,764,324 bp in length, 9721 putative gene sequences, and a GC content of 70.22 %.

Table 5.3. Summary of characteristics of complete KB16 genome assembly.

Topology	Linear
Size	10,764,324 bp
GC content	70.22 %
Coding gene sequences	9721
16S rRNA genes	6
tRNA genes	86

5.3.5. Genome-Wide Phylogenetic Typing of KB16

Genome-wide phylogenetic analysis of the completed KB16 genome assembly was then performed. A multiple alignment of 400 orthologous genes between KB16 and 102 *Streptomyces* genomes from the RefSeq database was performed (**Figure 5.6**). The results showed that KB16 was within the same clade as the RefSeq genome assembly of *Streptomyces canus* DSM 40275.

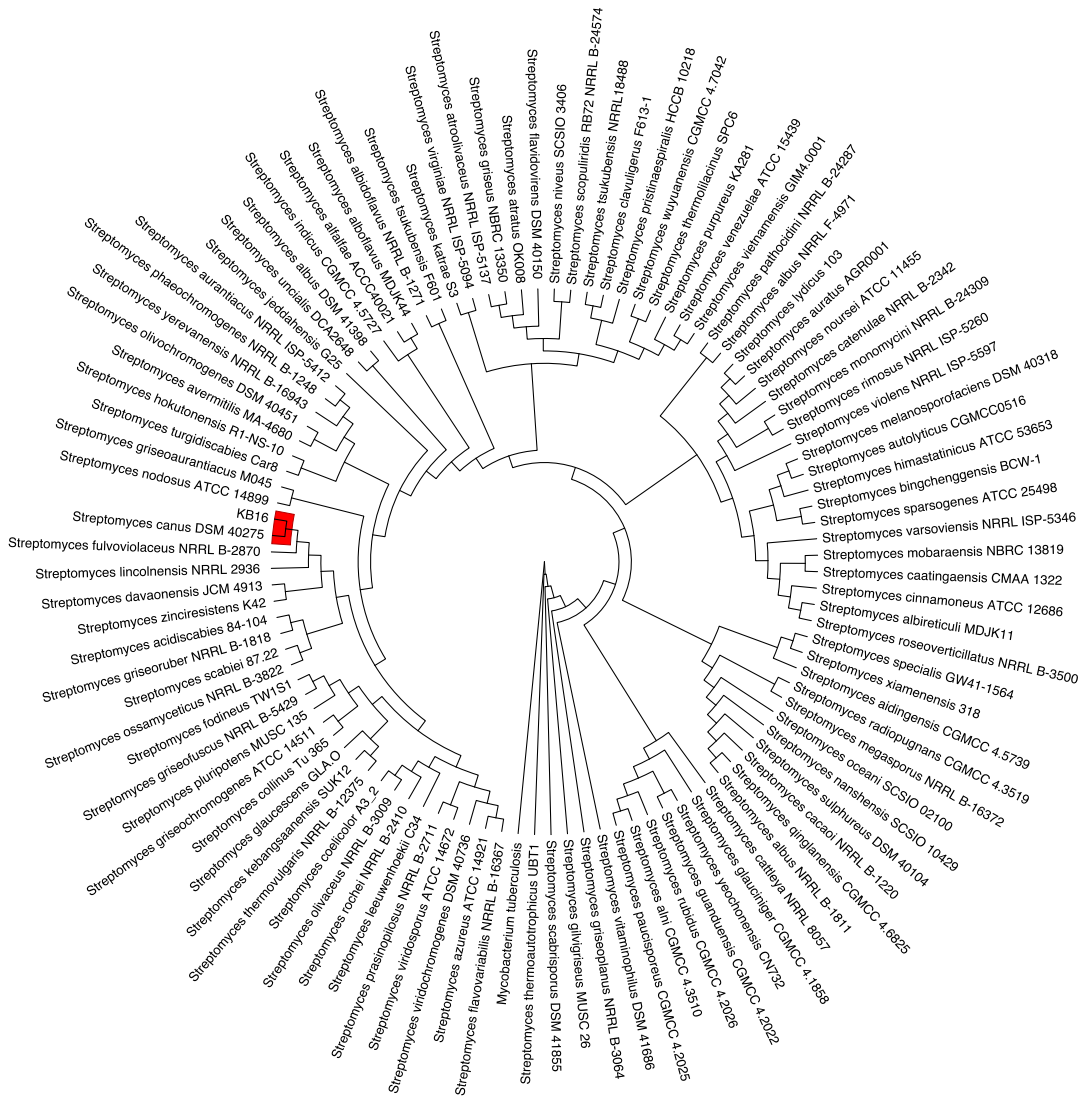


Figure 5.6. Cladogram of multiple alignments of 400 orthologs of KB16 genome against 102 RefSeq *Streptomyces* genomes. KB16 forms a clade with *Streptomyces canus* DSM 40275 which is highlighted in red.

Because of the large number of housekeeping genes that have been used in this analysis and because the analysis has included a large representative spread of the genus this phylogenetic tree is of a high resolution. Therefore, these findings add high confidence that KB16 is highly related to *Streptomyces canus* species and is also likely to be a member of this species group. This analysis was followed up by comparing the average nucleotide identity (ANI) between

the KB16 genome assembly and the RefSeq genome assembly of *Streptomyces canus* DSM 40275. This analysis showed a 97.4 % identity (1.95 % SD) between the two genomes. The Average Nucleotide Identity (ANI) is a pair-wise measurement of nucleotide similarity between the orthologs genes of two genomes. It is analogous to the traditional molecular biology approach of DNA-DNA hybridisation (DDH) as a means of species determination where a ~70 % DDH score was deemed to represent two strains of the same species (Goris et al., 2007). It has been demonstrated by researchers that have performed comparative studies that ANI scores >95 % is a reliable determination of intra-species relationship (Jain et al., 2018). These *in silico* findings further support the 16S rRNA and whole genome typing findings which suggest that KB16 is likely to be a member of the *Streptomyces canus* species. Formal determination of the species of a bacterial isolate also requires a full assessment of its morphology and metabolism using wet-lab techniques. These are time and labour-intensive assays which are beyond the scope of this bioprospecting investigation.

5.3.6. Genome-Mining of KB16 for Putative Natural Product BGCs

After the whole-genome phylogenetic analysis of KB16 was completed, the annotated genome was analysed for putative natural product BGCs using AntiSMASH. This was performed for purposes of dereplication of known antibiotic natural products and also to determine the likelihood that KB16 could be producing novel natural products. Analysis predicted 26 putative natural product biosynthetic gene clusters. Some of the clusters predicted have genetic homology with known gene clusters in the MIBiG database. **Table 5.4** summaries the output of the analysis of each putative cluster. Where homologue gene clusters have been identified, known functions for those compounds have been provided. None of these BGCs were predicted to be telomycin and amphormycin – the two antibiotic natural products reported to be produced by strains of *Streptomyces canus* (Yang et al., 2004; Fu et al., 2015; Liu, Li & Magarvey, 2016).

The lack of evidence for telomycin and amphormycin BGCs in KB16 demonstrates that reliance upon taxonomic classifications alone may not be a reliable dereplication strategy when bioprospecting microbes. This also validates the value of using a genome-mining approach to analyse the potential BGCs encoded in the genome. The finding is also in accord with the findings of other researchers, such as Spieke (2015), who had reported that there was not a strong correlation between the phylogenetic and secondary metabolite profiles of *Streptomyces albus* species. This observation is also in accord with recent work by Belknapp and co-workers (2020) who conducted both a genome-mining and phylogenetic analysis of 1,110 publicly available *Streptomyces* genomes and reported that BGC profiles can be highly

variable between strains of the same species. This finding is also in accord with the findings detailed in **Chapter 3**, whereby evidence of the BGC for AS-48 was seen in four *Bacillus* isolates.

Each cluster was inspected manually to validate analysis. In some instances, low homology matches appeared to be artefacts whereby the matches were to ubiquitous accessory proteins (such as transport proteins) as opposed to the core biosynthetic proteins that define the scaffold of the natural product - these homology matches were discounted. After manual analysis of the data, only 11 of the predicted biosynthetic gene clusters have a significant homology with known natural product biosynthetic gene clusters and are highlighted in grey in **Table 5.4**. These 11 BGCs were also analysed further for the purposes of dereplication and also to assess their likelihood to be novel antibiotics.

Table 5.4. Summary of the putative natural product biosynthetic gene clusters predicted in KB16. Homologous gene cluster matches >20 % or appeared to be an artefact upon manual inspection were discounted. Gene clusters with significant homology to a known BGC are highlighted in grey.

Cluster	Cluster type	Homologous cluster	Similarity (%)	Known function of homologous cluster product	References
1	t1pks	-			
2	terpene	Melanin	42	Protection against range of environmental stress factors.	Guo et al., 2014; El-Naggar & El-Ewasy, 2017; Martínez, Martinez & Gosset, 2019
3	terpene	-			
4	terpene	-			
5	other	-			
6	ectoine	Ectoine	100	Osmoprotectant	Reshetnikov et al., (2011)
7	t2pks	-			
8	other	-			
9	melanin	Melanin	100	Protection against range of environmental stress factors.	Guo et al., 2014; El-Naggar & El-Ewasy, 2017; Martínez, Martinez & Gosset, 2019
10	siderophore	Desferrioxamine B	100	Facilitate iron uptake from environment. Used as drug to treat iron toxicity.	Barona-Gómez et al., 2004; Barona-Gómez et al., 2006
11	t2pks	Spore pigment	83	Match to annotated gene cluster for spore pigment in *Streptomyces avermitilis.*	Omura et al., 2001
12	butyrolactone-nrps	Kutznerides	34	Antibacterial and antifungal agents	Broberg, Menkis & Vasiliauskas, 2006; Fujimori, et al., 2007, Jiang, et al., 2011, Zolova & Garneau-Tsodikova, 2014).
13	terpene	Albaflavenone	100	Sesquiterpene with antibiotic activity against Bacillus subtilis.	Gurtler et al., 1994
14	siderophore	-			
15	bacteriocin-t1pks	-			

16	terpene-butyrolactone	Gamma-butyrolactone	100	Signalling molecule involved in regulating secondary metabolism and morphological differentiation in response to environmental conditions.	Takano et al., 2006; Kong et al., 2019
17	other	-			
18	lantipeptide	-			
19	siderophore	-			
20	terpene	Hopene/Hopanoid	92	Modified circular triterpenes, hypothesised roles in protection against environmental stress and membrane fluidity.	Rohmer et al., 1984; Poralla, Muth & Härtner, 2000; Sáenz et al, 2005; Seipke & Loria, 2009; Schmerk, Bernards & Valvano, 2011
21	lantipeptide	Informatipeptin	100	Class III lantipeptide, function unknown.	Mohimani et al., 2014
22	terpene	-			
23	terpene	Isorenieratene	100	Carotenoid pigment. Hypothesised photo-induced anti-oxidant in Streptomyces.	(Takano, 2015)
24	t1pks-nrps	-			
25	bacteriocin	-			
26	terpene	-			

In contrast to the *Streptomyces coelicolor* genome annotated in **Chapter 4**, the KB16 genome shows fewer putative gene clusters for large natural products but instead has a higher proportion of putative clusters for small organic metabolites and small peptides. For example, ten of the putative clusters are identified as coding for terpene like compounds, whilst there are three putative clusters for bacteriocins. There are also 3 clusters for putative siderophores. Manual inspection of clusters identified to encode potentially large modular metabolites (Type 1 Polyketides or Non-Ribosomal Peptides) was performed. The Type 1 PKS enzymes predicted in clusters 1 and 15 both contain a single module with only two substrate domains. Cluster 24 is predicted to encode a Type 1 PK-NRP hybrid structure. Manual inspection of this pathway identified seven putative substrate domains (2 PKS and 5 NRP). **Figure 5.7** is the chemical structure scaffold that the PKS-NRPS hybrid pathway is predicted to synthesise. Two of the amino acid substrate domains were predicted as serine and threonine but the other three had no known matches. Manual inspection showed no evidence of truncated annotations which may frustrate substrate predictions, so these substrate domains may be novel. There are over 500 NRPS substrates currently known (Baranašić et al., 2014), and it is plausible that are still substrates yet to be characterised. There is no strong homology matches for this cluster in the MiBiG database, which is suggestive that this biosynthetic gene cluster may be novel.

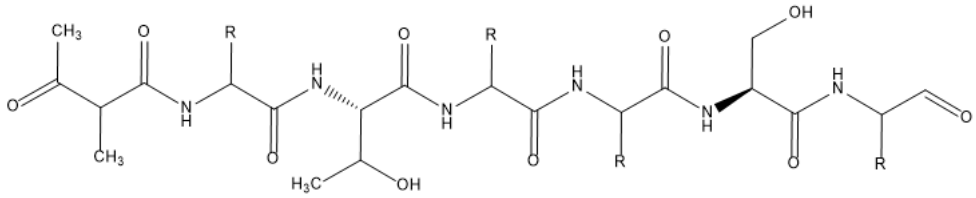


Figure 5.7. Predicted chemical scaffold structure produced by putative KB16 biosynthesis gene cluster 24. Some substrate residues could not be predicted. Structure drawn in ChemBioDraw Professional version 16.

The gene cluster predicted in cluster 12 contains seven substrate domains across three putative NRPS genes (BMLNJFMJ_06296, BMLNJFMJ_06297, BMLNJFMJ_06304) (**Figure 5.8**). Analysis also identified stand-alone termination (TE), reductase (KR), and carrier (PCP) which may act *in trans*. The putative gene cluster was flagged has having 34 % gene homology with the kutznerides biosynthetic gene cluster. Kutznerides are a family of cyclic hexadepsipeptides comprising non-proteinogenic amino acids with post-translational chlorination, hydroxylation and methylation (Fujimori, et al., 2007, Jiang, et al., 2011, Zolova & Garneau-Tsodikova, 2014). These compounds were first isolated by bioactivity guided isolation from an Actionmycetes in the genus *Kutzneria* that was cultivated from the roots of immature spruce trees (Broberg, Menkis & Vasiliauskas, 2006). The compounds are shown to exhibit antibiotic and antifungal activity, and therefore may have a role in maintaining the

rhizobia and mycorrhiza of these species. Fujimori and co-workers (2007) identified the biosynthetic pathway of the kutznerides to be comprised of three NRPS genes (*ktzE*, *ktzG*, *ktzH*) which contain a total of six substrate domains. The authors also identified two standalone adenlyation (A) domains (KtzB, KtzN) and a standalone termination domain (KtzC).

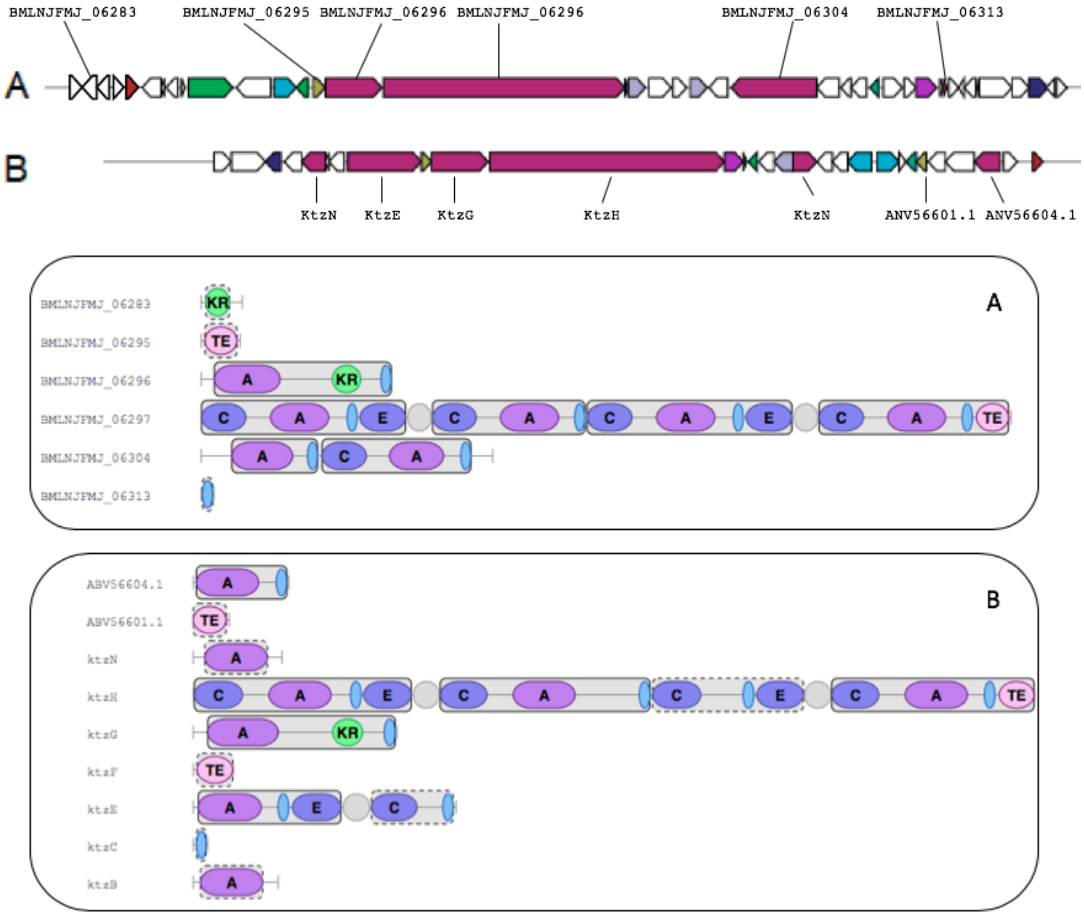


Figure 5.8. Comparison of KB16 cluster 12 (A) with kutznerides biosynthetic gene cluster (B). Top image shows a pairwise comparison of both pathways. Genes of the same colour show homology. Bottom images show catalytic domains of NRPS related genes.

Manual inspection of the homology between cluster 12 and kutznerides showed that two NRPS genes in cluster 12, BMLNJFMJ_06296 and BMLNJFMJ_06297, were homologous with *ktzH* and *ktzG* with 58 % and 61 % homology respectively. There was, however, no identified homology between the third NRPS gene BMLNJFMJ_06304 and *ktzE*. There are also structural differences between the NRPS genes of both clusters in terms of the arrangements of their catalytic domains. Further analysis using Multigene blast showed that there were also many putative halogenase and oxygenase enzymes identified in cluster 12 with reported homology with kutzneride tailoring enzymes. These findings are suggestive that

cluster 12 may encode novel compounds which are partially analogous to the kutznerides and could be a potential cause for the antimicrobial activity exhibited by KB16.

Some of the predicted biosynthetic gene clusters appear in accord with the environmental conditions that KB16 was isolated from and may give insights into the metabolism of the organisms. KB16 was isolated from the Roman Baths, UK which is a thermal spring with a high mineral content. More detail on the geochemistry of the site is provided in **Chapter 6**. Some of these putative metabolites may also have commercial interest. For example, cluster 6 has a 100 % homology match to the ectoine biosynthesis gene cluster. Ectoine is a known osmoprotectant which serves to protect cells from osmotic stress in high saline environments and has been identified in various methylotrophic and extremophilic genera of bacteria (Reshetnikov et al., 2011; Czech et al., 2018). Compounds of this type are commonly found in microorganisms adapted to survive in such environments (Czech et al., 2018) whilst other microorganisms have evolved to function with high accumulated levels of inorganic salts (Gunde-Cimerman, Plemenitaš, & Oren, 2018).

Whilst the synthesis of osmoprotectants such as ectoine is a more energetically expensive strategy than the former, it also gives greater flexibility to fluctuating salinity as osmoprotectant metabolism can be regulated in response to environmental conditions (Gunde-Cimerman, Plemenitaš, & Oren, 2018). This may partially explain the ability of KB16 to be cultivated in laboratory conditions that differ to the geochemistry of the site that it was isolated from. Ectoine has also been shown to serve as a molecular chaperone and has attracted commercial interest from the cosmetic and pharmaceutical industries due to its ability to stabilise biomolecules and whole cells against various stresses such as extreme temperature shifts, and UV radiation (Reshetnikov et al., 2011; Czech et al., 2018). This finding further supports the suggestion made in **Chapter 6** that the microbiome of this site could serve as a site for the bioprospecting of commercial biocatalysts.

Two of the biosynthetic pathways (clusters 2 and 9) were identified as associated with melanin production. Melanins are a diverse family of compounds that are produced by the oxidation of aromatics. They are often characterised by a dark brown or black colour and have been shown to protect microbes against various environmental stresses such as high temperatures, heavy metal oxidation, and UV radiation. Observations of KB16 culture on solid media showed a dark brown pigment diffused into the agar after 14 days of growth (**Figure 5.1 & Figure 5.2c**) which is most likely a melanin pigment. Melanins are of commercial value for their radiation absorbing properties and there is research focus on identification and

optimisation of high melanin production strains of *Streptomyces* (Guo et al., 2014; El-Naggar & El-Ewasy, 2017; Martínez, Martinez & Gosset, 2019).

Some microbially produced melanins have also been demonstrated to have antimicrobial activity. Sivaperumal and co-workers (2014) reported on the purification of a melanin pigment from a species *Streptomyces* that was isolated from marine sediments of the Versova coast in India. The authors claimed that the compound showed antimicrobial activity against *Pseudomonas aeruginosa* RMMH7 (MIC of 10 ± 0.02 μg/mL) and *Vibrio parahaemolytics* RMMH12 (MIC of 14 ± 0.02 μg/mL). Some melanin compounds may therefore have potential as antibiotic lead compounds.

Three gene clusters (10, 14, 19) are identified as encoding for siderophore compounds. With cluster 10 having 100 % homology with the siderophore desferrioxamine B. Desferrioxamines are a family of siderophores that are defined by their hydroxamate structures. First isolated from *Streptomyces pilosus* in 1950, members have also been identified in other *Streptomyces* species including *S. coelicolor* (Barona-Gómez et al., 2004; Barona-Gómez et al., 2006). Desferrioxamine B finds commercial use in medicine as a drug for the treatment of iron toxicity. Iron is an essential co-factor for many cellular metabolic functions and siderophores serve the function of facilitating iron uptake from the environment.

There some reports that claim siderophore families display antimicrobial activity. For example, van Asbeck and co-workers (1983) found that desferrioxamine at concentrations between 200–400 μg/mL plus ascorbic acid inhibit growth of 43 species of Gram-positive and Gram-negative bacteria including species in genera *Staphylococcus, Escherichia, Klebsiella, Proteus, Alcaligenes, Neisseria, Salmonella, Enterobacter, Pseudomonas* and *Providencia.* The authors of this study also noted that growth of these bacteria returned to normal once the medium was subsequently saturated with iron. They concluded from this that desferrioxamine demonstrate a bacteriostatic effect through depletion of iron available in the growth media. Thompson and co-workers (2012) tested the antimicrobial activity of desferrioxamine along with the siderophores deferiprone, Apo6619, and VK28 against ESKAPE pathogens using a similar methodology and reported growth inhibition by deferiprone, Apo6619, and VK28 but no effect by desferrioxamine. There are a further two putative biosynthetic gene clusters with predicted Pfam domains which suggest they encode siderophore molecules but have not been reliably matched to any known pathways. These suggests that KB16 could be producing novel siderophore compounds.

It may be possible that overproduction of siderophores may be contributing to the antimicrobial activity observed by KB16 in cross-streak assays (**Figures 5.1**). A simple experiment to test for this could be to repeat these assays using media supplemented with excess iron to determine if greater iron availability, and therefore saturation of the siderophores chelation effects, reduced the antimicrobial affects observed. However, the observed antimicrobial activities of siderophores, and the mode of action to explain this activity, is broad-spectrum against both Gram-positives and Gram-negatives. The antimicrobial activity observed by KB16 in the cross-streak assays performed showed it to be narrow-spectrum against only the Gram-positive bacteria tested. This suggests that the antimicrobial activity seen is not due to iron (or any other) essential nutrient depletion which would affect all organisms. The specific nature of the antimicrobial activity observed also suggests that it is being caused by a compound secreted by KB16 that the additional double membrane of the Gram-negative bacteria is preventing uptake.

Some of the putative gene clusters in KB16 encode hopanoid terpenes. These are a broad family of modified triterpenes derived from a precursor triterpene hopene. Biosynthesis of hopene involves the circularisation of linear triterpene squalene by the enzyme squalene-hopene cyclase (Hoshino & Sato, 2002). Cluster 20 contains an ORF with a putative conserved domain identifying it as such an enzyme. This class of compounds is found in a broad range of plant and prokaryotes species including *Streptomyces* (Rohmer et al., 1984). Hopanoids are hypothesised to help regulate a broad range of bacterial cellular functions, such as membrane fluidity and active transport (Sáenz et al, 2005), morphological differentiation (Poralla, Muth & Härtner, 2000; Seipke & Loria, 2009), pH tolerance, and mobility (Schmerk, Bernards & Valvano, 2011).

Cluster 13 has a 100% homology match to the biosynthetic gene cluster for the terpene albaflavenone. This compound was first isolated and characterised from *Streptomyces albidoflavus* and demonstrated to have antibiotic activity *in-vitro* (Gurtler et al., 1994). It has subsequently been shown that analogues of this compound are produced by several species of the genus. For example, Zhao and co-workers (2007) isolated and characterised the biosynthetic pathway for albaflavenone in *Streptomyces coelicolor* A3(2). And this biosynthetic pathway was also been identified in the genome-mining analysis detailed in **Chapter 4**.

Moody and co-workers (2011) built upon this study by Zhao and co-workers (2007) to identify *in silico* potential orthologous of albaflavenone biosynthesis genes in 10 different species of *Streptomyces*. They were able to then subsequently identify albaflavenone in the gas

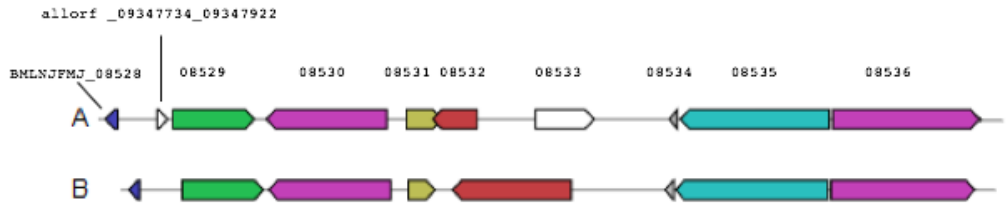
chromatography mass spectra of extracts in 5 of these species; *S. viridochromogenes, S. avermitilis, S. griseoflavus, S. ghanaensis,* and *S. albus*. The authors hypothesised that the albaflavenone biosynthesis pathway may play a conserved and essential function in *Streptomyces* metabolism. Genomic mining of *Streptomyces formicae* KY5 by Holmes and co-workers (2017) also identified a gene cluster homologous to albaflavenone. *S. formicase* KY5 is a taxonomically distinct species of Streptomyces isolated from leafcutter ants and is a noteworthy discovery as it has been shown to novel pentacyclic polyketides now known as formicamycins that have bioactivity against MRSA and VRE (Quin et al., 2017).

The albaflavenone gene cluster in KB16 could be responsible for the antibiotic activity seen. However, in other *Streptomyces* species where this compound has been identified, they also produce numerous other novel natural products which also have antibiotic activity. For example, *S. formicase* KY5 produces novel antibiotic polyketides while genomic analysis predicted up to 45 biosynthetic gene clusters which have not all yet been investigated (Holmes et al., 2017). Whilst secondary metabolites produced by *S. coeolicolor* with known antimicrobial activity in addition to albaflavenone include actinorhodin, prodigiosins, calcium-dependent antibiotic, and methylenomycin. Therefore, because of the fact that many of the other gene clusters do not have homology with known clusters it is possible that KB16 may also be encoding other, more novel, antibiotic compound structures. Further investigations into the antibiotic producing capabilities of KB16 would therefore first need to focus on a dereplication strategy to confirm if albaflavenone is the sole cause of the antibiotic activity observed.

Three putative gene clusters have been predicted to encode small ribosomal synthesized peptides (RiPPs). Cluster 25 was identified to encode a bacteriocin structural gene whilst cluster 18 was identified as containing a LAN-C like lantipeptide synthase gene. However, no structural gene was identified within this cluster. Cluster 21 was also predicted to encode a lantipeptide and was shown as having a 100 % homology match with the informatipeptin gene cluster – a class III lantipeptide identified *in silico* through the peptidogenomic assignment of tandem mass spectra of culture extracts *Streptomyces viridochromogenes* DSM 40736 against its annotated genome by Mohimani and co-workers (2014). There is no evidence in the literature of the purification of informatipeptin and the characterization of its biological activity.

Cluster 21 was manually inspected and compared to the annotated informatipeptin gene cluster (**Figure 5.9**). Cluster 21 annotation was found to contain all the essential genes that define a class III lantipeptide biosynthetic gene cluster. This includes a lanthionine synthase

(BMLNJFML_08535) with a LAN-C like domain lacking in zinc binding residues - BlastP analysis of this gene identified this putative domain with e-value of 2.38e-81. A transporter gene (BMLNJFML_08532) and a regulator (BMLNJFML_08531) show homology with similar genes in the informatipeptin gene cluster. AntiSMASH analysis identified two putative precursor genes for cluster 21, suggesting that this cluster has the potential to produce two lantipeptides. Comparison of the predicted modified sequences of the two lantipeptides against that of informatinpetin (**Figure 5.10**) showed differences in number and positions of modified residues, whilst Blast and Bactibase searches of both sequences returned no matches. These findings suggest that cluster 21 in KB16 may encode up to two potentially novel lantipeptides. This cluster should therefore be a candidate for further investigation as the cause of the antibiotic activity observed by KB16. The roles that lantipeptides have in bacterial metabolism are broad and there are many examples of lantipeptides with antibiotic activity as discussed in **Chapter 1**.



Gene	Predicted Function	Homology			
		Identity %	bit score	Coverage %	e-value
BMLNJFMJ_08528	Hypothetical peptide	76	100	100	6e-26
allorf _09347734_09347922	Lantipeptide precursor protein	-	-	-	-
BMLNJFMJ_08529	S1 family peptidase	84	614	78	2e-179
BMLNJFMJ_08530	SpoIIE family protein phosphatase	81	1129	100	0
BMLNJFMJ_08531	Response regulator transcription factor	80	247	77	2e-69
BMLNJFMJ_08532	ABC transporter ATP-binding protein	76	221	60	1e-61
BMLNJFMJ_08533	Hypothetical peptide	-	-	-	-
BMLNJFMJ_08534	Lantipeptide precursor protein	76	57	100	7e-13
BMLNJFMJ_08535	Lanthionine synthetase C-like protein	82	1343	100	0.0
BMLNJFMJ_08536	Hypothetical protein	73	1123	98	0.0

Figure 5.9. Comparison of putative biosynthetic gene cluster 21 from KB16 (A) against the annotated Informatipeptin biosynthesis gene cluster from the MIBiG database (B). Genes with the same colour are identified as having significant (=<70%) homology. The table below lists details of the homology of each gene in cluster 21 against the corresponding gene from the Informatipeptin biosynthesis gene cluster. The predicted function of each of the genes was determined from data from Blast analysis.

allorf _09347734_09347922
VVVAERLDRQAEVEDPRGEVV - GEDLVGFGGRHRHMGDCDhaRDhaRPARHGPWPNRRGNRGVKKRI

BMLNJFMJ_08534
MALLDLQTMESDEHTGGGGN – DhaDhbLDhaLLDhaCVDhaAADhaVDhbLCL

Informatipeptin
MTTDGHPMEGHTMALLDLQTIETEERTDGGGAS - DhbVDhaLLDhaCIDhaAADhaVLLCL

Legend
Dha = dehydroalanine
Dhb = dehydrobutyrine

Figure 5.10. Predicted primary structure of the two mature lantipeptides from the putative precursor genes (allorf_09347734_09347922 & BMLNJFMJ_08534). The leader peptide sequences are separated from the mature peptide sequence by a dash. The predicted leader peptide sequence is cleaved from the core peptide which is then dehydrated to contain dehydroalanine (Dha) and dehydrobutyrine (Dhb) residues which form cross-links with the cysteine residues. The predicted primary structure of Informatipeptin is given below for comparison.

In summary, the above genome-mining analysis of KB16 has revealed that the organism may have the potential to produce many natural products of a range of different chemical classes. Some of these putative gene clusters could be potential candidates for the antimicrobial activity seen by KB16. The most obvious example is cluster 13 which had a 100 % homology match with the albaflavenone biosynthetic gene cluster – a compound with known antimicrobial activity. The primary purpose of the genome-mining analysis was to gain a deeper insight into the biosynthetic potential of the isolate for the purposes of dereplication.

At this juncture it was considered if KB16 should be discounted from further investigations to avoid the possibility of isolating a known antibiotic compound such as albaflavenone. However, the albaflavenone BGC that has been identified in several *Streptomyces* genomes that also produce other antibiotic compounds, such as *S. coelicolor* A3(2) (**Chapter 4**). And the genome-mining analysis of KB16 has also predicted other putative BGCs which may be novel and responsible for the antimicrobial activity seen. Whilst the genome-mining analysis had given evidence that KB16 has the potential to produce albaflavenone, it does not attribute the antibiotic activity seen by KB16 to albaflavenone. For such as conclusion to be made, chemical analysis would be required.

Because albaflavenone is a well characterised compound; its structure and NMR peaks are known, as well as descriptions of its physical and chemical characteristics (Gurtler et al., 1994). This information could potentially allow for the presence of albaflavenone in bioactive

culture extracts taken from KB16 to be determined quickly by using resource-efficient fractionation approaches. Therefore, it was decided to use the observations made in the genome-mining analysis to employ a simple bioactivity-guided isolation strategy in an attempt to conclusively determine if albaflavenone was responsible for the antimicrobial activity of KB16.

5.3.7. Bioactivity Guided Isolation of Active Compound from KB16

The planned strategy for bioactivity-guided isolation taken was to create a crude fermentation extract of KB16 and fractionate this extract using RP-SPE. The active fraction would be identified by disk-diffusion assay and then the active component within the fraction would be identified by bioautography followed by purification by Prep-TLC. The purified compound would then be analysed by NMR to determine if it is albaflavenone by comparison of NMR spectra to the literature.

5.3.7.1. Generation of Crude Extract

Initially, KB16 was incubated in small volumes of nutrient broth and aliquots of this broth was tested for antimicrobial activity at different time points to determine if, and at what point, KB16 would produce antibiotic compounds in these conditions. The intention was to then upscale fermentation to produce a large extract. However, after 14 days of daily tests, no antimicrobial activity was seen (**data not shown**) and the experiment with nutrient broth was ended.

Other liquid media recipes or conditions could have been be tried to see if these produced the desired results. While it is preferable to find a liquid-based fermentation formula for ease of upscaling production, time considerations meant that the strategy was altered to focus on making culture extracts from solid media. The reason for focusing on solid media was because the cross-streak assay that demonstrated KB16's antibiotic activity (**Figure 5.1**) also demonstrates that the isolate secretes antibiotic compounds into the agar after 10 days incubation. For this reason, KB16 was cultivated on 100 Nutrient Agar plates and two solvent extractions (an ethyl acetate followed by a methanol extraction) of the agar were made and tested for antibiotic activity by disk diffusion assay. In addition to this an ethyl acetate solvent extraction of pure nutrient agar was also made as a control.

After each extract was dried by evaporation the masses were weighed. The ethyl acetate extractions of KB16 seed nutrient agar and pure nutrient agar was 165.4 mg and 11.0 mg

respectively – a difference of 154.4 mg. This suggests that the growth of KB16 had led to the increase of metabolite content in the agar. The methanol extraction of the KB16 was 13.6 mg – 151.8 mg less than the ethyl acetate extraction taken before. This suggests that a large proportion of the chemical compounds present were extracted by ethyl acetate. Because it was expected that any antibiotic compound produced by KB16 would be diffused into the agar and the agar is a water-based mixture – it was expected that any antibiotic metabolites inside the agar would have a relatively high polarity. Therefore, ethyl acetate was chosen as the first solvent system to use for the extraction as it is a solvent with a mid-range polarity would therefore target a broad range of metabolites in the agar. Ethyl acetate would also be likely to target the extraction of albaflavenone if it is present in the agar because it also has a mid-range polarity as a sesquiterpene and is a solvent which has been successfully used to extract albaflavenone from *Streptomyces* fermentations in other studies (Gurtler et al., 1994). Methanol, a solvent with high polarity, was then used as a second extraction system to extract any metabolites remaining in the agar with high polarity. It was not possible to resuspend each extraction to the same concentration; this was most likely due to the different extractions each containing different compounds at differing amounts which would cause differences in solubility. The final concentration of each resuspension is listed in **Table 5.5** below.

Table 5.5. Summary of KB16 agar solvent extractions. The dry mass of extracts A and B compared to C suggests that KB16 secreted over 150 mg of metabolites into the agar. The data also suggests that the ethyl acetate extracted more metabolites from the agar than methanol.

Extract	Description	Dry Mass (mg)	Resuspension Concentration (mg/mL)
A	KB16 seeded agar ethyl acetate	165.4	14.5
B	KB16 seeded agar methanol	13.6	4.5
C	Pure agar ethyl acetate	11.0	2.2

5.3.7.2. Antibiotic Activity of Crude Extract

Each extract was assayed by disk-diffusion assay to determine which, if any, contained the active antibiotic compound. **Figure 5.11.** shows the results of the disk diffusion assay of these solvent extractions. The extracts were tested for antibiotic activity against *Staphylococcus aureus* NCTC 12981. The reason that this microorganism was chosen as the indicator strain for this assay was because the cross-streak assay (**Figure 5.1**) previously performed had shown KB16 could inhibit the growth of this strain. Additionally, it is a characterised antibiotic-sensitive strain. The use of a sensitive strain as opposed to a resistant one is favourable for this assay because it would be more susceptible to lower concentrations of

antibiotic compounds, making it easier to observe if the extracts contained any antibiotics if they are at low concentration.

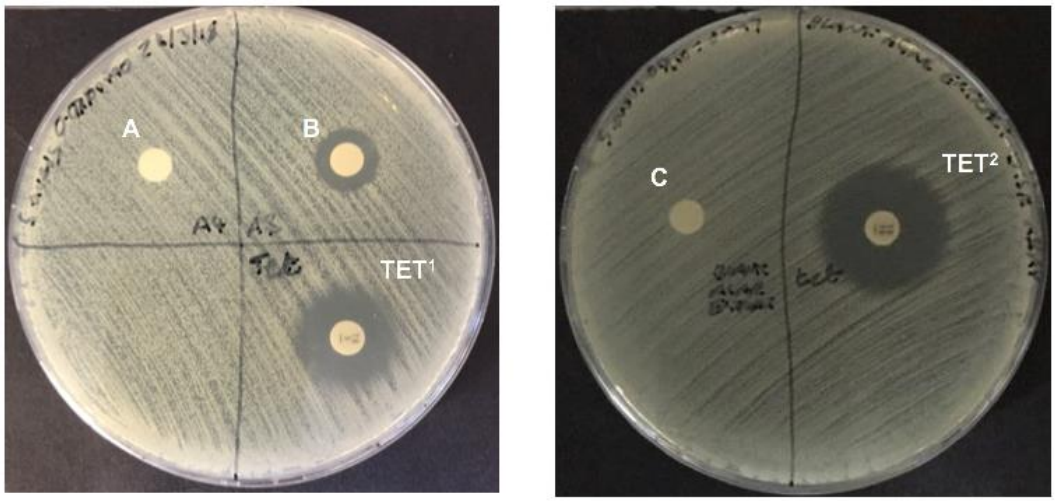


Figure 5.11. Results of disk diffusion assay of solvent extractions of Nutrient Agar seeded with KB16 and incubated for 14 days at room temperature. *Staphylococcus aureus* NCTC 12981 was used as the indicator strain. Disk 'A' was saturated with 20 µl of methanol solvent extraction of KB16 seeded agar. Disk 'B' was saturated with 20 µl of ethyl acetate solvent extraction of KB16 seeded agar. Disk 'C' is a control and was saturated with 20 µl of an ethyl acetate extraction of pure nutrient agar. Each plate also contained 5 µg tetracycline (TET^1 & TET^2) disks as controls. The zone of inhibition around disk 'B' demonstrates that this extract contains the antibiotic component.

The results of the disk diffusion assay (**Figure 5.11**) showed that there was a zone of inhibition around the disk containing the ethyl acetate extraction of the KB16 seeded agar (A), whilst there was no zones of inhibition around the disks containing the methanol extraction of KB16 seeded agar (B) or the ethyl acetate extraction of pure nutrient agar (C). There were also zones of inhibition around the tetracycline antibiotic disks. The presence of a zone of inhibition around disk A suggests that an antibiotic compound that was present in the KB16 seeded agar was extracted into the ethyl acetate solvent. The fact that no zone of inhibition was seen around the disk C suggests the compound that was causing the antibiotic effect was not something already present in the agar before seeding with KB16 and so is likely to be a compound produced by KB16. No zone of inhibition was observed in disk B, and this suggests that the methanol extract did not contain any antibiotic compound.

The tetracycline disks were used as a positive control to ensure that the indicator strain was showing the antibiotic sensitivity that was expected. The presence of zones around these disks was in accord with the expected sensitivity of this strain. This is an important control to use in a disk diffusion assay because of its qualitative nature which could lead to false negative results. For example, whilst the concentration of the indicator strain is controlled by adjusting

to a particular optical density range prior to inoculation onto the surface of the agar plate, there will still be some variations in the inoculation dosage because inoculation across the surface of the agar (either by a cotton swab or L-spreader) is not totally efficient with variances in pick up, dispensing, and coverage across replicates. Additionally, there may be slight variances in the composition of the nutrient agar (due to manufacturer batch differences for example) which may affect diffusion of compounds. The susceptibility of indicator strains could also in theory change with multiple passages in a laboratory through evolutionary selection pressures or cross-contamination. While all of these factors can and should be mitigated by good laboratory technique, the use of a tetracycline antibiotic disk as a control confirms consistency in the performance of the assay to ensure confidence in the results.

Based on the disk diffusion assay the ethyl acetate extraction of the KB16 seeded agar was taken forward for fractionation.

5.3.7.3. Fractionation of Crude Extract

The ethyl acetate extraction of the KB16 seeded agar was fractionated by reverse-phase solid phase extraction, and these fractions were tested for antibiotic activity by disk diffusion assay. **Table 5.6** and **Figure 5.12** summaries the fractions produced. Fifteen fractions were produced using elution solvents of differing polarities in order to separate the compounds in the mixture based on their polarities. Fraction 1 used an elution system with the highest polarity (100 % H_2O), with each subsequent fraction having a progressively lower polarity. As experienced before with the initial solvent extractions, resuspension of each fraction required the use of differing volumes and types of solvents due to the fractions differing polarities, which resulted in each resuspension having a different concentration. Visual inspection of each of these (**Figure 5.12**) showed that there was clearly a different in the content of each fraction through the fact that each fraction had different colour appearances. Each fraction was then tested by disk-diffusion assay to determine which, if any, retained the active component.

Table 5.6. Summary of each fraction produced by RP-SPE of the KB16 seeded agar ethyl acetate extraction. Dry mass of fractions eluted with relatively polar solvents is broadly greater than that of fractions eluted with relatively less polar solvents.

Fraction	Elution Solvent	Dry Mass (mg)	Resuspension Solvent	Resuspension Concentration (mg/mL)
1	100 % H_2O	24.0	2 mL MeOH	12.0
2	90 % H_2O : 10 % MeOH	33.4	1 mL MeOH	33.4
3	80 % H_2O : 20 % MeOH	22.4	1 mL MeOH	22.4
4	70 % H_2O : 30 % MeOH	15.1	1 mL MeOH	15.1
5	60 % H_2O : 40 % MeOH	6.3	1 mL MeOH	6.3
6	50 % H_2O : 50 % MeOH	17.4	1 mL MeOH	17.4
7	40 % H_2O : 40 % MeOH	4.5	1 mL MeOH	4.5
8	30 % H_2O : 30 % MeOH	3.5	1 mL MeOH	3.5
9	20 % H_2O : 20 % MeOH	2.9	1 mL MeOH	2.9
10	10 % H_2O : 90 % MeOH	5.3	1 mL MeOH	5.3
11	100 % MeOH	37.9	2 mL MeOH	19.0
12	100 % CH_3CN	0.6	1 mL $C_4H_8O_2$	0.6
13	100 % $C_4H_8O_2$	7.0	1 mL $C_4H_8O_2$	7.0
14	100 % C_3H_6O	2.4	1 mL $C_4H_8O_2$	2.4
15	100 % $CHCl_3$	3.2	1 mL $C_4H_8O_2$	3.2

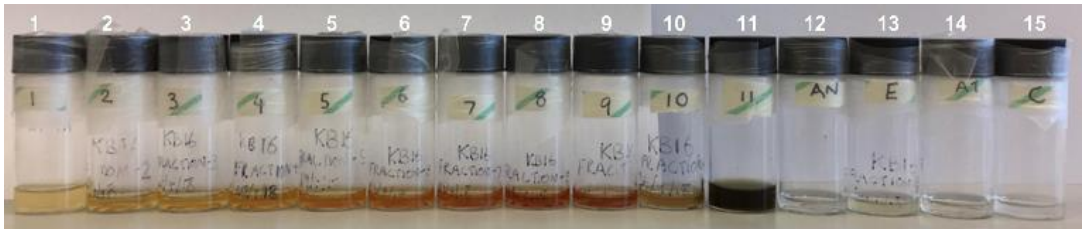


Figure 5.12. Each fraction after resuspension in solvent. Relative polarity of each fraction elute ranges from most polar (fraction 1) to least polar (fraction 15). The differences in colour across the fractions suggests that there may be a change in the content of the fractions.

5.3.7.4. Antibiotic Activity of Extract Fractions

The results of the disk diffusion assay (**Figure 5.13** and **Table 5.7**) showed that 11 of the 15 fractions had activity against *Staphylococcus aureus* NCTC 12981. That this many fractions would show activity was unexpected and could suggest that the crude extract had contained multiple active compounds of different polarities which have now been separated into different fractions. It is also possible that there was inefficient separation of active compound(s) which resulted in its elution across several fractions. This can occur either due to the active compound having variable polarity or due to suboptimal choice of elution solvents or solid phase matrix. There was observed differences in the appearance of some zones of inhibition (**Table 5.8**). Clear zones of inhibition contained an absence of *Staphylococcus aureus*, whilst intermediate zones contained a visible reduction growth. Zones also varied in their diameters. The variations of the size of zones around active fractions cannot be interpreted to give an indication of relative strength or mechanism of activity of each fraction. This is because each fraction is a complex mixture of unknown compounds of varying amounts which are also still unknown. Additionally, each fraction has differing total masses and solubilities. This means it is not possible to resuspend each fraction to ensure that concentrations of active compounds across each fraction are controlled. Therefore, each disk will inevitability receive a different amount of active compound(s). In addition to this, the relatively solubility of active compound(s) in the agar is unknown, and this can also influence the size of zone of inhibition they cause. This highlights a potential general limitation of this bioactivity-guided isolation strategy in that active compounds of low concentrations or with poor solubility in agar may not be detected and lead to false negative results.

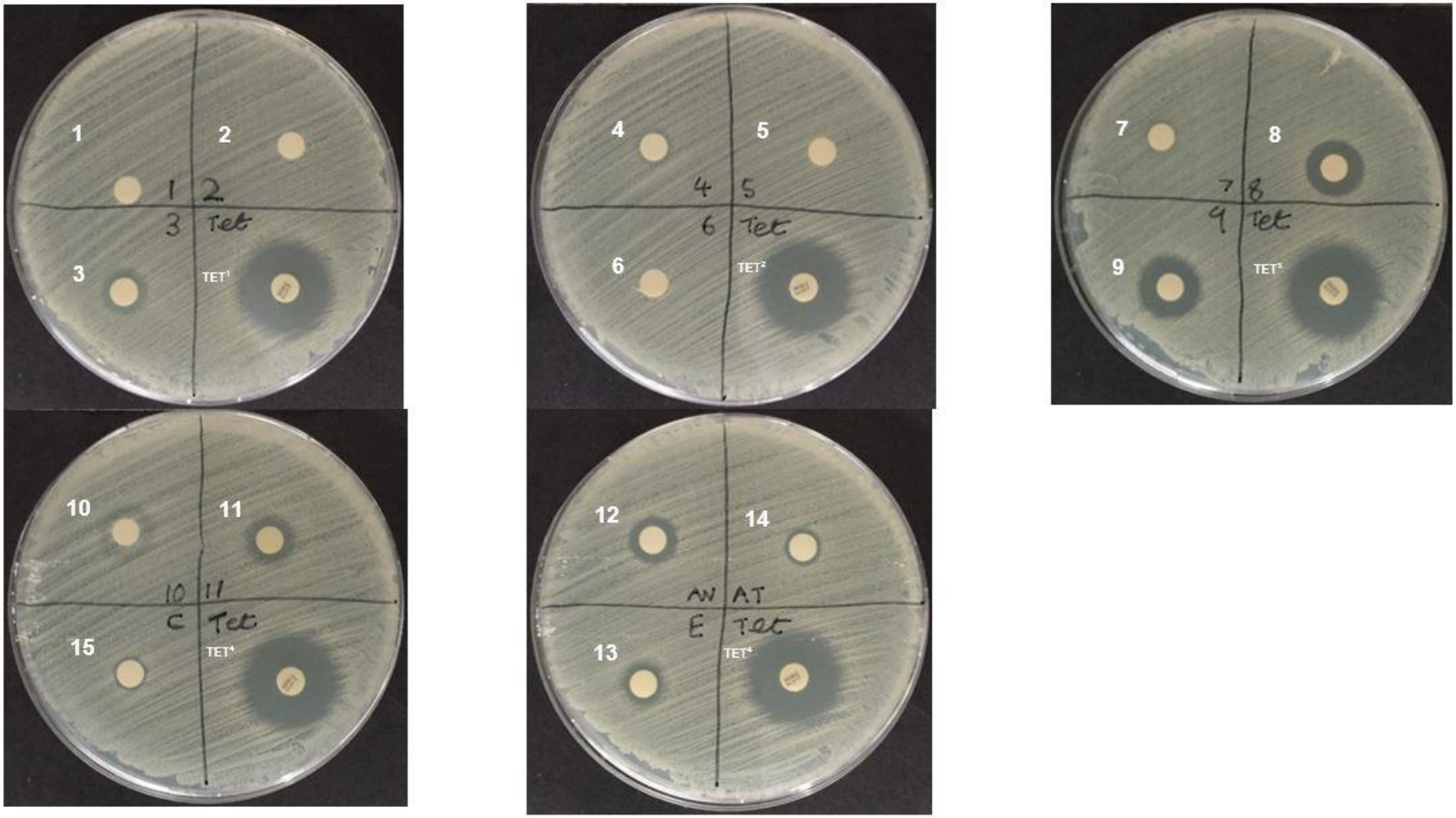


Figure 5.13. Results of disk diffusion assays of fractions (1-15 shown in Figure 5.12) produced by RP-SPE against *Staphylococcus aureus* NCTC 12981. Each plate also contained 5 µg tetracycline (TET^x) disks as controls. The images show that there are multiple zones of inhibition

Table 5.7. Description of dosage used of each fraction in disk diffusion assay and description of the appearance of each disk after assay

Fraction	Dosage on disk (ug)	Activity
1	240	No zone
2	668	No zone
3	448	Intermediate zone. 10 mm diameter
4	302	Intermediate zone. 8 mm diameter
5	126	No zone
6	348	No zone
7	90	Intermediate zone. 6.5 mm diameter
8	70	Clear zone. 13 mm diameter
9	58	Clear zone. 13.5 mm diameter
10	106	Intermediate zone. 13 mm diameter
11	380	Clear zone with small colonies in the outer edge of the zone. 10 mm diameter
12	Unknown	Clear zone. 7 mm diameter
13	Unknown	Clear zone. 8.5 mm diameter
14	Unknown	Clear zone. 11 mm diameter
15	Unknown	Clear zone. 8 mm diameter

The next step in the investigation was to analyse a subset of the fractions that had shown activity by both TLC and bioautography in order to separate and identify the active component(s) within the fraction. This analysis may also provide insight into whether the activity of the different fractions was due to multiple compounds through comparison of Rf values.

5.3.7.5. Analytical Thin-layer Chromatography and Bioautography of Bioactive Fractions

Figures 5.14-5.17 shows the results of the analytical TLC and bioautography of some of the active fractions. Fractions 8, 9, 10, and 11 were chosen for initial investigation rather than to attempt to analyse all 10 active fractions for ease of operation. Analytical TLC was chosen as a method to visualise the composition of fractions because it is a low-cost and simple method to perform which could also be paired easily with bioautography to understand where the active compounds were within the chromatogram of each fraction. This assay works on the principle that the compounds now separated along the analytical TLC plate will desorb into the agar. The active compound within the chromatogram can then be identified by the presence of a zone of inhibition in the agar which corresponds to the Rf value. This quick comparison may give an indication if there are multiple active compounds present based on zones of inhibition at different Rf values in each fraction.

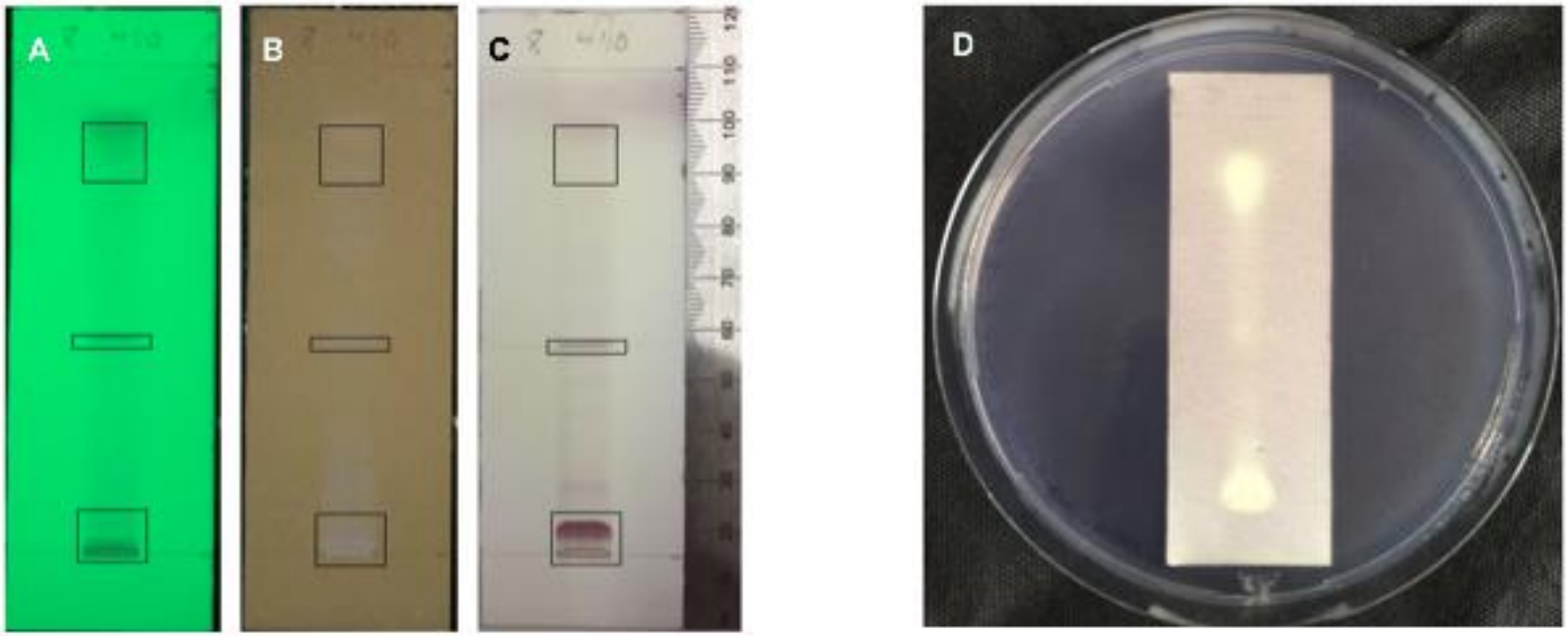


Figure 5.14. Thin-layer chromatography of active fraction 8 (a-c) and bioautography against *Staphylococcus aureus* NCTC 12981 (d). The extract was spot applied across a 1 cm section of baseline ten times and developed in triplicate in a mobile phase mixture containing 95 % ethyl acetate and 5 % acetonitrile that had been acidified with acetic acid. Developed plates were visualised at a) 254 nm, b) 366 nm, c) vanillin (1 %) in sulphuric acid. Duplicate plates were used for bioautography assay and zones of inhibition were visualised with MTT stain. The areas on the TLC plate highlighted in rectangular boxes correspond to the zones of antibiotic activity seen on the bioautography.

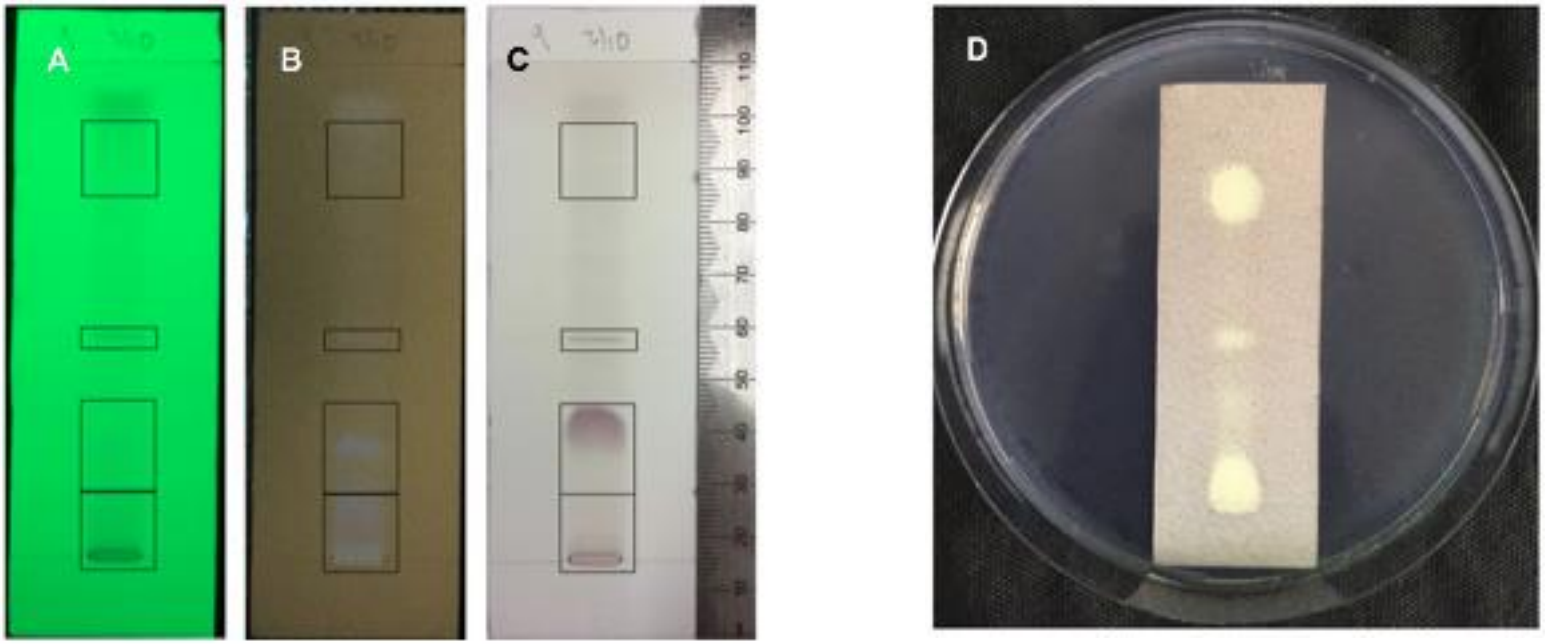


Figure 5.15. Thin-layer chromatography of active fraction 9 (a-c) and bioautography against *Staphylococcus aureus* NCTC 12981 (d). The extract was spot applied across a 1 cm section of baseline ten times and developed in triplicate in a mobile phase mixture containing 95 % ethyl acetate and 5 % acetonitrile that had been acidified with acetic acid. Developed plates were visualised at a) 254 nm, b) 366 nm, c) vanillin (1 %) in sulphuric acid. Duplicate plates were used for bioautography assay and zones of inhibition were visualised with MTT stain. The areas on the TLC plate highlighted in rectangular boxes correspond to the zones of antibiotic activity seen on the bioautography.

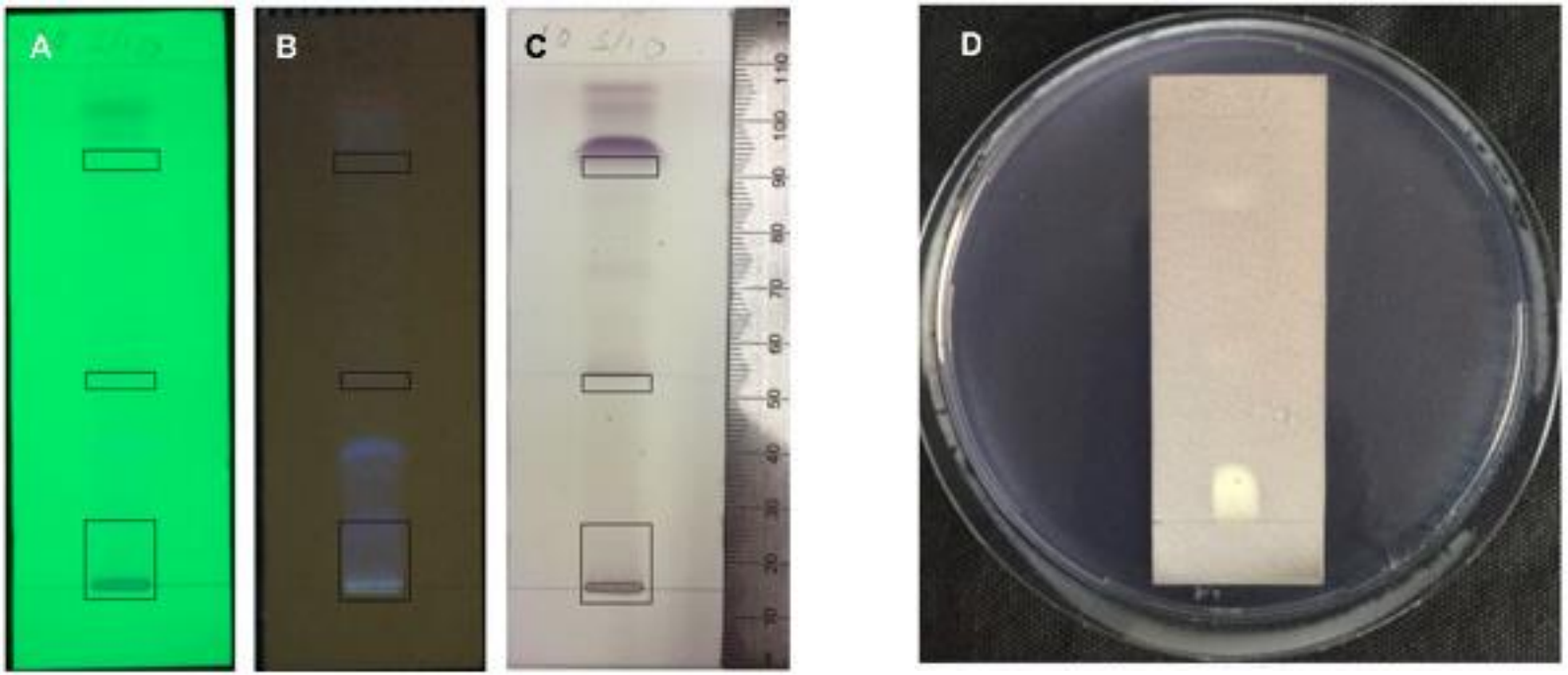


Figure 5.16. Thin-layer chromatography of active fraction 10 (a-c) and bioautography against *Staphylococcus aureus* NCTC 12981 (d). The extract was spot applied across a 1 cm section of baseline ten times and developed in triplicate in a mobile phase mixture containing 95 % ethyl acetate and 5 % acetonitrile that had been acidified with acetic acid. Developed plates were visualised at a) 254 nm, b) 366 nm, c) vanillin (1 %) in sulphuric acid. Duplicate plates were used for bioautography assay and zones of inhibition were visualised with MTT stain. The areas on the TLC plate highlighted in rectangular boxes correspond to the zones of antibiotic activity seen on the bioautography.

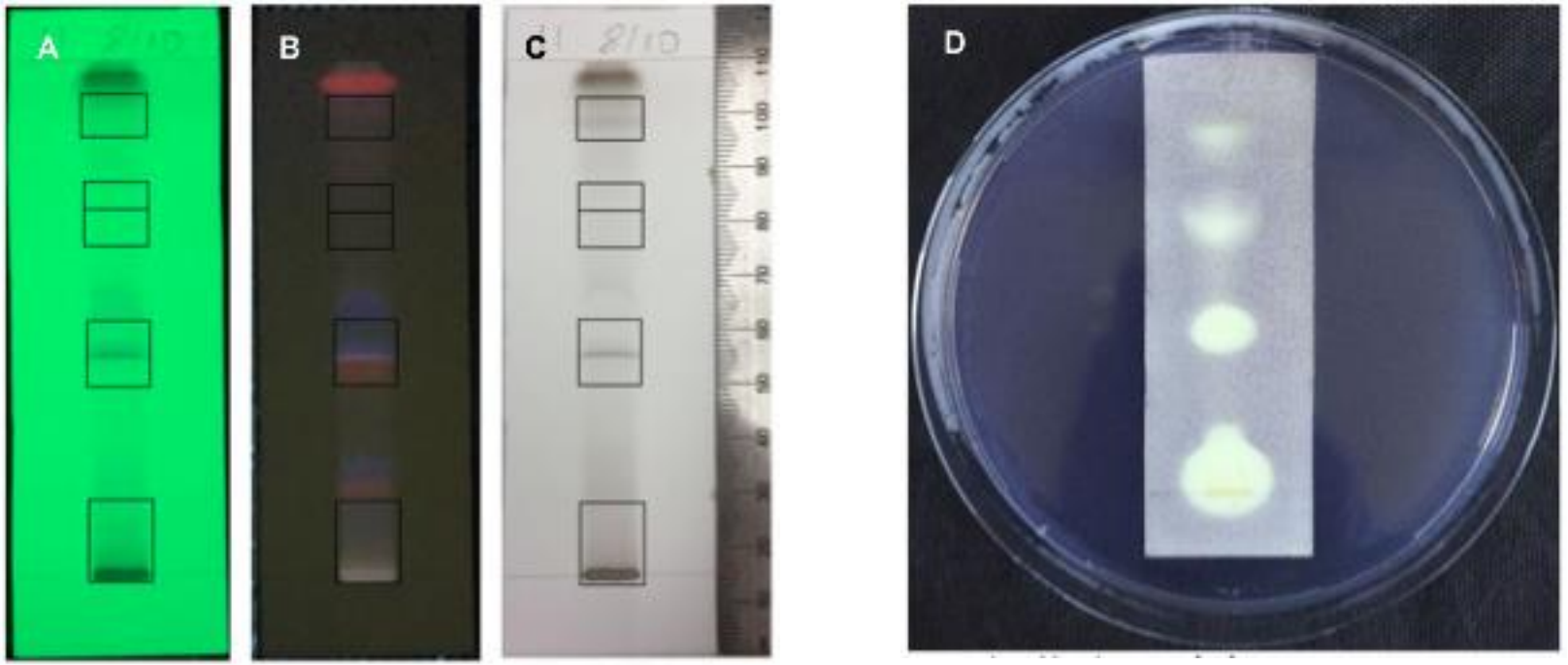


Figure 5.17. Thin-layer chromatography of active fraction 11 (a-c) and bioautography against *Staphylococcus aureus* NCTC 12981 (d). The extract was spot applied across a 1 cm section of baseline ten times and developed in triplicate in a mobile phase mixture containing 95 % ethyl acetate and 5 % acetonitrile that had been acidified with acetic acid. Developed plates were visualised at a) 254 nm, b) 366 nm, c) vanillin (1 %) in sulphuric acid. Duplicate plates were used for bioautography assay and zones of inhibition were visualised with MTT stain. The areas on the TLC plate highlighted in rectangular boxes correspond to the zones of antibiotic activity seen on the bioautography.

As the results in **Figures 5.14-5.17** show, the chromatograms of each fraction were highly complex with many different bands representing different compounds present with differing polarities. This complexity could in part be due to components of the Nutrient Agar being co-extracted. The bioautographies also revealed that these fractions had multiple compounds with antibiotic activities with different Rf values.

For example, the bioautography of fraction 9 (**Figure 5.15**) showed a clear zone of inhibition at baseline to Rf 0.14. This is immediately followed by a lighter "smear" of inhibited growth up to Rf 0.32. Suggestive that there is inefficient separation of compounds along the solid phase. There were multiple compounds observed on the analytical TLC plates under the different visualisation conditions used. The presence of compounds on the baseline and smearing above that is seen broadly under all visualisation methods. However, different patterns of compounds within these regions is seen under each visualisation method. This makes it difficult to determine which compounds may be responsible for antibiotic activity. A discrete zone of inhibition at Rf 0.44 which tracks with a discrete band is seen on the analytical TLC plate under all visualisation conditions. This is further followed by another, large and clear zone of inhibition centred at Rf 0.81, but no discrete compound is easily visible on the analytical TLC plate in this region. These results suggest that there are multiple compounds within the fraction with antibiotic activity. The different Rf values of the compounds along the solid phase show that these compounds have differing properties.

A similar set of results is observed in the bioautography of fraction 8 (**Figure 5.14**) where there were three zones of inhibition. One on the baseline which covered an area where multiple compounds could be seen on the plates. Another at Rf 0.44 which could be tracked to a discrete compound band on the plate, and a large zone at Rf 0.82. There was also smearing seen across the plates. This observation also suggests that the active compounds in both fractions 8 and 9 may be the same and that there was inefficient fractionation.

Fraction 10 bioautography (**Figure 5.16**) also had a similar pattern of results but with a notable difference in the size and intensity of the zones of inhibition. A large and clear zone can be seen just above the baseline, this suggests that the compound(s) responsible have all successfully eluted from the baseline, compared to fractions 8 and 9. A zone is still seen at the mid region of the plate (Rf 0.40) but is far weaker than in fractions 8 and 9, and no compounds that track to this region are observed on the plate. There is also a faint zone at Rf 0.80 which couldn’t be clearly matched to any discrete compound on the plates. It is possible that these zones, which are in similar positions to those in fraction 8 and 9 but showing weaker activity

and visualisation could be the same compounds present in fractions 8 and 9 but at a lower concentration.

The bioautography of fraction 11 (**Figure 5.17**) had a different profile to the other fractions. There are four regions along the plate with zones of inhibition as opposed to the three seen in the other fractions. There was a large zone of inhibition at baseline to Rf 0.14 with a small enclave to Rf 0.18. This pattern is suggestive of multiple compounds with activity within this region eluting at differing rates but have failed to be fully separated. There is also another large zone centred at Rf 0.42 along with irregularly shaped zone between Rf 0.63-0.71 and a further weaker enclave up to Rf 0.76. And finally, there is a clear zone of inhibition at Rf 0.84. The different profile in this fraction compared to the others examined suggests it contains different active compounds.

It is clear from the observations of the bioautographies of these four active fractions that these solutions are still highly complex, and the presence of so many active compounds with different Rf values is suggestive that the KB16 extract may be producing multiple antibiotic compounds. Therefore, the antibiotic activity of KB16 is unlikely to be solely caused by albaflavenone. Enquires were made to see if a pure albaflavenone could be obtained to determine its characteristics on the TLC study, but it was not possible to obtain this compound during this study. However, as albaflavenone is a volatile sesquiterpene with a polar nature it is expected to have a high Rf value. The presence of active compounds at the baseline and at relatively low Rf values suggest that these compounds may not be albaflavenone. It is also worth noting that as albaflavenone is a volatile compound, efficiency of retention on the solid phase may be low. Therefore, the presence of prominent compound bands with activity on the solid phase is encouraging news that these compounds are unlikely to be albaflavenone. It is important to also note that the evidence that KB16 maybe producing multiple antibiotic compounds would have been missed if further chemical studies had not been performed after genome-mining. Integration of both genome and chemical analysis pipelines is therefore important.

In order to determine if any of the compounds were albaflavenone, preparative TLC was performed on the fractions to purify the active component(s) for spectrometry analysis.

5.3.7.6. Purification and Analytical Analysis (NMR & LESA-MS) of Active Compounds

Using the Rf values of the active components identified by bioautography, preparative TLC was performed on the fractions to attempt to purify them. In the first instance these purified components were analysed using NMR to attempt to identify them. NMR was chosen as the spectroscopic technique because it can produce reproducible spectra for a certain compound structure which can be interpreted with reliability between different devices. This could allow for the spectra of the isolated compound(s) to be compared to the known spectra of albaflavenone. Other spectrometry techniques may not provide as conclusive an identification. Additionally, NMR is a non-destructive spectrometric technique which would allow for the analysed compounds to be retained for further characterisation, such as mechanism of action investigations, if deemed necessary.

NMR spectra could not be obtained for the active components purified by preparative TLC. This could be due to the lower sensitivity of NMR compared to other spectroscopy methods. The masses of fractions 8-11 was between 2.9-37.9 mg whilst the sensitivity of H^1 NMR may require material mass in the range of 5-25 mg. It is probable that preparative TLC did not yield sufficient mass for NMR analysis.

To address this an attempt was made to upscale the production of KB16 crude fermentation from the 100 plates initially made to ~600 plates in order to generate sufficient mas for NMR analysis. However, the passage of this work was disrupted due to external circumstances beyond the researcher's control which necessitated a sudden cessation of the project. This led to the loss of this upscaled sample which could not be recovered due to remaining time constraints and other limiting factors upon return to the project.

An alternative approach was then adopted to use LESA-MS to pair the analytical TLC and bioautography methods which had already been utilised on the low mass fractions to isolate the active component(s) and pair this with high-sensitivity mass spectroscopy in order to determine the nature of compounds.

LESA-MS is a relatively recent innovation upon standard electrospray mass spectrometry that allows for analytes on a solid matrix to be analysed. A robotically controlled pipette tip places a small droplet of solvent on the surface of the sample to be analysed. Analytes on the surface are extracted from the sample surface and then passed up through the pipette for mass spectrometry. This method has been adapted for many applications. For example, to

characterise polymer materials such as the degradation of food packaging (Issart et al., 2019), or to determine the distribution of drug metabolites within organ tissue cryo-sections (Eikel et al., 2011), and identifying surface peptides on live bacterial (Kocurek et al., 2017) and yeast (Kocurek et al., 2020) colonies.

In principle, LESA-MS could be applied to analyse the active compounds that were identified by bioautography by focusing the extraction solvent pipette on the area of a duplicate analytical TLC plate that corresponds with the Rf values of the zones of inhibition. This may have the advantages of allowing for dereplication of the current extracts obtained and so avoid the need for further upscaling of production or the optimisation of a new HPLC-based chromatography method. Thus, providing the potential for integrating the genome-mining pipeline (developed in **Chapter 4**) utilised in this chapter with a resource efficient bioactivity-guided isolation approach.

Duplicate analytical TLC plates of fractions 8-11 were made and the presence and Rf values of the active compounds were confirmed by bioautography. The results of the duplicate bioautographies all corresponded to the results previously observed (**data not shown**).

The Rf values corresponding to the active areas were analysed by LESA-MS. However, LESA-MS analysis provided inconclusive results as to the structure of any of the compounds located at these Rf values. This could potentially be due to the extraction method used to desorb the analytes from the surface of the analytical TLC plates not being well optimised to obtain sufficient mass. It was not possible to attempt any further optimisation due to time constraints at this point in the project.

5.3.8. Conclusions

An uncharacterised bacterium which had previously been isolated from the water of the King's Spring of the Roman Baths, UK was demonstrated to have antibiotic activity against multidrug resistant Gram-positive bacteria but no activity against Gram-negative bacteria. The isolate was identified morphologically and by 16S rRNA gene typing to in the genera *Streptomyces*, which is a genus of bacteria that is characterised as a prolific producer of natural products with antimicrobial activity. The partial 16S rRNA gene aligned most closely with those from species *Streptomyces canus*. The Oxford Nanopore long-read genome sequencing pipeline optimised in **Chapter 4** was utilised to obtain a high-resolution phylogenetic analysis of KB16. The genome assembly was also analysed to obtain a high-quality annotation of potential BGCs for the purposes of dereplication and also to inform bioactivity-guided isolation strategies to obtain the active antibiotic compound(s). The genome assembly pipeline produced a contiguous assembly which has a 99.7 % completeness score as measured by BUSCO. This score is the same as that obtained for the assembly of *Streptomyces coelicolor* A3(2) detailed in **Chapter 4** and so was deemed to be of sufficient quality for accurate phylogenetic and biosynthetic gene cluster analysis, and this also validates the pipeline for producing quality assemblies of environmental isolates.

A whole-genome phylogenetic analysis was performed by aligning the sequences of 400 housekeeping genes from the KB16 genome against those from 102 *Streptomyces* RefSeq genome sequences. This analysis identified the genome of KB16 to most closely related with that of *Streptomyces canus* DSM 40275, whilst the average nucleotide identity of both was found to be 97.4 %. These results suggest KB16 to be a member of the *Streptomyces canus* clade and very likely a member of this species. However, further metabolism and biochemical tests would be necessary to fully confirm this.

The genome mining analysis predicted 26 secondary metabolite gene clusters of a variety of different chemical classes including NRPS, PKS, terpenes, and RiPPs. Few of the BGCs showed significant homology with known secondary metabolite compounds in the MIBiG database which suggests that KB16 may be producing multiple novel secondary metabolites. There was also no homology with antibiotics reported to be produced by *Streptomyces canus* species, suggesting that these compounds are not produced by KB16. The lack of evidence for antibiotic BGCs associated with *Streptomyces canus* is further evidence that taxonomic classifications may not be a reliable dereplication strategy, as also evidenced in **Chapter 3** with the discovery of *Enterococcus* associated AS-48 pathways in *Bacillus* isolates.

The genome-mining analysis revealed several potentially novel BGCs of interest in addition to a BGC with 100 % homology to the antibiotic terpene albaflavenone. In order to confirm if albaflavenone was the cause of the antibiotic activity seen, a bioactivity-guided isolation strategy was employed to isolate and identify the active compound. Bioautography revealed multiple active components which is suggestive that KB16 may be producing multiple compounds with antibiotic activity and so, by extension, if albaflavenone is contributing to the antibiotic activity seen it may not be solely responsible.

Attempts to isolate these active components for NMR analysis were unsuccessful due to the low mass of the fractions. LESA-MS analysis was therefore attempted but was also unsuccessful and may need optimisation, but this was not possible due to time constraints of the project.

In summary, the work in this chapter demonstrates the value of genome-mining for dereplication in bioprospecting and validates the Nanopore sequencing pipeline developed in **Chapter 4** for this purpose. KB16 maybe producing multiple antibiotic compounds of interest, and this observation would not have been made without utilising a bioactivity-guided isolation strategy in addition to genomic analysis. The use of genome-mining to inform the decisions taken in bioactivity-guided isolation is therefore important and development of strategies to effectively integrate both approaches is required. The use of LESA-MS analysis may have potential as a relatively resource-efficient approach but would require optimisation to be realised.

Chapter 6.

Assessment of the Microbiome of the Roman Baths, UK by Oxford Nanopore 16S rRNA Gene and Shotgun Metagenomic Sequencing

6.1. Introduction

6.1.1. The Roman Baths, UK

The city of Bath in the United Kingdom is unique in the British Isles for the presence of thermal springs with temperatures above 40 °C. There are three major springs in the city: The King's, Hetling, and Cross Springs, as well as several minor springs around the greater Avon Valley area. These springs have played a major role in the development of the region and are of historical and scientific interest. Evidence of human interaction with the spring waters of the region date back to the Mesolithic period. During the early period of the Roman occupation of Britain a stone reservoir was built upon the King's spring which fed a temple spa dedicated to the Goddess Sinus-Minera. Since then, Baths on the site were used near continuously for general bathing and the treatment of rheumatism and other illnesses until 1977 when the site was closed to bathers due to public health concerns (Kellaway, 1991). Most of the historic complex, now known as The Roman Baths, was built in the 19th Century and is a UNESCO world heritage site of major archaeological significance and a popular tourist attraction. The complex contains many baths that are fed by the King's Spring - with the King's Bath sitting directly above the spring (**Figures 6.1 & 6.2**).

The source of the thermal springs is believed to be rainfall on the nearby Mendips hills which percolates through limestone aquifers to a depth of up to 4,300 meters. The water is heated and pressurised underground and rises back up to surface along fissures and faults in the limestone. The oxygenated water undergoes chemical reactions with the stone it travels through, leading to changes in its mineral content. Oxygen in the water reacts with iron sulphides in the rock to produce sulphuric acid which attacks calcium carbonates and other minerals in the rock. Once the water reaches the surface, oxidation with air leads to the development of bright orange deposits of iron oxide that line the walls of the ancient baths and drainage network, and is a defining characteristic of the thermal waters of Bath (**Figure 6.3**) (Kellaway, 1991)

Figure 6.1. Bathers in the King's Bath c.1800. Painting by John Dixon, printed in Hayward (1991).

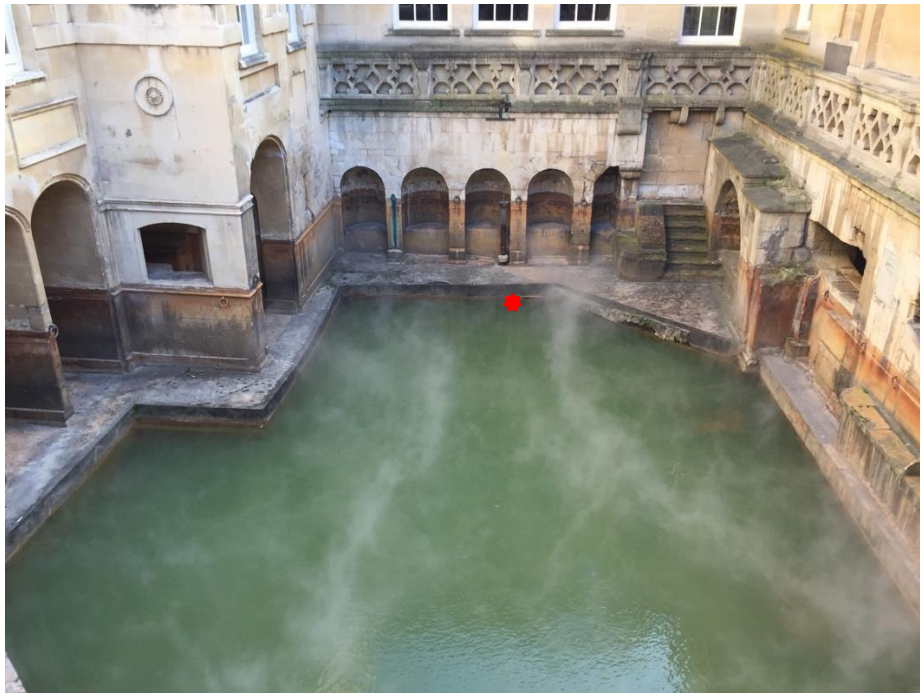


Figure 6.2. The King's Bath in 2018. Image by Tim Walker. Location that the water sample was collected is marked by red star. The King's Spring is directly below the bath complex.

Figure 6.3. Images of the ancient Roman drainage network of the Roman Baths, UK. The images show the orange iron oxide deposits. Images by Tim Walker, 2018.

6.1.1.1. The Geochemistry of The Roman Baths

Investigations into the chemical composition of the King's Spring have shown it to have a total mineralisation of 2.18 g/L with the dominant components being sulphate and calcium – which make up 65 % of the total. Additionally, the most abundant singular metal element is iron (**Table 6.1**) (Edmunds and Miles, 1991). Exsolved gas content of the water is dominated by N_2 at 93.5 % followed by CO_2 at 3.5 %. The content of O_2 is 0.5 % (**Table 6.2**). Radioactive isotopes of Radon, Uranium, Radium, and Lead have also been detected in the water (**Table 6.3**) (Andrews et al., 1982).

Table 6.1. Geochemical measurements of the King's Spring. Measurements taken in 1979 and reproduced from Edmunds & Miles (1991)

Measurement	Value
Temp (°C)	45.30
pH	6.65
Dissolved O_2 (mg/L)	< 0.20
U (mg/L)	0.00005
Hg (mg/L)	< 0.0001
HS (mg/L)	< 0.001
Sc (mg/L)	< 0.001
Y (mg/L)	< 0.001
Sb (mg/L)	< 0.002
Cu (mg/L)	0.002
Cd (mg/L)	< 0.0025
Se (mg/L)	< 0.004
La (mg/L)	< 0.01
Zr (mg/L)	< 0.015
Rb (mg/L)	< 0.02
Co (mg/L)	< 0.02
Zn (mg/L)	< 0.02
Ni (mg/L)	0.022
Ba (mg/L)	0.024
Cr (mg/L)	< 0.03
I (mg/L)	0.043
Al (mg/L)	< 0.05
Total Mn (mg/L)	0.068
Mo (mg/L)	< 0.1
Pb (mg/L)	< 0.2
V (mg/L)	< 0.2
Li (mg/L)	0.242
B (mg/L)	0.59
Total Fe (mg/L)	0.88
NO_3 (mg/L)	< 1
Br (mg/L)	2.02
F (mg/L)	2.08
Sr (mg/L)	5.92
K (mg/L)	17.4
Si (mg/L)	20.6
Mg (mg/L)	53
Na (mg/L)	183
HCO_2 (mg/L)	192
Cl (mg/L)	287
Ca (mg/L)	382
SO_4 (mg/L)	1032
Total Mineralisation (mg/L)	2179

Table 6.2. Exsolved gas composition of the King's Spring from Andrews et al., (1982)

Exsolved Gas	Percentage (%)
N_2	93.5
CO_2	3.5
Ar	1
C_3H_8	0.8
CH_4	0.6
O_2	0.5
4He	0.1

Table 6.3. Radioelement composition of King's Spring from Andrews et al., (1982)

Radioelement	Mode between 1977-1979 (pCi/kg)
^{222}Rn	2350
^{210}Pb	< 1.0
^{226}Ra	11
^{238}U	0.017

6.1.1.2. Current Knowledge of the Microbiology of The Roman Baths

Microorganisms are known to inhabit the waters of the spring. One such discovery being the pathogenic amoeba *Naegleria fowleri* which unfortunately forced the closure of the Roman Baths to bathers in 1977 after nearly two millennia (Kellaway, 1991). There are only three other examples in the literature of microorganisms isolated from the Roman Baths, and they all have novel biochemical properties that may be of future biotechnological application. The first is *Methylococcus capsulantus* (Bath) which is an obligate methanotroph that is also predicted from its genome annotations of having the ability to scavenge copper from the environment (Ward et al., 2004). This microbe therefore has potential uses in bioremediation of industrial greenhouse gases and metal contamination. The other two microorganisms were reported by Wood and Kelly in the 1980s; Originally both classified as *Thiobacillus* spp. but now known as *Annwoodia aquaesulis* and *Thermithiobacillus tepidarius* (Wood and Kelly, 1986, 1988; Kelly and Wood, 2000; Williams and Kelly, 2013; Boden, Hutt and Rae, 2017). The two isolates are thermophilic with optimum growth temperatures of 43-44 °C, and are both capable of oxidising sulphur containing compounds. *Annwoodia aquaesulis* is also capable of nitrification under anaerobic conditions. Genome sequencing of *Thermithiobacillus tepidarius* reveals genes predicted to confer resistance to the metallic elements arsenic tellurite, cadmium, cobalt, zinc, copper and silver. The genome sequence also revealed evidence that suggests horizontal gene transfer with *Methylococcus capsulantus* (Bath) and *Thiobacillus* spp. (Boden et al., 2016). These traits give these two bacteria potential usages in the recovery of heavy metals and sulphur from environmental waste. Despite these examples, no studies have yet been published that examine the microbiome of the Roman Baths.

6.1.2. Profiling Microbiomes for Bioprospecting

The discovery of antibiotic producing isolate KB16, and its characterisation detailed in **Chapter 5**, gives an indication that the microbiome of the Roman Baths, UK maybe a lucrative site for bioprospecting for microbially-derived compounds of medicinal and commercial value. This would also be in keeping with the precedent set by researchers who have prospected other hot spring sites around the world (detailed in **Section 1.6.4.3**).

Metagenomic techniques can be used to gain an understanding of the biosynthetic potential of a particular microbiome (as detailed in **Section 1.6.2**). Often these approaches have involved selectively targeting common biosynthetic genes of interest for sequencing and analysis. Such as in the work detailed in **Chapter 2** to assess a human oral microbiome for PK and NRPS genes. While this approach does provide insights into the presence of these genes, it can also be limited by the fact that it targets only a small subset of very long BGCs and removes the gene from its wider context. This prevents any analysis into the full gene cluster and the potential product it produces. Additionally, other biosynthetic gene types are not detected. High-throughput direct shotgun sequencing and assembly of metagenomes could in part address these issues by capturing all BGC classes. However, assembling long and repetitive BGC pathways from complex metagenomes is highly complex, computationally and financially expensive, and unreliable (Meleshko et al., 2019).

As is the case with single species genome sequencing, long-read sequencing of metagenomes may improve assembly and detection of BGCs. Long-read Nanopore shotgun sequencing has been demonstrated to assemble the whole genomes of microbial organisms within a microbiome (known as metagenomic assembled genomes – "MAGs") allowing for identification and characterisation *in silico* of novel species that are otherwise currently inaccessible through culturing methods (Nicholls et al., 2019). Nicholls and co-workers (2019) demonstrated a proof-of-concept pipeline for the accurate assembly and annotation of MAGs from a defined mock microbial community using hybrid approach of high-throughput Nanopore and Illumina shotgun sequencing. Whilst Stewart and co-workers (2018) demonstrated a similar approach to investigate the metabolism of the bovine gut microbiome using both Oxford Nanopore technology and Illumina short reads. The researchers reported on the identification of 913 microbial genomes and over 69,000 predicted carbohydrate metabolic enzymes, the vast majority of which (over 90 %) appeared to be novel.

Long-read sequencing and assembly of contiguous MAGs from microbiomes could be used for detection and analysis of complex multigene natural product pathways, such as for NRPS

and PKS genes. However, both the Stewart and Nicholls models required high sequencing data throughputs (768 Gbp and 330 Gbp respectively) and two different platforms, and high computational resources, all of which can serve as a barrier to the utilisation of these approaches.

Typically, assembly of metagenomic genomes requires greater computational resource than single genome assembly due to additional factors posed by the heterogeneous nature of the sample. These factors include unknown variation in species diversity and abundances, and the separation of genomes of closely related species or those which share closely related or conserved sequences stretches (Martin, Clark, Leggett, 2020).

It has been shown that direct sequencing of environmental metagenomes by long-read Nanopore sequencing alone can be used for taxonomic profiling and also for annotation of genes of interest within a microbiome without the requirement for contig assembly. Therefore, requiring far less sequencing throughput and computational resource which puts it into reach of more researchers. This approach has been applied in clinical applications such as epidemiological surveillance (Quick et al., 2016; Taxt, et al., 2020), AMR gene detection (Lemon et al., 2017; Schmidt et al., 2017; González-Escalona et al., 2019; Taxt, et al., 2020), and also in taxonomic profiling of extremophile microbiomes (Gowers et al., 2019).

6.1.4. Aims and Objectives

The aims of the work detailed in this chapter are to utilise long-read Oxford Nanopore sequencing to both gain an initial survey of the taxonomy of the microbiome of the hot spring water of the Roman Baths and to detect and annotate BGCs, especially long BGCs. The rationale for this investigation is that understanding the taxonomic diversity of the site may help inform sampling and culturing conditions for future bioprospecting of isolates from the site. In addition to this, detection of BGCs could allow bioprospectors to understand the abundancy, variety, and potential novelty of natural products that may be obtainable from the site.

6.2. Material and Methods

6.2.1. Collection of Sample

In February 2018, approximately 5.75 L of water from the King's Bath at the Roman Baths, UK was collected from a depth of 6-10 cm using twenty-three 250 mL sterile tissue culture flasks. The flasks were sealed and transported to London that same day at room temperature. DNA was extracted from the water on the same day as it was collected once in London.

6.2.2. DNA Extraction of Water Sample

The water was filtered through a sterile PVDF 0.22 µm membrane under vacuum using a HPLC filtration system (**Figure 6.4**). Prior to use, the system was cleaned with detergent and a 1 % bleach solution before autoclaving. As a negative control to test for sterility – 250 mL of DNA-free molecular grade water was filtered through the system before filtration of the test sample.

Between 1.5-2.5 L of sample was filtered through a membrane before the membrane was replaced. Three membranes were used in total for the test sample and one membrane was used for the negative control.

Figure 6.4. Picture of the HPLC filtration system used to filter water sample

The concentrated biomass collected on the membranes was resuspended into 10 ml of DNA-free molecular grade water. Half of this suspension was used to make 50 % glycerol stocks for storage at -80 °C whilst the other half was used for DNA extraction.

DNA was extracted from the suspension using the Qiagen DNeasy Blood & Tissue Kit (Qiagen GmbH, Germany) according to the manufacturer's protocol, but with the modification that the heat incubation step was increased from 70 °C for 10 mins to 85 °C for 25 mins.

6.2.3. Gel Electrophoresis

Samples (1 µl) were analysed on a 1 % agarose gel containing 0.5 % GelRed™ (Biotium, Inc., USA) in TAE buffer (40mM Tris, 20mM acetic acid, and 1mM EDTA) run for 3 hours at 65 V. Prior to loading, each sample was mixed with NEB Purple Loading Dye (New England Biolabs Inc., USA) according to the manufacturer's protocol. As a molecular size marker, NEB λ DNA-*Hin*dIII Digest (New England Biolabs Inc., USA) was used according to manufacturer's protocol. The gel bands were visualised on an UV transilluminator.

6.2.4. Preparation of 16S rRNA Gene Sequencing Library

The 16S rRNA gene amplification and library preparation was performed using the Nanopore 16S Barcoding Kit (SQK-RAB204) (Oxford Nanopore Technologies Ltd, UK) according to the manufacturer's protocol. Barcoded primers from this kit are based on degenerate primers 27F and 1492R (Lane et al., 1991). PCR conditions used: 95 °C for 1 min, 35 cycles [95 °C for 20 secs, 55 °C for 30 secs, 65 °C for 2 mins], 65 °C for 5 mins.

6.2.5. Multiple displacement Amplification (MDA) of King's Bath Metagenome

Approximately 10 ng of sample was amplified using REPLI-g Midi Kit (QAIGEN GmbH, Germany) according to manufacturer's protocol, using incubation time of 16 hours. In addition to metagenome, 2.5 µl blank DNA extraction, 10 ng of ZymoBIOMICS Microbial Community DNA Standard (ZYMO Research, Inc., U.S.A.) (subsequently referred to as the "mock microbiome"), and 2.5 µl pure molecular grade water were processed as controls.

Amplified samples sample was purified by SPRI bead size selection using Ampure XP beads (Beckman Coulter, Inc, U.S.A.) according to manufacturer's protocol at a ratio of 1.8:1.0 of beads to sample and eluted into 90 µL elution buffer.

Sample was digested using T7 Endonuclease in the following reaction: 1 ug DNA, 2 µL NEBuffer (New England Biolabs, Inc., U.S.A.), 1 µl T7 Endonuclease I (New England Biolabs, Inc., U.S.A.), water up to 20 µL, at 37 °C for 15 mins.

Digests were purified and concentrated by SPRI bead size selection using Ampure XP beads (Beckman Coulter, Inc, U.S.A.) according to manufacturer's protocol at a ratio of 1.8:1.0 of beads to sample and eluted into 40 µL elution buffer. The samples were assessed for quality by Nanodrop and gel electrophoresis.

6.2.6. Preparation of Shotgun Sequencing Library

6.2.6.1. Multiple Displacement Amplified Samples

Samples were prepared using the Nanopore Rapid Barcoding Sequencing Kit (SQK-RBK004) (Oxford Nanopore Technologies Ltd, UK) according to the manufacturer's protocol.

6.2.6.2. PCR Amplified Samples

Sample fragmentation, amplification, and library preparation was performed using the Nanopore rapid PCR barcoding kit (SQK-RPB004) (Oxford Nanopore Technologies Ltd, UK) according to the manufacturer's protocol. PCR conditions used: 95 °C for 3 min, 14 cycles [95 °C for 15 secs, 56 °C for 15 secs, 65 °C for 6 mins], 65 ºC for 1 mins.

6.2.7. Sequencing of All Libraries

The samples were loaded into a MinION™ Mk1B Sequencer (SKU-MIN101B) containing a 'SpotON' flow cell with R9 chemistry (SKU-FLOMIN106) (Oxford Nanopore Technologies Ltd, UK) according to the manufacturer's instructions and were sequenced without live basecalling.

6.2.8. Bioinformatics Analysis

6.2.8.1. Processing and Analysis of 16S rRNA Gene Sequencing Reads

After sequencing, 3,000,000 FAST5 raw read files were based called using Albacore v2.2 (Oxford Nanopore Technologies Ltd, UK) with a q-score threshold setting of 10. The resultant FASTQ files were then demultiplexed, trimmed of their adapters, and classified against the NCBI 16S RefSeq database (O'Leary et al., 2016) using the EPI2ME 16S Classification Workflow v.2.1.1 (Oxford Nanopore Technologies Ltd, UK). Settings: q-score 10, minimum horizontal coverage 30 %, minimum identity match 77 %.

The resultant trimmed FASTQ reads and CSV files containing sequencing metadata (q-score, length in bp (before trimming of barcodes and adapters ~200 bp), barcode, and taxa assignments (NCBI taxon id and identity match) were downloaded from the EPI2ME server and analysed using Python Pandas (McKinney, 2012), Biophyton (Cock et al., 2009), MEGAN v6 (Huson et al., 2016), NCBI BLAST (Altschul et al., 1990), Ycard (Marijon, Chikhi & Varré, 2020), and SankeyMATIC (Bogart, 2018).

Python Pandas was used to analyse and make bar plots of the sequencing metadata and taxon assignments. Pandas was also used to reformat data so it could be parsed into MEGAN which was used to visualise the metagenome as a taxonomic tree for manual inspection. Google Charts was used to plot Sankey diagrams from reformatted taxon data. Pandas and bespoke Biopython scripts were used to select for FASTQ reads of interest that were then further analysed by BLAST searching against the RefSeq 16S database. The results of these searches were then analysed using Python Pandas.

6.2.8.2. Processing and Analysis of Shotgun Sequencing Reads

FAST5 raw read files were based called, trimmed, and demultiplexed using Guppy v2.3.1 (Oxford Nanopore Technologies Ltd, UK) with a q-score threshold setting of 10. NCBI taxonomy ranks of resultant were assigned to the reads using the Centrifuge engine (Kim et al., 2016) against the NCBI RefSeq database via the EPI2ME WIMP workflow v 3.2.1 (Oxford Nanopore Technologies Ltd, UK). Data was reformatted for plotting using Pyton Pandas (McKinney, 2012) and visualised using Pavian (Breitwieser & Salzberg, 2020). Analysis for the detection of secondary metabolite BGCs was performed using AntiSMASH v3.0.5 (Weber et al., 2015). Functional gene annotations of metal and biocide resistance genes of fastq files were obtained using the NanoARG Server (Arango-Argoty et al., 2019).

Analysis was carried out on local machines and also using the MRC Cloud Infrastructure for Microbial Bioinformatics (CLIMB) facility (Connor et al., 2016).

6.3. Results and Discussion

6.3.1. Analysis of DNA Extraction of King's Bath

The result of the DNA extractions of the King's Bath water sample and the molecular water negative control were measured using a Nanodrop and by gel electrophoresis. The result of the gel electrophoresis (**Figure 6.5**) showed that the extraction of the molecular water control did not detect any DNA. While the King's Bath water sample contains DNA concentrated near the 23 kbp marker of the ladder. This indicates that there was no significant DNA contamination introduced into the test sample by the handing and DNA extraction processes. The fact that the DNA of the test sample is concentrated around near the 23 kbp marker of the ladder is also suggests that the sample is HMW. However, there is some smearing below this concentration which suggests that there is some sample degradation. This analysis was followed-up by quantifying the sample on a nanopore where the total amount of DNA obtained was recorded as being ~60 ng which equates to ~10.5 ng/L of water collected.

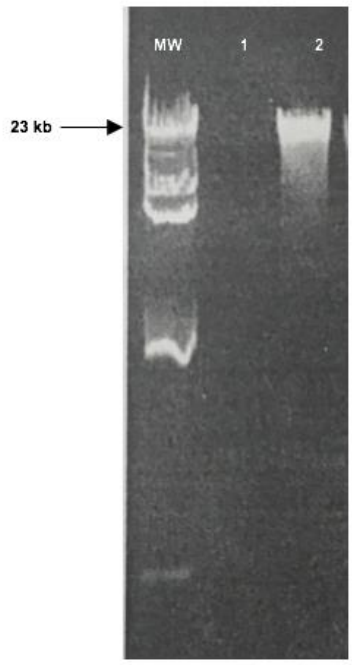


Figure 6.5. 1 % agarose gel of 1 µL (~ 3 ng) of the DNA extraction of the King's Bath water sample and molecular water control. MW = λ DNA-*Hin*dIII digest molecular size marker, 1 = molecular water extraction. 2 = King's Bath water sample. Results suggest HMW DNA.

The amount of DNA collected from the extraction was below the minimal input requirements recommended (400 ng) for direct shotgun sequencing. This could risk resulting in suboptimal library preparations and low yields, while also compromising the

performance of the sequencer flow cell which would reduce read quality and subsequent analysis. Therefore, an amplification step of the input metagenome was deemed necessary before sequencing. Prior to performing this step, the metagenome was profiled using 16S rRNA profiling in order to profile the taxonomy of the microbiome and to also offer a basis for comparison of the taxonomy profiling performance of the shotgun sequenced samples. Profiling the taxonomy of the site would also give insight into what potential genera inhabit the waters which could give indications of the bioprospecting potential and required culturing strategies to selectively cultivate microorganisms of interest.

6.3.2. 16S rRNA Gene Profiling

6.3.2.1. Analysis of Sequencing

6.3.2.1.1. Comparison of the Sample Against the Negative Control

The EPI2ME workflow returned 1,224,329 classified reads from the King's Bath metagenome sample whilst the negative control returned 13 reads. Metagenomic samples can become contaminated with DNA present on laboratory equipment, from dust and other particles in the air, or even from commercial DNA extraction kits and PCR reagents (Glassing et al., 2016). The reason for sequencing the negative control was to test for the presence of any contaminating DNA that the sample may have come into contact with during its preparation which could distort the taxonomic distributions recorded. However, the low abundance of reads returned from the negative control in comparison to the test sample suggest that environmental contamination is not at a level high enough to compromise the results for the test sample.

6.3.2.1.2. Analysis of Read Lengths

The mean average length of the reads before removal of adapter sequences was 1578 bp. **Figure 6.6** shows the distribution of read lengths in the sample. The highest abundancy was in the 1500-1600 bp range with over 800,000 reads being in this range. The second highest abundancy was in the 1600-1700 bp range with approximately 350,000 reads. Therefore over 90 % of the reads returned were in the size range 1500-1700 bp. The histogram also shows that there was a small number of reads ranging in size from 400-1500 bp and also between 3000-3300 bp. The observation that most reads were in the region 1500-1700 bp is expected because the primers used in this study are designed to amplify an approximate 1500 bp region of the 16S gene, and the addition of adapter and barcode sequences to the primers will increase the observed length of the reads by 100-200 bp. Whilst the length of the 16S rRNA gene can vary between bacterial species due to variations in the hypervariable regions of the gene, the reads outside of the expected size range could also be artefacts caused by, in the case of shorter lengths, fragmented of amplicons. While in the case of the larger lengths, they could be chimeras formed by primer dimers.

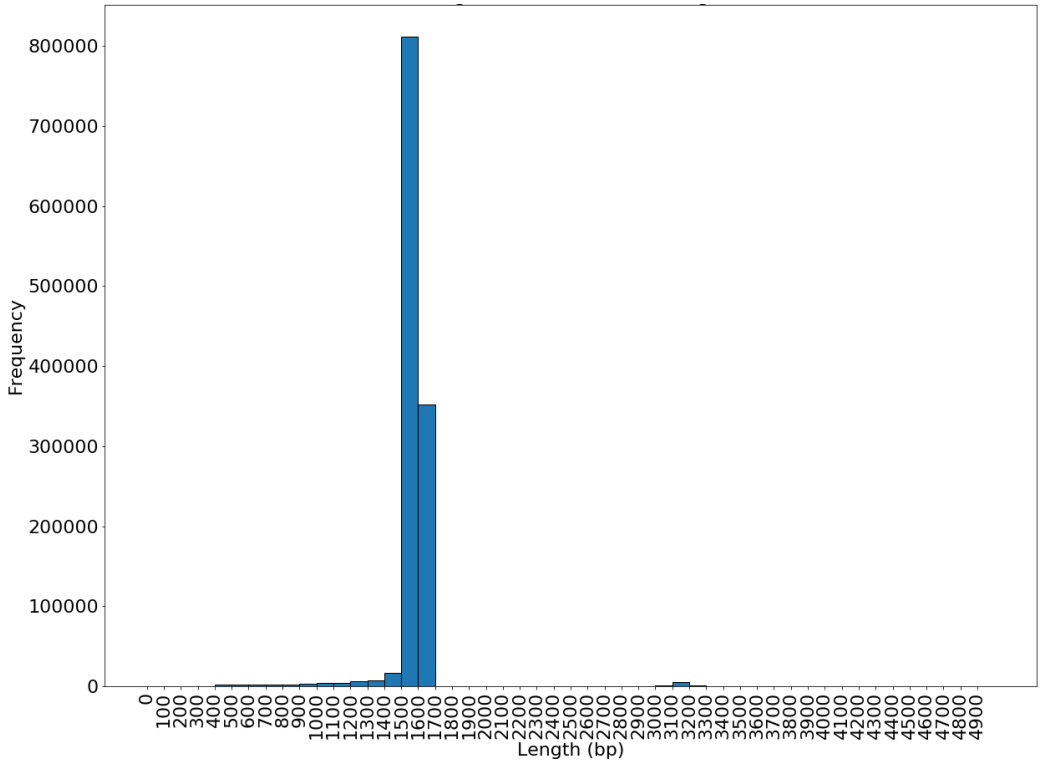


Figure 6.6. Histogram of read lengths for the King's Bath metagenome sample. Most reads are within the expected size range 1500-1700 bp. There is a small proportion of reads ranging from ~400 bp to 1500 bp and 3000-3300 bp.

Fragmented amplicons could be caused by the target gene sequence failing to fully amplify during PCR, or by mechanical shearing of the amplicons during its preparation for sequencing. Primer dimers are a common artefact in PCR which are formed when primer pairs hybridise due to the complementary nature of their sequences, these dimers can then serve as competing templates for the polymerase (Rychlik, 1995). This phenomenon can be exacerbated when the primers also contain complementary barcode sequences. If newly formed amplicons hybridise at these barcode regions, chimeric amplicons containing two 16S rRNA genes are formed. Additionally, bacteria can contain multiple copies of the 16S rRNA gene (Větrovský and Baldrian, 2013) and if two copies are in close proximity to each other on the chromosome, and the primer pair each hybridise to a different copy, a larger amplicon containing two 16S genes maybe formed.

Manual inspection of some of the larger and smaller reads using MEGAN (**Figure 6.7**) appeared to support the hypothesis that these reads are artefacts. This is because, as shown in **Figure 6.7**, the larger read has alignments with NCBI hits at two distinct regions and with no overlaps. These artefacts could distort the taxonomic classifications recorded for the sample, therefore the reads were filtered by size so that only those within the size range 1500-1700 bp were used for further analysis. This removed 60,451 reads leaving 1,163,873 for further analysis. After this, experimental software Ycard was used as a further chimera check. This returned 117 putative chimeric reads (0.01 % of total reads). However, manual inspection of these reads after alignment to the RefSeq 16S database did not confirm them as chimeric and so these reads were retained.

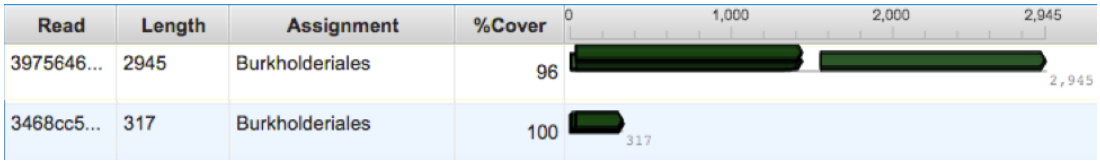


Figure 6.7. Screenshot showing example of alignment against NCBI hits of two reads outside of the expected size range in MEGAN. The above read is larger than the expected at 2945 bp whilst the below read is below expected size at 317 bp. The areas and position along each read that align to a NCBI hit are shown in green. There are two full-length alignments adjacent to each other for the larger read, suggesting that this read is a chimera of two full-length reads. The smaller read below is fully aligned with NCBI hits and is likely a fragmented read.

6.3.2.1.3. Analysis of Read Q-Scores

The basecalling and classification workflows were both set to a minimum quality score cut off of phred 10, meaning that each read has a minimum average per-base accuracy of 90 % across the entirety of its length (Ewing and Green, 1998). The histogram (**Figure 6.8**) shows that all of the reads returned met this criteria, with over half being between phred 10.0-10.6. In fact, the mean quality score of the reads (1500-1700 bp size) was calculated to be 10.52. This phred score is lower than that of other sequencing platforms commonly used in 16S profiling studies, such as Illumina. However, other researchers have reported comparable taxonomic assignment results between both platforms and have also observed higher taxonomic resolution using Oxford Nanopore 16S rRNA profiling alone. This is possibility due to the nearly full-length 16S rRNA gene reads produced by the Nanopore platform (Benítez-Páez, Portune and Sanz, 2016; Shin et al., 2016).

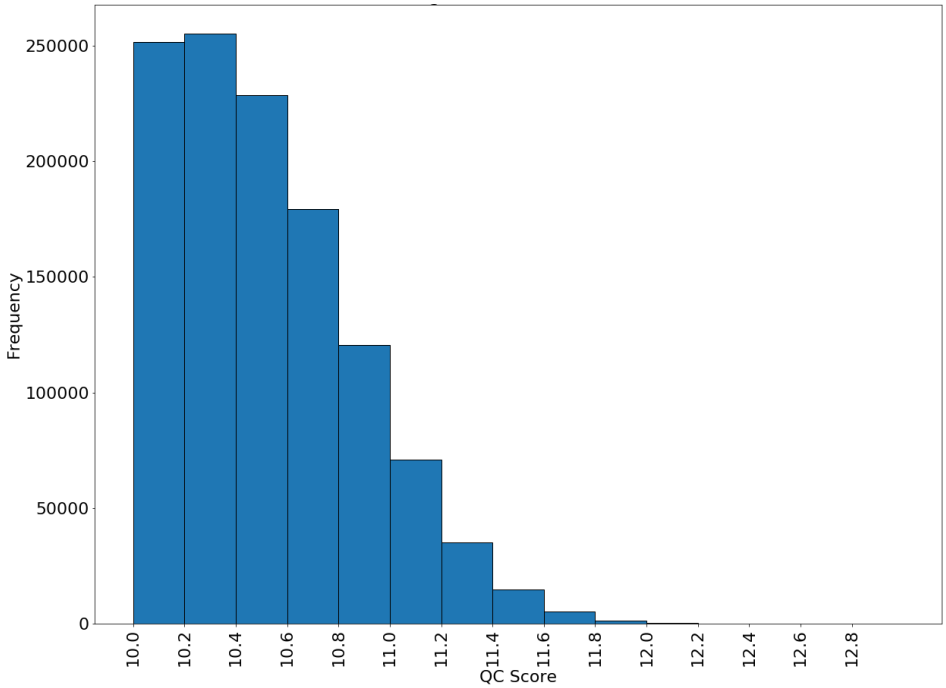


Figure 6.8. Histogram of read Q-scores for the King's Bath metagenome sample.

6.3.2.2 Taxonomy of King's Bath Metagenome

After filtering the data to select for reads in the size range 1500-1700 bp, the taxonomic classification of these reads by the EPI2ME workflow was analysed.

6.3.2.2.1. Taxonomic Resolution of Reads

Figure 6.9 shows that nearly all reads were resolved to the phylum (99.68 % of all reads) and class (99.40 %) ranks. However, there are large decreases in resolution at the order (42.94 %) and genus (5.65 %) ranks. Manual inspection of the data in MEGAN revealed that there were two major bottlenecks at the class Betaproteobacteria and the family *Rhodocycaeae* – which are both in the same lineage. At Betaproteobacteria, 58.75 % of the reads mapped to this rank did not resolve to lower ranks. Whilst at *Rhodocycaeae*, 93.87 % of reads did not resolve to lower ranks. Whilst loss of resolution at each successive taxonomic rank is to be expected, such large decreases in resolution may distort the reported distributions of lower taxonomic ranks. Therefore, it was important to further analyse these drops to determine what the cause may be. This analysis is detailed in **Section 6.3.2.2.4**.

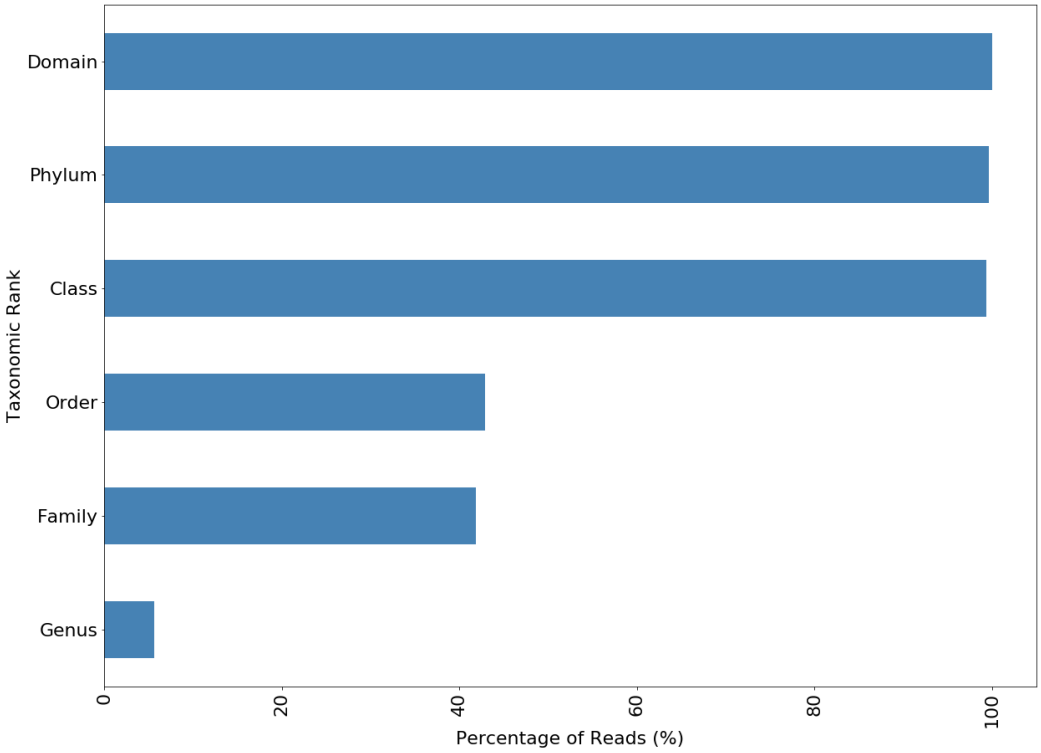


Figure 6.9. Bar chart showing percentage of classified reads mapped to each taxonomic rank. Each read is mapped to the rank of its NCBI classification and all higher taxonomic ranks. The figure shows significant reduction in read classifications between Class to Order and Family to Genus ranks.

6.3.2.2.2. Taxonomic Distributions of Reads

The data in **Table 6.4** and **Figures 6.10** and **6.11** show that the King's Bath metagenome returned a large number of taxonomic classifications, suggesting a rich and diverse microbiome is present in the water. No archaea classifications were returned, which is to be expected because the primer pair used in this study (27F & 1492R) are designed to amplify bacterial 16S rRNA genes. The 1492R primer is universal to both bacteria and archaea domains, but the 27F primer is specific to bacteria which means archaeal 16S rRNA genes are poorly amplified with this primer pair (Reysenbach, Longnecker & Kirshtein, 2000). Some studies into the microflora of thermal springs using 16S rRNA gene analysis have reported archaea to be both diverse and abundant (Barns et al., 1994; Meyer-Dombard, Shock and Amend, 2005; Mardanov et al., 2011). However, comparison of many such studies has shown the microbiome of thermal springs to be highly distinctive and some studies, such as one into the coastal thermal springs in Iceland, also found no archaea (Hobel et al., 2005).

The sample is dominated by Proteobacteria (**Figure 6.10**) which accounts for 98.06 % of all phylum mapped reads. Nitrospirae is the second most abundant phylum with 1.26 % of reads, whilst the other 21 phyla in the sample account for the remaining 0.68 %. Firmicutes and Actinobacteria, the phyla that Roman Bath isolates KB11 and KB16 belong (C**hapter 5**), account for 0.17 % and 0.05 % of reads respectively. The bacterium known as KB11 was isolated from the same swab sample as KB16 and was identified by another member of the lab group to be of the genus *Paenibacillus*.

Table 6.4. Table showing number of taxonomic classifications within each rank

Taxon Rank	Number of Classification
Phylum	23
Class	52
Order	102
Family	186
Genus	345

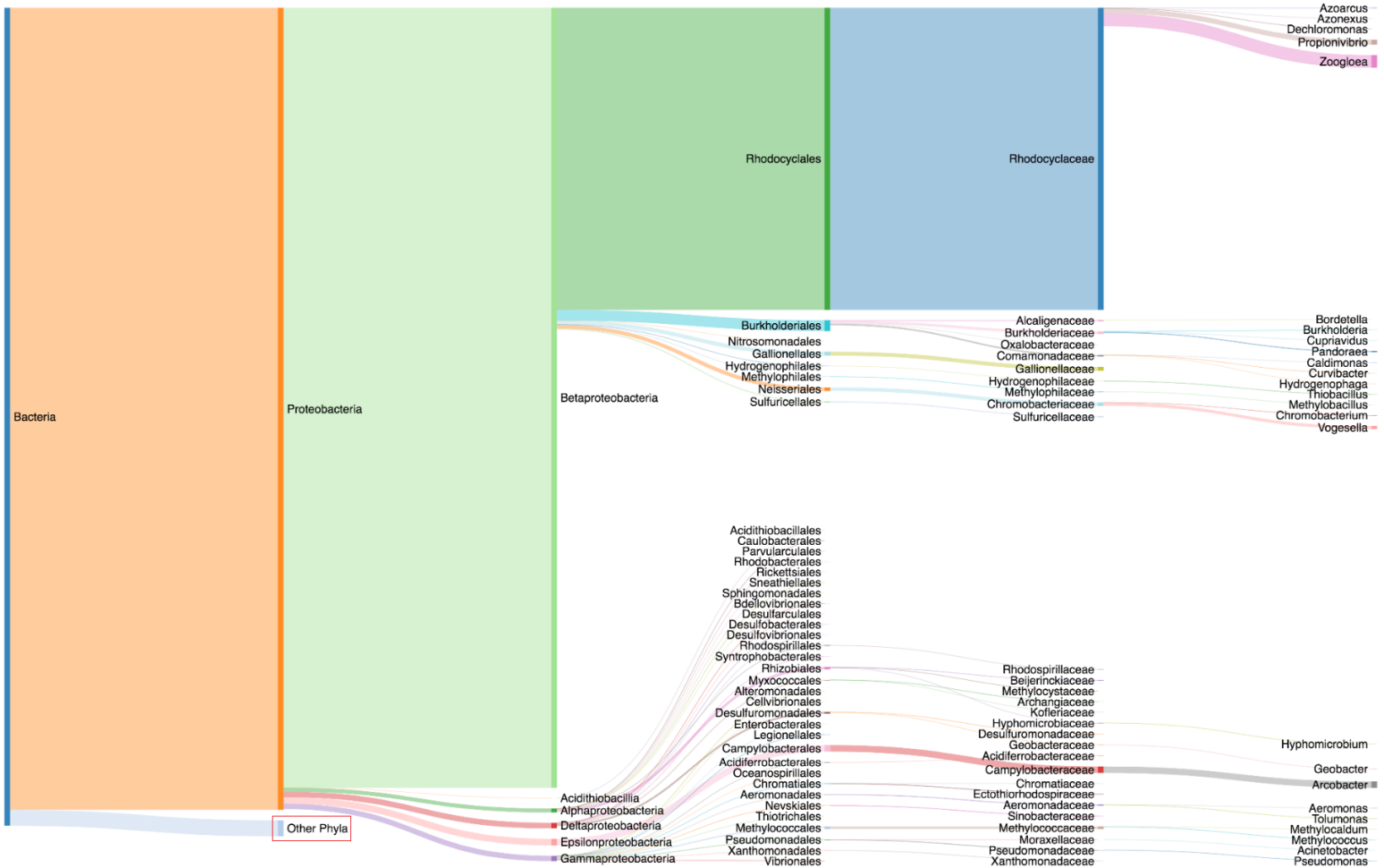


Figure 6.10. Sankey diagram of 16S rRNA gene taxonomic assignments that are ≥ 1% of the total of the King's Bath microbiome, with focus on the phylum Proteobacteria. The other 'Other Phyla' (highlighted by red box) are detailed in figure 6.11.

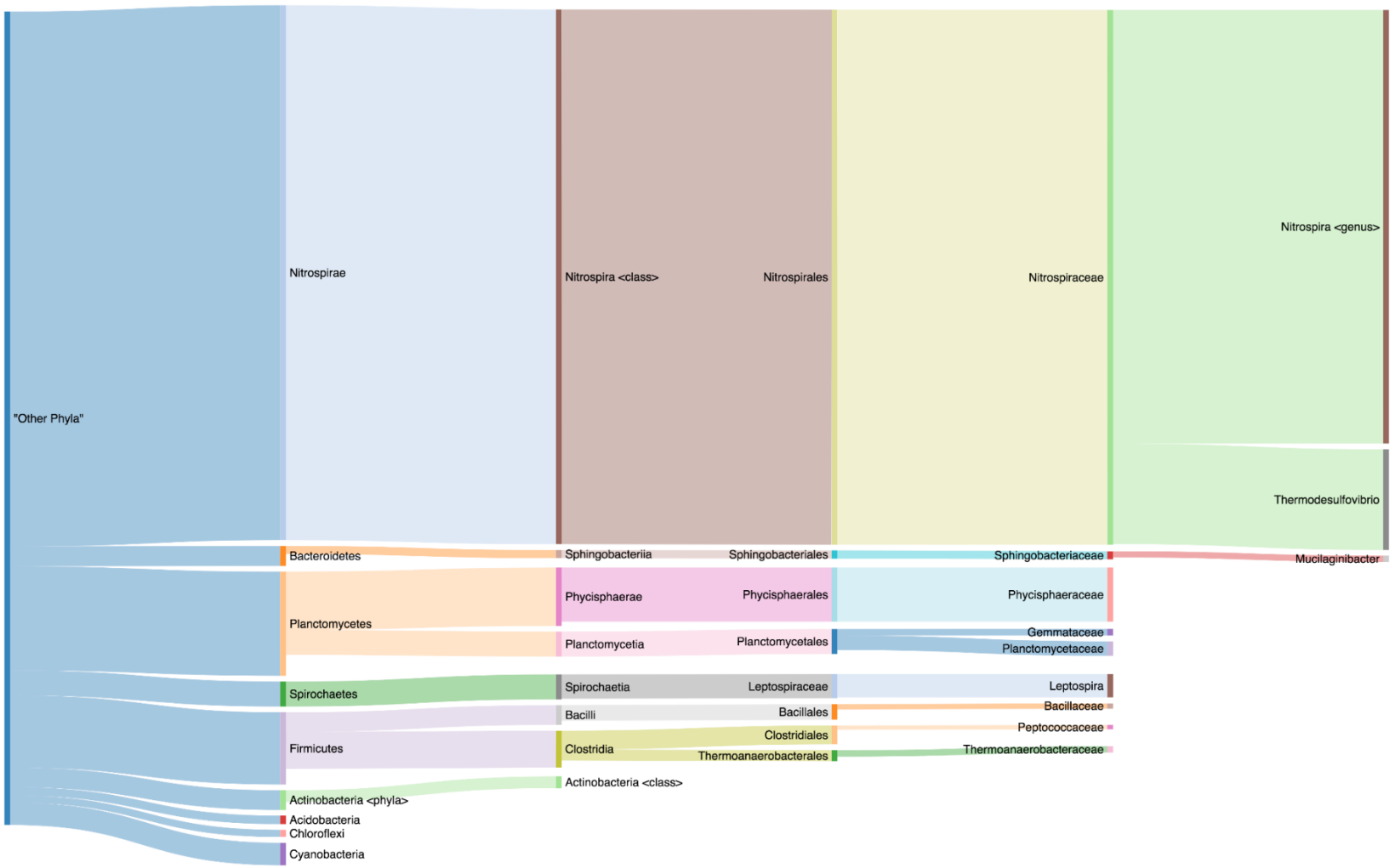


Figure 6.11. Sankey diagram of 16S rRNA gene taxonomic assignments that are ≥ 1% of the total of the King's Bath microbiome, focusing on 'Other Phyla' besides Proteobacteria.

The Proteobacteria phylum is dominated by the class Betaproteobacteria with 97.22 % of Proteobacteria reads mapped here. The remaining is made up of the classes Epsilonproteobacteria (0.81 %), Deltaproteobacteria (0.70 %), Gammaproteobacteria (0.66 %), Alphaproteobacteria (0.49 %). *Methylococcus capsulatus* (Bath) – isolated from the Roman Baths, is a member of the class Gammaproteobacteria. There was also a very low abundance of reads mapped to Acidithiobacillia (0.0003 %), which is the class that *Annwoodia Aquaesulis* and *Thermithiobacillus Tepidarius* – both isolated from the Roman Baths, belong (Wood and Kelly, 1986, 1988; Kelly and Wood, 2000; Williams and Kelly, 2013; Boden et al., 2016; Boden, Hutt and Rae, 2017).

Within the class Betaproteobacteria, the majority (58.75 %) of reads resolved at this rank, whilst 38.72 % mapped to the order Rhodocycales. Other orders mapped within this class are Burkholderiales (1.38 %), Gallonellales (0.48 %), Neisseriales (0.48 %), Sulfuricellales (0.12 %), Methylophilales (0.05 %), Hydrogenophilales (0.02 %), and a low abundance of Nitrosomonadles (0.002 %). The other Proteobacteria classes are divided into a further 31 orders. These orders contain numerous genera with diverse metabolic and phenotypic traits, further suggesting that the King's Bath contains a diverse microbial ecosystem.

6.3.2.2.3. Most Abundant Genera in King's Bath Metagenome

The most abundant genera, classified as those that mapped to ≥1 % of the total reads at Genus rank, are shown in **Figure 6.12**. Of the 345 genera returned from this sample, only 9 were classed as abundant with the remaining 336 making up 13.84 % of the total. Unsurprisingly considering its relative abundance (**Figures 6.10 & 6.11**), five of the listed genera - *Zoogloea*, *Azoarcus*, *Vogesella* and *Pandorea* – are Betaproteobacteria. While Arcobacteria and Geobacter are Epsilonproteobacteria and Deltaprotecobacteria respectively. Additionally, two of the genera, *Thermodesulfovibrio* and *Nitrospira*, are from the phylum Nitrospirae.

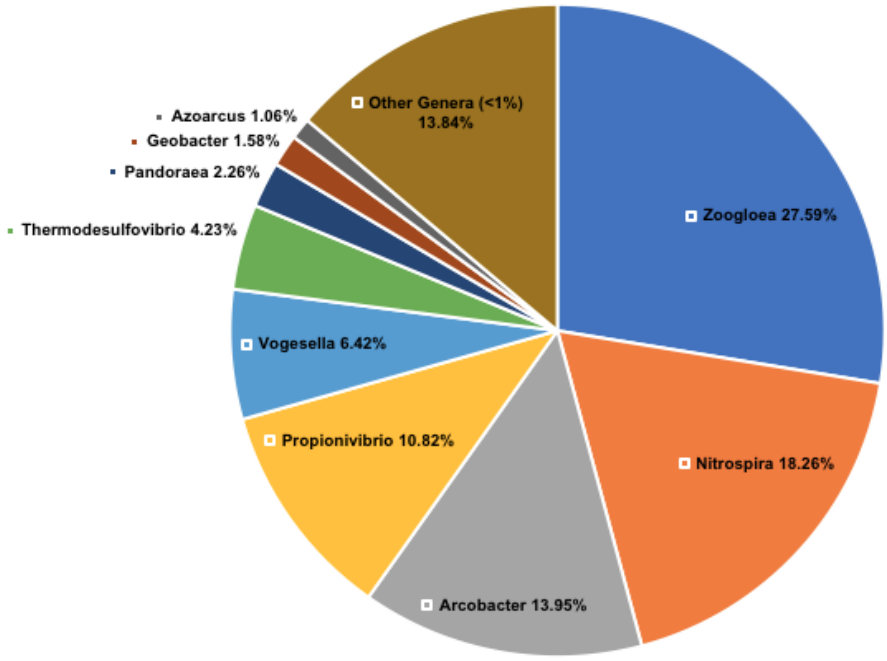


Figure 6.12. Pie chart showing the most abundant genera in the King's Bath metagenome. Percentages shown are reads mapped to this genus against total reads mapped at Genus rank.

Zoogloea is the most abundant genera accounting for 27.59 % of the reads from this rank. The genus is from the family *Rhodocyclaceae* within the order Rhodocyclales. *Zoogloea* are chemoorganotrophic and are straight to slightly curved rods that are 1.0-1.3 x 2.1-3.6 μm in size and contain a single polar flagella. The cells are free living and actively mobile and are found in waters that are organically polluted (Unz, 2015), and also in waste water treatment facilities (Madigan and Martinko, 1997; Unz, 2015). The presence of *Zoogloea* as the largest genera in this metagenomic sample therefore suggests that the composition of the water in the King's Bath waters may contain significant amount of organic chemical species. At present, published data on the waters has focused mainly on the composition of inorganic chemical species (**Tables 6.1-6.3**) (Andrews et al., 1982; Edmunds and Miles, 1991). The lifecycle and metabolism of *Zoogleoa* give it a biotechnological value in both bioremediation and the production of high value chemicals from waste water. *Zoogloea* cells amass into flocs known as zoogloeae through the production of a polysaccharide matrix. This matrix is able to absorb organic matter to lower biological oxygen demand of wastewaters (Madigan and Martinko, 1997). It has also been shown that zoogloeae are capable of absorbing heavy metals (Sağ and Kutsal, 1995; Ahn, Chung and Pak, 1998). Zoogloea also accumulates intracellular granules of poly-β-hydroxybutyrate during its growth. These granules could be used as precursors for bioplastics or biofuels (Muller et al., 2017). Some strains of *Zoogloea* have also been shown to fix nitrogen (Sağ and Kutsal, 1995). Considering the high abundance of nitrogen detected in the waters of the Roman Baths (**Tables 6.1-6.3**) (Andrews et al., 1982; Edmunds and Miles,

1991) it is possible that *Zoogloea* in the King's Bath may also be able to metabolise nitrogen. *Azoarcus* is from the same family as *Zoogloea* and is the ninth most abundant genera, accounting for 1.06 % of reads. Members from this genus have also been reported as capable of accumulating poly-β-hydroxybutyrate granules and fixing nitrogen (Reinhold-Hurek, Tan and Hurek, 2015).

Propionivibrio is the fourth most abundant genera found in the metagenome sample, accounting for 10.82 % of genus mapped reads. It is from the same family as *Zoogloea* and *Azoarcus*. Cell morphology is curved or straight rods with a single polar flagellum (Whitman et al., 2015b). The genus is found in aquatic environments and has an anaerobic metabolism through a novel pathway that degrades hydroareomatic compounds to produce Propionate and acetate (Brune, Ludwig and Schink, 2002). Species in the genus have been discovered that are capable of accumulating glycogen (Albertsen et al., 2016) and reduce perchlorate (Thrash et al., 2010).

Nitrospira from the family *Nitrospiraceae* in the order Nitrospirales is the second most abundant genus with 18.26 % of reads. Cell morphology is vibrio or spiral shaped rods between 0.2-0.4 x 0.9-2.2 μm in size. They are nonmotile and found in diverse aquatic environments such as oceans, freshwater, wastewaters, and heating systems. They have also been found in soil samples (Spieck and Bock, 2015). *Nitrospira* are aerobic and their metabolism involves the oxidative breakdown of ammonia via nitrite to nitrate. Most known *Nitrospira* species catalyse the second step of this process from nitrite to nitrate only, however some *Nitrospira* species have been recently discovered that can catalyse the complete nitrification of ammonia to nitrate (Daims et al., 2015; Daims and Wagner, 2018). These strains could have a biotechnological value in bioremediation of ammonia. Also, from the same family as *Nitrospira* is *Thermodesulfovibrio*. It is the sixth most abundant genus, accounting for 4.23 % of the total genus reads in the metagenome. The cell morphology is curved or vibrio rods that may appear singly or in chains, and have a single polar flagellum. The genus is defined as a thermophile able to grow in the range 40-70 °C with an anaerobic metabolism that involves reduction of sulphates (Oliver, Wells and Maxwell, 2014). Edmunds & Miles (1991) (**Table 6.1**) reported the average temperature of the waters of the King's Spring to be 46 °C, and that the water contains high abundancies of both nitrogen and sulfates. Such conditions are well suited to support the growth of genera such as *Nitrospora* and *Thermodesulfovibrio*, and it is therefore unsurprising to find these two genera to be major components of the microbiome of the sample.

Arcobacter from the family *Campylobacteraceae* in the order Campylobacterales is the third most abundant genera with 13.95 % of reads. Cell morphology is curved rods between 0.2–0.9 × 0.5–3 μm in size and containing a single polar flagellum (Vandamme et al., 2015). They are found in a diverse range of habitats including freshwater, sewage, plants, and animal reproductive tracts, and some have been found to be human and animal pathogens (Madigan and Martinko, 1997). Species in the genus have been shown to exhibit diverse metabolisms, some of which are rare. For example, *Arcobacter* species have been characterised that are nitrogen-fixing (McClung, Patriquin and David, 1983; Vandamme et al., 1991), that can oxidize sulphides and produce sulphur (Wirsen et al., 2002), that are halophilic (Donachie et al., 2005), and that can reduce perchlorate (Carlström et al., 2013). The finding of a genus with such metabolic diversity as abundant in the sample suggests the potential to find novel natural product chemistry and biocatalysts in the King's Bath.

Vogesella, from the family *Neisseriaceae* in the order Neisseriales, accounts for 6.42 % of genus mapped reads. Cell morphology is straight rods, 0.5 x 3.5 μm in size, with a single polar flagellum. *Vogella* has an aerobic chemoorganotrophic metabolism and they is capable of converting nitrate into nitrogen gas (Thrash et al., 2010). The genus is found in aquatic environments (Grimes et al., 1997; Chou et al., 2008; Jørgensen et al., 2010; Sheu et al., 2013) and strains have been characterised that are capable of degrading biopolymers such as peptidoglycan and chitin (Jørgensen et al., 2010). Chitin is a polysaccharide that is a component of fungal cell walls, insect exoskeletons, and crustacean shells. Chitinases therefore have an economical value as potential antifungal agents, and in bioremediation of waste from the seafood industry (Kim, 2013). While novel enzymes involved in degradation of peptidoglycan could have a value as alternative antibiotics in the food and agricultural industries (Oliver, Wells and Maxwell, 2014).

Pandoraea is the seventh most abundant genera, accounting for 2.26 % of reads. It is in the family *Burkholderiaceae* in the order Burkholderiales and is noted as an emerging opportunistic respiratory pathogen of Cystic Fibrosis patients (Martina et al., 2017). The cells are rods between 0.5–0.7 × 1.5–4.0 μm in size and are motile by means of a single polar flagellum (Whitman et al., 2015a).

Geobacter is the from the family *Geobacteraceae* in the order Desulfuromonales and accounts for 1.58 % of genus reads, which makes it the eight most abundant genera reported in the metagenome. Cells are rod-shaped, 1-4 x 0.6 μm, nonmotile, and can appear either singularly, in pairs, or in chains. The genus are anaerobes that utilise ferric or other available metals as their electron acceptors (Coates and Lovley, 2015). Aspects of their metabolism make them

of high biotechnological value. For example, *Geobacter* can oxidise a range of compounds such as metals, radioactive elements, and petroleum by-products into CO_2. This ability makes them useful in the bioremediation of toxic industrial wastes (Cologgi et al., 2014) and is in keeping with chemical analysis of the waters of the Roman Baths which have reported the presence of radioactive ions, metals, and hydrocarbons (**Tables 6.1-6.3**) (Andrews et al., 1982; Edmunds and Miles, 1991). Some strains of *Geobacter* can form pili that have conductive capability. It's been proposed that these conductive pili serve as nanowires for accumulation and sharing of metal ions. Some species also form conductive biofilms and some researchers are currently focusing on how these can be utilised in microbial fuel cells (Yi et al., 2009).

6.3.2.2.4. Analysis of Reads Resolved at Betaproteobacteria and *Rhodocycaeae*

As previously stated in **Section 6.3.2.2.1.**, there are two major bottlenecks in the taxonomic classification of the reads. At the rank Betaproteobacteria 58.75 % of the mapped reads were not mapped to a lower rank, and this was also the case with 93.87 % of *Rhodocycaeae* reads. This could be taken to suggest that there are insufficient identity matches in the database used. This could suggest there is an abundance of uncharacterised bacterial species in the King's Bath. However, it is also possible that this is an artefact due to the workflow used and if that is the case then it means that the distributions of the taxonomic classifications reported could be distorted. To investigate this, the reads that were resolved to the ranks Betaproteobacteria and *Rhodocycaeae* were analysed using a BLASTn search against the RefSeq 16S gene database – which is the same databased used by the EPI2ME workflow.

The BLAST searches returned hits for Betaproteobacteria reads across 123 genera. The lowest percentage identity score returned was 80.06 %, and the highest was 87.44 %, and the mean across all the hits was 83.37 %. For *Rhodocycaeae* reads, 73 genera were returned, with the lowest identity score 81.37 %, the highest 88.68 %, and the mean 84.01 %. These percentage identity scores are comparable to the scores returned by the EPI2ME workflow, which suggested that the reason the reads were not classified to lower ranks was not because there were insufficient identity matches in the database.

Figure 6.13 reveals that the most abundant genera retuned from both searches is *Sideroxydans*, while *Dechlorobacter* is the fourth and second most abundant in for Betaproteobacteria and *Rhodocycaeae* respectively. Neither of these genera feature in the results returned from the EPI2ME workflow. *Sideroxydans* and *Dechlorobacter* are two proposed but as yet unvalidated taxa which are reported to oxidise Fe(II) and reduce

perchlorate respectively (Weiss et al., 2007; Thrash et al., 2010). As proposed taxa, information is scarce and there are only two entries for each in the RefSeq 16S gene database. The EPI2ME workflow is designed to assign each read to the lowest common ancestor (LCA) of the top three matches from the RefSeq database. This will therefore be the reason that these two taxa were not reported by the EPI2ME workflow.

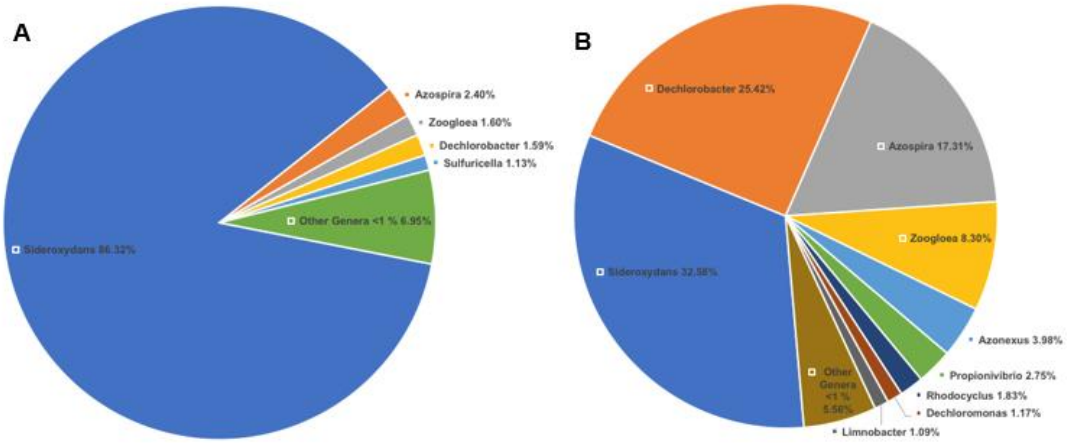


Figure 6.13. Pie charts showing the top hits by genus returned from the BLASTn searches of A) the Betaproteobacteria resolved reads and B) the *Rhodocycaeae* resolved reads against the RefSeq 16S gene database. Percentage values are the number of reads assigned to each genus. Genera with abundance <1 % are grouped as 'Other Genera'.

These findings suggest that the King's Bath metagenome sample may contain a lot more taxonomic diversity than what has been reported by the EPI2ME workflow. Not just in terms of under characterised taxa, but also in potential new taxa.

6.3.2.3. Discussion of 16S rRNA Gene Profiling of King's Bath Metagenome

The prokaryotic diversity of the King's Bath was measured using 16S rRNA gene sequencing. Over 1.2 million reads were classified across 345 genera and 23 phyla, which suggests that the King's Bath contains a diverse microbial ecosystem. The most dominant taxa in the sample was the phylum Proteobacteria which accounted for 98.06 % of phylum reads, with Nitrospirae a distant second (1.26 %).

The most abundant genera found in the sample were *Zoogloea*, *Nitrospira*, *Arcobacter*, *Propionivibrio*, *Vogesella*, *Thermodesulfovibrio*, *Pandoraea*, *Geobacter*, and *Azoarcus*. This is a diverse group of genera which are reported to have differing metabolisms. However, the metabolic characteristics that have been associated with these genera are in keeping with the

reported geochemistry of the waters. For example, some of the genera (*Zoogloea, Nitrospira*, *Vogesella*) have been reported to utilise nitrogen chemical species in their metabolism - which is abundant in the waters. Sulphate is abundant in the water at over 1 g/L and two of the genera (*Thermodesulfonvbrio*, *Arcobacter*) are reported to utilise this in their metabolism. The waters also contain a wide variety of metallic elements and many of the genera are reported to utilise different metals to a different extent. The known preferred temperature conditions of these genera are also in keeping with the reported temperature of the waters.

These findings are in keeping with other studies into the microbiomes of thermal springs which have found the most dominant taxa to have metabolic traits that utilise the most abundant chemical components of the water (Meyer-Dombard, Shock and Amend, 2005; Mardanov et al., 2011). The presence of certain genera may suggest additional information about the geochemistry of the waters. For example, the abundance of *Zoogloea*, which is reported to utilise organic compounds, suggests such compounds may be abundance in the waters. Perchlorate is reported to be utilised by *Arcobacter*, but is not reported to be present in the published chemical analysis (**Tables 6.1-6.3**) (Andrews et al., 1982; Edmunds and Miles, 1991).

However, it is important to note that phylogenetic diversity alone is not an accurate presentation of phenotypic diversity or secondary metabolite potential. As exemplified in the work of **Chapters 3 & 5** which revealed BGCs in genera previously unassociated with their phylogeny, and is also in accord with reports from other researchers (Speike, 2015; Belknapp et al., 2020). Horizontal gene transfer between unrelated species in different taxonomic clades can occur, meaning traits not previously seen in some genera can be discovered (Doroghazi & Buckley, 2010). Furthermore, there are examples of phenotypic traits not usually associated with some of these genera being discovered in members - such as members of *Zoogloea* being shown capable of fixing nitrogen.

Therefore, assumptions should not be made about the potential metabolism of the microbiome of the King's Bath without further investigations. However, the diversity of genera recorded in this 16S rRNA gene profiling analysis, and the fact that many have previously been earmarked as having potential biotechnological uses, does suggest that the King's Bath merits further investigation.

The phyla Actinobacteria and Firmicutes, from which the two antibiotic producing isolates previously isolated in 2015 came from (the Firmicute isolate is not featured in this thesis), make very small proportions of the sample. This finding demonstrates how the culturing

methods can play a major role in the isolation of microorganisms from the environment. These two isolates were recovered by culturing on nutrient agar at room temperature. Such conditions will favour nutritionally versatile microbes over ones with metabolisms more specialised to the conditions of the King's Bath. The two sulphur bacteria isolated from the Roman Baths in the 1980s (Wood and Kelly, 1986, 1988; Kelly and Wood, 2000; Williams and Kelly, 2013; Boden, Hutt and Rae, 2017) belong to the class Acidithiobacillia which also makes a small proportion of the microflora of sample. These isolates were cultivated using conditions that selected for bacteria with these desired traits. Now that an understanding of the microbial composition of the water has been gathered, this information can be used to support future efforts to cultivate microorganisms from the water. Different cultivation conditions could be used to select for organisms from particular taxa of interest that are shown as being present in the sample based on past precedent.

Analysis of read assignments revealed two large bottlenecks in read assignments at the ranks Betaproteobacteria and *Rhodocycaeae*. Further investigation of these reads revealed the majority of them to match to the proposed genera *Sideroxydans* and *Dechlorobacter*. These two genera also have reported metabolisms that are in keeping with the chemical composition of the King's Bath waters. The finding that large numbers of reads could not be assigned to verified lower taxonomic levels suggests that there could be some, as yet, uncharacterised taxa in the sample. However, it also reveals a potential limitation in the EPI2ME assignment workflow which could distort the taxonomic distributions recorded. This workflow is designed to assign each read to the lowest common ancestor (LCA) from its the top three matches to the RefSeq 16S rRNA gene database, irrespective of the relative strength of the three matches. This approach results in a more conservative assignment of reads which can be advantageous in ensuring high accuracy, but also has a disadvantage in that poorly characterised taxa that have two or less entries in the RefSeq 16S rRNA gene database are lost entirely from the final results. Modifications to the LCA assignment algorithm which applies a weighting to each match based on its relative strength may improve accuracy of assignment and avoid taxonomic distortions. However, despite the bottlenecks seen in this study, the most abundant genera found came from the most abundant higher taxa – suggesting that there is consistency in classifications.

Another potential limitation in this workflow is the lack of third-party chimera detection software optimised and verified to process Nanopore data. Currently, third party software chimera detection software optimised to process long-reads is in its infancy and accuracy is unvalidated. A packaged called Ycard was tested on this dataset, which flagged 117 reads (0.01 % of total) as potentially chimeric. However, manual inspection of the these reads in

BLAST showed that the reads were not chimeric. To mitigate this issue, the reads were selected by size within the 1500-1700 bp range before further analysis after manual inspection. While this would remove the vast majority of suspected chimeric reads, it may not be a perfect solution as some chimeras may be within that size range. This can happen due to partial amplicons from previous PCR cycles being used as primers for the amplification of other templates in the sample during subsequent cycles. However, in the case of this particular study, the minimum identity match threshold was 77 % whilst the average match of the reads used in this study is 86.14 %, which further reduces the likelihood that classified reads in the selected size region are chimeric. Also, some reads may have been unduly excluded as the size selection makes that assumption that all 16S rRNA genes would conform to this size range. There is a possibility that there may be some that are outside of this size range due to very large or small hypervariable regions.

The results give an indication of the relative abundances of taxa present. However, it is important to note that the results do not provide a definitive assessment of the microbiome of the water as a whole. This is because of aspects of the study, which are true for all 16S metagenomic studies, that may introduce basis into the results recorded. For example, water was taken that the bath from a shallow depth of 6-10 cm because it is not possible to enter the water due to public health concerns. The composition of bacteria may appear different at different locations and depths of the bath.

The method to extract DNA from the microbes could also bias results. The method chosen in this study used a mechanical lysis method to break the cells open. It is believed that this method, being more aggressive than chemical or heat treatment alone would ensure that DNA was extracted from as wide a range of organisms as possible. However, as some bacteria will be more susceptible to lysis than others (Han et al., 2018), there is a possibility that more of those cells are lysed, which will distort the relative abundances of taxa recorded. A mock microbiome standard has been used by some researchers to attempt to optimise DNA extraction procedures (Fouhy et al, 2016). However, environmental microbiomes are extremely complex and variable communities are often unknown. This makes it extremely difficult to have confidence that any DNA extraction approach optimised for a control sample would be optimal for any test sample.

Additionally, the PCR reaction may also introduce bias as some templates may be easier to amplify that others due to factors such as GC content and secondary structure formation (Laursen, Dalgaard and Bahl, 2017).

Another factor to consider is 16S gene copy number. Many bacterial species contain multiple copies of the 16S rRNA gene, and this can distort the taxa distributions recorded as the species with more 16S copies will have more templates that can be amplified by PCR (Kembel et al., 2012; Větrovský and Baldrian, 2013).

Enumerating the different bacterial abundances present within the water would not be possible from this dataset because of the reasons of DNA extraction and PCR bias, and variable 16S rRNA gene copy numbers mentioned above. It is also not possible to accurately quantify the concentration of bacteria present in the water from the total mass of DNA extracted from the water sample. This is because it cannot be assumed that the DNA extraction procedure recovered only bacterial DNA with 100 % efficiency. In addition to this, the genome sizes of different bacterial species vary.

Also, by their nature, 16S rRNA gene metaprofiling studies focus only on prokaryotic organisms and exclude fungi and viruses which also play an import role in the microbial ecosystem and may also produce novel natural products of interest. However, while these aspects of the study may bias the results obtained, it is important to stress that these issues are true of any 16S rRNA gene metaprofiling investigation.

Nonetheless, despite the limitations discussed above, the results do provide a comprehensive indication of what prokaryotic genera maybe present in the waters. The results also provide a basis for comparison against the taxonomic data returned from shotgun metagenomic sequencing after amplification of the metagenome sample in order to determine if this approach influences the taxonomic profile seen. This is because shotgun sequencing could potentially offer more accurate taxonomic classification through the direct profiling of all the sequenced material (Segata et al., 2003; Handelsman, 2004). Also, it may remove the bias of requiring annotated 16S rRNA marker genes to identify species potentially allowing for the identification of eukaryotic and virial members of the microbiome. However, the metagenome will require amplification to increase input mass first which may itself bias the results produced.

6.3.3. Shotgun Metagenomic Sequencing of King's Bath Metagenome

Long-read shotgun sequencing of the metagenome was performed in order to attempt to identify *in silico* long BGCs such as PKS and NRPS pathways to get a more accurate understanding of the bioprospecting potential of the site.

In order to generate sufficient input mass for shotgun sequencing, a multiple displacement amplification (MDA) protocol was performed on the King's Bath metagenome sample. The reason for attempting MDA as the method to amplify the DNA extraction is because it can offer the potential to generate large amounts of DNA rapidly and without theoretical limits to the length of DNA fragments that are generated. Thus, fully leveraging the benefits of Nanopore long-read sequencing for identifying long BGCs by shotgun sequencing.

6.3.3.1. Analysis of Multiple Displacement Amplification of King's Bath Metagenome Sample

Multiple displacement amplification of the King's Bath metagenome sample was performed to increase DNA content for sufficient shotgun sequencing. The nature of MDA results in hyperbranched DNA structures (Lasken & Stockwell, 2007). Branched DNA strands would not be conducive to sequencing on the MinION platform as these branches may cause blockages in the flowcell when passed through its pores. Therefore, the amplified sample was digested using T7 endonuclease to attempt to remove those branched structures, and the linearised HMW DNA purified from the digest using magnetic SPRI beads size selection. The success of this process was monitored by gel electrophoresis and Nanodrop.

Initial results showed that while the DNA content of each sample had increased, the concentrations and purities of the samples were below recommended levels for efficient library preparation. Therefore, two further clean-up and concentration steps were performed using a 1:1 ratio of SPRI beads until concentrations were sufficient. **Table 6.5** and **Figure 6.14** details the results of the Nanodrop and gel electrophoresis analysis of those final concentrated products. The results showed that DNA concentration was broadly the same for all the samples, including the blank DNA extraction that had been made at the time of the initial DNA extraction of the King's Bath biomass as a purity control, as well as in the pure molecular water which had been used as a sterility control for the MDA procedure.

Any low-level contaminating DNA introduced into a sample by the equipment or reagents used may also be amplified, and this can also affect the reliability of subsequent downstream analysis (Binga, Lasken & Neufeld, 2008). The fact DNA is present at the same concentrations in both purity control samples suggests that this had indeed occurred. The fact that it is in both the pure water and blank DNA extraction control further suggests that contamination could have been introduced by the MDA reagents themselves as DNA contamination of pre-packaged molecular biology reagents has been reported by researchers (Glassing et al., 2016). Later enquires made to the reagent manufactures confirmed that the reagents used are not certified free of DNA (private communication).

The initial input amounts of the King's Bath and mock microbiome metagenome (**Section 6.2.5**; ZymoBIOMICS Microbial Community DNA Standard) samples was approximately ~10 ng. With elution volumes of 20 µl, the approximate amount of DNA in these samples is ~1 µg. It can be concluded from this that the DNA material in these samples had been successfully amplified and that there was sufficient material for shotgun sequencing. The non-specific amplification and hyperbranching caused in the MDA process can produce chimeric structures (Lasken & Stockwell, 2007), which may affect subsequent taxonomic classifications, genome assembly, and annotation of the sequenced material. The use of the mock microbiome metagenome which contains DNA from known bacteria and known amounts was used to assess if MDA amplification may affect taxonomic profiling.

Purity scores for all the samples were broadly similar too, with the 260/280 slightly higher than the optimal value of ~1.8. This may be due to residual carry over of SPRI beads which had been used in the clean-ups. Analysis by gel electrophoresis (**Figure 6.14**) showed that all the samples contained a strong band concentrated at the 25 kbp marker which is suggestive of HMW DNA, however, there was also smearing present that continued to 100 bp. This suggests that short linearised fragments caused by the T7 digestion had not been fully removed by the size-selection SPRI bead clean-ups. However, after three clean-up cycles it was decided to proceed with library preparation and sequencing.

Table 6.5. Results of Nanodrop analysis of samples after multiple displacement amplification, T7 endonuclease digestion, and after three SPRI bead size selection and clean-ups

Sample	Concentration (ng/µl)	260/280	260/230
King's Bath metagenome	55.1	2.00	2.82
Blank DNA extraction	58.1	1.99	2.77
Molecular water MDA control sample	57.9	1.99	2.84
Mock microbiome metagenome	57.2	1.96	3.05

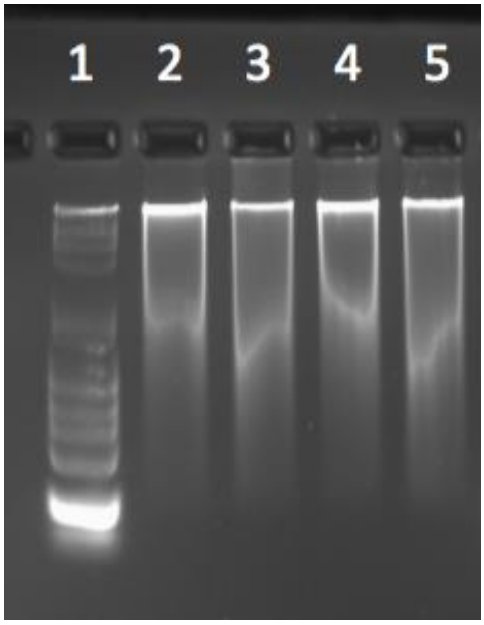


Figure 6.14. 1 % agarose gel of 1 µL of the products of the samples after multiple displacement amplification, T7 endonuclease digestions, and SPRI bead size selection and concentration x 3. 1 = molecular marker, 2= King's Bath metagenome, 3 = blank DNA extraction, 4 = molecular water MDA control sample, 5 = mock microbiome metagenome (ZymoBIOMICS Microbial Community DNA Standard)

6.3.3.2. Analysis of Sequencing Performance of MDA Bath Metagenome Sample

All MDA samples and native mock microbiome metagenome were sequenced using two flowcells each for approximately 22 hours making 44 hours of sequencing in total. **Table 6.6** summarised the throughput of the sequencing run. The throughput for the sequencing run was low, with a total of 1,849,023 reads generated with an average length of 3,044 bp. This resulted in a yield of 5.6 Gbp. Of these reads only 31.2 % passed the quality threshold (q10). Also, 1,093,491 reads could not be assigned to one of the barcoded samples during demultiplexing which is a very high attrition rate of 59 %. Of the reads that passed filtering, the percentages of reads that could be assigned to a taxa varied between the samples. Taxa assigned reads were lowest in the King's Bath metagenome sample (31.9 %) whilst the reads assigned in the purity

controls were both over double this (64.2 % & 70.1 %). Assigned reads was highest for the mock metagenome samples with nearly all assigned (96.4 % and 97.4 %). This suggests that the King's Bath metagenome sample may contain taxa that are not present in the NCBI RefSeq database. Especially when compared to the higher percentages of reads assigned in the purity controls.

These results were below expectations and seen after sequencing the samples across two flow cells. This suggests that an issue with the samples may have affected sequencing performance. This could have been caused by contaminants in the samples affecting the quality of the library prep and sequencing. Another possible explanation is that T7 endonuclease digestion may have failed to fully linearise the branched DNA. These branched structures could have led to strands being caught in the flowcell pores and impeding further sequencing. Additionally, the endonuclease digestion may have introduced nicks in the ends and along the lengths of linearised DNA strands. Inadequate repair of these nicks can affect quality of sequenced strands.

Table 6.6. Key metrics of the throughput of the reads sequenced from the MDA prepared samples. Table shows high yield of reads assigned to the control samples in comparison to the King's Bath and mock samples. Additionally, a significant number of reads were not assigned to a sample.

Sample	Reads sequenced	Yield (Gbp)	Average length (bp)	% reads passed quality threshold <q10	% passed reads taxon classified
King's Bath metagenome	209,167	0.7	3,348	45.3	31.9
Blank DNA extraction	220,319	0.8	3,470	36.7	64.2
Molecular water MDA control sample	153,215	0.5	3,118	35.5	70.1
MDA prepped mock microbiome metagenome	166,916	0.5	2,790	52.9	96.4
Native mock microbiome metagenome	5,863	0.008	1,374	58.8	97.4
No barcode*	1,093,491	3.2	2,938	23.3	68.9
Total	1,849,023	5.6	3,044	31.2	66.6

To better understand how the MDA process may affect taxonomic classification, the mock microbiome metagenomes where analysed and compared to the expected taxonomic profile from the reference file (**Figure 6.15**).

The results showed that the native mock microbiome metagenome sample which had not been amplified by MDA prior to sequencing had a taxa profile very similar to the expected profile seen in the reference file with all the expected genera reported at relative abundancies which were similar to the reference. This suggests that shotgun sequencing can provide a reliable taxonomic profile of microbiomes and that the library prep methodology did not introduce significant bias. Also, because the number of reads obtained for the native sample was lower than for the others (3,356), these findings may also suggest that a high read depth is not required to obtain a reliable taxonomic profile from shotgun metagenomic sequencing. These findings are significant as it demonstrates that the use of Nanoprobe sequencing can provide a resource-efficient means of profiling the microbiome of potential bioprospecting sites.

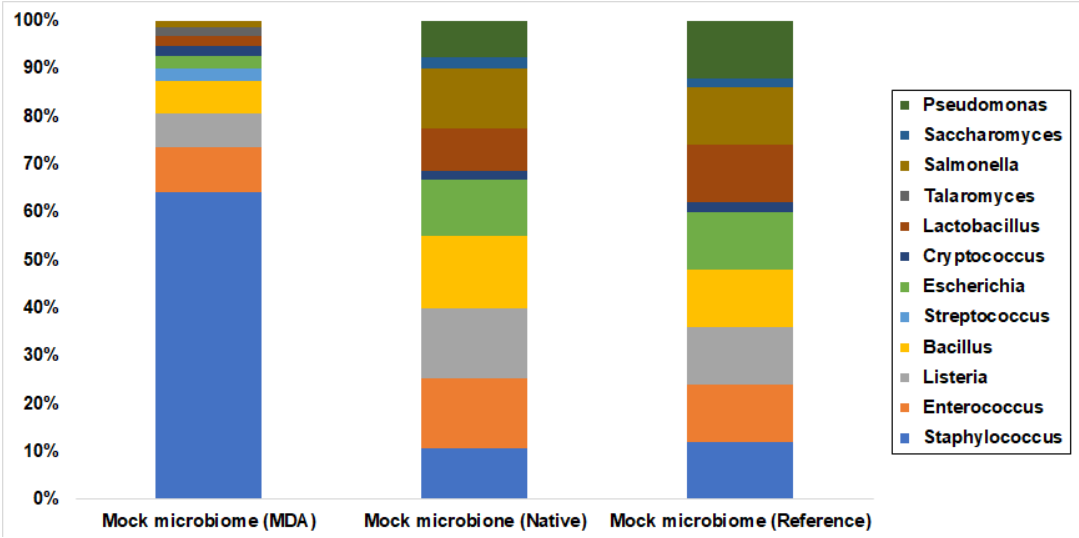


Figure 6.15. Comparison of the distribution of assigned genera of the MDA mock microbiome shotgun metagenomes in comparison with the mock microbiome reference taxonomy. Taxonomy profiling was based on reads that mapped to the top ten genera at ≥ 1% abundancy.

The mock microbiome metagenome which had been amplified by MDA prior to shotgun sequencing had a taxonomy profile which differed to both the native and reference. Most notable was that *Staphylococcus* was heavily overrepresented. Additionally, two genera present in the reference and reported in the native metagenome; *Saccharomyces* and *Pseudomonas* were not amongst the top ten genera reported whilst two erroneous taxa; *Streptococcus* and *Talaromyces* were reported as present. The misreporting of the genera could be due to bias caused by the amplification reactions and also due to contamination which may have been introduced into the sample by the MDA process.

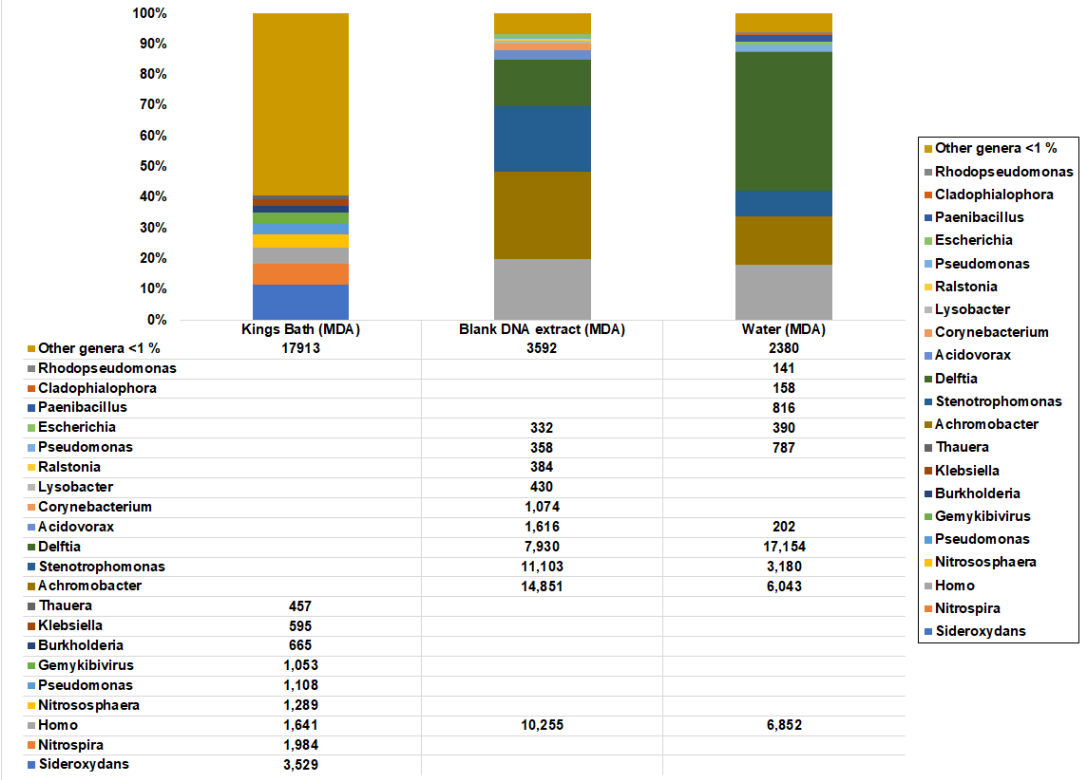


Figure 6.16. Comparison taxonomy profiles of the distribution of assigned genera of the MDA King's Bath, Blank DNA extract, and water shotgun metagenomes. Reported taxa of purity controls are similar to each other differ from the King's Bath.

The top classified genera in the MDA King's Bath metagenome were compared to those of blank DNA extract and pure water samples in order to understand the DNA contamination that may be present. **Figure 6.16** shows the results of this comparison. The results showed that the two control samples contained similar taxa profiles which differed to the taxa profile of the King's Bath metagenome. A consistent trend in all three samples was the presence of human (homo) DNA.

The similar and diverse taxonomic profiles reported for the two purity control samples does suggest that there was contaminating DNA present which has been amplified during the MDA process. This contaminating DNA may have been introduced by the reagents or equipment that was used in the MDA process. While it does not appear in the most abundant taxa reported for the MDA King's Bath sample, some of these contaminating genera do appear at lower abundancies. However, upon manual inspection of the full dataset the two most abundant genera reported in the test sample in both the shotgun sequencing and the 16S rRNA gene typing, *Sideroxydans* and *Nitrospira*, are not reported in either control samples. This gives confidence in the findings that these two genera are present in the King's Bath microbiome and that the findings from the 16S rRNA typing dataset are reliable as the purity controls of the 16S rRNA dataset were clean.

The MDA King's Bath is also shown to have a more complex taxonomy profile with a greater percentage of 'other genera < 1 %' beyond the top ten summarised in the chart. The single most abundant genera in the King's Bath is reported to be *Sideroxydans*, followed by *Nitrospira*. This is consistent with the results of the 16S profiling which reported these two genera as abundant. However, there are also notable differences such as the absence of *Zooloega*. Other genera which did not feature in high abundance in the 16S typing such as *Pseudomonas* and *Klebseilla* appear more prominent in this data. Also, the shotgun metagenome has reported an archaea genus – *Nitrososphaera* as the third most prominent microbial genera in the sample. This highlights the potential of the shotgun sequencing approach to capture a fuller picture of the microbiome than is possible with 16S typing alone.

The reason for attempting MDA as the method to amplify the DNA extraction is because it can offer the potential to generate large amounts of DNA rapidly, and also no theoretical limits to the length of DNA fragments that are generated. These factors can help ensure more accurate taxonomy profiling and long contig assemblies after sequencing with the MinION. However, the potential limitations of this approach, such as the amplification of exogenous DNA and amplification bias, have been highlighted in the analysis of the taxonomy profiles of the samples sequenced. Additionally, the throughput of the sequencing was poor and the average read lengths (~3,000 kbp: **Table 6.6**) were shorter than expected, which limits the extent to which any functional annotations of long BGCs from the microbiome could be made.

Therefore, it was decided to attempt to amplify the metagenome using a different method in the hope that it may produce a cleaner result which could be conducive to further functional analysis. The approach taken was to use a library kit protocol (SQK-RPB004; Oxford Nanopore Technologies Ltd, UK) which fragments the metagenome DNA using a transposase which attaches primer binding sequences onto the ends of the fragments. These fragments are then amplified by standard PCR. The approach may avoid contamination as the amplification is more specific and runs for a shorter time than MDA approach. The reason for avoiding this approach initially was the expectation that read lengths would only be within the region ~2-3 kbp which would fail to fully utilise the benefits of long-read sequencing. However, the average read lengths reported for the MDA samples proved to only be within this range.

6.3.3.3. Analysis of Sequencing Performance of PCR Amplified King's Bath Metagenome Sample

The throughput of the sequencing of the PCR amplified samples was superior to the MDA samples (**Table 6.7**). As before, the samples were sequenced using two flowcells but the flowcells retained viability for approximately 140 hours of sequencing compared to 44 hours with the MDA samples. In total 11,488,907 reads were generated yielding 38.1 Gbp, with 85.4 % passing quality thresholding. In addition to this, the blank DNA extraction and molecular water control samples returned very few reads in comparison to the King's Bath and mock microbiome metagenome samples. This suggests that there is no significant contamination present in these samples. It also further suggests that the source of the previous contamination was due to the MDA reagents used rather than the wider environment or equipment being used.

Table 6.7. Key metrics of the throughput of the reads sequenced from the PCR prepared samples. The data shows significantly higher read yields and that there were significantly fewer reads classified to the controls than to the King's Bath and mock samples. Additionally, there was still a significant number of reads unassigned to a sample.

Sample	Reads sequenced	Yield (Gbp)	Average length (bp)	% reads passed quality threshold <q10	% passed reads taxon classified
PCR amplified King's Bath metagenome	7,126,212	27.1	3,802	95.7	64.7
PCR amplified blank DNA extraction	212	5.7E-4	2,699	83.0	72.2
PCR amplified molecular water	794	1.9E-3	2,382	91.9	97.1
PCR amplified mock microbiome metagenome	2,406,709	6.7	2,767	96.6	99.3
No barcode*	1,954,958	4.3	2,213	34.2	62.9
Total	11,488,907	38.1	3,315	85.4	72.8

Comparison of the PCR amplified mock microbiome to the reference showed that accuracy of taxonomic profiling was better than with the MDA prepped samples (**Figure 6.17**). As with MDA sample, *Staphylococcus* did appear overrepresented, but the difference was not as great. The remaining taxa were all present in approximately the correct proportions and there were no erroneous taxa reported. This suggests that this PCR amplification method may not bias or contaminate the results of the King's Bath sample to such a large extent as the MDA method.

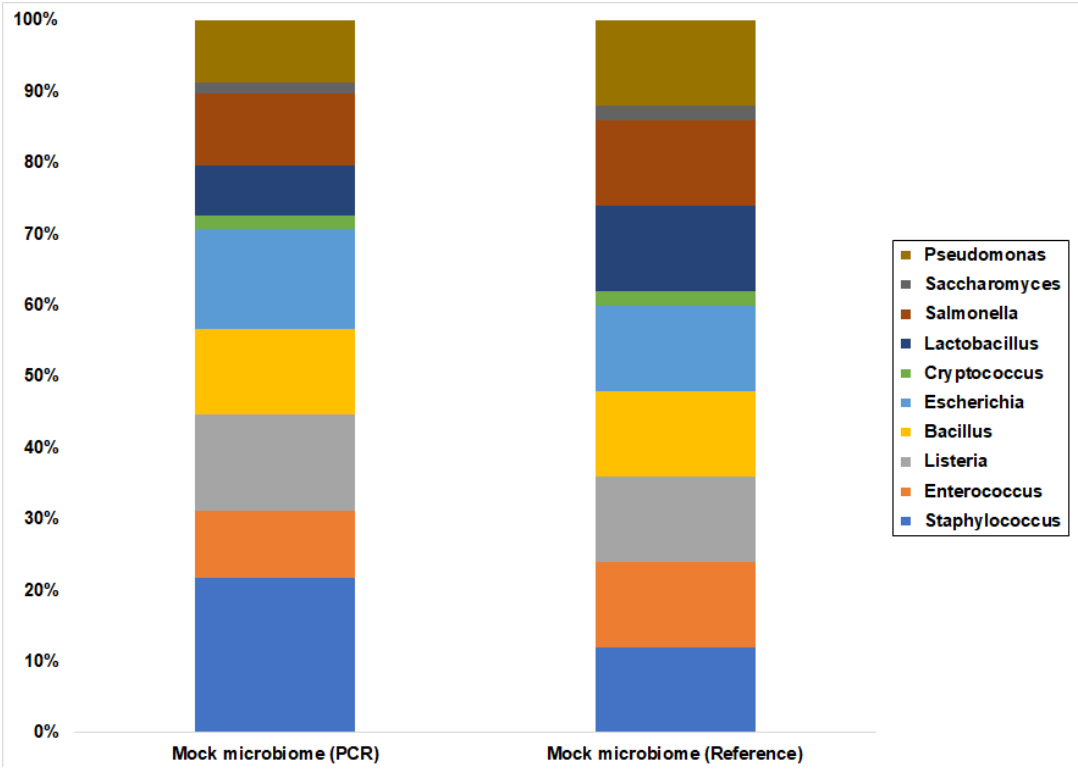


Figure 6.17. Comparison of the distribution of assigned genera of the PCR mock microbiome shotgun metagenomes in comparison with the mock microbiome reference taxonomy. Taxonomy profiling was based on reads that mapped to the top ten genera at ≥ 1% abundancy.

6.3.3.4. Taxonomic Profile King's Bath Shotgun Metagenome

Based upon the better sequencing throughputs, lower contamination, and more accurate taxonomic profiling of the mock microbiome seen with the PCR amplified samples, it was decided to use this sample for analysis of the King's Bath microbiome rather the MDA sample.

Figure 6.18 shows the taxonomic distribution of the assigned reads from the PCR amplified shotgun sequencing of the King's Bath metagenome. The data is in accord with the main observations taken from the 16S rRNA typing dataset (**Figures 6.10 & 6.11**). As in the 16S rRNA gene dataset, Proteobacteria is the dominant phyla reported, with Betaproteobacteria being the dominant class within this phylum. However, Gamma- and Alphaproteobacteria classes now have greater prominence. Nitrospirae is the next most abundant phylum, as it also is in the 16S rRNA typing dataset.

In accord with the MDA processed shotgun sample (**Figure 6.16**), the archaea in the genus *Nitrososphaera* feature prominently. Studies into this genus have characterized it to have an ammonia-oxidizing metabolism which is in accord with the geochemistry of the water and tallies with the abundance of other nitrogen metabolising genera such as *Nitrospira*. No

archaeal species were identified in the 16S rRNA gene dataset. Whilst archaeal species are known to contain 16S rRNA genes (Stahl & Torre, 2012), and have been detected in 16S rRNA typing studies of thermal springs (Barns et al., 1994; Meyer-Dombard, Shock and Amend, 2005; Mardanov et al., 2011). It has also been reported by some researchers that archaea, especially thermophilic kinds, can be under-reported in such studies due to the positioning of introns within the hypervariable regions of the 16S rRNA gene which affect the annealing of universal 16S rRNA gene primers (Jay & Inskeep, 2015). This may explain why *Nitrososphaera* was not reported in the 16S rRNA typing dataset and also demonstrates the utility of a shotgun metagenomic sequencing approach for capturing a fuller diversity of the microbiome.

Sideroxydans is also a prominent genus in the shotgun dataset, which corresponds with the 16S dataset. A notable difference from the 16S rRNA gene dataset is *Pseudomonas* genus having a greater abundance. Overall, the taxonomy profile of the shotgun metagenomic dataset appears broadly in accord with the findings of the 16S rRNA typing dataset which gives confidence in the findings of the types of genera present in the waters of the King's Bath. The finding also demonstrates that long-read shotgun sequencing can provide reliable taxonomic profiling.

However, it doesn't provide a definitive assessment of the taxonomic content and relative abundances. Also, it is worth noting the datasets are intrinsically different which necessitates differences in their processing. Additionally, factors known to bias metagenomic profiling studies (such as PCR amplification bias, variations in gene copy number or genome size, and nuances in individual profiling databases and algorithms) can influence the results of both datasets to differing extents.

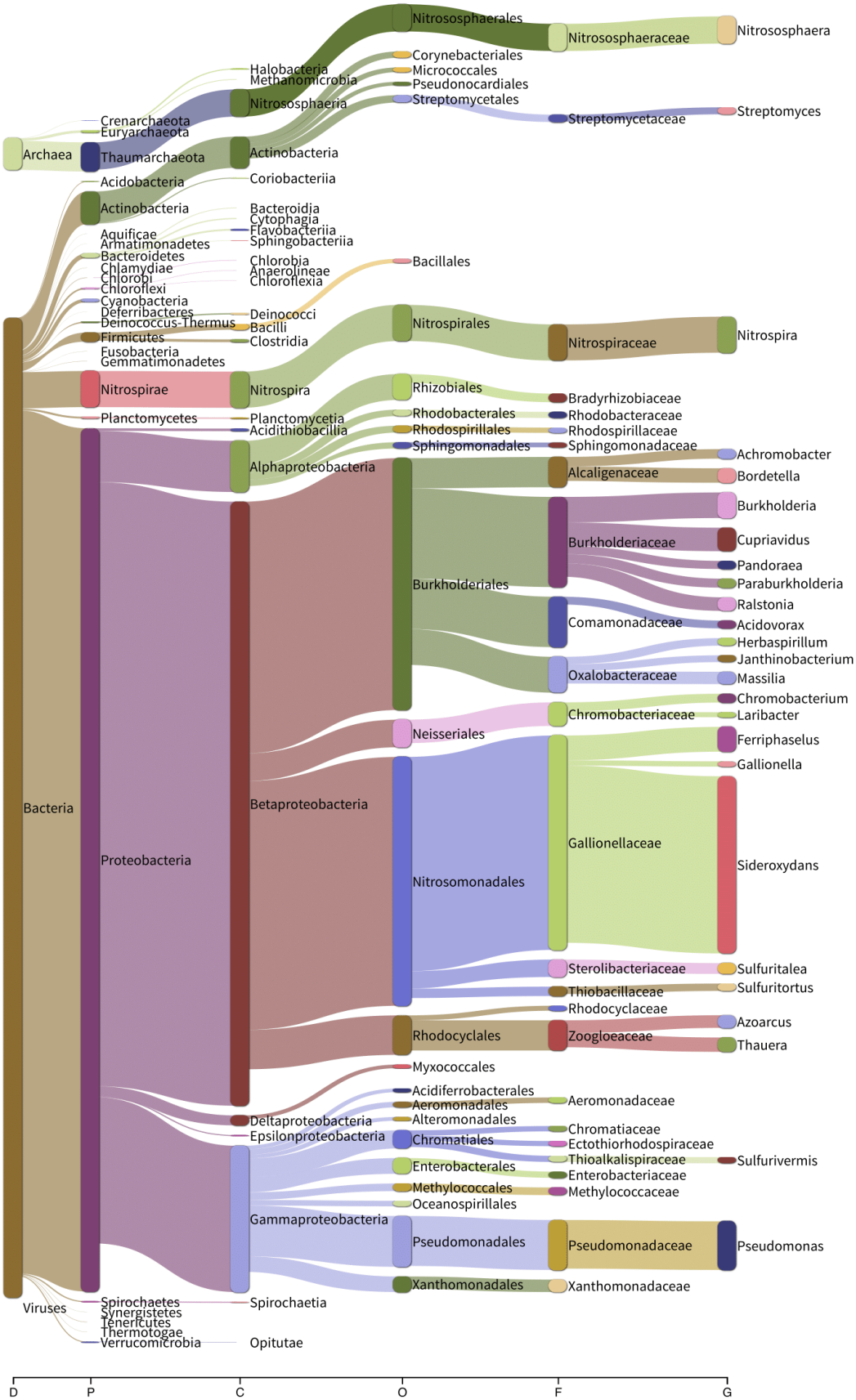


Figure 6.18. Sankey diagram of PCR shotgun metagenome taxonomic assignments that are ≥ 1% of the total of the King's Bath microbiome. Taxa profile is in accord with the main observations taken from the 16S rRNA typing dataset (**Figures 6.10 & 6.11**). But presence of archaea genus *Nitrososphaera* now notable.

6.3.3.5. Functional Gene Profiling of King's Bath Shotgun Metagenome

Analysis using AntiSMASH was performed on the PCR amplified King's Bath metagenome to attempt to identify putative secondary metabolite BGCs. However, the analysis failed to return any results. Possible reasons for could be that the average read length of the dataset (3,802 bp) was not sufficient for BGCs which are often comprised to multiple genes. Additionally, while the data yield generated from the shotgun sequencing of the PCR amplified King's Bath (27.1 Gbp) was higher than the yield of the MDA sample, it is far lower than many metagenomic sequencing studies. It is likely that the sequencing lacked the necessary depth to capture functional genes in sufficient quantity to be annotated when considering the wide spread of millions of differing genomes that the metagenome comprises. Assembling the reads to attempt to reconstruct longer gene pathways was not considered a viable option due to high computation resource required and the low chance of success based on read depth. One of the aims of this study was to determine if low-throughput long-read shotgun sequencing could be sufficient to capture long multigene BGCs in isolation and without the requirement to assemble the genomes first. The results of this analysis suggest that this is not possible with the methodology adopted in this study.

To determine what other functional gene information could be extracted from the dataset which may inform future bioprospecting, it was decided to attempt to analyse the microbiome for presence of metal resistance genes. The taxonomy profiling identified genera with potential applications in bioremediation of toxic substances such as heavy metals or biocide pollutants. Such bacteria have adapted to survive and metabolise such substances through the acquisition of genes that confer resistance to these toxins (Das, Dash & Chakraborty, 2016). The abundance of these resistance genes within the metagenome could be used as a co-marker for the bioremediation potential of the microbiome. Additionally, because these resistance mechanisms are encoded on single genes it was hypothesed that the average read length of ~3.8 kbp would not limit their detection. Detection of metal resistance genes in the dataset would also validate the use of Nanopore long-read shotgun sequencing for functional gene profiling of microbiomes because the geochemistry of the water means these are genes which are expected to be found.

Analysis of a subset (476275) of the reads identified a metal or biocide resistance gene in 9304 of these reads, which is 1.95 % of those analysed (**Figure 6.19**). The resistome included genes for resistance against 22 metals. The largest proportion of resistance genes was against arsenic (1271 / 14 %), with large proportions also for copper (1030 / 11 %) and zinc (923 / 10 %). The high abundance of copper resistance genes is notable as one of the few bacteria isolated

from the Roman Baths, *Methylococcus capsulantus* (Bath) is thought to have metabolic pathways for the scavenging of copper from the environment (Ward et al., 2004). Biocide resistance genes also feature highly with 1716 genes identified which is 18 % of the total genes returned. This category includes genes for resistance against 30 chemical toxins.

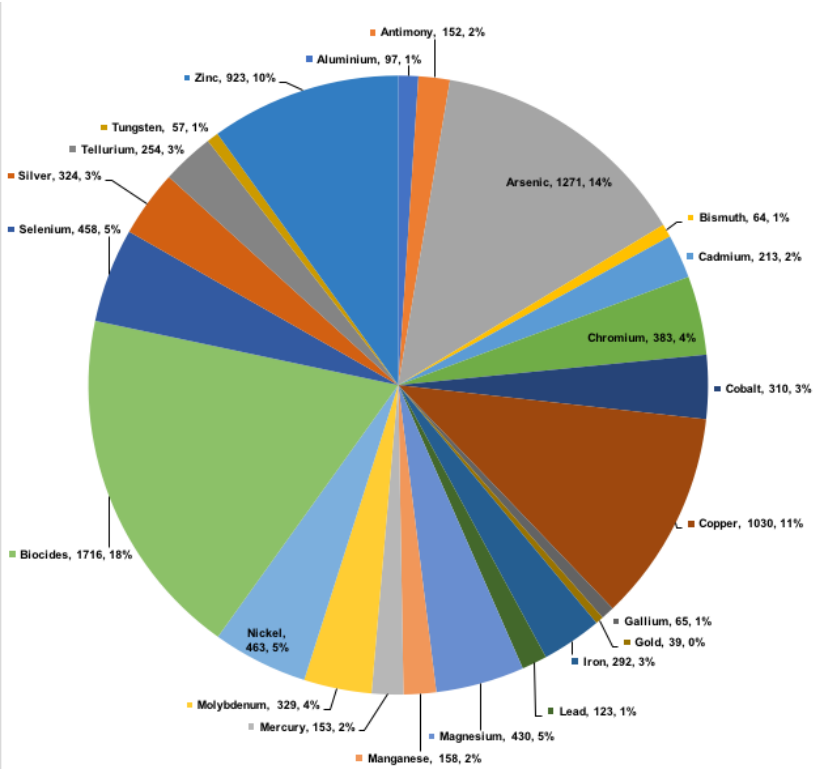


Figure 6.19. Results of screen for presence of metal and biocide resistance genes in King's Bath shotgun metagenome. Subset (476275) of reads were screened and a total of 9304 genes were identified. Data labels state: substance, total genes identified, percentage of total.

Arsenic is a highly toxic heavy metal and a major environmental contaminant which threatens biodiversity and public health. The threat posed by arsenic has led to research focusing on identifying microorganisms for potential arsenic bioremediation is occurring around the world (Lim, Shukor & Wasoh 2014; Kapahi & Sachdeva, 2019; Shukla, Sarim & Singh, 2020). The reported geochemistry of the Roman Baths did not identify arsenic (**Table 6.1**). However, due to the high toxicity of arsenic, even at very low abundance, to all domains of life, there is a strong natural selection pressure for the evolution of arsenic resistance mechanisms. This strong selection pressure has been seen through the presence of arsenic detoxification pathways distributed widely across the bacterial domain (Fekih, at al., 2018). This may explain the large number of arsenic resistance genes within the King's Bath metagenome. Other toxic heavy metal environmental contaminants include lead, mercury, and cadmium. There are relatively few resistance genes identified against these metals in the sample in comparison to arsenic (**Figure 6.20**).

Analysis of the distribution of the heavy metal resistance genes reveals they are distributed widely amongst reads classified to different genera (**Figure 6.20**). The largest proportion of these genes were identified in reads classified as *Pseudomonas*. This finding is consistent with a review by Chellaiah (2018) who concluded that multi-resistant *Pseudomonas* is often identified as a prominent member of the microbiome of heavy-metal contaminated sites. While *Pseudomonas* isolates are also being investigated for heavy metal bioremediation (Singh, Bishnoi & Kirrolia, 2013; Lampis et al., 2015; Satyapal et al., 2018; Al-Dhabi et al., 2019).

The findings of metal and biocide resistance genes distributed widely amongst different genera within this metagenome gives a further indication that the King's Bath metagenome may be a lucrative site for bioprospecting for bioremediation species. This finding also serves to validate the use of Nanopore long-read shotgun sequencing for functional gene profiling of microbiomes for bioprospecting.

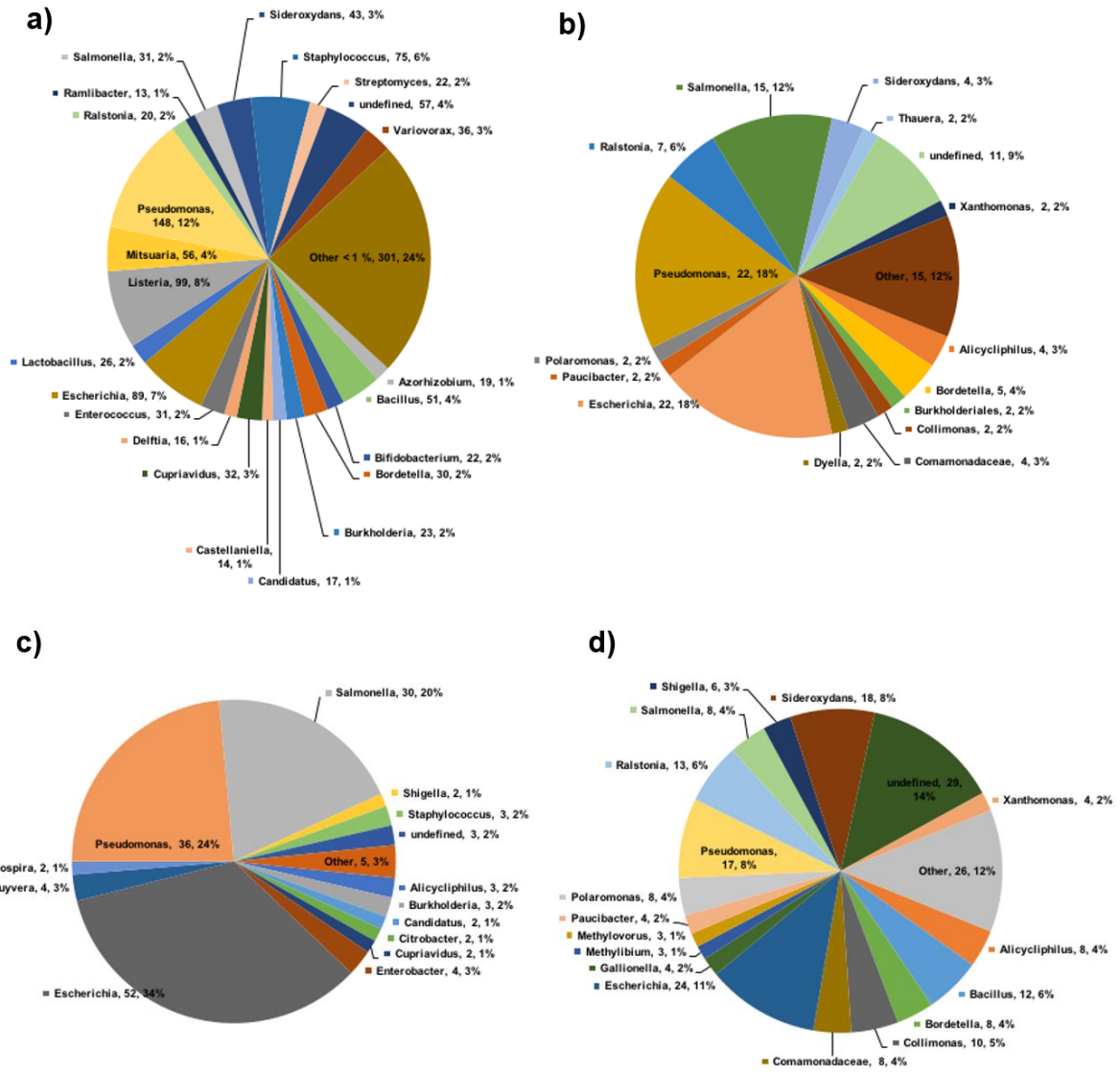


Figure 6.20. Breakdown of distribution of a) arsenic b) lead c) mercury and d) cadmium resistance genes per genus in King's Bath shotgun metagenome. Data labels: genus, total genes, percentage of total. Genes appear widely distributed amongst genera

6.3.3.6. Summary of Shotgun Metagenome Sequencing and Future Directions

Two approaches were utilised to attempt to amplify the extracted metagenome to generate sufficient mass for shotgun sequencing; multiple displacement amplification and a PCR based approach. The MDA approach had been favoured in the first instance in the hope that it would provide the longest read lengths which would be conducive to potentially detecting long, multigene secondary metabolite BGCs. However, the sequencing yield and quality obtained from the MDA processed samples were relatively poor. Additionally, there was evidence of significant contamination occurring which is most likely due to environmental DNA present in the reagents used. In contrast, the PCR approach produced far higher yields and average read quality, with no evidence of significant environmental contamination. The average read lengths of both samples were also comparable.

Analysis of the taxonomy of the PCR shotgun metagenomic reads corroborated the findings of the previous 16S rRNA profiling in revealing a diverse microbiome of the King's Bath and in its reporting of major taxa present and their relative abundances. There were some notable variations such as the discovery of *Nitrososphaera* archaea. This is most likely due to the archaea 16S rRNA gene not amplifying well and demonstrates how the taxonomy reported in these studies can be affected by biases introduced from the experimental design. Many of these potential biases were discussed in the summary of the 16S analysis above (**Section 6.3.2.3**.) and continue to apply to the shotgun metagenomic dataset as well. In this case, as the input had to be amplified by PCR first before sequencing then this too will have added an amplification bias which could distort the relative abundances of species reported.

The issue of amplification bias was demonstrated by comparison of the taxonomy of the mock microbiome after MDA (**Figure 6.15**). This comparison also demonstrated that without a prior amplification step, the library prep and shotgun sequencing alone of the metagenome on the Oxford Nanopore MinION did not introduce any significant bias to the taxonomy profile. This suggests that direct shotgun sequencing of native metagenome extracts may offer the best approach to obtaining the most reliable taxonic profiles. But this would still be subject to the influence of the DNA extraction methodology used. Purity controls assured that the PCR amplified shotgun metagenome did not contain any significant contamination from exogenous DNA, which gives confidence to the taxonomy results obtained.

The average read lengths obtained from both MDA and PCR shotgun sequenced libraries was ~3 kbp, which may have contributed to the failure to detect long BGCs in the dataset. Therefore, the methodologies detailed in this chapter may not be best suited for bioprospecting of such long pathways. If sufficient DNA mass could be obtained directly from the source, then libraries could be used which may produce longer read lengths which are more conductive to the detection of such BGCs. Collecting greater biomass for DNA extraction through the use of a pump-powered filtration system at site of collection to filter water through a membrane would facilitate greater collection of biomass for DNA extraction.

Despite the failure to detect long BGCs in the metagenome sample, detection of metal and biocide resistance serves to validate the use of Nanopore long-read shotgun sequencing for functional gene profiling of microbiomes for bioprospecting. This technique could therefore support small-scale bioprospecting studies of novel ecological niches.

Analysis of a subset of the shotgun metagenome reads revealed metal and biocide resistance genes for 22 metals and up to 30 biocides in 1.95 % of the reads analysed. These genes appeared in a broad cross section of genera with high abundancies of heavy metal resistance genes seen in *Pseudomonas*, *Salmonella*, and *Escherichia* reads. The presence of toxin resistance can be used as a co-marker for pathways for bioremediation of these toxins. So, the widespread presence of these genes within the metagenome further supports the suggestion that this environment may harbour species with potential biotechnological value.

The pipeline used to perform this analysis was chosen as it is optimised for higher indel occurrences which are common in Nanopore datasets. While a subset of reads was chosen for analysis as it requires less computational resource. The pipeline grouped 30 different biocide genes into a single grouping which limited more detailed investigation into which specific biocides these genes are active against. To address this issue future studies could involve the development of bespoke databases of biocide resistance or other specific genes of interest (such as terpene synthases, or lantipeptide synthases) for searches. Candidate hits could be targeted for cloning and functional characterisation. Such approaches were taken by both Baud and co-workers (2017) and Leipold and co-workers (2019) into the identification and characterisation of transaminases biocatalysis enzymes from the microbiomes of a domestic drain and oral cavity respectively. In those studies, metagenomes were shotgun sequenced to generate short reads, which were assembled into larger contigs, and then screened against bespoke Pfam transaminase domain databases to identify contigs with candidate genes. Primers were then designed to amplified and characterise the function of these putative genes. Alternatively, a targeted PCR based approach using degenerate primers for known

biocatalysis genes of interest could be utilised without prior database searches – similar to the methodology used in **Chapter 2**. However, a reliance on degenerate primers may reduce the resolution of genes which can be detected and increase false positive results. Or a functional metagenomic approach could also be used by cloning the metagenome into expression vectors and screening this library for desired phenotypic traits.

6.3.4. Conclusions

The work detailed in this chapter utilises Oxford Nanopore long-read sequencing for taxonomic and functional gene profiling of the microbiome of the King's Bath spring water in the Roman Baths, UK. The purpose of this was to develop a pipeline which would support resource-efficient microbiome profiling for small-scale bioprospecting.

Amplification of metagenome by MDA introduced significant exogenous DNA contamination, and poor sequencing performance which made it an unsuitable methodology. PCR based amplification did produce better results and its taxonomic profiling was in accord with the 16s rRNA profiling. This demonstrates that this methodology could provide a reliable taxonomic profiling of microbiome sites.

Taxonomic profiles report abundant genera whose metabolisms are consistent with the geochemistry of the waters. Many of the abundant genera have been reported to be sources of biocatalysis such as those for bioremediation of environmental toxins. Which could suggest that the waters of the King's Bath could be a lucrative bioprospecting site for novel secondary metabolites and biocatalysis such as those for bioremediation of environmental toxins.

Detection of multigene BGCs such as PKS or NRPS was unsuccessful, which is most likely due to the relatively short read-lengths and sequencing throughput. Generation of more biomass at site of collection could mitigate this by allowing for library prep methods which could generate longer reads, though not necessarily a significant increase in yield.

Putative genes for metal and biocide resistance were detected in a broad range of the genera in the metagenome. This finding is in keeping with the geochemistry of the waters and also with the reported metabolic traits associated with many of the abundant genera reported in the taxonomic profiling. The successful detection of these genes compared to multigene BGCs is likely due to the fact they are single genes and so the shorter read-lengths did not impede their detection.

The datasets created present a basis by which further investigations into discovery of microbial natural products from this source could be made while also validating the use of Nanopore long-read sequencing for the taxonomic profiling and detection of genes of interest for bioprospecting.

Chapter 7.

Concluding Remarks

The work detailed in this thesis utilised technologies and approaches from the fields of microbiology, metagenomics, genome-mining, and natural product chemistry to identify microbes from underexplored microbiomes which are producing novel antibiotic compounds. Underexplored microbiomes were targeted for prospecting based upon the hypothesis that these novel environments would harbour novel microorganisms which would, in turn, have evolved novel secondary metabolite structure and functions.

The work in this thesis characterised five bacterial isolates with antibiotic activity, four from a honey source and one from a hot spring water source. Genome-mining of the isolates revealed multiple secondary metabolite gene clusters that were potentially novel. Thus, the work in this thesis has identified potentially novel secondary metabolites leads for further characterisation. This demonstrates the value in bioprospecting novel or underexplored environmental sources as a means of discovering potentially novel chemistry.

A consistent theme discovered during the course of this work was the disconnect between the secondary metabolite profiles of the isolates and their phylogenetics. For example, the four *Bacillus* isolates (**Chapter 3**) from honey were shown to have 16S rRNA gene alignments between 99.9-100%. However, they were also shown to be distinct organisms by RFLP and genome-mining of the isolates showed different secondary metabolite profiles. In addition to this, all the isolates had a gene cluster homologous with that of the bacteriocin AS-48 which had previously only been reported in *Enterococcus* species before. The *Streptomyces* isolate (**Chapter 5**) from the hot spring water of the Roman Baths was shown to be highly related to *Streptomyces canus* but no evidence of antibiotic compounds reported from this species was found in genomic analysis of this isolate. These findings demonstrate that a reliance on taxonic or phylogenetic classification of environmental isolates is not a safe dereplication strategy in bioprospecting, and is consistent with reports by other researchers (Spieke, 2015; Belknapp et al., 2020). This finding would not have been observed if the genomes of the isolates had not been sequenced and analysed. This therefore highlights the importance that such analysis can play in ensuring that lead isolates with potentially novel chemistry are not overlooked.

Novel environmental sites were targeted for bioprospecting on the hypothesis that they would contain novel microorganisms that would encode novel chemistry. However, it is not clear from the work in this thesis if novel environments are a strong source of novel biology and, in turn, novel chemistry. Whilst it is clear new isolates of antibiotic activity have been found in these environmental sites, they all showed a genomic similarity to species which have also been isolated in other environments, so it cannot be claimed that the isolates were taxonomically unique to the environmental site. Additionally, the work presented in this thesis

highlighted that taxonomy does not track closely with secondary metabolite potential, which calls into question the link between novel biology and novel chemistry. The final point to make is that there is no evidence to suggest that any of the novel BGCs predicted to be produced by the isolates detailed in this thesis are unique to the environments that they were sourced. Infact, all of the isolates were shown to also have genes for known compounds found in isolates from other environmental sites. Additional work to isolate and elucidate the structures and functions of the compounds these isolates are producing may provide evidence to access their novelty.

In addition to focusing on novel environmental sites to improve the bioprospecting pipeline, another strategy that was focused upon was to utilise genome-mining as an early stage dereplication strategy. The value of genome-mining as an early dereplication strategy was demonstrated in **Chapter 3** with the identification of putative AS-48 gene clusters in the *Bacillus* isolates. However, the genome assemblies used in this study were fragmented and some of the putatively novel BGCs had low resolution annotations which limited structure prediction – an artefact which is attributed to short-read assemblies.

To address the limitations of the genome assemblies used in **Chapter 3**, a particular focus was made in developing genome-mining pipelines using Oxford Nanopore long-read sequencing (**Chapter 4**). The rationale for this being that long-read assemblies would more accurately assemble contiguous genomes than short-reads, and therefore provide more confident annotations of long and repetitive biosynthetic gene clusters. Additionally, the lower costs of Nanopore sequencing compared to other providers would remove a barrier that may prevent smaller-scale bioprospecting studies from engaging with genome-mining approaches. If such barriers can be lifted, then it may encourage more researchers to begin bioprospecting for antibiotic leads which may help accelerate novel antibiotic discovery. Or even encourage private companies that found antibiotic bioprospecting no longer economical to consider re-entering the market.

A pipeline was developed in this thesis for the sequencing, assembly and annotation of bacterial genomes for bioprospecting purposes using Oxford Nanopore long read sequencing. The long-read sequencing was shown to produce a contiguous genome assembly of *Streptomyces coelicolor* A3(2) – a model organism for the genera. The pipeline also demonstrated that with Nanopore long-reads alone, accurate identification of the expected BGCs was made. In contrast, an Illumina only assembly produced a highly fragmented assembly of 177 contigs which was not conducive to BGC annotation. These findings demonstrated the value of the pipeline for dereplication analysis.

However, even after extensive processing of the raw assembly, the Nanopore-only assembly retained a high number of indels which caused fragmented and truncated gene annotations. This affected the quality of alignments of phylogenetic marker genes and also the structure predictions of the products of identified BGCs. But these issues were seen to be resolved by polishing the Nanopore assembly with Illumina reads. Therefore, the work in this thesis has shown that Nanopore long-read sequencing could provide a resource-efficient means to analyse isolates for dereplication of known BGCs, which can support small-scale bioprospecting studies where higher-cost sequencing platforms are maybe unattainable. However, Illumina sequencing is still required to enable other analysis such as whole-genome wide phylogenetic analysis, or to make predictions of chemical structures of novel BGCs to help inform bioactivity-guided isolation techniques.

The pipeline developed in **Chapter 4** was validated after being utilised as part of a strategy to identify antibiotic compounds produced by KB16 isolated from the Roman Baths (**Chapter 5**). The pipeline produced a contiguous assembly and high-quality annotations which allowed for genome-wide phylogenetic analysis and genome-mining of the secondary metabolite biosynthetic potential of the microorganism. The results of this, in particular, the identification of genes analogous to those for the production of a known antibiotic – albaflavenone, then informed the strategy for bioactivity-guided isolation.

Genome-mining is an important tool that offers potential in accelerating bioprospecting pipelines and so, in increasing the identification of novel antibiotic leads. However, it is important to be aware of the limitations that a genome-mining analysis pipeline has. In terms of dereplication, it is dependent upon known BGCs having well annotated database entries. While predicting novel BGCs are limited to BGCs of known biosynthetic logic which means that potential new classes of microbial secondary metabolites are overlooked. The point of how genome-mining is limited to a reliance upon known biosynthetic logic can be seen in the annotation of cluster 24 in KB16, whereby substrates for three NRPS A domains could not be predicted not due to sequences quality but due to a lack of sufficient database matches.

Genome-mining is also reliant upon genome assembly. The *de novo* genomic assemblies of unknown isolates, by their nature, cannot be fully depended upon to be accurate due to the effect of factors such as sequencing depth, read length distributions, and read coverage, all of which can be influenced by both the experimental procedures used and intrinsic aspects of the isolates genomic sequence which researchers cannot control for in practice.

It is therefore important that there is effective integration of genome-mining and bioactivity-guided isolation pipelines to fully realise the benefits of each. The importance of the need for both chemical and genomic data was highlighted in **Chapter 5** upon the discovery of multiple active components in the bioautographies of KB16 which suggests KB16 may be producing multiple antibiotic compounds. From genome-mining analysis alone, only one known antibiotic had been identified and if the isolates had been discounted at that stage then this discovery would have been missed.

The bioactivity-guided isolation strategy described in **Chapter 5** was informed by the genomic analysis but frustrated by difficulties in obtaining sufficient compound mass. This highlights how this can serve as a bottleneck in effective bioprospecting and how methods to expediate chemical dereplication need to be developed in tandem to genomic analysis in order to create more efficient bioprospecting pipelines. The use of LESA-MS coupled with bioautography could offer a potential pipeline to achieve this by allowing researchers to analyse small quantities of the active compound and cross reference against the genomic analysis. Additionally, the comparison of both genome and chemical data may highlight discrepancies by which novel biosynthetic logic can be discovered. However, this method was not proven in this thesis and would require further development. Other strategies such as the use of recombinant genetic techniques or cross-analysis of chemical spectra with genome annotations have also been pursued by other research groups to achieve the broadly similar aim of integration of the genomic predictions with experimental analysis. However, as previously discussed throughout this thesis, these methods are also resource and labour intensive and require significant optimisation. Additionally, such approaches may also rely upon only the known biosynthetic logic of the genomic analysis and forgo serendipity.

In addition to the characterisation of single isolates, the work in this thesis also sought to characterise the biosynthetic potential of novel microbiomes using metagenomic techniques. A PCR-based screen was used to investigate the human oral microbiome (**Chapter 2**). The rationale of this approach is to attempt to identify *in silico* novel BGCs without the bottleneck of culturing. By bypassing the requirement of culturing microorganism, it may be possible to access a wider verity of BGCs which might not be found in more tractable microorganisms. The work of this thesis successfully identified PK and NRP fragment genes from a human oral microbiome which may be novel. The work, however, did not progress further due to practical limitations from a laboratory move. But despite this, the findings form a basis by which further studies could be employed to isolate these gene clusters in a culture independent manner, which has been detailed in **Chapter 2**. However, these proposed approaches would require significant development to realise without guarantee of success. In addition, the PK and NRPS

fragments identified in this study are devoid of wider context of their full BGCs with no indication of their function or if they are inaccessible through standard culturing techniques. This makes these proposed approaches a high-risk and resource-intensive strategy which may not contribute to making bioprospecting more accessible.

In the case of the Roman Baths, Nanopore long-read sequencing was utilised to profile the taxonomy and functional profile of the microbiome (**Chapter 6**). The rationale for using Nanopore long-read sequencing was to determine if it could be used to annotate long and repetitive BGCs of interest in antibiotic discovery without the resource bottlenecks that such methods usually incur. This could in part address some of the limitations of the PCR-screen methodology which only isolated fragments of pathways, whilst also being potentially more cost and resource efficient. The work in this chapter produced a first report into the taxonomic profile of microbiome of the Roman Baths, whilst also provided information of metal resistant gene content.

Longer BGCs such as PKS or NRPS gene clusters could not be detected, most likely due to limitations of sequencing throughput and read size that the methodology used produced. However, the taxonomic profiles were consistent in both 16S and genomic shotgun profiling of the Roman Baths and the discovery of high incidences of metal resistance is in keeping with the known geochemistry of the site. These findings validated the approach of using Nanopore long-read sequencing for the profiling the microbiome and obtaining functional gene profiles, although further optimisation is necessary to enable longer BGC detection.

The approach of using Nanopore sequencing of the Roman Baths produced more informative data than the PCR-screen approach of the human oral microbiome. However, both approaches, still focused upon known biosynthetic logic at the expense of potentially discovering greater novelty. Furthermore, the knowledge of the potential genera in the waters of the bath could allow researchers to attempt to selectively target their cultivation, the findings from the work in **Chapters 3 and 5** that the link between taxonomy and secondary metabolism is not strong means that this approach may not guarantee the discovery of novel compounds of interest.

Culturing strategies based upon the geochemistry and physical conditions of the environmental site may therefore offer greater verity in cultivating the microbiome. However, it is important not to discount the role that serendipity may play in discovery. For example, the phyla of KB16 which was isolated from Roman Baths was shown to be a very small component of the taxonomic profile returned, whilst the conditions used to isolate the organism were different to the geochemistry of the site.

There are no magic bullets to the issue of antibiotic discovery, and in additional to the technical challenges, there are also many socio-economic aspects which require resolution too. But at its heart is a need for a determined effort to discover putative antibiotic leads to be discovered to serve as a foundation for development, and any strategies and technologies which can remove bottlenecks in the process and thus encourage wider adoption of bioprospecting is vitally important. The work in this thesis has exemplified such approaches which have led to the characterisation of multiple novel isolates and microbiomes which serve as leads for further exploration and development.

References

Ahn, D.-H., Chung, Y.-C., & Pak, D. (1998). Biosorption of heavy metal ions by immobilized zoogloea and zooglan. *Applied Biochemistry and Biotechnology*, *73*(1), 43–50. https://doi.org/10.1007/BF02788832

Albertsen, M., McIlroy, S. J., Stokholm-Bjerregaard, M., Karst, S. M., & Nielsen, P. H. (2016). "Candidatus Propionivibrio aalborgensis": A Novel Glycogen Accumulating Organism Abundant in Full-Scale Enhanced Biological Phosphorus Removal Plants. *Frontiers in Microbiology*, *7*. https://doi.org/10.3389/fmicb.2016.01033

Al-Dhabi, N. A., Esmail, G. A., Mohammed Ghilan, A.-K., & Valan Arasu, M. (2019). Optimizing the Management of Cadmium Bioremediation Capacity of Metal-Resistant Pseudomonas sp. Strain Al-Dhabi-126 Isolated from the Industrial City of Saudi Arabian Environment. *International Journal of Environmental Research and Public Health*, *16*(23), 4788. https://doi.org/10.3390/ijerph16234788

Alhashash, F., Weston, V., Diggle, M., & McNally, A. (2013). Multidrug-resistant *Escherichia coli* bacteremia. *Emerging infectious diseases*, *19*(10), 1699–1701. https://doi.org/10.3201/eid1910.130309

Alkhalili, R., Bernfur, K., Dishisha, T., Mamo, G., Schelin, J., Canbäck, B., … Hatti-Kaul, R. (2016). Antimicrobial Protein Candidates from the Thermophilic Geobacillus sp. Strain ZGt-1: Production, Proteomics, and Bioinformatics Analysis. *International Journal of Molecular Sciences*, *17*(8), 1363. https://doi.org/10.3390/ijms17081363

Allen, N. E., Hobbs, J. N., & Alborn Jr, W. E. (1987). Inhibition of peptidoglycan biosynthesis in gram-positive bacteria by LY146032. *Antimicrobial Agents and Chemotherapy*, *31*(7), 1093–1099. https://doi.org/10.1128/aac.31.7.1093

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Amos, G. C. A., Borsetto, C., Laskaris, P., Krsek, M., Berry, A. E., Newsham, K. K., … Wellington, E. M. H. (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), e0138327. Retrieved from https://doi.org/10.1371/journal.pone.0138327

Amos, G. C., Borsetto, C., Laskaris, P., Krsek, M., Berry, A. E., Newsham, K. K., Calvo-Bado, L., Pearce, D. A., Vallin, C., & Wellington, E. M. (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), e0138327. https://doi.org/10.1371/journal.pone.0138327

Andrews, J. N., Burgess, W. G., Edmunds, W. M., Kay, R. L. F., & Lee, D. J. (1982). The thermal springs of Bath. *Nature*, *298*(5872), 339–343. https://doi.org/10.1038/298339a0

Anesti, V., McDonald, I. R., Ramaswamy, M., Wade, W. G., Kelly, D. P., & Wood, A. P. (2005). Isolation and molecular detection of methylotrophic bacteria occurring in the human mouth. *Environmental Microbiology*, *7*(8), 1227–1238. https://doi.org/10.1111/j.1462-2920.2005.00805.x

Antipov, D., Korobeynikov, A., McLean, J. S., & Pevzner, P. A. (2015). hybridSPAdes: an algorithm for hybrid assembly of short and long reads. *Bioinformatics*, *32*(7), 1009–1015. https://doi.org/10.1093/bioinformatics/btv688

Antony-Babu, S., Stien, D., Eparvier, V., Parrot, D., Tomasi, S., & Suzuki, M. T. (2017). Multiple Streptomyces species with distinct secondary metabolomes have identical 16S rRNA gene sequences. *Scientific Reports*, *7*(1), 11089. https://doi.org/10.1038/s41598-017-11363-1

Arango-Argoty, G. A., Dai, D., Pruden, A., Vikesland, P., Heath, L. S., & Zhang, L. (2019). NanoARG: a web service for detecting and contextualizing antimicrobial resistance genes from nanopore-derived metagenomes. *Microbiome*, *7*(1), 88. https://doi.org/10.1186/s40168-019-0703-9

Arnison P. G, Bibb M. J., Bierbaum G., Bowers A. A., Bugni T. S., Bulaj G., Camarero J. A., Campopiano D. J., Challis G. L., Clardy J., Cotter P. D., Craik D. J., Dawson M., Dittmann E., Donadio S., Dorrestein P. C., Entian K. D., Fischbach M. A., Garavelli J. S., Göransson U., Gruber C. W., Haft D. H., Hemscheidt T. K., Hertweck C., Hill C., Horswill A. R., Jaspars M., Kelly W. L., Klinman J. P., Kuipers O. P., Link A. J., Liu W., Marahiel M. A., Mitchell D. A., Moll G. N., Moore B. S., Müller R., Nair S. K., Nes I. F., Norris G. E., Olivera B. M., Onaka H., Patchett M. L., Piel J., Reaney M. J., Rebuffat S., Ross R. P., Sahl H. G., Schmidt E. W., Selsted M. E., Severinov K., Shen B., Sivonen K., Smith L., Stein T., Süssmuth R. D., Tagg J. R., Tang G. L., Truman A. W., Vederas J. C., Walsh C. T., Walton J. D., Wenzel S. C., Willey J. M., van der Donk W. A. (2013) Ribosomally synthesized and post-translationally modified peptide natural products: overview and recommendations for a universal nomenclature. *Natural Product Reports. 30*(1):108-60. https://10.1039/c2np20085f.

August, P. R., Tang, L., Yoon, Y. J., Ning, S., Müller, R., Yu, T. W., … Floss, H. G. (1998). Biosynthesis of the ansamycin antibiotic rifamycin: deductions from the molecular analysis of the rif biosynthetic gene cluster of Amycolatopsis mediterranei S699. *Chemistry & Biology*, *5*(2), 69–79. https://doi.org/10.1016/s1074-5521(98)90141-7

Aunpad, R., & Na-Bangchang, K. (2007). Pumilicin 4, a novel bacteriocin with anti-MRSA and anti-VRE activity produced by newly isolated bacteria Bacillus pumilus strain WAPB4. *Current Microbiology*, *55*(4), 308–313. https://doi.org/10.1007/s00284-006-0632-2

Ayling, M., Clark, M. D., & Leggett, R. M. (2020). New approaches for metagenome assembly with short reads. *Briefings in Bioinformatics*, *21*(2), 584–594. https://doi.org/10.1093/bib/bbz020

Ayuso-Sacido, A., & Genilloud, O. (2005). New PCR primers for the screening of NRPS and PKS-I systems in actinomycetes: detection and distribution of these biosynthetic gene sequences in major taxonomic groups. *Microbial Ecology*, *49*(1), 10–24. https://doi.org/10.1007/s00248-004-0249-6

Bala, S., Khanna, R., Dadhwal, M., Prabagaran, S. R., Shivaji, S., Cullum, J., & Lal, R. (2004). Reclassification of Amycolatopsis mediterranei DSM 46095 as Amycolatopsis rifamycinica sp. nov. *International Journal of Systematic and Evolutionary Microbiology*, *54*(Pt 4), 1145–1149. https://doi.org/10.1099/ijs.0.02901-0

Balouiri, M., Sadiki, M., & Ibnsouda, S. K. (2016). Methods for *in vitro* evaluating antimicrobial activity: A review. *Journal of pharmaceutical analysis*, *6*(2), 71–79. https://doi.org/10.1016/j.jpha.2015.11.005

Baltz, R. H. (2008). Renaissance in antibacterial discovery from actinomycetes. *Current Opinion in Pharmacology*, *8*(5), 557–563. https://doi.org/10.1016/j.coph.2008.04.008

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., … Pevzner, P. A. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology : A Journal of Computational Molecular Cell Biology*, *19*(5), 455–477. https://doi.org/10.1089/cmb.2012.0021

Baranašić, D., Zucko, J., Diminic, J., Gacesa, R., Long, P. F., Cullum, J., … Starcevic, A. (2014). Predicting substrate specificity of adenylation domains of nonribosomal peptide synthetases and other protein properties by latent semantic indexing. *Journal of Industrial Microbiology & Biotechnology*, *41*(2), 461–467. https://doi.org/10.1007/s10295-013-1322-2

Barnhart, C. E., Robertson, J. C., & Miller, H. W. (1960). Virginiamycin, A New Antibiotic For Growing Swine. *Journal of Animal Science.*, *19*(4), 1247.

Barns, S. M., Fundyga, R. E., Jeffries, M. W., & Pace, N. R. (1994). Remarkable archaeal diversity detected in a Yellowstone National Park hot spring environment. *Proceedings of the National Academy of Sciences of the United States of America*, *91*(5), 1609–1613. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/7510403

Barona-Gómez, F., Lautru, S., Francou, F.-X., Leblond, P., Pernodet, J.-L., & Challis, G. L. (2006). Multiple biosynthetic and uptake systems mediate siderophore-dependent iron acquisition in Streptomyces coelicolor A3(2) and Streptomyces ambofaciens ATCC 23877. *Microbiology (Reading, England)*, *152*(Pt 11), 3355–3366. https://doi.org/10.1099/mic.0.29161-0

Barona-Gómez, F., Wong, U., Giannakopulos, A. E., Derrick, P. J., & Challis, G. L. (2004). Identification of a cluster of genes that directs desferrioxamine biosynthesis in Streptomyces coelicolor M145. *Journal of the American Chemical Society*, *126*(50), 16282–16283. https://doi.org/10.1021/ja045774k

Baud, D., Jeffries, J. W. E., Moody, T. S., Ward, J. M., & Hailes, H. C. (2017). A metagenomics approach for new biocatalyst discovery: application to transaminases and the synthesis of allylic amines. *Green Chemistry*, *19*(4), 1134–1143. https://doi.org/10.1039/C6GC02769E

Bauman, R. W. (2015). *Microbiology: With Diseases by Body System* (4th ed.). Boston: Pearson.

Begley, M., Cotter, P. D., Hill, C., & Ross, R. P. (2009). Identification of a novel two-peptide lantibiotic, lichenicidin, following rational genome mining for LanM proteins. *Applied and Environmental Microbiology*, *75*(17), 5451–5460. https://doi.org/10.1128/AEM.00730-09

Belknap, K. C., Park, C. J., Barth, B. M., & Andam, C. P. (2020). Genome mining of biosynthetic and chemotherapeutic gene clusters in Streptomyces bacteria. *Scientific Reports*, *10*(1), 2003. https://doi.org/10.1038/s41598-020-58904-9

Ben Fekih, I., Zhang, C., Li, Y. P., Zhao, Y., Alwathnani, H. A., Saquib, Q., … Cervantes, C. (2018). Distribution of Arsenic Resistance Genes in Prokaryotes. *Frontiers in Microbiology*, *9*. https://doi.org/10.3389/fmicb.2018.02473

Benítez-Páez, A., Portune, K. J., & Sanz, Y. (2016). Species-level resolution of 16S rRNA gene amplicons sequenced through the MinION™ portable nanopore sequencer. *GigaScience*, *5*(1), 4. https://doi.org/10.1186/s13742-016-0111-z

Bentley, S. D., Chater, K. F., Cerdeño-Tárraga, A.-M., Challis, G. L., Thomson, N. R., James, K. D., … Hopwood, D. A. (2002). Complete genome sequence of the model actinomycete Streptomyces coelicolor A3(2). *Nature*, *417*(6885), 141–147. https://doi.org/10.1038/417141a

Binga, E. K., Lasken, R. S., & Neufeld, J. D. (2008). Something from (almost) nothing: the impact of multiple displacement amplification on microbial ecology. *The ISME Journal*, *2*(3), 233–241. https://doi.org/10.1038/ismej.2008.10

Blair, J. M. A., Webber, M. A., Baylay, A. J., Ogbolu, D. O., & Piddock, L. J. V. (2015). Molecular mechanisms of antibiotic resistance. *Nature Reviews Microbiology*, *13*(1), 42–51. https://doi.org/10.1038/nrmicro3380

Blomquist, P. B., Miari, V. F., Biddulph, J. P., & Charalambous, B. M. (2014). Is gonorrhea becoming untreatable?. *Future microbiology*, *9*(2), 189–201. https://doi.org/10.2217/fmb.13.155

Blondeau, J. M. (2004). Fluoroquinolones: mechanism of action, classification, and development of resistance. *Survey of Ophthalmology*, *49 Suppl 2*, S73-8. https://doi.org/10.1016/j.survophthal.2004.01.005

Boden, R., Hutt, L. P., & Rae, A. W. (2017). Reclassification of Thiobacillus aquaesulis (Wood & Kelly, 1995) as Annwoodia aquaesulis gen. nov., comb. nov., transfer of Thiobacillus (Beijerinck, 1904) from the Hydrogenophilales to the Nitrosomonadales, proposal of Hydrogenophilalia class. nov. w. *International Journal of Systematic and Evolutionary Microbiology*, *67*(5), 1191–1205. https://doi.org/10.1099/ijsem.0.001927

Boden, R., Hutt, L. P., Huntemann, M., Clum, A., Pillay, M., Palaniappan, K., … Kyrpides, N. (2016). Permanent draft genome of Thermithiobacillus tepidarius DSM 3134T, a moderately thermophilic, obligately chemolithoautotrophic member of the Acidithiobacillia. *Standards in Genomic Sciences*, *11*(1), 74. https://doi.org/10.1186/s40793-016-0188-0

Bogart, S. (2018). SankeyMATIC. Retrieved August 1, 2018, from https://github.com/nowthis/sankeymatic

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Borsetto, C., Amos, G., da Rocha, U. N., Mitchell, A. L., Finn, R. D., Laidi, R. F., Vallin, C., Pearce, D. A., Newsham, K. K., & Wellington, E. (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

Brady, S. F., & Clardy, J. (2004). Palmitoylputrescine, an Antibiotic Isolated from the Heterologous Expression of DNA Extracted from Bromeliad Tank Water. *Journal of Natural Products*, *67*(8), 1283–1286. https://doi.org/10.1021/np0499766

Brady, S. F., & Clardy, J. (2005). Cloning and heterologous expression of isocyanide biosynthetic genes from environmental DNA. *Angewandte Chemie (International Ed. in English)*, *44*(43), 7063–7065. https://doi.org/10.1002/anie.200501941

Brady, S. F., Chao, C. J., & Clardy, J. (2004). Long-chain N-acyltyrosine synthases from environmental DNA. *Applied and Environmental Microbiology*, *70*(11), 6865–6870. https://doi.org/10.1128/AEM.70.11.6865-6870.2004

Branquinho, R., Klein, G., Kämpfer, P., & Peixe, L. V. (2015). The status of the species Bacillus aerophilus and Bacillus stratosphericus. Request for an Opinion. *International Journal of Systematic and Evolutionary Microbiology*, *65*(Pt 3), 1101. https://doi.org/10.1099/ijs.0.000004

Breitwieser, F. P., & Salzberg, S. L. (2019). Pavian: interactive analysis of metagenomics data for microbiome studies and pathogen identification. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btz715

Broberg, A., Menkis, A., & Vasiliauskas, R. (2006). Kutznerides 1-4, depsipeptides from the actinomycete Kutzneria sp. 744 inhabiting mycorrhizal roots of Picea abies seedlings. *Journal of Natural Products*, *69*(1), 97–102. https://doi.org/10.1021/np050378g

Brock, T. D., & Freeze, H. (1969). Thermus aquaticus gen. n. and sp. n., a nonsporulating extreme thermophile. *Journal of Bacteriology*, *98*(1), 289–297. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/5781580

Brown, E. D., & Wright, G. D. (2016). Antibacterial drug discovery in the resistance era. *Nature*, *529*(7586), 336–343. https://doi.org/10.1038/nature17042

Brune, A., Ludwig, W., & Schink, B. (2002). Propionivibrio limicola sp. nov., a fermentative bacterium specialized in the degradation of hydroaromatic compounds, reclassification of Propionibacter pelophilus as Propionivibrio pelophilus comb. nov. and amended description of the genus Propionivibrio. *International Journal of Systematic and Evolutionary Microbiology*, *52*(2), 441–444. https://doi.org/10.1099/00207713-52-2-441

Bush, K., & Jacoby, G. A. (2010). Updated functional classification of beta-lactamases. *Antimicrobial Agents and Chemotherapy*, *54*(3), 969–976. https://doi.org/10.1128/AAC.01009-09

Butler, M. S., & Buss, A. D. (2006). Natural products--the future scaffolds for novel antibiotics? *Biochemical Pharmacology*, *71*(7), 919–929. https://doi.org/10.1016/j.bcp.2005.10.012

Campbell, E. A., Korzheva, N., Mustaev, A., Murakami, K., Nair, S., Goldfarb, A., & Darst, S. A. (2001). Structural mechanism for rifampicin inhibition of bacterial rna polymerase. *Cell*, *104*(6), 901–912. https://doi.org/10.1016/s0092-8674(01)00286-0

Cantón, R., Morosini, M. I., Martin, O., De La Maza, S., & De La Pedrosa, E. G. G. (2008). IRT and CMT β-lactamases and inhibitor resistance. *Clinical Microbiology and Infection*, *14*(s1), 53–62. https://doi.org/10.1111/j.1469-0691.2007.01849.x

Carlström, C. I., Wang, O., Melnyk, R. A., Bauer, S., Lee, J., Engelbrektson, A., & Coates, J. D. (2013). Physiological and Genetic Description of Dissimilatory Perchlorate Reduction by the Novel Marine Bacterium Arcobacter sp. Strain CAB. *MBio*, *4*(3). https://doi.org/10.1128/mBio.00217-13

CDC. *Antibiotic Resistance Threats in the United States*, 2019. Atlanta, GA: U.S. Department of Health and Human Services, CDC; 2019.

Challinor, V. L., & Bode, H. B. (2015). Bioactive natural products from novel microbial sources. *Annals of the New York Academy of Sciences*, *1354*, 82–97. https://doi.org/10.1111/nyas.12954

Challis, G. L. (2014). Exploitation of the Streptomyces coelicolor A3(2) genome sequence for discovery of new natural products and biosynthetic pathways. *Journal of Industrial Microbiology & Biotechnology*, *41*(2), 219–232. https://doi.org/10.1007/s10295-013-1383-2

Charlop-Powers, Z., Milshteyn, A., & Brady, S. F. (2014). Metagenomic small molecule discovery methods. *Current Opinion in Microbiology*, *19*, 70–75. https://doi.org/10.1016/j.mib.2014.05.021

Charlop-Powers, Z., Pregitzer, C. C., Lemetre, C., Ternei, M. A., Maniko, J., Hover, B. M., … Brady, S. F. (2016). Urban park soil microbiomes are a rich reservoir of natural product biosynthetic diversity. *Proceedings of the National Academy of Sciences*, *113*(51), 14811 LP – 14816. https://doi.org/10.1073/pnas.1615581113

Chellaiah, E. R. (2018). Cadmium (heavy metals) bioremediation by Pseudomonas aeruginosa: a minireview. *Applied Water Science*, *8*(6), 154. https://doi.org/10.1007/s13201-018-0796-5

Chen, T., Yu, W.-H., Izard, J., Baranova, O. V, Lakshmanan, A., & Dewhirst, F. E. (2010). The Human Oral Microbiome Database: a web accessible resource for investigating oral microbe taxonomic and genomic information. *Database*, *2010*. https://doi.org/10.1093/database/baq013

Chen, X. H., Koumoutsi, A., Scholz, R., Eisenreich, A., Schneider, K., Heinemeyer, I., … Borriss, R. (2007). Comparative analysis of the complete genome sequence of the plant growth-promoting bacterium Bacillus amyloliquefaciens FZB42. *Nature Biotechnology*, *25*(9), 1007–1014. https://doi.org/10.1038/nbt1325

Chen, Y.-C., Liu, T., Yu, C.-H., Chiang, T.-Y., & Hwang, C.-C. (2013). Effects of GC Bias in Next-Generation-Sequencing Data on De Novo Genome Assembly. *PLOS ONE*, *8*(4), e62856. Retrieved from https://doi.org/10.1371/journal.pone.0062856

Chou, Y.-J., Chou, J.-H., Lin, M.-C., Arun, A. B., Young, C.-C., & Chen, W.-M. (2008). Vogesella perlucida sp. nov., a non-pigmented bacterium isolated from spring water. *I International Journal of Systematic and Evolutionary Microbiology*, *58*(12), 2677–2681. https://doi.org/10.1099/ijs.0.65766-0

Chuanchuen, R., Beinlich, K., Hoang, T. T., Becher, A., Karkhoff-Schweizer, R. R., & Schweizer, H. P. (2001). Cross-resistance between triclosan and antibiotics in Pseudomonas aeruginosa is mediated by multidrug efflux pumps: exposure of a susceptible mutant strain to triclosan selects nfxB mutants overexpressing MexCD-OprJ. *Antimicrobial Agents and Chemotherapy*, *45*(2), 428–432. https://doi.org/10.1128/AAC.45.2.428-432.2001

Ciofu, O., & Tolker-Nielsen, T. (2019). Tolerance and Resistance of Pseudomonas aeruginosa Biofilms to Antimicrobial Agents-How P. aeruginosa Can Escape Antibiotics. *Frontiers in Microbiology*, *10*, 913. https://doi.org/10.3389/fmicb.2019.00913

Claesen, J., & Bibb, M. (2010). Genome mining and genetic analysis of cypemycin biosynthesis reveal an unusual class of posttranslationally modified peptides. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(37), 16297–16302. https://doi.org/10.1073/pnas.1008608107

Clamp, M., Cuff, J., Searle, S. M., & Barton, G. J. (2004). The Jalview Java alignment editor. *Bioinformatics*, *20*(3), 426–427. https://doi.org/10.1093/bioinformatics/btg430

Clardy, J., Fischbach, M. A., & Walsh, C. T. (2006). New antibiotics from bacterial natural products. *Nature Biotechnology*, *24*(12), 1541–1550. https://doi.org/10.1038/nbt1266

Coates, J. D., & Lovley, D. R. (2015). Geobacter. In *Bergey's Manual of Systematics of Archaea and Bacteria* (pp. 1–6). Chichester, UK: John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118960608.gbm01043

Cocito, C., & Chinali, G. (1985). Molecular mechanism of action of virginiamycin-like antibiotics (synergimycins) on protein synthesis in bacterial cell-free systems. *The Journal of Antimicrobial Chemotherapy*, *16 Suppl A*, 35–52. https://doi.org/10.1093/jac/16.suppl_a.35

Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., … de Hoon, M. J. L. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, *25*(11), 1422–1423. https://doi.org/10.1093/bioinformatics/btp163

Coker, J. A. (2016). Extremophiles and biotechnology: current uses and prospects. *F1000Research*, *5*, 396. https://doi.org/10.12688/f1000research.7432.1

Cologgi, D. L., Speers, A. M., Bullard, B. A., Kelly, S. D., & Reguera, G. (2014). Enhanced Uranium Immobilization and Reduction by Geobacter sulfurreducens Biofilms. *Applied and Environmental Microbiology*, *80*(21), 6638–6646. https://doi.org/10.1128/AEM.02289-14

Connor, T. R., Loman, N. J., Thompson, S., Smith, A., Southgate, J., Poplawski, R., … Pallen, M. J. (2016). CLIMB (the Cloud Infrastructure for Microbial Bioinformatics): an online resource for the medical microbiology community. *Microbial Genomics*, *2*(9). https://doi.org/10.1099/mgen.0.000086

Cornaglia, G., Mazzariol, A., Fontana, R., & Satta, G. (1996). Diffusion of carbapenems through the outer membrane of enterobacteriaceae and correlation of their activities with their periplasmic concentrations. *Microbial Drug Resistance (Larchmont, N.Y.)*, *2*(2), 273–276. https://doi.org/10.1089/mdr.1996.2.273

Cox, G., & Wright, G. D. (2013). Intrinsic antibiotic resistance: mechanisms, origins, challenges and solutions. *International Journal of Medical Microbiology : IJMM*, *303*(6–7), 287–292. https://doi.org/10.1016/j.ijmm.2013.02.009

Cuervo, G., Camoez, M., Shaw, E., Dominguez, M. Á., Gasch, O., Padilla, B., Pintado, V., Almirante, B., Molina, J., López-Medrano, F., Ruiz de Gopegui, E., Martinez, J. A., Bereciartua, E., Rodriguez-Lopez, F., Fernandez-Mazarrasa, C., Goenaga, M. Á., Benito, N., Rodriguez-Baño, J., Espejo, E., Pujol, M., & REIPI/GEIH Study Group (2015). Methicillin-resistant Staphylococcus aureus (MRSA) catheter-related bacteraemia in haemodialysis patients. *BMC Infectious Diseases*, *15*, 484. https://doi.org/10.1186/s12879-015-1227-y

Cui, L., Iwamoto, A., Lian, J.-Q., Neoh, H., Maruyama, T., Horikawa, Y., & Hiramatsu, K. (2006). Novel mechanism of antibiotic resistance originating in vancomycin-intermediate Staphylococcus aureus. *Antimicrobial Agents and Chemotherapy*, *50*(2), 428–438. https://doi.org/10.1128/AAC.50.2.428-438.2006

Cui, L., Ma, X., Sato, K., Okuma, K., Tenover, F. C., Mamizuka, E. M., … Hiramatsu, K. (2003). Cell wall thickening is a common feature of vancomycin resistance in Staphylococcus aureus. *Journal of Clinical Microbiology*, *41*(1), 5—14. https://doi.org/10.1128/jcm.41.1.5-14.2003

Czech, L., Hermann, L., Stöveken, N., Richter, A. A., Höppner, A., Smits, S. H. J., … Bremer, E. (2018). Role of the Extremolytes Ectoine and Hydroxyectoine as Stress Protectants and Nutrients: Genetics, Phylogenomics, Biochemistry, and Structural Analysis. *Genes*, *9*(4). https://doi.org/10.3390/genes9040177

Daims, H., & Wagner, M. (2018). Nitrospira. *Trends in Microbiology*, *26*(5), 462–463. https://doi.org/10.1016/j.tim.2018.02.001

Daims, H., Lebedeva, E. V., Pjevac, P., Han, P., Herbold, C., Albertsen, M., … Wagner, M. (2015). Complete nitrification by Nitrospira bacteria. *Nature*, *528*(7583), 504–509. https://doi.org/10.1038/nature16461

Das, S., Dash, H. R., & Chakraborty, J. (2016). Genetic basis and importance of metal resistant genes in bacteria for bioremediation of contaminated environments with toxic metal pollutants. *Applied Microbiology and Biotechnology*, *100*(7), 2967–2984. https://doi.org/10.1007/s00253-016-7364-4

de Bruin, O. M., & Birnboim, H. C. (2016). A method for assessing efficiency of bacterial cell disruption and DNA release. *BMC Microbiology*, *16*(1), 197. https://doi.org/10.1186/s12866-016-0815-3

De Coster, W., D'Hert, S., Schultz, D. T., Cruts, M., & Van Broeckhoven, C. (2018). NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics*, *34*(15), 2666–2669. https://doi.org/10.1093/bioinformatics/bty149

de Kraker, M. E., Jarlier, V., Monen, J. C., Heuer, O. E., van de Sande, N., & Grundmann, H. (2013). The changing epidemiology of bacteraemias in Europe: trends from the European Antimicrobial Resistance Surveillance System. *Clinical microbiology and infection : the official publication of the European Society of Clinical Microbiology and Infectious Diseases*, *19*(9), 860–868. https://doi.org/10.1111/1469-0691.12028

De Maio, N., Shaw, L. P., Hubbard, A., George, S., Sanderson, N. D., Swann, J., … Consortium, O. B. O. T. R. (2019). Comparison of long-read sequencing technologies in the hybrid assembly of complex bacterial genomes. *Microbial Genomics*, *5*(9). https://doi.org/10.1099/mgen.0.000294

de Vos, M., Müller, B., Borrell, S., Black, P. A., van Helden, P. D., Warren, R. M., … Victor, T. C. (2013). Putative compensatory mutations in the rpoC gene of rifampin-resistant Mycobacterium tuberculosis are associated with ongoing transmission. *Antimicrobial Agents and Chemotherapy*, *57*(2), 827–832. https://doi.org/10.1128/AAC.01541-12

Dervan, A., & Shendure, J. (2017). Chapter 3 - The State of Whole-Genome Sequencing. In G. S. Ginsburg & H. F. B. T.-G. and P. M. (Third E. Willard (Eds.) (pp. 45–62). Boston: Academic Press. https://doi.org/https://doi.org/10.1016/B978-0-12-800681-8.00003-7

Dewanjee, S., Gangopadhyay, M., Bhattacharya, N., Khanra, R., & Dua, T. K. (2015). Bioautography and its scope in the field of natural product chemistry. *Journal of Pharmaceutical Analysis*, *5*(2), 75–84. https://doi.org/10.1016/j.jpha.2014.06.002

Dewhirst, F. E., Chen, T., Izard, J., Paster, B. J., Tanner, A. C. R., Yu, W.-H., … Wade, W. G. (2010). The Human Oral Microbiome. *Journal of Bacteriology*, *192*(19), 5002 LP – 5017. https://doi.org/10.1128/JB.00542-10

Dias, D. A., Urban, S., & Roessner, U. (2012). A historical overview of natural products in drug discovery. *Metabolites*, *2*(2), 303–336. https://doi.org/10.3390/metabo2020303

Dobson, A., Cotter, P. D., Ross, R. P., & Hill, C. (2012). Bacteriocin Production: a Probiotic Trait? *Applied and Environmental Microbiology*, *78*(1), 1 LP – 6. https://doi.org/10.1128/AEM.05576-11

Donachie, S. P., Bowman, J. P., On, S. L. W., & Alam, M. (2005). Arcobacter halophilus sp. nov., the first obligate halophile in the genus Arcobacter. *International Journal of Systematic and Evolutionary Microbiology*, *55*(3), 1271–1277. https://doi.org/10.1099/ijs.0.63581-0

Doroghazi, J. R., & Buckley, D. H. (2010). Widespread homologous recombination within and between Streptomyces species. *The ISME Journal*, *4*(9), 1136–1143. https://doi.org/10.1038/ismej.2010.45

Dubos, R. J. (1939). Studies on a bactericidal agent extracted from a soil *Bacillus*: I Preparation of the agent. Its activity *in vitro*. *The Journal of Experimental Medicine*, *70*(1), 1–10. https://doi.org/10.1084/jem.70.1.1

Duggar, B. M. (1948). Aureomycin; a product of the continuing search for new antibiotics. *Annals of the New York Academy of Sciences*, *51*(Art. 2), 177–181. https://doi.org/10.1111/j.1749-6632.1948.tb27262.x

Dunlap, C. A. (2015). The status of the species Bacillus aerius. Request for an Opinion. *International Journal of Systematic and Evolutionary Microbiology*, *65*(7), 2341. https://doi.org/10.1099/ijs.0.000271

Edmunds, W. M., & Miles, D. L. (1991). The geochemistry of the Bath thermal waters. In G. A. Kellaway (Ed.), *Hot springs of Bath : investigations of the thermal waters of the Avon Valley* (pp. 143–156). Bath: Bath City Council.

Ehrlich, J., Bartz, Q. R., Smith, R. M., Joslyn, D. A., & Burkholder, P. R. (1947). Chloromycetin, a New Antibiotic From a Soil Actinomycete. *Science (New York, N.Y.)*, *106*(2757), 417. https://doi.org/10.1126/science.106.2757.417

Eikel, D., Vavrek, M., Smith, S., Bason, C., Yeh, S., Korfmacher, W. A., & Henion, J. D. (2011). Liquid extraction surface analysis mass spectrometry (LESA-MS) as a novel profiling tool for drug distribution and metabolism analysis: the terfenadine example. *Rapid Communications in Mass Spectrometry : RCM*, *25*(23), 3587–3596. https://doi.org/10.1002/rcm.5274

El-Naggar, N. E.-A., & El-Ewasy, S. M. (2017). Bioproduction, characterization, anticancer and antioxidant activities of extracellular melanin pigment produced by newly isolated microbial cell factories Streptomyces glaucescens NEAE-H. *Scientific Reports*, *7*(1), 42129. https://doi.org/10.1038/srep42129

Esikova, T. Z., Temirov, Y. V, Sokolov, S. L., & Alakhov, Y. B. (2002). Secondary antimicrobial metabolites produced by thermophilic Bacillus spp. strains VK2 and VK21. *Applied Biochemistry and Microbiology*, *38*(3), 226–231. https://doi.org/10.1023/A:1015463122840

Ewing, B., & Green, P. (1998). Base-Calling of Automated Sequencer Traces Using Phred. II. Error Probabilities. *Genome Research*, *8*(3), 186–194. https://doi.org/10.1101/gr.8.3.186

Fazal, A., Thankachan, D., Harris, E., & Seipke, R. F. (2020). A chromatogram-simplified Streptomyces albus host for heterologous production of natural products. *Antonie van Leeuwenhoek*, *113*(4), 511–520. https://doi.org/10.1007/s10482-019-01360-x

Feng, Z., Chakraborty, D., Dewell, S. B., Reddy, B. V. B., & Brady, S. F. (2012). Environmental DNA-Encoded Antibiotics Fasamycins A and B Inhibit FabF in Type II Fatty Acid Biosynthesis. *Journal of the American Chemical Society*, *134*(6), 2981–2987. https://doi.org/10.1021/ja207662w

Feng, Z., Kallifidas, D., & Brady, S. F. (2011). Functional analysis of environmental DNA-derived type II polyketide synthases reveals structurally diverse secondary metabolites. *Proceedings of the National Academy of Sciences*, *108*(31), 12629 LP – 12634. https://doi.org/10.1073/pnas.1103921108

Finlay, A. C., Hobby, G. L., Hoschstein, F., Lees, T. M., Lenert, T. F., Means, J. A., P'an, S. Y., Regna, P. P., Routien, J. B., Sobin, B. A., Tate, K. B., & Kane, J. H. (1951). Viomycin a new antibiotic active against Mycobacteria. *American Review of Tuberculosis*, *63*(1), 1–3. https://doi.org/10.1164/art.1951.63.1.1

Fleming, A. (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), 226-236.

Flinspach, K., Rückert, C., Kalinowski, J., Heide, L., & Apel, A. K. (2014). Draft Genome Sequence of Streptomyces niveus NCIMB 11891, Producer of the Aminocoumarin Antibiotic Novobiocin. *Genome Announcements*, *2*(1). https://doi.org/10.1128/genomeA.01146-13

Fouhy, F., Clooney, A. G., Stanton, C., Claesson, M. J., & Cotter, P. D. (2016). 16S rRNA gene sequencing of mock microbial populations- impact of DNA extraction method, primer choice and sequencing platform. *BMC Microbiology*, *16*(1), 123. https://doi.org/10.1186/s12866-016-0738-z

Fu, C., Keller, L., Bauer, A., Brönstrup, M., Froidbise, A., Hammann, P., … Müller, R. (2015). Biosynthetic Studies of Telomycin Reveal New Lipopeptides with Enhanced Activity. *Journal of the American Chemical Society*, *137*(24), 7692–7705. https://doi.org/10.1021/jacs.5b01794

Fujimori, D. G., Hrvatin, S., Neumann, C. S., Strieker, M., Marahiel, M. A., & Walsh, C. T. (2007). Cloning and characterization of the biosynthetic gene cluster for kutznerides. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(42), 16498–16503. https://doi.org/10.1073/pnas.0708242104

Fuller, A. T., Mellows, G., Woolford, M., Banks, G. T., Barrow, K. D., & Chain, E. B. (1971). Pseudomonic acid: an antibiotic produced by Pseudomonas fluorescens. *Nature*, *234*(5329), 416–417. https://doi.org/10.1038/234416a0

Gao, S. S., Hothersall, J., Wu, J., Murphy, A. C., Song, Z., Stephens, E. R., Thomas, C. M., Crump, M. P., Cox, R. J., Simpson, T. J., & Willis, C. L. (2014). Biosynthesis of mupirocin by *Pseudomonas fluorescens* NCIMB 10586 involves parallel pathways. *Journal of the American Chemical Society, 136*(14), 5501–5507. https://doi.org/10.1021/ja501731p

George, S., Pankhurst, L., Hubbard, A., Votintseva, A., Stoesser, N., Sheppard, A. E., … Phan, H. T. T. (2017). Resolving plasmid structures in Enterobacteriaceae using the MinION nanopore sequencer: assessment of MinION and MinION/Illumina hybrid data assembly approaches. *Microbial Genomics*, *3*(8), e000118. https://doi.org/10.1099/mgen.0.000118

Gerner-Smidt, P., Keiser-Nielsen, H., Dorsch, M., Stackebrandt, E., Ursing, J., Blom, J., Christensen, A. C., Christensen, J. J., Frederiksen, W., & Hoffmann, S. (1994). *Lautropia mirabilis* gen. nov., sp. nov., a gram-negative motile coccus with unusual morphology isolated from the human mouth. *Microbiology*, *140*(7), 1787–1797. https://doi.org/10.1099/13500872-140-7-1787

Ghrairi, T., Braiek, O. Ben, & Hani, K. (2015). Detection and characterization of a bacteriocin, putadicin T01, produced by Pseudomonas putida isolated from hot spring water. *APMIS*, *123*(3), 260–268. https://doi.org/10.1111/apm.12343

Gill, M. J., Simjee, S., Al-Hattawi, K., Robertson, B. D., Easmon, C. S., & Ison, C. A. (1998). Gonococcal resistance to beta-lactams and tetracycline involves mutation in loop 3 of the porin encoded at the penB locus. *Antimicrobial Agents and Chemotherapy*, *42*(11), 2799–2803.

Gillespie, D. E., Brady, S. F., Bettermann, A. D., Cianciotto, N. P., Liles, M. R., Rondon, M. R., Clardy, J., Goodman, R. M., & Handelsman, J. (2002). Isolation of antibiotics

turbomycin a and B from a metagenomic library of soil microbial DNA. *Applied and Environmental Microbiology, 68*(9), 4301–4306.

Gilliam, M., & Prest, D. B. (1987). Microbiology of feces of the larval honey bee, Apis mellifera. *Journal of Invertebrate Pathology*, *49*(1), 70–75. https://doi.org/https://doi.org/10.1016/0022-2011(87)90127-3

Glassing, A., Dowd, S. E., Galandiuk, S., Davis, B., & Chiodini, R. J. (2016). Inherent bacterial DNA contamination of extraction and sequencing reagents may affect interpretation of microbiota in low bacterial biomass samples. *Gut Pathogens*, *8*(1), 24. https://doi.org/10.1186/s13099-016-0103-7

Godiska, R., Mead, D., Dhodda, V., Wu, C., Hochstein, R., Karsi, A., … Ravin, N. (2010). Linear plasmid vector for cloning of repetitive or unstable sequences in Escherichia coli. *Nucleic Acids Research*, *38*(6), e88. https://doi.org/10.1093/nar/gkp1181

Godtfredsen, W. O., Lorck, H., van Tamelen, E. E., Willett, J. D., & Clayton, R. B. (1968). Biosynthesis of fusidic acid from squalene 2,3-oxide. *Journal of the American Chemical Society*, *90*(1), 208–209. https://doi.org/10.1021/ja01003a036

Godtfredsen, W., Roholt, K., & Tybring, L. (1962). Fucidin: a new orally active antibiotic. *Lancet*, *1*(7236), 928–931. https://doi.org/10.1016/s0140-6736(62)91968-2

Goldstein, S., Beka, L., Graf, J., & Klassen, J. L. (2019). Evaluation of strategies for the assembly of diverse bacterial genomes using MinION long-read sequencing. *BMC Genomics*, *20*(1), 23. https://doi.org/10.1186/s12864-018-5381-7

Goodwin, S., McPherson, J. D., & McCombie, W. R. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, *17*(6), 333–351. https://doi.org/10.1038/nrg.2016.49

Goris, J., Konstantinidis, K. T., Klappenbach, J. A., Coenye, T., Vandamme, P., & Tiedje, J. M. (2007). DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. *International Journal of Systematic and Evolutionary Microbiology*, *57*(Pt 1), 81–91. https://doi.org/10.1099/ijs.0.64483-0

Goto, Y., Li, B., Claesen, J., Shi, Y., Bibb, M. J., & van der Donk, W. A. (2010). Discovery of unique lanthionine synthetases reveals new mechanistic and evolutionary insights. *PLoS biology*, *8*(3), e1000339. https://doi.org/10.1371/journal.pbio.1000339

Griffith, R. S., & Peck, F. B. J. (n.d.). Vancomycin, a new antibiotic. III. Preliminary clinical and laboratory studies. *Antibiotics Annual*, *3*, 619–622.

Grimes, D. J., Woese, C. R., MacDonell, M. T., & Colwell, R. R. (1997). Systematic study of the genus *Vogesella* gen. nov. and its type species, *Vogesella indigofera* comb. nov. *International Journal of Systematic Bacteriology, 47*(1), 19–27. https://doi.org/10.1099/00207713-47-1-19

Gromadski, K. B., & Rodnina, M. V. (2004). Streptomycin interferes with conformational coupling between codon recognition and GTPase activation on the ribosome. *Nature Structural & Molecular Biology*, *11*(4), 316–322. https://doi.org/10.1038/nsmb742

Gudkov, A. T. (2001). Structure and Functions of the Prokaryotic Elongation Factor G. *Molecular Biology*, *35*(4), 552–558. https://doi.org/10.1023/A:1010570909693

Gunde-Cimerman, N., Plemenitaš, A., & Oren, A. (2018). Strategies of adaptation of microorganisms of the three domains of life to high salt concentrations. *FEMS Microbiology Reviews*, *42*(3), 353–375. https://doi.org/10.1093/femsre/fuy009

Guo, J., Rao, Z., Yang, T., Man, Z., Xu, M., & Zhang, X. (2014). High-level production of melanin by a novel isolate of Streptomyces kathirae. *FEMS Microbiology Letters*, *357*(1), 85–91. https://doi.org/10.1111/1574-6968.12497

Guo, Y., Zheng, W., Rong, X., & Huang, Y. (2008). A multilocus phylogeny of the Streptomyces griseus 16S rRNA gene clade: use of multilocus sequence analysis for streptomycete systematics. *International Journal of Systematic and Evolutionary Microbiology*, *58*(Pt 1), 149–159. https://doi.org/10.1099/ijs.0.65224-0

Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. *Bioinformatics*, *29*(8), 1072–1075. https://doi.org/10.1093/bioinformatics/btt086

Gürtler, H., Pedersen, R., Anthoni, U., Christophersen, C., Nielsen, P. H., Wellington, E. M., Pedersen, C., & Bock, K. (1994). Albaflavenone, a sesquiterpene ketone with a zizaene skeleton produced by a streptomycete with a new rope morphology. *The Journal of Antibiotics, 47*(4), 434–439. https://doi.org/10.7164/antibiotics.47.434

Gurusamy, K. S., Koti, R., Toon, C. D., Wilson, P., & Davidson, B. R. (2013). Antibiotic therapy for the treatment of methicillin-resistant Staphylococcus aureus (MRSA) infections in surgical wounds. *The Cochrane database of systematic reviews*, (8), CD009726. https://doi.org/10.1002/14651858.CD009726.pub2

Hammami, R., Zouhir, A., Le Lay, C., Ben Hamida, J., & Fliss, I. (2010). BACTIBASE second release: a database and tool platform for bacteriocin characterization. *BMC Microbiology*, *10*(1), 22. https://doi.org/10.1186/1471-2180-10-22

Han, Z., Sun, J., Lv, A., & Wang, A. (2019). Biases from different DNA extraction methods in intestine microbiome research based on 16S rDNA sequencing: a case in the koi carp, Cyprinus carpio var. Koi. *MicrobiologyOpen*, *8*(1), e00626. https://doi.org/10.1002/mbo3.626

Hancock, R. E. (1997). Peptide antibiotics. *Lancet*, *349*(9049), 418–422. https://doi.org/10.1016/S0140-6736(97)80051-7

Handelsman, J. (2004). Metagenomics: Application of Genomics to Uncultured Microorganisms. *Microbiology and Molecular Biology Reviews*, *68*(4), 669–685. https://doi.org/10.1128/MMBR.68.4.669-685.2004

Hayward, A. (1991). Lead, gout and Bath Spa therapy. In G. A. Kellaway (Ed.), *Hot springs of Bath : investigations of the thermal waters of the Avon Valley* (pp. 77–88). Bath: Bath City Council.

He, J., Magarvey, N., Piraee, M., & Vining, L. C. (2001). The gene cluster for chloramphenicol biosynthesis in Streptomyces venezuelae ISP5230 includes novel shikimate pathway homologues and a monomodular non-ribosomal peptide synthetase gene. *Microbiology*, *147*(10), 2817–2829. https://doi.org/10.1099/00221287-147-10-2817

Hegemann, J. D., & Süssmuth, R. D. (2020) Matters of class: coming of age of class III and IV lanthipeptides. *RSC Chemical Biology, 1*, 110-127. https://doi.org/10.1039/D0CB00073F

Heinrich, M., Barnes, J., Gibbons, S., & Williamson, E. M. (2021). *Fundamentals of Pharmacognosy and Phytotherapy* (2nd ed.). Edinburgh: Churchill Livingstone.

Hendlin, D., Stapley, E. O., Jackson, M., Wallick, H., Miller, A. K., Wolf, F. J., … Mochales, S. (1969). Phosphonomycin, a new antibiotic produced by strains of streptomyces. *Science*, *166*(3901), 122–123. https://doi.org/10.1126/science.166.3901.122

Hertweck C. (2009). The biosynthetic logic of polyketide diversity. *Angewandte Chemie (International ed. in English)*, *48*(26), 4688–4716. https://doi.org/10.1002/anie.200806121

Hobel, C. F. V., Marteinsson, V. T., Hreggvidsson, G. O., & Kristjánsson, J. K. (2005). Investigation of the Microbial Ecology of Intertidal Hot Springs by Using Diversity Analysis of 16S rRNA and Chitinase Genes. *Applied and Environmental Microbiology*, *71*(5), 2771–2776. https://doi.org/10.1128/AEM.71.5.2771-2776.2005

Hoeksema, H., Johnson, J. L., & Hinman, J. W. (1955). Structural studies on streptonivicin, 1 a new antibiotic. *Journal of the American Chemical Society*, *77*(24), 6710–6711. https://doi.org/10.1021/ja01629a129

Hojati, Z., Milne, C., Harvey, B., Gordon, L., Borg, M., Flett, F., Wilkinson, B., Sidebottom, P. J., Rudd, B. A., Hayes, M. A., Smith, C. P., & Micklefield, J. (2002). Structure, biosynthetic origin, and engineered biosynthesis of calcium-dependent antibiotics from *Streptomyces coelicolor*. *Chemistry & Biology*, *9*(11), 1175–1187. https://doi.org/10.1016/s1074-5521(02)00252-1

Holmes, N. A., Devine, R., Qin, Z., Seipke, R. F., Wilkinson, B., & Hutchings, M. I. (2018). Complete genome sequence of Streptomyces formicae KY5, the formicamycin producer. *Journal of Biotechnology*, *265*, 116–118. https://doi.org/10.1016/j.jbiotec.2017.11.011

Holmes, N. A., Innocent, T. M., Heine, D., Bassam, M. A., Worsley, S. F., Trottmann, F., Patrick, E. H., Yu, D. W., Murrell, J. C., Schiøtt, M., Wilkinson, B., Boomsma, J. J., & Hutchings, M. I. (2016). Genome Analysis of Two *Pseudonocardia* Phylotypes Associated with *Acromyrmex* Leafcutter Ants Reveals Their Biosynthetic Potential. *Frontiers in microbiology*, *7*, 2073. https://doi.org/10.3389/fmicb.2016.02073

Hopwood, D. A. (1999). Forty years of genetics with Streptomyces: from in vivo through in vitro to in silico. *Microbiology*, *145* (9), 2183–2202. https://doi.org/10.1099/00221287-145-9-2183

Horn, H., Cheng, C., Edrada-Ebel, R., Hentschel, U., & Abdelmohsen, U. R. (2015). Draft genome sequences of three chemically rich actinomycetes isolated from Mediterranean sponges. *Marine Genomics*, *24*, 285–287. https://doi.org/https://doi.org/10.1016/j.margen.2015.10.003

Hoshino, T., & Sato, T. (2002). Squalene-hopene cyclase: catalytic mechanism and substrate recognition. *Chemical Communications*, (4), 291–301. https://doi.org/10.1039/b108995c

Hotchkiss, R. D., & Dubos, R. J. (1940). Bactericidal fractions from an aerobic sporulating *Bacillus*. *Journal of Biological Chemistry*, *136*(3), 803–804. https://doi.org/10.1016/S0021-9258(18)73041-X

Hover, B. M., Kim, S.-H., Katz, M., Charlop-Powers, Z., Owen, J. G., Ternei, M. A., … Brady, S. F. (2018). Culture-independent discovery of the malacidins as calcium-dependent antibiotics with activity against multidrug-resistant Gram-positive pathogens. *Nature Microbiology*, *3*(4), 415–422. https://doi.org/10.1038/s41564-018-0110-1

Huang, T., Geng, H., Miyyapuram, V. R., Sit, C. S., Vederas, J. C., & Nakano, M. M. (2009). Isolation of a variant of subtilosin A with hemolytic activity. *Journal of bacteriology*, *191*(18), 5690–5696. https://doi.org/10.1128/JB.00541-09

Hubert, J., Nuzillard, JM. & Renault, JH. Dereplication strategies in natural product research: How many tools and methodologies behind the same concept?. *Phytochemistry Reviews 16*, 55–95 (2017). https://doi.org/10.1007/s11101-015-9448-7

Hughes, J., & Mellows, G. (1978). On the mode of action of pseudomonic acid: inhibition of protein synthesis in Staphylococcus aureus. *The Journal of Antibiotics*, *31*(4), 330–335. https://doi.org/10.7164/antibiotics.31.330

Huson, D. H., Beier, S., Flade, I., Górska, A., El-Hadidi, M., Mitra, S., Ruscheweyh, H. J., & Tappu, R. (2016). MEGAN Community Edition - Interactive Exploration and Analysis of Large-Scale Microbiome Sequencing Data. *PLoS Computational Biology, 12*(6), e1004957. https://doi.org/10.1371/journal.pcbi.1004957

Ibrahim, A., Yang, L., Johnston, C., Liu, X., Ma, B., & Magarvey, N. A. (2012). Dereplicating nonribosomal peptides using an informatic search algorithm for natural products (iSNAP) discovery. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(47), 19196–19201. https://doi.org/10.1073/pnas.1206376109

Ibrahim, M. E., Bilal, N. E., & Hamid, M. E. (2012). Increased multi-drug resistant Escherichia coli from hospitals in Khartoum state, Sudan. *African health sciences*, *12*(3), 368–375. https://doi.org/10.4314/ahs.v12i3.19

Issart, A., Godin, S., Preud'homme, H., Bierla, K., Allal, A., & Szpunar, J. (2019). Direct screening of food packaging materials for post-polymerization residues, degradation products and additives by liquid extraction surface analysis nanoelectrospray mass spectrometry (LESA-nESI-MS). *Analytica Chimica Acta*, *1058*, 117–126. https://doi.org/10.1016/j.aca.2019.01.028

Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T., & Aluru, S. (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

Jang, K. H., Nam, S.-J., Locke, J. B., Kauffman, C. A., Beatty, D. S., Paul, L. A., & Fenical, W. (2013). Anthracimycin, a Potent Anthrax Antibiotic from a Marine-Derived Actinomycete. *Angewandte Chemie International Edition*, *52*(30), 7822–7824. https://doi.org/10.1002/anie.201302749

Jay, Z. J., & Inskeep, W. P. (2015). The distribution, diversity, and importance of 16S rRNA gene introns in the order Thermoproteales. *Biology Direct*, *10*(1), 35. https://doi.org/10.1186/s13062-015-0065-6

Johnson, B. A., Anker, H., & Meleney, F. L. (1945). Bacitracin: A new antibiotic produced by a member of the *B. subtilis* group. *Science*, *102*(2650), 376–377. https://doi.org/10.1126/science.102.2650.376

Jørgensen, N. O. G., Brandt, K. K., Nybroe, O., & Hansen, M. (2010). *Vogesella mureinivorans* sp. nov., a peptidoglycan-degrading bacterium from lake water. *International Journal of Systematic and Evolutionary Microbiology*, *60*(10), 2467–2472. https://doi.org/10.1099/ijs.0.018630-0

Judge, K., Hunt, M., Reuter, S., Tracey, A., Quail, M. A., Parkhill, J., & Peacock, S. J. (2016). Comparison of bacterial genome assembly software for MinION data and

their applicability to medical microbiology. *Microbial Genomics*, *2*(9), e000085. https://doi.org/10.1099/mgen.0.000085

Just-Baringo, X., Albericio, F., & Álvarez, M. (2014). Thiopeptide antibiotics: retrospective and recent advances. *Marine drugs*, *12*(1), 317–351. https://doi.org/10.3390/md12010317

Kämpfer, P., Rückert, C., Blom, J., Goesmann, A., Wink, J., Kalinowski, J., & Glaeser, S. P. (2018). Streptomyces ciscaucasicus Sveshnikova et al. 1983 is a later subjective synonym of Streptomyces canus Heinemann et al. 1953. *International Journal of Systematic and Evolutionary Microbiology*, *68*(1), 42–46. https://doi.org/10.1099/ijsem.0.002418

Kapahi, M., & Sachdeva, S. (2019). Bioremediation Options for Heavy Metal Pollution. *Journal of Health and Pollution*, *9*(24), 191203. https://doi.org/10.5696/2156-9614-9.24.191203

Kapoor, G., Saigal, S., & Elongavan, A. (2017). Action and resistance mechanisms of antibiotics: A guide for clinicians. *Journal of Anaesthesiology Clinical Pharmacology*, *33*(3), 300–305. https://doi.org/10.4103/joacp.JOACP_349_15

Kawahara, S., Utsunomiya, C., Ishikawa, S., & Sekiguchi, J. (1997). Purification and characterization of an autolysin of *Bacillus polymyxa* var. colistinus which is most active at acidic pH. *Journal of Fermentation and Bioengineering*, *83*(5), 419–422. https://doi.org/https://doi.org/10.1016/S0922-338X(97)82994-7

Kellaway, G. A. (1991). Prelude. In G. A. Kellaway (Ed.), *Hot springs of Bath : investigations of the thermal waters of the Avon Valley* (pp. 13–24). Bath: Bath City Council.

Kelly, D. P., & Wood, A. P. (2000). Reclassification of some species of *Thiobacillus* to the newly designated genera *Acidithiobacillus* gen. nov., *Halothiobacillus* gen. nov. and *Thermithiobacillus* gen. nov. *International Journal of Systematic and Evolutionary Microbiology*, *50*(2), 511–516. https://doi.org/10.1099/00207713-50-2-511

Kembel, S. W., Wu, M., Eisen, J. A., & Green, J. L. (2012). Incorporating 16S Gene Copy Number Information Improves Estimates of Microbial Diversity and Abundance. *PLoS Computational Biology*, *8*(10), e1002743. https://doi.org/10.1371/journal.pcbi.1002743

Kersten, R. D., Lane, A. L., Nett, M., Richter, T. K. S., Duggan, B. M., Dorrestein, P. C., & Moore, B. S. (2013). Bioactivity-guided genome mining reveals the lomaiviticin biosynthetic gene cluster in Salinispora tropica. *Chembiochem : A European Journal of Chemical Biology*, *14*(8), 955–962. https://doi.org/10.1002/cbic.201300147

Kersten, R. D., Yang, Y. L., Xu, Y., Cimermancic, P., Nam, S. J., Fenical, W., Fischbach, M. A., Moore, B. S., & Dorrestein, P. C. (2011). A mass spectrometry-guided genome mining approach for natural product peptidogenomics. *Nature Chemical Biology, 7*(11), 794–802.

Kieleczawa, J. (2006). Fundamentals of sequencing of difficult templates--an overview. *Journal of Biomolecular Techniques : JBT*, *17*(3), 207–217.

Kieser, T., Bibb, M. J., Buttner, M. J., Chater, K. F., & Hopwood, D. A. (2000). *Practical Streptomyces Genetics*. Norwich: John Innes Foundation.

Kim, D., Song, L., Breitwieser, F. P., & Salzberg, S. L. (2016). Centrifuge: rapid and sensitive classification of metagenomic sequences. *Genome Research*, *26*(12), 1721–1729. https://doi.org/10.1101/gr.210641.116

Kim, J. H., Feng, Z., Bauer, J. D., Kallifidas, D., Calle, P. Y., & Brady, S. F. (2010). Cloning large natural product gene clusters from the environment: piecing environmental DNA gene clusters back together with TAR. *Biopolymers*, *93*(9), 833–844. https://doi.org/10.1002/bip.21450

Kim, S.-K. (Ed.). (2013). *Marine Microbiology-Bioactive Compounds and Biotechnological Applications*. John Wiley & Sons, Ltd.

Klassen, J. L., & Currie, C. R. (2012). Gene fragmentation in bacterial draft genomes: extent, consequences and mitigation. *BMC Genomics*, *13*(1), 14. https://doi.org/10.1186/1471-2164-13-14

Koberská, M., Kopecký, J., Olsovská, J., Jelínková, M., Ulanova, D., Man, P., … Janata, J. (2008). Sequence analysis and heterologous expression of the lincomycin biosynthetic cluster of the type strain Streptomyces lincolnensis ATCC 25466. *Folia Microbiologica*, *53*(5), 395–401. https://doi.org/10.1007/s12223-008-0060-8

Kocurek, K. I., Stones, L., Bunch, J., May, R. C., & Cooper, H. J. (2017). Top-Down LESA Mass Spectrometry Protein Analysis of Gram-Positive and Gram-Negative Bacteria. *Journal of the American Society for Mass Spectrometry*, *28*(10), 2066–2077. https://doi.org/10.1007/s13361-017-1718-8

Koller, K.-P. (2014). Antibiotics. Targets, Mechanisms and Resistance. Edited by Claudio O. Gualerzi, Letizia Brandi, Attilio Fabbretti, and Cynthia L. Pon. *Angewandte Chemie International Edition*, *53*(12), 3062. https://doi.org/10.1002/anie.201400593

Kolmogorov, M., Yuan, J., Lin, Y., & Pevzner, P. A. (2019). Assembly of long, error-prone reads using repeat graphs. *Nature Biotechnology*, *37*(5), 540–546. https://doi.org/10.1038/s41587-019-0072-8

Kong, D., Wang, X., Nie, J., & Niu, G. (2019). Regulation of Antibiotic Production by Signaling Molecules in *Streptomyces*. *Frontiers in Microbiology*, *10*, 2927. https://doi.org/10.3389/fmicb.2019.02927

Koren, S., Harhay, G. P., Smith, T. P., Bono, J. L., Harhay, D. M., Mcvey, S. D., Radune, D., Bergman, N. H., & Phillippy, A. M. (2013). Reducing assembly complexity of microbial genomes with single-molecule sequencing. *Genome Biology*, *14*(9), R101. https://doi.org/10.1186/gb-2013-14-9-r101

Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., & Phillippy, A. M. (2017). Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Research*, *27*(5), 722–736. https://doi.org/10.1101/gr.215087.116

Kotowska, M., & Pawlik, K. (2014). Roles of type II thioesterases and their application for secondary metabolite yield improvement. *Applied Microbiology and Biotechnology*, *98*(18), 7735–7746. https://doi.org/10.1007/s00253-014-5952-8

Kourtis, A. P., Sheriff, E. A., Weiner-Lastinger, L. M., Elmore, K., Preston, L. E., Dudeck, M., & McDonald, L. C. (2020). Antibiotic multi-drug-resistance of Escherichia coli causing device- and procedure-related infections in the United States reported to the National Healthcare Safety Network (NHSN), 2013-2017. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America*, ciaa1031. Advance online publication. https://doi.org/10.1093/cid/ciaa1031

Krismer, B., Weidenmaier, C., Zipperer, A., & Peschel, A. (2017). The commensal lifestyle of Staphylococcus aureus and its interactions with the nasal microbiota. *Nature reviews. Microbiology*, *15*(11), 675–687. https://doi.org/10.1038/nrmicro.2017.104

Kudo, F., & Eguchi, T. (2009). Biosynthetic genes for aminoglycoside antibiotics. *The Journal of Antibiotics*, *62*(9), 471–481. https://doi.org/10.1038/ja.2009.76

Kumagai, T., Koyama, Y., Oda, K., Noda, M., Matoba, Y., & Sugiyama, M. (2010). Molecular cloning and heterologous expression of a biosynthetic gene cluster for the antitubercular agent D-cycloserine produced by Streptomyces lavendulae. *Antimicrobial Agents and Chemotherapy*, *54*(3), 1132–1139. https://doi.org/10.1128/AAC.01226-09

Kumar, S., Stecher, G., & Tamura, K. (2016). MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Molecular Biology and Evolution*, *33*(7), 1870–1874. https://doi.org/10.1093/molbev/msw054

Kurosawa, H. (1952). The isolation of an antibiotic produced by a strain of *Streptomyces* K-300. *Yokohama Medical Bulletin*, *3*(6), 386–399.

Labeda, D. P. (1987). Transfer of the Type Strain of *Streptomyces erythraeus* (Waksman 1923) Waksman and Henrici 1948 to the Genus *Saccharopolyspora* Lacey and Goodfellow 1975 as *Saccharopolyspora erythraea* sp. nov., and Designation of a Neotype Strain for *Streptomyces erythraeus*. *International Journal of Systematic Bacteriology*, *37*(1), 19–22. https://doi.org/10.1099/00207713-37-1-19

Lambert, P. A. (2002). Cellular impermeability and uptake of biocides and antibiotics in Gram-positive bacteria and mycobacteria. *Journal of Applied Microbiology*, *92 Suppl*, 46S-54S.

Lampis, S., Santi, C., Ciurli, A., Andreolli, M., & Vallini, G. (2015). Promotion of arsenic phytoextraction efficiency in the fern Pteris vittata by the inoculation of As-resistant bacteria: a soil bioremediation perspective. *Frontiers in Plant Science*, *6*. https://doi.org/10.3389/fpls.2015.00080

Larsen, L., Nielsen, P., & Ahring, B. K. (1997). *Thermoanaerobacter mathranii* sp. nov., an ethanol-producing, extremely thermophilic anaerobic bacterium from a hot spring in Iceland. *Archives of Microbiology*, *168*(2), 114–119. https://doi.org/10.1007/s002030050476

Lasken, R. S., & Stockwell, T. B. (2007). Mechanism of chimera formation during the Multiple Displacement Amplification reaction. *BMC Biotechnology*, *7*(1), 19. https://doi.org/10.1186/1472-6750-7-19

Laursen, M. F., Dalgaard, M. D., & Bahl, M. I. (2017). Genomic GC-Content Affects the Accuracy of 16S rRNA Gene Sequencing Based Microbial Profiling due to PCR Bias. *Frontiers in Microbiology*, *8*. https://doi.org/10.3389/fmicb.2017.01934

Lawrence, D. P., Kroken, S., Pryor, B. M., & Arnold, A. E. (2011). Interkingdom Gene Transfer of a Hybrid NPS/PKS from Bacteria to Filamentous Ascomycota. *PLOS ONE*, *6*(11), e28231. Retrieved from https://doi.org/10.1371/journal.pone.0028231

Leipold, L., Dobrijevic, D., Jeffries, J. W. E., Bawn, M., Moody, T. S., Ward, J. M., & Hailes, H. C. (2019). The identification and use of robust transaminases from a domestic drain metagenome. *Green Chemistry*, *21*(1), 75–86. https://doi.org/10.1039/C8GC02986E

Lemetre, C., Maniko, J., Charlop-Powers, Z., Sparrow, B., Lowe, A. J., & Brady, S. F. (2017). Bacterial natural product biosynthetic domain composition in soil correlates with changes in latitude on a continent-wide scale. *Proceedings of the National Academy of Sciences*, *114*(44), 11615 LP – 11620. https://doi.org/10.1073/pnas.1710262114

Li, H. (2016). Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. *Bioinformatics*, *32*(14), 2103–2110. https://doi.org/10.1093/bioinformatics/btw152

Liem, M., Jansen, H. J., Dirks, R. P., Henkel, C. V, van Heusden, G. P. H., Lemmers, R. J. L. F., Omer, T., Shao, S., Punt., P. J., & Spaink, H. P. (2017). De novo whole-genome assembly of a wild type yeast isolate using nanopore sequencing. *F1000Research*, *6*, 618. https://doi.org/10.12688/f1000research.11146.2

Lim, H. K., Chung, E. J., Kim, J. C., Choi, G. J., Jang, K. S., Chung, Y. R., Cho, K. Y., & Lee, S. W. (2005). Characterization of a forest soil metagenome clone that confers indirubin and indigo production on *Escherichia coli*. *Applied and Environmental Microbiology*, *71*(12), 7768–7777. https://doi.org/10.1128/AEM.71.12.7768-7777.2005

Lim, K. T., Shukor, M. Y., & Wasoh, H. (2014). Physical, Chemical, and Biological Methods for the Removal of Arsenic Compounds. *BioMed Research International*, *2014*, 1–9. https://doi.org/10.1155/2014/503784

Ling L. L., Schneider T., Peoples A. J., Spoering A. L., Engels I., Conlon B. P., Mueller A., Schäberle T. F., Hughes D. E., Epstein S., Jones M., Lazarides L., Steadman V. A., Cohen D. R., Felix C. R., Fetterman K. A., Millett W. P., Nitti A. G., Zullo A. M., Chen C., & Lewis K. (2015). Erratum: A new antibiotic kills pathogens without detectable resistance. *, 520(7547), 388.* https://doi.org/10.1038/nature14303

Liu, D. Y., Li, Y., & Magarvey, N. A. (2016). Draft Genome Sequence of *Streptomyces canus* ATCC 12647, a Producer of Telomycin. *Genome Announcements*, *4*(2). https://doi.org/10.1128/genomeA.00173-16

Liu, W., Sun, F., & Hu, Y. (2018). Genome Mining-Mediated Discovery of a New Avermipeptin Analogue in *Streptomyces actuosus* ATCC 25421. *ChemistryOpen*, *7*(7), 558–561. https://doi.org/10.1002/open.201800130

Liu, Y., Lai, Q., Dong, C., Sun, F., Wang, L., Li, G., & Shao, Z. (2013). Phylogenetic Diversity of the *Bacillus pumilus* Group and the Marine Ecotype Revealed by Multilocus Sequence Analysis. *PLoS ONE*, *8*(11), e80097. https://doi.org/10.1371/journal.pone.0080097

Loman, N. J., Quick, J., & Simpson, J. T. (2015). A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nature Methods*, *12*(8), 733–735. https://doi.org/10.1038/nmeth.3444

Lynch, M., & Conery, J. S. (2003). The origins of genome complexity. *Science*, *302*(5649), 1401–1404. https://doi.org/10.1126/science.1089370

Ma, S., & Zhang, Q. (2020) Linaridin natural products. *Nat Prod Reports, 37*(9), 1152-1163. doi: 10.1039/c9np00074g.

MacLeod, A. J., Ross, H. B., Ozere, R. L., Digout, G., & van Rooyenc (1964). Linomycin: A new antibiotic active against Staphylococci and other Gram-positive cocci: clinical and laboratory studies. *Canadian Medical Association Journal, 91*(20), 1056–1060.

MacNeil, I. A., Tiong, C. L., Minor, C., August, P. R., Grossman, T. H., Loiacono, K. A., Lynch, B. A., Phillips, T., Narula, S., Sundaramoorthi, R., Tyler, A., Aldredge, T., Long, H.,

Gilman, M., Holt, D., & Osburne, M. S. (2001). Expression and isolation of antimicrobial small molecules from soil DNA libraries. *Journal of Molecular Microbiology and Biotechnology*, *3*(2), 301–308.

Madigan, M. T., & Martinko, J. (1997). *Brock Biology of Microorganism*. Pearson.

Mahenthiralingam, E., Song, L., Sass, A., White, J., Wilmot, C., Marchbank, A., Boaisha, O., Paine, J., Knight, D., & Challis, G. L. (2011). Enacyloxins are products of an unusual hybrid modular polyketide synthase encoded by a cryptic *Burkholderia ambifaria* Genomic Island. *Chemistry & Biology*, *18*(5), 665–677. https://doi.org/10.1016/j.chembiol.2011.01.020

Maksimov, M. O., & Link, A. J. (2014). Prospecting genomes for lasso peptides. *Journal of industrial microbiology & biotechnology*, *41*(2), 333–344. https://doi.org/10.1007/s10295-013-1357-4

Mardanov, A. V., Gumerov, V. M., Beletsky, A. V., Perevalova, A. A., Karpov, G. A., Bonch-Osmolovskaya, E. A., & Ravin, N. V. (2011). Uncultured archaea dominate in the thermal groundwater of Uzon Caldera, Kamchatka. *Extremophiles*, *15*(3), 365–372. https://doi.org/10.1007/s00792-011-0368-1

Margalith, P., & Beretta, G. (1960). Rifomycin. XI. taxonomic study on streptomyces mediterranei nov. sp. *Mycopathologia et Mycologia Applicata*, *13*(4), 321–330. https://doi.org/10.1007/BF02089930

Marijon, P., Chikhi, R., & Varré, J.-S. (2020). yacrd and fpa: upstream tools for long-read genome assembly. *Bioinformatics*, *36*(12), 3894–3896. https://doi.org/10.1093/bioinformatics/btaa262

Marques da Silva, R., Caugant, D. A., Eribe, E. R., Aas, J. A., Lingaas, P. S., Geiran, O., Tronstad, L., & Olsen, I. (2006). Bacterial diversity in aortic aneurysms determined by 16S ribosomal RNA gene analysis. *Journal of Vascular Surgery*, *44*(5), 1055–1060. https://doi.org/10.1016/j.jvs.2006.07.021

Martina, P. F., Martínez, M., Frada, G., Alvarez, F., Leguizamón, L., Prieto, C., … Von Specht, M. (2017). First time identification of *Pandoraea sputorum* from a patient with cystic fibrosis in Argentina: a case report. *BMC Pulmonary Medicine*, *17*(1), 33. https://doi.org/10.1186/s12890-017-0373-y

Martínez, L. M., Martinez, A., & Gosset, G. (2019). Production of Melanins With Recombinant Microorganisms. *Frontiers in Bioengineering and Biotechnology*, *7*, 285. https://doi.org/10.3389/fbioe.2019.00285

Martínez-Bueno, M., Gálvez, A., Valdivia, E., & Maqueda, M. (1990). A transferable plasmid associated with AS-48 production in *Enterococcus faecalis*. *Journal of Bacteriology*, *172*(5), 2817 LP – 2818. https://doi.org/10.1128/jb.172.5.2817-2818.1990

May, J. J., Wendrich, T. M., & Marahiel, M. A. (2001). The dhb operon of *Bacillus subtilis* encodes the biosynthetic template for the catecholic siderophore 2,3-dihydroxybenzoate-glycine-threonine trimeric ester bacillibactin. *The Journal of Biological Chemistry*, *276*(10), 7209–7217. https://doi.org/10.1074/jbc.M009140200

McClerren, A. L., Cooper, L. E., Quan, C., Thomas, P. M., Kelleher, N. L., & van der Donk, W. A. (2006). Discovery and *in vitro* biosynthesis of haloduracin, a two-component lantibiotic. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(46), 17243–17248. https://doi.org/10.1073/pnas.0606088103

McClung, C. R., Patriquin, D. G., & Davis, R. E. (1983). *Campylobacter nitrofigilis* sp. nov., a Nitrogen-Fixing Bacterium Associated with Roots of Spartina alterniflora Loisel. *International Journal of Systematic Bacteriology*, *33*(3), 605–612. https://doi.org/10.1099/00207713-33-3-605

McGuire, J. M., Bunch, R. L., Anderson, R. C., Boaz, H. E., Flynn, E. H., Powell, H. M., & Smith, J. W. (1952). Ilotycin, a new antibiotic. *Antibiotics & Chemotherapy*, *2*(6), 281–283.

McKinney, W. (2012). *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython*. O'Reilly Media, Inc.

Medema, M. H., & Fischbach, M. A. (2015). Computational approaches to natural product discovery. *Nature Chemical Biology*, *11*(9), 639–648. https://doi.org/10.1038/nchembio.1884

Medema, M. H., Blin, K., Cimermancic, P., de Jager, V., Zakrzewski, P., Fischbach, M. A., Weber, T., Takano, E., & Breitling, R. (2011). antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. *Nucleic Acids Research*, *39*(Web Server issue), W339-46. https://doi.org/10.1093/nar/gkr466

Medema, M. H., Kottmann, R., Yilmaz, P., Cummings, M., Biggins, J. B., Blin, K., … Glöckner, F. O. (2015). Minimum Information about a Biosynthetic Gene cluster. *Nature Chemical Biology*, *11*(9), 625–631. https://doi.org/10.1038/nchembio.1890

Medema, M. H., Paalvast, Y., Nguyen, D. D., Melnik, A., Dorrestein, P. C., Takano, E., & Breitling, R. (2014). Pep2Path: automated mass spectrometry-guided genome mining of peptidic natural products. *PLoS Computational Biology*, *10*(9), e1003822. https://doi.org/10.1371/journal.pcbi.1003822

Medema, M. H., Takano, E., & Breitling, R. (2013). Detecting sequence homology at the gene cluster level with MultiGeneBlast. *Molecular Biology and Evolution*, *30*(5), 1218–1223. https://doi.org/10.1093/molbev/mst025

Menzel, P., Ng, K. L., & Krogh, A. (2016). Fast and sensitive taxonomic classification for metagenomics with Kaiju. *Nature Communications*, *7*, 11257. https://doi.org/10.1038/ncomms11257

Meyer-Dombard, D. R., Shock, E. L., & Amend, J. P. (2005). Archaeal and bacterial communities in geochemically diverse hot springs of Yellowstone National Park, USA. *Geobiology*, *3*(3), 211–227. https://doi.org/10.1111/j.1472-4669.2005.00052.x

Miao, V., Coëffet-LeGal, M. F., Brian, P., Brost, R., Penn, J., Whiting, A., Martin, S., Ford, R., Parr, I., Bouchard, M., Silva, C. J., Wrigley, S. K., & Baltz, R. H. (2005). Daptomycin biosynthesis in *Streptomyces roseosporus*: cloning and analysis of the gene cluster and revision of peptide stereochemistry. *Microbiology*, *151*(5), 1507–1523. https://doi.org/10.1099/mic.0.27757-0

Michalopoulos, A. S., Livaditis, I. G., & Gougoutas, V. (2011). The revival of fosfomycin. *International Journal of Infectious Diseases : IJID : Official Publication of the International Society for Infectious Diseases*, *15*(11), e732-9. https://doi.org/10.1016/j.ijid.2011.07.007

Miller, D. E., Staber, C., Zeitlinger, J., & Hawley, R. S. (2018). Highly Contiguous Genome Assemblies of 15 Drosophila Species Generated Using Nanopore Sequencing. *G3*, *8*(10), 3131–3141. https://doi.org/10.1534/g3.118.200160

Miller, S. I. (2016). Antibiotic Resistance and Regulation of the Gram-Negative Bacterial Outer Membrane Barrier by Host Innate Immune Molecules. *MBio*, *7*(5). https://doi.org/10.1128/mBio.01541-16

Mills, S., Stanton, C., Hill, C., & Ross, R. P. (2011). New developments and applications of bacteriocins and peptides in foods. *Annual Review of Food Science and Technology*, *2*, 299–329. https://doi.org/10.1146/annurev-food-022510-133721

Milne, I., Stephen, G., Bayer, M., Cock, P. J., Pritchard, L., Cardle, L., Shaw, P. D., & Marshall, D. (2013). Using Tablet for visual exploration of second-generation sequencing data. *Briefings in Bioinformatics*, *14*(2), 193–202. https://doi.org/10.1093/bib/bbs012

Mo, T., Liu, W. Q., Ji, W., Zhao, J., Chen, T., Ding, W., Yu, S., & Zhang, Q. (2017). Biosynthetic Insights into Linaridin Natural Products from Genome Mining and Precursor Peptide Mutagenesis. *ACS chemical biology*, *12*(6), 1484–1488. https://doi.org/10.1021/acschembio.7b00262

Mohimani, H., Gurevich, A., Shlemov, A., Mikheenko, A., Korobeynikov, A., Cao, L., Shcherbin, E., Nothias, L. F., Dorrestein, P. C., & Pevzner, P. A. (2018). Dereplication of microbial metabolites through database search of mass spectra. *Nature Communications*, *9*(1), 4035. https://doi.org/10.1038/s41467-018-06082-8

Mohimani, H., Kersten, R. D., Liu, W. T., Wang, M., Purvine, S. O., Wu, S., Brewer, H. M., Pasa-Tolic, L., Bandeira, N., Moore, B. S., Pevzner, P. A., & Dorrestein, P. C. (2014). Automated genome mining of ribosomal peptide natural products. *ACS Chemical Biology*, *9*(7), 1545–1551. https://doi.org/10.1021/cb500199h

Mohimani, H., Liu, W.-T., Kersten, R. D., Moore, B. S., Dorrestein, P. C., & Pevzner, P. A. (2014). NRPquest: Coupling Mass Spectrometry and Genome Mining for Nonribosomal Peptide Discovery. *Journal of Natural Products*, *77*(8), 1902–1909. https://doi.org/10.1021/np500370c

Molohon, K. J., Melby, J. O., Lee, J., Evans, B. S., Dunbar, K. L., Bumpus, S. B., Kelleher, N. L., & Mitchell, D. A. (2011). Structure determination and interception of biosynthetic intermediates for the plantazolicin class of highly discriminating antibiotics. *ACS chemical biology*, *6*(12), 1307–1313. https://doi.org/10.1021/cb200339d

Morris, R. P., Leeds, J. A., Naegeli, H. U., Oberer, L., Memmert, K., Weber, E., LaMarche, M. J., Parker, C. N., Burrer, N., Esterow, S., Hein, A. E., Schmitt, E. K., & Krastel, P. (2009). Ribosomally synthesized thiopeptide antibiotics targeting elongation factor Tu. *Journal of the American Chemical Society*, *131*(16), 5946–5955. https://doi.org/10.1021/ja900488a

Mullai, V., & Menon, T. (2007). Bactericidal activity of different types of honey against clinical and environmental isolates of *Pseudomonas aeruginosa*. *Journal of Alternative and Complementary Medicine (New York, N.Y.)*, *13*(4), 439–441. https://doi.org/10.1089/acm.2007.6366

Muller, E., Narayanasamy, S., Zeimes, M., Laczny, C. C., Lebrun, L. A., Herold, M., Hicks, N. D., Gillece, J. D., Schupp, J. M., Keim, P., & Wilmes, P. (2017). First draft genome sequence of a strain belonging to the *Zoogloea* genus and its gene expression *in situ*. *Standards in Genomic Sciences*, *12*(1), 64. https://doi.org/10.1186/s40793-017-0274-y

Mutters, N. T., Mampel, A., Kropidlowski, R., Biehler, K., Günther, F., Bălu, I., Malek, V., & Frank, U. (2018). Treating urinary tract infections due to MDR E. coli with Isothiocyanates - a phytotherapeutic alternative to antibiotics?. *Fitoterapia*, *129*, 237–240. https://doi.org/10.1016/j.fitote.2018.07.012

Nelson, M. L., & Levy, S. B. (2011). The history of the tetracyclines. *Annals of the New York Academy of Sciences*, *1241*, 17–32. https://doi.org/10.1111/j.1749-6632.2011.06354.x

Nelson, R. E., Hatfield, K. M., Wolford, H., Samore, M. H., Scott, R. D., Reddy, S. C., Olubajo, B., Paul, P., Jernigan, J. A., & Baggs, J. (2021). National Estimates of Healthcare Costs Associated With Multidrug-Resistant Bacterial Infections Among Hospitalized Patients in the United States. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America*, *72*(Supplement_1), S17–S26. https://doi.org/10.1093/cid/ciaa1581

Neyraud, E., Palicki, O., Schwartz, C., Nicklaus, S., & Feron, G. (2012). Variability of human saliva composition: possible relationships with fat perception and liking. *Archives of Oral Biology*, *57*(5), 556–566. https://doi.org/10.1016/j.archoralbio.2011.09.016

Nicholls, S. M., Quick, J. C., Tang, S., & Loman, N. J. (2019). Ultra-deep, long-read nanopore sequencing of mock microbial community standards. *GigaScience*, *8*(5). https://doi.org/10.1093/gigascience/giz043

Nichols, D., Cahoon, N., Trakhtenberg, E. M., Pham, L., Mehta, A., Belanger, A., Kanigan, T., Lewis, K., & Epstein, S. S. (2010). Use of ichip for high-throughput *in situ* cultivation of "uncultivable" microbial species. *Applied and Environmental Microbiology*, *76*(8), 2445 – 2450. https://doi.org/10.1128/AEM.01754-09

Nongkhlaw, F. M. W., & Joshi, S. R. (2016). Horizontal Gene Transfer of the Non-ribosomal Peptide Synthetase Gene Among Endophytic and Epiphytic Bacteria Associated with Ethnomedicinal Plants. *Current Microbiology*, *72*(1), 1–11. https://doi.org/10.1007/s00284-015-0910-y

O'Neill, J. (2016). Tackling Drug-Resistant Infections Globally: final report and recommendations. Retrieved June 1, 2016, from https://amr-review.org

Olaitan, P. B., Adeleke, O. E., & Ola, I. O. (2007). Honey: a reservoir for microorganisms and an inhibitory agent for microbes. *African Health Sciences*, *7*(3), 159–165. https://doi.org/10.5555/afhs.2007.7.3.159

O'Leary N. A., Wright M. W., Brister J. R., Ciufo S., Haddad D., McVeigh R., Rajput B., Robbertse B., Smith-White B., Ako-Adjei D., Astashyn A., Badretdin A., Bao Y., Blinkova O., Brover V., Chetvernin V., Choi J., Cox E., Ermolaeva O., Farrell C. M., Goldfarb T., Gupta T., Haft D., Hatcher E., Hlavina W., Joardar V. S., Kodali V. K., Li W., Maglott D., Masterson P., McGarvey K. M., Murphy M. R., O'Neill K., Pujar S., Rangwala S. H., Rausch D., Riddick L. D., Schoch C., Shkeda A., Storz S. S., Sun H., Thibaud-Nissen F., Tolstoy I., Tully R. E., Vatsan A. R., Wallin C., Webb D., Wu W., Landrum M. J., Kimchi A., Tatusova T., DiCuccio M., Kitts P., Murphy T. D., & Pruitt K. D. (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*, *44*(D1), D733–D745. https://doi.org/10.1093/nar/gkv1189

Oliver, W. T., Wells, J. E., & Maxwell, C. V. (2014). Lysozyme as an alternative to antibiotics improves performance in nursery pigs during an indirect immune challenge1,2. *Journal of Animal Science*, *92*(11), 4927–4934. https://doi.org/10.2527/jas.2014-8033

Olofsson, T. C., Butler, È., Markowicz, P., Lindholm, C., Larsson, L., & Vásquez, A. (2016). Lactic acid bacterial symbionts in honeybees - an unknown key to honey's antimicrobial and therapeutic activities. *International Wound Journal*, *13*(5), 668–679. https://doi.org/10.1111/iwj.12345

Omura, S., Ikeda, H., Ishikawa, J., Hanamoto, A., Takahashi, C., Shinose, M., Takahashi, Y., Horikawa, H., Nakazawa, H., Osonoe, T., Kikuchi, H., Shiba, T., Sakaki, Y., & Hattori, M. (2001). Genome sequence of an industrial microorganism *Streptomyces avermitilis*: Deducing the ability of producing secondary metabolites. *Proceedings of the National Academy of Sciences*, *98*(21), 12215 – 12220. https://doi.org/10.1073/pnas.211433198

Owen, R. J. (1989). Chromosomal DNA fingerprinting--a new method of species and strain identification applicable to microbial pathogens. *Journal of Medical Microbiology*, *30*(2), 89–99. https://doi.org/10.1099/00222615-30-2-89

Pal, C., Bengtsson-Palme, J., Rensing, C., Kristiansson, E., & Larsson, D. G. J. (2014). BacMet: antibacterial biocide and metal resistance genes database. *Nucleic Acids Research*, *42*(D1), D737–D743. https://doi.org/10.1093/nar/gkt1252

Papp-Wallace, K. M., Endimiani, A., Taracila, M. A., & Bonomo, R. A. (2011). Carbapenems: Past, Present, and Future. *Antimicrobial Agents and Chemotherapy*, *55*(11), 4943 LP – 4960. https://doi.org/10.1128/AAC.00296-11

Pfang, B. G., García-Cañete, J., García-Lasheras, J., Blanco, A., Auñón, Á., Parron-Cambero, R., Macías-Valcayo, A., & Esteban, J. (2019). Orthopedic Implant-Associated Infection by Multidrug Resistant *Enterobacteriaceae*. *Journal of clinical medicine*, *8*(2), 220. https://doi.org/10.3390/jcm8020220

Pogliano, J., Pogliano, N., & Silverman, J. A. (2012). Daptomycin-mediated reorganization of membrane architecture causes mislocalization of essential cell division proteins. *Journal of Bacteriology*, *194*(17), 4494–4504. https://doi.org/10.1128/JB.00011-12

Poole, K. (2000). Efflux-mediated resistance to fluoroquinolones in Gram-negative bacteria. *Antimicrobial Agents and Chemotherapy*, *44*(9), 2233–2241. https://doi.org/10.1128/aac.44.9.2233-2241.2000

Poralla, K., Muth, G., & Härtner, T. (2000). Hopanoids are formed during transition from substrate to aerial hyphae in *Streptomyces coelicolor* A3(2). *FEMS Microbiology Letters*, *189*(1), 93–95. https://doi.org/10.1111/j.1574-6968.2000.tb09212.x

Portmann, C., Blom, J. F., Kaiser, M., Brun, R., Jüttner, F., & Gademann, K. (2008). Isolation of aerucyclamides C and D and structure revision of microcyclamide 7806A: heterocyclic ribosomal peptides from *Microcystis aeruginosa* PCC 7806 and their antiparasite evaluation. *Journal of Natural Products*, *71*(11), 1891–1896. https://doi.org/10.1021/np800409z

Prosser, G. A., & de Carvalho, L. P. S. (2013). Kinetic mechanism and inhibition of *Mycobacterium tuberculosis* D-alanine:D-alanine ligase by the antibiotic D-cycloserine. *The FEBS Journal*, *280*(4), 1150–1166. https://doi.org/10.1111/febs.12108

Pulsawat, N., Kitani, S., & Nihira, T. (2007). Characterization of biosynthetic gene cluster for the production of virginiamycin M, a streptogramin type A antibiotic, in *Streptomyces virginiae*. *Gene*, *393*(1–2), 31–42. https://doi.org/10.1016/j.gene.2006.12.035

Putri, S. P., Kinoshita, H., Ihara, F., Igarashi, Y., & Nihira, T. (2010). Ophiosetin, a new tetramic acid derivative from the mycopathogenic fungus *Elaphocordyceps ophioglossoides*. *The Journal of Antibiotics*, *63*(4), 195–198. https://doi.org/10.1038/ja.2010.8

Qin, Z., Munnoch, J. T., Devine, R., Holmes, N. A., Seipke, R. F., Wilkinson, K. A., … Hutchings, M. I. (2017). Formicamycins, antibacterial polyketides produced by *Streptomyces formicae* isolated from African Tetraponera plant-ants. *Chemical Science*, *8*(4), 3218–3227. https://doi.org/10.1039/C6SC04265A

Queenan, A. M., & Bush, K. (2007). Carbapenemases: the versatile beta-lactamases. *Clinical Microbiology Reviews*, *20*(3), 440–458, table of contents. https://doi.org/10.1128/CMR.00001-07

Quick J., Loman N. J., Duraffour S., Simpson J. T., Severi E., Cowley L., Bore J. A., Koundouno R., Dudas G., Mikhail A., Ouédraogo N., Afrough B., Bah A., Baum J. H., Becker-Ziaja B., Boettcher J. P., Cabeza-Cabrerizo M., Camino-Sanchez A., Carter L. L., Doerrbecker J., Enkirch T., Dorival I. G. G., Hetzelt N., Hinzmann J., Holm T., Kafetzopoulou L. E., Koropogui M., Kosgey A., Kuisma E., Logue C. H., Mazzarelli A., Meisel S., Mertens M., Michel J., Ngabo D., Nitzsche K., Pallash E., Patrono L. V., Portmann J., Repits J.G., Rickett N. Y., Sachse A., Singethan K., Vitoriano I., Yemanaberhan R. L., Zekeng E. G., Trina R., Bello A., Sall A. A., Faye O., Faye O., Magassouba N., Williams C. V. , Amburgey V., Winona L., Davis E., Gerlach J., Washington F., Monteil V., Jourdain M., Bererd M., Camara A., Somlare H., Camara A., Gerard M., Bado G., Baillet B., Delaune D., Nebie K. Y., Diarra A., Savane Y., Pallawo R. B., Gutierrez G. J., Milhano N., Roger I., Williams C. J., Yattara F., Lewandowski K., Taylor J., Rachwal P., Turner D., Pollakis G., Hiscox J. A., Matthews D. A., O'Shea M. K., Johnston A. M., Wilson D., Hutley E., Smit E., Di Caro A., Woelfel R., Stoecker K., Fleischmann E., Gabriel M., Weller S. A., Koivogui L., Diallo B., Keita S., Rambaut A., Formenty P., Gunther S., & Carroll M. W. (2016). Real-time, portable genome sequencing for Ebola surveillance. *Nature*, *530*(7589), 228–232. https://doi.org/10.1038/nature16996

Ramirez, M. S., & Tolmasky, M. E. (2010). Aminoglycoside modifying enzymes. *Drug Resistance Updates : Reviews and Commentaries in Antimicrobial and Anticancer Chemotherapy*, *13*(6), 151–171. https://doi.org/10.1016/j.drup.2010.08.003

Rawat, D., & Nair, D. (2010). Extended-spectrum β-lactamases in Gram Negative Bacteria. *Journal of Global Infectious Diseases*, *2*(3), 263–274. https://doi.org/10.4103/0974-777X.68531

Rawlings, B. J. (2001). Type I polyketide biosynthesis in bacteria (Part A--erythromycin biosynthesis). *Natural Product Reports*, *18*(2), 190–227. https://doi.org/10.1039/b009329g

Reddy, B. V. B., Kallifidas, D., Kim, J. H., Charlop-Powers, Z., Feng, Z., & Brady, S. F. (2012). Natural Product Biosynthetic Gene Diversity in Geographically Distinct Soil Microbiomes. *Applied and Environmental Microbiology*, *78*(10), 3744 LP – 3752. https://doi.org/10.1128/AEM.00102-12

Redgrave, L. S., Sutton, S. B., Webber, M. A., & Piddock, L. J. V. (2014). Fluoroquinolone resistance: mechanisms, impact on bacteria, and role in evolutionary success. *Trends in Microbiology*, *22*(8), 438–445. https://doi.org/10.1016/j.tim.2014.04.007

Reinhold-Hurek, B., Tan, Z., & Hurek, T. (2015). Azoarcus. In *Bergey's Manual of Systematics of Archaea and Bacteria* (pp. 1–19). Chichester, UK: John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118960608.gbm00994

Reshetnikov, A. S., Khmelenina, V. N., Mustakhimov, I. I., & Trotsenko, Y. A. (2011). Chapter Two - Genes and Enzymes of Ectoine Biosynthesis in Halotolerant Methanotrophs. In A. C. Rosenzweig & S. W. B. T.-M. in E. Ragsdale (Eds.), *Methods in Methane Metabolism, Part B: Methanotrophy* (Vol. 495, pp. 15–30). Academic Press. https://doi.org/https://doi.org/10.1016/B978-0-12-386905-0.00002-4

Reynolds W. F. (2017) Natural Product Structure Elucidation by NMR Spectroscopy. In Badal S., Delgoda, R. (eds) *Pharmacognosy* (pp 567-596). Academic Press). https://doi.org/10.1016/B978-0-12-802104-0.00029-9.

Reysenbach, A. L., Longnecker, K., & Kirshtein, J. (2000). Novel bacterial and archaeal lineages from an in situ growth chamber deployed at a Mid-Atlantic Ridge hydrothermal

vent. *Applied and environmental microbiology*, *66*(9), 3798–3806. https://doi.org/10.1128/aem.66.9.3798-3806.2000

Rohmer, M., Bouvier-Nave, P., & Ourisson, G. (1984). Distribution of Hopanoid Triterpenes in Prokaryotes. *Microbiology*, *130*(5), 1137–1150. https://doi.org/10.1099/00221287-130-5-1137

Roongsawang, N., Washio, K., & Morikawa, M. (2010). Diversity of nonribosomal peptide synthetases involved in the biosynthesis of lipopeptide biosurfactants. *International Journal of Molecular Sciences*, *12*(1), 141–172. https://doi.org/10.3390/ijms12010141

Roskoski, R. J., Gevers, W., Kleinkauf, H., & Lipmann, F. (1970). Tyrocidine biosynthesis by three complementary fractions from *Bacillus brevis* (ATCC 8185). *Biochemistry*, *9*(25), 4839–4845. https://doi.org/10.1021/bi00827a002

Ross, J. N., Fields, F. R., Kalwajtys, V. R., Gonzalez, A. J., O'Connor, S., Zhang, A., Moran, T. E., Hammers, D. E., Carothers, K. E., & Lee, S. W. (2020). Synthetic Peptide Libraries Designed From a Minimal Alpha-Helical Domain of AS-48-Bacteriocin Homologs Exhibit Potent Antibacterial Activity. *Frontiers in microbiology*, *11*, 589666. https://doi.org/10.3389/fmicb.2020.589666

Rutledge, P. J., & Challis, G. L. (2015). Discovery of microbial natural products by activation of silent biosynthetic gene clusters. *Nature Reviews Microbiology*, *13*(8), 509–523. https://doi.org/10.1038/nrmicro3496

Rychlik, W. (1995). Selection of primers for polymerase chain reaction. *Molecular Biotechnology*, *3*(2), 129–134. https://doi.org/10.1007/BF02789108

Sáenz, J. P., Grosser, D., Bradley, A. S., Lagny, T. J., Lavrynenko, O., Broda, M., & Simons, K. (2015). Hopanoids as functional analogues of cholesterol in bacterial membranes. *Proceedings of the National Academy of Sciences*, *112*(38), 11971 LP – 11976. https://doi.org/10.1073/pnas.1515607112

Sağ, Y., & Kutsal, T. (1995). Biosorption of heavy metals by *Zoogloea ramigera*: use of adsorption isotherms and a comparison of biosorption characteristics. *The Chemical Engineering Journal and the Biochemical Engineering Journal*, *60*(1–3), 181–188. https://doi.org/10.1016/0923-0467(95)03014-X

Saha, S., Kapoor, S., Tariq, R., Schuetz, A. N., Tosh, P. K., Pardi, D. S., & Khanna, S. (2019). Increasing antibiotic resistance in Clostridioides difficile: A systematic review and meta-analysis. *Anaerobe*, *58*, 35–46. https://doi.org/10.1016/j.anaerobe.2019.102072

Saiki, R. K., Gelfand, D. H., Stoffel, S., Scharf, S. J., Higuchi, R., Horn, G. T., Mullis, K. B., & Erlich, H. A. (1988). Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science*, *239*(4839), 487–491. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/2448875

Saikia, R., Gogoi, D. K., Mazumder, S., Yadav, A., Sarma, R. K., Bora, T. C., & Gogoi, B. K. (2011). *Brevibacillus laterosporus* strain BPM3, a potential biocontrol agent isolated from a natural hot water spring of Assam, India. *Microbiological Research*, *166*(3), 216–225. https://doi.org/10.1016/j.micres.2010.03.002

Sambrook, J. F., & Russell, D. W. (2001). *Molecular Cloning: A Laboratory Manual* (3rd ed.). Cold Spring Harbor: Cold Spring Harbor Laboratory Press.

Sánchez-Hidalgo, M., Montalbán-López, M., Cebrián, R., Valdivia, E., Martínez-Bueno, M., & Maqueda, M. (2011). AS-48 bacteriocin: close to perfection. *Cellular and Molecular Life Sciences : CMLS*, *68*(17), 2845–2857. https://doi.org/10.1007/s00018-011-0724-4

Sato, K., Okazaki, T., Maeda, K., & Okami, Y. (1978). New antibiotics, aplasmomycins B and C. *The Journal of Antibiotics*, *31*(6), 632–635. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/681248

Satyapal, G. K., Mishra, S. K., Srivastava, A., Ranjan, R. K., Prakash, K., Haque, R., & Kumar, N. (2018). Possible bioremediation of arsenic toxicity by isolating indigenous bacteria from the middle Gangetic plain of Bihar, India. *Biotechnology Reports*, *17*, 117–125. https://doi.org/10.1016/j.btre.2018.02.002

Schatz, A., Bugie, E., & Waksman, S. A. (2005). Streptomycin, a substance exhibiting antibiotic activity against gram-positive and gram-negative bacteria. 1944. *Clinical Orthopaedics and Related Research*, (437), 3–6. https://doi.org/10.1097/01.blo.0000175887.98112.fe

Schirmer, A., Gadkari, R., Reeves, C. D., Ibrahim, F., DeLong, E. F., & Hutchinson, C. R. (2005). Metagenomic analysis reveals diverse polyketide synthase gene clusters in microorganisms associated with the marine sponge *Discodermia dissoluta*. *Applied and Environmental Microbiology*, *71*(8), 4840–4849. https://doi.org/10.1128/AEM.71.8.4840-4849.2005

Schmerk, C. L., Bernards, M. A., & Valvano, M. A. (2011). Hopanoid Production Is Required for Low-pH Tolerance, Antimicrobial Resistance, and Motility in *Burkholderia cenocepacia. Journal of Bacteriology*, *193*(23), 6712 LP – 6723. https://doi.org/10.1128/JB.05979-11

Schulz zur Wiesch, P., Engelstädter, J., & Bonhoeffer, S. (2010). Compensation of fitness costs and reversibility of antibiotic resistance mutations. *Antimicrobial Agents and Chemotherapy*, *54*(5), 2085–2095. https://doi.org/10.1128/AAC.01460-09

Schwarzer, D., Finking, R., & Marahiel, M. A. (2003). Nonribosomal peptides: from genes to products. *Natural Product Reports*, *20*(3), 275–287. https://doi.org/10.1039/b111145k

Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, *30*(14), 2068–2069. https://doi.org/10.1093/bioinformatics/btu153

Segata, N., Boernigen, D., Tickle, T. L., Morgan, X. C., Garrett, W. S., & Huttenhower, C. (2013). Computational meta'omics for microbial community studies. *Molecular Systems Biology*, *9*(1), 666. https://doi.org/10.1038/msb.2013.22

Segata, N., Börnigen, D., Morgan, X. C., & Huttenhower, C. (2013). PhyloPhlAn is a new method for improved phylogenetic and taxonomic placement of microbes. *Nature Communications*, *4*, 2304. https://doi.org/10.1038/ncomms3304

Seipke, R. F. (2015). Strain-level diversity of secondary metabolism in Streptomyces albus. *PloS One*, *10*(1), e0116457. https://doi.org/10.1371/journal.pone.0116457

Seipke, R. F., & Loria, R. (2009). Hopanoids are not essential for growth of Streptomyces scabies 87-22. *Journal of Bacteriology*, *191*(16), 5216–5223. https://doi.org/10.1128/JB.00390-09

Sentausa, E., & Fournier, P.-E. (2013). Advantages and limitations of genomics in prokaryotic taxonomy. *Clinical Microbiology and Infection : The Official Publication of the*

European Society of Clinical Microbiology and Infectious Diseases, *19*(9), 790–795. https://doi.org/10.1111/1469-0691.12181

Seung, K. J., Keshavjee, S., & Rich, M. L. (2015). Multidrug-Resistant Tuberculosis and Extensively Drug-Resistant Tuberculosis. *Cold Spring Harbor perspectives in medicine*, *5*(9), a017863. https://doi.org/10.1101/cshperspect.a017863

Severinov, K., & Nair, S. K. (2012). Microcin C: biosynthesis and mechanisms of bacterial resistance. *Future Microbiology*, *7*(2), 281–289. https://doi.org/10.2217/fmb.11.148

Seymour, G. J., Ford, P. J., Cullinan, M. P., Leishman, S., & Yamazaki, K. (2007). Relationship between periodontal infections and systemic disease. *Clinical Microbiology and Infection : The Official Publication of the European Society of Clinical Microbiology and Infectious Diseases*, *13 Suppl 4*, 3–10. https://doi.org/10.1111/j.1469-0691.2007.01798.x

Sherlock, O., Dolan, A., Athman, R., Power, A., Gethin, G., Cowman, S., & Humphreys, H. (2010). Comparison of the antimicrobial activity of Ulmo honey from Chile and Manuka honey against methicillin-resistant *Staphylococcus aureus*, *Escherichia coli* and *Pseudomonas aeruginosa*. BMC Complementary and Alternative Medicine, 10(1), 47. https://doi.org/10.1186/1472-6882-10-47

Sheu, S.-Y., Chen, J.-C., Young, C.-C., & Chen, W.-M. (2013). *Vogesella fluminis* sp. nov., isolated from a freshwater river, and emended description of the genus *Vogesella*. *International Journal of Systematic and Evolutionary Microbiology*, *63*(Pt_8), 3043–3049. https://doi.org/10.1099/ijs.0.048629-0

Shimamura, H., Gouda, H., Nagai, K., Hirose, T., Ichioka, M., Furuya, Y., Kobayashi, Y., Hirono, S., Sunazuka, T., & Omura, S. (2009). Structure determination and total synthesis of bottromycin A2: a potent antibiotic against MRSA and VRE. *Angewandte Chemie (International ed. in English)*, *48*(5), 914–917. https://doi.org/10.1002/anie.200804138

Shin, J., Lee, S., Go, M.-J., Lee, S. Y., Kim, S. C., Lee, C.-H., & Cho, B.-K. (2016). Analysis of the mouse gut microbiome using full-length 16S rRNA amplicon sequencing. *Scientific Reports*, *6*(1), 29681. https://doi.org/10.1038/srep29681

Shin, S. C., Kim, H., Lee, J. H., Kim, H. W., Park, J., Choi, B. S., Lee, S. C., Kim, J. H., Lee, H., & Kim, S. (2019). Nanopore sequencing reads improve assembly and gene annotation of the *Parochlus steinenii* genome. *Scientific Reports*, *9*(1), 5095. https://doi.org/10.1038/s41598-019-41549-8

Shivaji, S., Chaturvedi, P., Suresh, K., Reddy, G. S. N., Dutt, C. B. S., Wainwright, M., … Bhargava, P. M. (2006). *Bacillus aerius* sp. nov., *Bacillus aerophilus* sp. nov., *Bacillus stratosphericus* sp. nov. and *Bacillus altitudinis* sp. nov., isolated from cryogenic tubes used for collecting air samples from high altitudes. *International Journal of Systematic and Evolutionary Microbiology*, *56*(7), 1465–1473. https://doi.org/10.1099/ijs.0.64029-0

Shukla, R., Sarim, K. M., & Singh, D. P. (2020). Microbe-mediated management of arsenic contamination: current status and future prospects. *Environmental Sustainability*, *3*(1), 83–90. https://doi.org/10.1007/s42398-019-00090-0

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V, & Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, *31*(19), 3210–3212. https://doi.org/10.1093/bioinformatics/btv351

Singh, R., Bishnoi, N. R., & Kirrolia, A. (2013). Evaluation of Pseudomonas aeruginosa an innovative bioremediation tool in multi metals ions from simulated system using multi

response methodology. *Bioresource Technology*, *138*, 222–234. https://doi.org/10.1016/j.biortech.2013.03.100

Sivaperumal, P., Kamala, K., Rajaram, R., & Mishra, S. S. (2014). Melanin from marine Streptomyces sp. (MVCS13) with potential effect against ornamental fish pathogens of *Carassius auratus* (Linnaeus, 1758). *Biocatalysis and Agricultural Biotechnology*, *3*(4), 134–141. https://doi.org/https://doi.org/10.1016/j.bcab.2014.09.007

Sivonen, K., Leikoski, N., Fewer, D. P., & Jokela, J. (2010). Cyanobactins-ribosomal cyclic peptides produced by cyanobacteria. *Applied microbiology and biotechnology*, *86*(5), 1213–1225. https://doi.org/10.1007/s00253-010-2482-x

Skerman, V. B. D., McGowan, V., & Sneath, P. H. A. (Eds.). (1989). *Approved Lists of Bacterial Names (Amended Edition)*. *American Society for Microbiology*, Washington, D.C., 188 pp

Sköld, O. (2001). Resistance to trimethoprim and sulfonamides. *Veterinary Research*, *32*(3–4), 261–273. https://doi.org/10.1051/vetres:2001123

Smith, D. J., Burnham, M. K., Edwards, J., Earl, A. J., & Turner, G. (1990). Cloning and heterologous expression of the penicillin biosynthetic gene cluster from *Penicillum chrysogenum*. *Bio/Technology*, *8*(1), 39–41. https://doi.org/10.1038/nbt0190-39

Spieck, E., & Bock, E. (2015). *Nitrospira*. In *Bergey's Manual of Systematics of Archaea and Bacteria* (pp. 1–4). Chichester, UK: John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118960608.gbm00780

Stahl, D. A., & de la Torre, J. R. (2012). Physiology and Diversity of Ammonia-Oxidizing Archaea. *Annual Review of Microbiology*, *66*(1), 83–101. https://doi.org/10.1146/annurev-micro-092611-150128

Stanley, R. E., Blaha, G., Grodzicki, R. L., Strickler, M. D., & Steitz, T. A. (2010). The structures of the anti-tuberculosis antibiotics viomycin and capreomycin bound to the 70S ribosome. *Nature Structural & Molecular Biology*, *17*(3), 289–293. https://doi.org/10.1038/nsmb.1755

Steffensky, M., Mühlenweg, A., Wang, Z. X., Li, S. M., & Heide, L. (2000). Identification of the novobiocin biosynthetic gene cluster of Streptomyces spheroides NCIB 11891. *Antimicrobial Agents and Chemotherapy*, *44*(5), 1214–1222. https://doi.org/10.1128/aac.44.5.1214-1222.2000

Sterner O. (2012) Isolation of Microbial Natural Products. In: Sarker S., Nahar L. (eds) *Natural Products Isolation. Methods in Molecular Biology (Methods and Protocols),* vol 864 (pp 393-413). Humana Press. https://doi.org/10.1007/978-1-61779-624-1_15

Stewart, R. D., Auffret, M. D., Warr, A., Walker, A. W., Roehe, R., & Watson, M. (2019). Compendium of 4,941 rumen metagenome-assembled genomes for rumen microbiome biology and enzyme discovery. *Nature Biotechnology*, *37*(8), 953–961. https://doi.org/10.1038/s41587-019-0202-3

Stewart, R. D., Auffret, M. D., Warr, A., Wiser, A. H., Press, M. O., Langford, K. W., Liachko, I., Snelling, T. J., Dewhurst, R. J., Walker, A. W., Roehe, R., & Watson, M. (2018). Assembly of 913 microbial genomes from metagenomic sequencing of the cow rumen. *Nature Communications*, *9*(1), 870. https://doi.org/10.1038/s41467-018-03317-6

Takano, E. (2006). Gamma-butyrolactones: Streptomyces signalling molecules regulating antibiotic production and differentiation. *Current Opinion in Microbiology*, *9*(3), 287–294. https://doi.org/10.1016/j.mib.2006.04.003

Takano, H. (2016). The regulatory mechanism underlying light-inducible production of carotenoids in nonphototrophic bacteria. *Bioscience, Biotechnology, and Biochemistry*, *80*(7), 1264–1273. https://doi.org/10.1080/09168451.2016.1156478

Tambadou, F., Caradec, T., Gagez, A. L., Bonnet, A., Sopéna, V., Bridiau, N., Thiéry, V., Didelot, S., Barthélémy, C., & Chevrot, R. (2015). Characterization of the colistin (polymyxin E1 and E2) biosynthetic gene cluster. *Archives of Microbiology*, *197*(4), 521–532. https://doi.org/10.1007/s00203-015-1084-5

Tamura, K., Stecher, G., Peterson, D., Filipski, A., & Kumar, S. (2013). MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Molecular Biology and Evolution*, *30*(12), 2725–2729. https://doi.org/10.1093/molbev/mst197

Tan, S., Moore, G., & Nodwell, J. (2019). Put a Bow on It: Knotted Antibiotics Take Center Stage. *Antibiotics, 8*(3), 117. https://doi.org/10.3390/antibiotics8030117

Tenson, T., Lovmar, M., & Ehrenberg, M. (2003). The mechanism of action of macrolides, lincosamides and streptogramin B reveals the nascent peptide exit path in the ribosome. *Journal of Molecular Biology*, *330*(5), 1005–1014. https://doi.org/10.1016/s0022-2836(03)00662-4

The Editorial Board. Pandoraea. (2015). In *Bergey's Manual of Systematics of Archaea and Bacteria* (pp. 1–3). Chichester, UK: John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118960608.gbm00938

The Editorial Board. Propionivibrio. (2015). In *Bergey's Manual of Systematics of Archaea and Bacteria* (pp. 1–2). Chichester, UK: John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118960608.gbm01002

Theobald, S., Vesth, T. C., & Andersen, M. R. (2019). Genus level analysis of PKS-NRPS and NRPS-PKS hybrids reveals their origin in Aspergilli. *BMC Genomics*, *20*(1), 847. https://doi.org/10.1186/s12864-019-6114-2

Thomas, M. G., Chan, Y. A., & Ozanick, S. G. (2003). Deciphering tuberactinomycin biosynthesis: isolation, sequencing, and annotation of the viomycin biosynthetic gene cluster. *Antimicrobial Agents and Chemotherapy*, *47*(9), 2823–2830. https://doi.org/10.1128/aac.47.9.2823-2830.2003

Thompson, M. G., Corey, B. W., Si, Y., Craft, D. W., & Zurawski, D. V. (2012). Antibacterial activities of iron chelators against common nosocomial pathogens. *Antimicrobial Agents and Chemotherapy*, *56*(10), 5419–5421. https://doi.org/10.1128/AAC.01197-12

Thrash, J. C., Pollock, J., Torok, T., & Coates, J. D. (2010). Description of the novel perchlorate-reducing bacteria Dechlorobacter hydrogenophilus gen. nov., sp. nov. and Propionivibrio militaris, sp. nov. *Applied Microbiology and Biotechnology*, *86*(1), 335–343. https://doi.org/10.1007/s00253-009-2336-6

Tindall, B. J., Kämpfer, P., Euzéby, J. P., & Oren, A. (2006). Valid publication of names of prokaryotes according to the rules of nomenclature: past history and current practice. *International Journal of Systematic and Evolutionary Microbiology*, *56*(Pt 11), 2715–2720. https://doi.org/10.1099/ijs.0.64780-0

Tooke, C. L., Hinchliffe, P., Bragginton, E. C., Colenso, C. K., Hirvonen, V. H. A., Takebayashi, Y., & Spencer, J. (2019). β-Lactamases and β-Lactamase Inhibitors in the 21st Century. *Journal of Molecular Biology*, *431*(18), 3472–3500. https://doi.org/https://doi.org/10.1016/j.jmb.2019.04.002

Toscano, W. A. J., & Storm, D. R. (1982). Bacitracin. *Pharmacology & Therapeutics*, *16*(2), 199–210. https://doi.org/10.1016/0163-7258(82)90054-7

Tse-Dinh, Y.-C. (2007). Exploring DNA topoisomerases as targets of novel therapeutic agents in the treatment of infectious diseases. *Infectious Disorders Drug Targets*, *7*(1), 3–9. https://doi.org/10.2174/187152607780090748

Tseng, B. S., Zhang, W., Harrison, J. J., Quach, T. P., Song, J. L., Penterman, J., Singh, P. K., Chopp, D. L., Packman, A. I., & Parsek, M. R. (2013). The extracellular matrix protects *Pseudomonas aeruginosa* biofilms by limiting the penetration of tobramycin. *Environmental Microbiology*, *15*(10), 2865–2878. https://doi.org/10.1111/1462-2920.12155

Unemo, M., & Shafer, W. M. (2014). Antimicrobial resistance in Neisseria gonorrhoeae in the 21st century: past, evolution, and future. *Clinical microbiology reviews*, *27*(3), 587–613. https://doi.org/10.1128/CMR.00010-14

Unz, R. F. (2015). Zoogloea. In *Bergey's Manual of Systematics of Archaea and Bacteria* (pp. 1–13). Chichester, UK: John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118960608.gbm01005

van Asbeck, B. S., Marcelis, J. H., Marx, J. J., Struyvenberg, A., van Kats, J. H., & Verhoef, J. (1983). Inhibition of bacterial multiplication by the iron chelator deferoxamine: potentiating effect of ascorbic acid. *European Journal of Clinical Microbiology*, *2*(5), 426–431. https://doi.org/10.1007/BF02013899

Van Epps, H. L. (2006). René Dubos: unearthing antibiotics. *The Journal of Experimental Medicine*, *203*(2), 259. https://doi.org/10.1084/jem.2032fta

van Santen J. A., Jacob G., Singh A. L., Aniebok V., Balunas M. J., Bunsko D., Neto F. C., Castaño-Espriu L., Chang C., Clark T. N., Cleary Little J. L., Delgadillo D. A., Dorrestein P. C., Duncan K. R., Egan J. M., Galey M. M., Haeckl F. P. J., Hua A., Hughes A. H., Iskakova D., Khadilkar A., Lee J. H., Lee S., LeGrow N., Liu D. Y., Macho J. M., McCaughey C. S., Medema M. H., Neupane R. P., O'Donnell T. J., Paula J. S., Sanchez L. M., Shaikh A. F., Soldatou S., Terlouw B. R., Tran T. A., Valentine M., van der Hooft J. J. J., Vo D. A., Wang M., Wilson D., Zink K. E., & Linington R. G. (2019). The Natural Products Atlas: An Open Access Knowledge Base for Microbial Natural Products Discovery. *ACS Central Science*, *5*(11), 1824–1833. https://doi.org/10.1021/acscentsci.9b00806

Vandamme, P., Dewhirst, F. E., Paster, B. J., & On, S. L. W. (2015). *Arcobacter*. In *Bergey's Manual of Systematics of Archaea and Bacteria* (pp. 1–8). Chichester, UK: John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118960608.gbm01070

Vandamme, P., Falsen, E., Rossau, R., Hoste, B., Segers, P., Tytgat, R., & De Ley, J. (1991). Revision of *Campylobacter*, *Helicobacter*, and *Wolinella* taxonomy: emendation of generic descriptions and proposal of *Arcobacter* gen. nov. *International Journal of Systematic Bacteriology, 41*(1), 88–103. https://doi.org/10.1099/00207713-41-1-88

Vaser, R., Sović, I., Nagarajan, N., & Šikić, M. (2017). Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Research*, *27*(5), 737–746. https://doi.org/10.1101/gr.214270.116

Ventola, C. L. (2015). The antibiotic resistance crisis: part 1: causes and threats. *P & T : A Peer-Reviewed Journal for Formulary Management*, *40*(4), 277–283.

Větrovský, T., & Baldrian, P. (2013). The Variability of the 16S rRNA Gene in Bacterial Genomes and Its Consequences for Bacterial Community Analyses. *PLoS ONE*, *8*(2), e57923. https://doi.org/10.1371/journal.pone.0057923

Vollmer, W., Blanot, D., & De Pedro, M. A. (2008). Peptidoglycan structure and architecture. *FEMS Microbiology Reviews*, *32*(2), 149–167. https://doi.org/10.1111/j.1574-6976.2007.00094.x

Vrancken, K., Van Mellaert, L., & Anné, J. (2010). Cloning and expression vectors for a Gram-positive host, *Streptomyces lividans*. *Methods in Molecular Biology (Clifton, N.J.)*, *668*, 97–107. https://doi.org/10.1007/978-1-60761-823-2_6

Waksman, S. A., & Woodruff, H. B. (1940). The Soil as a Source of Microorganisms Antagonistic to Disease-Producing Bacteria. *Journal of Bacteriology*, *40*(4), 581–600. https://doi.org/10.1128/JB.40.4.581-600.1940

Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C. A., Zeng, Q., Wortman, J., Young, S. K., & Earl, A. M. (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

Wang, L.-T., Lee, F.-L., Tai, C.-J., & Kasai, H. (2007). Comparison of gyrB gene sequences, 16S rRNA gene sequences and DNA-DNA hybridization in the Bacillus subtilis group. *International Journal of Systematic and Evolutionary Microbiology*, *57*(Pt 8), 1846–1850. https://doi.org/10.1099/ijs.0.64685-0

Wang, M. , Carver, J. J., Phelan, V. V., Sanchez, L. M., Garg, N., Peng, Y., Nguyen, D. D., Watrous, J., Kapono, C. A., Luzzatto-Knaan, T., Porto, C., Bouslimani, A., Melnik, A. V., Meehan, M. J., Liu, W. T., Crüsemann, M., Boudreau, P. D., Esquenazi, E., Sandoval-Calderón, M., Kersten, R. D., Pace, L. A., Quinn, R. A., Duncan, K. R., Hsu, C. C., Floros, D. J., Gavilan, R. G., Kleigrewe, K., Northern, T., Dutton, R. J., Parrot, D., Carlson, E. E., Aigle, B., Michelsen, C. F., Jelsbak, L., Sohlenkamp, C., Pevzner, P., Edlund, A., McLean, J., Piel, J., Murphy, B. T., Gerwick, L., Liaw, C. C., Yang, Y. L., Humpf, H. U., Maansson, M., Keyzers, R. A., Sims, A. C., Johnson, A. R., Sidebottom, A. M., Sedio, B. E., Klitgaard, A., Larson, C. B. P. C. A. B., Torres-Mendoza, D., Gonzalez, D. J., Silva, D. B., Marques, L. M., Demarque, D.P., Pociute, E., O'Neill, E.C., Briand, E., Helfrich, E. J. N., Granatosky, E. A., Glukhov, E., Ryffel, F., Houson, H., Mohimani, H., Kharbush, J. J., Zeng, Y., Vorholt, J. A. , Kurita, K. L., Charusanti, P., McPhail, K. L., Nielsen, K. F., Vuong, L., Elfeki, M., Traxler, M. F., Engene, N., Koyama, N., Vining, O. B., Baric, R., Silva, R. R., Mascuch, S. J., Tomasi, S., Jenkins, S., Macherla, V., Hoffman, T., Agarwal, V., Williams, P. G., Dai, J., Neupane, R., Gurr, J., Rodríguez, A. M. C., Lamsa, A., Zhang, C., Dorrestein, K., Duggan, B.M., Almaliti, J., Allard, P. M., Phapale, P., Nothias, L. F., Alexandrov, T., Litaudon, M., Wolfender, J. L., Kyle, J. E., Metz, T. O., Peryea, T., Nguyen, D. T., VanLeer, D., Shinn, P., Jadhav, A., Müller, R., Waters, K. M., Shi, W., Liu, X., Zhang, L., Knight, R., Jensen, P.R., Palsson, B. O., Pogliano, K., Linington, R. G., Gutiérrez, M., Lopes, N. P., Gerwick, W. H., Moore, B. S., Dorrestein, P. C., & Bandeira, N. (2016). Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nature Biotechnology*, *34*(8), 828–837. https://doi.org/10.1038/nbt.3597

Ward, N., Larsen, Ø., Sakwa, J., Bruseth, L., Khouri, H., Durkin, A. S., Dimitrov, G., Jiang, L., Scanlan, D., Kang, K. H., Lewis, M., Nelson, K. E., Methé, B., Wu, M., Heidelberg, J. F., Paulsen, I. T., Fouts, D., Ravel, J., Tettelin, H., Ren, Q., Read, T., DeBoy, R. T., Seshadri, R., Salzberg, S. L., Jensen, H. B., Birkeland, N. K., Nelson, W. C., Dodson, R. J., Grindhaug, S. H., Holt, I., Eidhammer, I., Jonasen, I., Vanaken, S., Utterback, T.,

Feldblyum, T. V., Fraser, C. M., Lillehaug, J. R, & Eisen, J. A. (2004). Genomic Insights into Methanotrophy: The Complete Genome Sequence of *Methylococcus capsulatus* (Bath). *PLoS Biology*, *2*(10), e303. https://doi.org/10.1371/journal.pbio.0020303

Watanakunakorn, C. (1981). The antibacterial action of vancomycin. *Reviews of Infectious Diseases*, *3 suppl*, S210-5.

Watanakunakorn, C. (1984). Mode of action and *in-vitro* activity of vancomycin. *The Journal of Antimicrobial Chemotherapy*, *14 Suppl D*, 7–18. https://doi.org/10.1093/jac/14.suppl_d.7

Watson, M., & Warr, A. (2019). Errors in long-read assemblies can critically affect protein prediction. *Nature Biotechnology*, *37*(2), 124–126. https://doi.org/10.1038/s41587-018-0004-z

Webber, M. A., & Piddock, L. J. V. (2003). The importance of efflux pumps in bacterial antibiotic resistance. *Journal of Antimicrobial Chemotherapy*, *51*(1), 9–11. https://doi.org/10.1093/jac/dkg050

Weber, T., Blin, K., Duddela, S., Krug, D., Kim, H. U., Bruccoleri, R., Lee, S. Y., Fischbach, M. A., Müller, R., Wohlleben, W., Breitling, R., Takano, E., & Medema, M. H. (2015). antiSMASH 3.0-a comprehensive resource for the genome mining of biosynthetic gene clusters. *Nucleic Acids Research*, *43*(W1), W237-43. https://doi.org/10.1093/nar/gkv437

Weiss, J. V., Rentz, J. A., Plaia, T., Neubauer, S. C., Merrill-Floyd, M., Lilburn, T., Bradburne, C., Megonigal, J. P., & Emerson, D. (2007). Characterization of Neutrophilic Fe(II)-Oxidizing Bacteria Isolated from the Rhizosphere of Wetland Plants and Description of *Ferritrophicum radicicola* gen. nov. sp. nov., and *Sideroxydans paludicola* sp. nov. *Geomicrobiology Journal*, *24*(7–8), 559–570. https://doi.org/10.1080/01490450701670152

Weissman, K. J., & Müller, R. (2008). Protein-protein interactions in multienzyme megasynthetases. *Chembiochem : A European Journal of Chemical Biology*, *9*(6), 826–848. https://doi.org/10.1002/cbic.200700751

Whitman, W. B., Coleman, D. C., & Wiebe, W. J. (1998). Prokaryotes: the unseen majority. *Proceedings of the National Academy of Sciences of the United States of America*, *95*(12), 6578–6583. https://doi.org/10.1073/pnas.95.12.6578

Wick, R. R., & Holt, K. E. (2019). Benchmarking of long-read assemblers for prokaryote whole genome sequencing. *F1000Research*, *8*, 2138. https://doi.org/10.12688/f1000research.21782.2

Wick, R. R., Judd, L. M., Gorrie, C. L., & Holt, K. E. (2017). Completing bacterial genome assemblies with multiplex MinION sequencing. *Microbial Genomics*, *3*(10), e000132. https://doi.org/10.1099/mgen.0.000132

Wick, R. R., Judd, L. M., Gorrie, C. L., & Holt, K. E. (2017). Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. *PLOS Computational Biology*, *13*(6), e1005595. Retrieved from https://doi.org/10.1371/journal.pcbi.1005595

Wielders, C. L. C., Fluit, A. C., Brisse, S., Verhoef, J., & Schmitz, F. J. (2002). mecA gene is widely disseminated in Staphylococcus aureus population. *Journal of Clinical Microbiology*, *40*(11), 3970–3975. https://doi.org/10.1128/jcm.40.11.3970-3975.2002

Williams, K. P., & Kelly, D. P. (2013). Proposal for a new class within the phylum Proteobacteria , *Acidithiobacillia* classis nov., with the type order Acidithiobacillales , and emended description of the class Gammaproteobacteria. *International Journal of Systematic and Evolutionary Microbiology*, *63*(8), 2901–2906. https://doi.org/10.1099/ijs.0.049270-0

Wirsen, C. O., Sievert, S. M., Cavanaugh, C. M., Molyneaux, S. J., Ahmad, A., Taylor, L. T., DeLong, E. F., & Taylor, C. D. (2002). Characterization of an Autotrophic Sulfide-Oxidizing Marine *Arcobacter* sp. That Produces Filamentous Sulfur. *Applied and Environmental Microbiology*, *68*(1), 316–325. https://doi.org/10.1128/AEM.68.1.316-325.2002

Wolfe, A. D., & Hahn, F. E. (1965). Mode of action of chloramphenicol IX effects of chloramphenicol upon a ribosomal amino acid polymerization system and its binding to bacterial ribosome. *Biochimica et Biophysica Acta*, *95*, 146–155. https://doi.org/10.1016/0005-2787(65)90219-4

Wood, A. P., & Kelly, D. P. (1986). Chemolithotrophic metabolism of the newly-isolated moderately thermophilic, obligately autotrophic *Thiobacillus tepidarius*. *Archives of Microbiology*, *144*(1), 71–77. https://doi.org/10.1007/BF00454959

Wood, A. P., & Kelly, D. P. (1988). Isolation and physiological characterisation of *Thiobacillus aquaesulis* sp. nov., a novel facultatively autotrophic moderate thermophile. *Archives of Microbiology*, *149*(4), 339–343. https://doi.org/10.1007/BF00411653

Wood, D. E., & Salzberg, S. L. (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biology*, *15*(3), R46. https://doi.org/10.1186/gb-2014-15-3-r46

Woodyer, R. D., Shao, Z., Thomas, P. M., Kelleher, N. L., Blodgett, J. A., Metcalf, W. W., van der Donk, W. A., & Zhao, H. (2006). Heterologous production of fosfomycin and identification of the minimal biosynthetic gene cluster. *Chemistry & Biology*, *13*(11), 1171–1182. https://doi.org/10.1016/j.chembiol.2006.09.007

Xu, L., Huang, H., Wei, W., Zhong, Y., Tang, B., Yuan, H., Zhu, L., Huang, W., Ge, M., Yang, S., Zheng, H., Jiang, W., Chen, D., Zhao, G. P., & Zhao, W. (2014). Complete genome sequence and comparative genomic analyses of the vancomycin-producing *Amycolatopsis orientalis. BMC Genomics*, *15*(1), 363. https://doi.org/10.1186/1471-2164-15-363

Yamanaka, K., Reynolds, K. A., Kersten, R. D., Ryan, K. S., Gonzalez, D. J., Nizet, V., Dorrestein, P. C., & Moore, B. S. (2014). Direct cloning and refactoring of a silent lipopeptide biosynthetic gene cluster yields the antibiotic taromycin A. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(5), 1957–1962. https://doi.org/10.1073/pnas.1319584111

Yang, H. J., Huang, X. Z., Zhang, Z. L., Wang, C. X., Zhou, J., Huang, K., Zhou, J. M., & Zheng, W. (2014). Two novel amphomycin analogues from *Streptomyces canus* strain FIM-0916. *Natural Product Research*, *28*(12), 861–867. https://doi.org/10.1080/14786419.2014.886210

Yang, J. Y., Karr, J. R., Watrous, J. D., & Dorrestein, P. C. (2011). Integrating "-omics" and natural product discovery platforms to investigate metabolic exchange in microbiomes. *Current Opinion in Chemical Biology*, *15*(1), 79–87. https://doi.org/10.1016/j.cbpa.2010.10.025

Yi, H., Nevin, K. P., Kim, B.-C., Franks, A. E., Klimes, A., Tender, L. M., & Lovley, D. R. (2009). Selection of a variant of *Geobacter sulfurreducens* with enhanced capacity for

current production in microbial fuel cells. *Biosensors and Bioelectronics*, *24*(12), 3498–3503. https://doi.org/10.1016/j.bios.2009.05.004

Yocum, R. R., Rasmussen, J. R., & Strominger, J. L. (1980). The mechanism of action of penicillin. Penicillin acylates the active site of Bacillus stearothermophilus D-alanine carboxypeptidase. *The Journal of Biological Chemistry*, *255*(9), 3977–3986.

Yoneyama, H., & Katsumata, R. (2006). Antibiotic resistance in bacteria and its future for novel antibiotic development. *Bioscience, Biotechnology, and Biochemistry*, *70*(5), 1060–1075. https://doi.org/10.1271/bbb.70.1060

Zhalnina, K. V., Dias, R., Leonard, M. T., Dorr de Quadros, P., Camargo, F. A., Drew, J. C., Farmerie, W. G., Daroub, S. H., & Triplett, E. W. (2014). Genome Sequence of Candidatus *Nitrososphaera evergladensis* from Group I.1b Enriched from Everglades Soil Reveals Novel Genomic Features of the Ammonia-Oxidizing Archaea. *PLoS ONE*, *9*(7), e101648. https://doi.org/10.1371/journal.pone.0101648

Zhang, Q., Liu, W. (2013) Biosynthesis of thiopeptide antibiotics and their pathway engineering. *Natural Product Reports, 30*(2), 218-26. doi: 10.1039/c2np20107k. PMID: 23250571.

Zhang, Q., Yu, Y., Vélasquez, J. E., & van der Donk, W. A. (2012). Evolution of lanthipeptide synthetases. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(45), 18361–18366. https://doi.org/10.1073/pnas.1210393109

Zhao, B., Gao, Z., Shao, Y., Yan, J., Hu, Y., Yu, J., Liu, Q., & Chen, F. (2012). Diversity analysis of type I ketosynthase in rhizosphere soil of cucumber. *Journal of Basic Microbiology*, *52*(2), 224–231. https://doi.org/10.1002/jobm.201000455

Zhao, B., Moody, S. C., Hider, R. C., Lei, L., Kelly, S. L., Waterman, M. R., & Lamb, D. C. (2012). Structural analysis of cytochrome P450 105N1 involved in the biosynthesis of the zincophore, coelibactin. *International Journal of Molecular Sciences*, *13*(7), 8500–8513. https://doi.org/10.3390/ijms13078500

Zheng, Q., Wang, Q., Wang, S., Wu, J., Gao, Q., & Liu, W. (2015). Thiopeptide Antibiotics Exhibit a Dual Mode of Action against Intracellular Pathogens by Affecting Both Host and Microbe. *Chemistry and biology*, *22*(8), 1002–1007. https://doi.org/10.1016/j.chembiol.2015.06.019

Zhu, T., Cheng, X., Liu, Y., Deng, Z., & You, D. (2013). Deciphering and engineering of the final step halogenase for improved chlortetracycline biosynthesis in industrial *Streptomyces aureofaciens*. *Metabolic Engineering*, *19*, 69–78. https://doi.org/10.1016/j.ymben.2013.06.003

Ziemert, N., Ishida, K., Quillardet, P., Bouchier, C., Hertweck, C., de Marsac, N. T., & Dittmann, E. (2008). Microcyclamide Biosynthesis in Two Strains of *Microcystis aeruginosa* from Structure to Genes and Vice Versa. *Applied and Environmental Microbiology*, *74*(6), 1791 – 1797. https://doi.org/10.1128/AEM.02392-07

Zipperer, A., Konnerth, M. C., Laux, C., Berscheid, A., Janek, D., Weidenmaier, C., Burian, M., Schilling, N. A., Slavetinsky, C., Marschal, M., Willmann, M., Kalbacher, H., Schittek, B., Brötz-Oesterhelt, H., Grond, S., Peschel, A., & Krismer, B. (2016). Human commensals producing a novel antibiotic impair pathogen colonization. *Nature*, *535*(7613), 511–516. https://doi.org/10.1038/nature18634