Genome-guided bioprospecting for novel antibiotic lead compounds.

AWOLOPE, O.K.

2023

Genome-guided Bioprospecting for Novel Antibiotic Lead Compounds.

AWOLOPE, O.K.

PhD 2023

Genome-guided Bioprospecting for Novel Antibiotic Lead Compounds

OPEYEMI K. AWOLOPE

A thesis submitted in partial fulfilment of the requirements of the Robert Gordon University for the degree of Doctor of Philosophy

January 2023

DECLARATION

I declare that this work is my own original research. This thesis has not been previously presented at this university or any other academic or professional institution in pursuance of any other qualification. All materials and data contained therein that are attributable to other sources have been acknowledged and/or cited.

Opeyemi K. Awolope.

"For precept must be upon precept...
Line upon line...
Here a little, and there a little."

ACKNOWLEDGEMENTS

I would like to extend my sincere thanks to my supervisory team - Prof Andrew Lamb, Dr Noelle O'Driscoll, and Dr Alberto Di Salvo - for the opportunity to undertake this research and for their support, and to Dr Colin Thompson for his support throughout my studies.

I would like to thank Hannah Jebb, Jannath Nilma, and Prof J. Moody for being part of this journey. I would also like to thank colleagues in N626 - Len Montgomery, Qasim Ali, Joshua Burns, Aakash Welgamage Don, Sylvia Soldatou, and Prof Christine Edwards - for their assistance and kind words. I am also very grateful for all the help and support I received from PALS technical team. Thank you all for going the extra mile on my behalf, I am truly humbled by your kindness.

A debt of gratitude is owed to Dr David McGuinness (Glasgow Polyomics), Prof Chambers Hughes (University of Tübingen), and Dr Kyari Yates (Robert Gordon University) for being extremely generous with their time and expertise. It would have been impossible to undertake this research without your input and guidance. Thank you.

I would like to thank the staff at DNA Sequencing and Services (University of Dundee), Glasgow Polyomics (University of Glasgow), and MicrobesNG (University of Birmingham) for the first-class sequencing services provided.

I am grateful to Tenovus Scotland, Microbiology Society, and Federation of European Microbiological Societies for providing grants to support various aspects of this work.

Lastly, I would like to acknowledge every scientist whose work continues to push the boundaries and advance science. Time and space will not permit me to mention every name that springs to mind, but I am compelled to mention Prof James Taylor (1979-2020). Yours was a short life, but it was impactful, revolutionary, and refreshingly selfless. You were in a league of your own, but you slowed down to carry others along. You gave so much to the extent that some of us are almost fearful; because to whom much has been given, much will be required. The world of computational biology is poorer without you, but your legacy lives on. Thank you.

ABSTRACT

Antimicrobial resistance continues to pose a threat to health and wellbeing. Unmitigated, it is predicted to be the leading cause of death by 2050. Hence, the sustained development of novel antibiotics is crucial. As over 60% of licensed antibiotics are based on scaffolds derived from less than 1% of all known bacterial species, bacterial secondary metabolites constitute an untapped source of novel antibiotics. The aim of this project therefore was to expand the chemical space of bacteria-derived antibiotic lead compounds, using genomics approach.

To that end, a topsoil sample was collected from the rhizosphere in which antibiosis occurs naturally. Using starvation stress, sixty-five isolates were recovered from the sample, out of which four were selected based on morphology and designated A13BB, A23BA, A13AA, and A23AA. A13BB was identified by 16S rRNA gene sequence comparison as a *Pseudomonas* spp. and the other three isolates as *Hafnia*/*Obesumbacterium* spp. A database search showed that species belonging to these genera have genomes larger than the 3 Mb size above which increasing proportion of a bacterial genome is dedicated to secondary metabolism. Given their ecological origin, ability to withstand starvation stress, and expected genome size, these four isolates were presumed to harbour antibiotic-encoding gene clusters. Therefore, isolates A13BB and A23BA were selected for genome mining in the first instance. Illumina and GridION/MinION sequencing data were obtained for both isolates and assembled into high-quality genomes. Isolates' identities were confirmed by FastANI analysis as strains of *P. fragi* and *H. alvei*, with 4.94 and 4.77 Mb genomes, respectively.

Assembled genomes were mined with antiSMASH. Amongst other secondary metabolite biosynthetic gene clusters (smBGCs) detected, the β-lactone smBGCs in both genomes were selected for activation as their end products bear the hallmarks of an 'ideal antibiotic' that can inhibit several bacteria-specific enzymes simultaneously. Analysis of these smBGCs revealed genes encoding two core enzymes - 2-isopropylmalate synthase (2-IPMS) and acyl CoA ligase homologues. In the biosynthetic pathway, 2-IPMS catalyses the condensation of acetyl CoA with the degradation product of valine or isoleucine to form 2-IPM. 2-IPM is isomerised to 3-IPM which then forms the β-lactone warhead through reactions catalysed by acyl CoA ligase. It was speculated that the β-lactone compound is biosynthesised

to efficiently rid the organism of potentially harmful metabolic intermediates as it grows on poor carbon and nitrogen sources.

Therefore, strain fermentation was performed with 10.8 mM acetate as the main carbon source and 5 mM L-valine or L-isoleucine as the nitrogen source. Fermentation extracts were analysed by LC-MS with at least thirty-seven metabolite ions detected. Many of these ions have masses in the range *m/z* 230-750, which is an ideal mass range for antibiotic molecules. As β-lactone compounds are difficult to identify in crude extracts, especially when utilising single-stage mass spectrometry, reactivity-guided screening of extracts with cysteine thiol probe was performed as the probe forms UV- and MS-visible adducts with β-lactone compounds. However, complete dimerization of probe at a faster-than-expected rate in extract matrices hindered successful screening. This meant that it was not possible, without further analysis, to determine if any crude extract components were β-lactone compounds. Measures to limit or eliminate probe dimerization are proposed, together with molecular networking strategies that can afford global visualisation and rapid dereplication of extract components, using tandem mass spectrometry fragmentation patterns of parent ions.

This project provides an original and robust workflow that serves as a strong starting point in the isolation of novel β-lactone compounds from crude extracts, followed by structural optimisation and bioactivity profiling. The hitherto unrecognised potential of β-lactone natural compounds as 'ideal antibiotics' is highlighted, and several structural optimisation strategies required to harness this potential are proposed. The genomes assembled here, and associated data have been deposited in the repositories of the International Nucleotide Sequence Database Collaboration for repurposing by other researchers. Likewise, the hidden metabolic and biosynthetic potentials of *P. fragi* and *H. alvei* species uncovered by RASTtk and antiSMASH analyses have been catalogued and placed in the public domain, with many of these attributes reported for the first time.

Keywords: β-Lactone natural compounds, genome mining, bioprospecting, antibiotic lead compounds, antimicrobial resistance, *Pseudomonas fragi*, *Hafnia alvei*.

RESEARCH OUTPUTS

Awolope *et al*. (2018) Genome-guided screening of bacterial isolates to identify potential antibiotic producers. School of Pharmacy and Life Sciences Research Symposium. Robert Gordon University, Aberdeen, UK.

Awolope *et al*. (2019) Genome-guided screening of bacterial isolates to identify potential antibiotic producers. Proceedings of the Microbiology Society Annual Conference. International Conference Centre, Belfast, UK.

Awolope *et al*. (2019) Genome-guided screening of bacterial isolates to identify potential antibiotic producers. *Access Microbiology*. 2019; 1(1A):0062 https://doi.org/10.1099/acmi.ac2019.po0062

Awolope *et al.* (2020) *Hafnia alvei* strain A23BA chromosome, complete genome*. GenBank.* https://identifiers.org/insdc:CP050150*.*

Awolope *et al.* (2020) *Pseudomonas fragi* strain A13BB chromosome, complete genome*. GenBank.* https://identifiers.org/insdc:CP065202*.*

National Center for Biotechnology Information. Sequence Read Archive*.* (2020) https://identifiers.org/ncbi/insdc.sra:SRP251948

Awolope *et al*. (2021) The complete genome sequence of *Hafnia alvei* A23BA; a potential antibiotic-producing rhizobacterium. *BMC Research Notes.* 2021; 14:8. https://doi.org/10.1186/s13104-020-05418-2

Awolope *et al*. (2021) De novo genome assembly and analysis unveil biosynthetic and metabolic potentials of *Pseudomonas fragi* A13BB. *BMC Genomic Data*. 2021; 22(1):15.
https://doi.org/10.1186/s12863-021-00969-0

Awolope OK. (2021) Activation of the β-Lactone Biosynthetic Gene Clusters in *Pseudomonas fragi* A13BB. School of Pharmacy and Life Sciences Research Symposium. Robert Gordon University, Aberdeen, UK.

LIST OF ABBREVIATIONS

ACS	Acetyl coenzyme A synthetase
AMP	Adenosine monophosphate
cAMP	Cyclic adenosine monophosphate
AMR	Antimicrobial resistance
ANI	Average nucleotide identity
antiSMASH	Antibiotics and secondary metabolites analysis shell
ARG	Antibiotic resistance gene
ASCII	American standard code for information interchange
ATP	Adenosine triphosphate
BGC	Biosynthetic gene cluster
BLAST	Basic logical alignment search tool
BLASTN	Basic logical alignment search tool-nucleotide
BLASTP	Basic logical alignment search tool-protein
BUSCO	Benchmarking universal single copy orthologs
CCS	Circular consensus sequencing
CDS	Coding sequences
CoA	Coenzyme A
CRE	Carbapenem-resistant *Enterobacteriaceae*
CRISPRs	Clustered regularly interspaced short palindromic repeats
DAD	Diode array detector
DNA	Deoxyribonucleic acid
gDNA	Genomic deoxyribonucleic acid
EMBL	European molecular biology laboratory

EPIs	Efflux pump inhibitors
ESI	Electrospray ionisation
ESBL	Extended spectrum beta-lactamase
ESBL-E	Extended spectrum beta-lactamase-*Enterobacteriaceae*
FASTA	Fast-all
FastANI	Fast average nucleotide identity
FDA	Food and drug administration
FIGfams	Fellowship for interpretation of genomes protein families
GNPS	Global natural products social molecular networking
HGT	Horizontal gene transfer
HiFi	High fidelity
HPLC	High performance liquid chromatography
HR-MS/MS	High resolution-mass spectrometry/mass spectrometry
HTA	Homoserine transacetylase
HTS	High throughput screening
IPM	Isopropylmalate
IPMS	Isopropylmalate synthase
LC	Liquid chromatography
LC-MS	Liquid chromatography-mass spectrometry
MD	Multidrug
MDR	Multidrug resistant
MIBiG	Minimum information about a biosynthetic gene cluster
MIC	Minimum inhibitory concentration
mRNA	Messenger ribonucleic acid
MRSA	Methicillin-resistant *Staphylococcus aureus*

MS	Mass spectrometry
MS/MS	Mass spectrometry/mass spectrometry
MS-TIC	Mass spectrometry-total ion chromatogram
m/z	Mass to charge ratio
NMR	Nuclear magnetic resonance
NRPS	Non-ribosomal polyketide synthetase
OM	Outer membrane
ONS	Oxford nanopore sequencing
ONT	Oxford nanopore technology
OSMAC	One strain many compounds
PBP	Penicillin-binding protein
PCR	Polymerase chain reaction
PDR	Pan-drug resistant
pHMMs	Profile hidden Markov models
PKS	Polyketide synthase
PP_i	Inorganic pyrophosphate
QUAST	Quality assessment tool for genome assemblies
R&D	Research and development
RAST	Rapid annotation using subsystem technology
RASTtk	Rapid annotation using subsystem technology tool kit
RDD	Rational drug design
RiPPs	Ribosomally synthesized and post-translationally modified peptides
rRNA	Ribosomal ribonucleic acid
SAM	S-adenosylmethionine

SBS	Sequencing by synthesis
SDS	Sodium dodecyl sulphate
smBGC	Secondary metabolite biosynthetic gene cluster
smCOG	Specialised metabolite clusters of orthologous genes
SMRT	Single molecule real-time
SMS	Single molecule sequencing
TBE	Tris-borate-ethylenediaminetetraacetic acid
TCA	Tricarboxylic acid
TE	Tris-ethylenediaminetetraacetic acid
tRNA	Transfer ribonucleic acid
UV	Ultraviolet
VGT	Vertical gene transfer
VRE	Vancomycin-resistant Enterococcus
WHO	World health organisation
XDR	Extensively drug-resistant

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE

INTRODUCTION

1 Introduction

Deployment of antibiotics to treat or prevent infectious diseases has undoubtedly revolutionised modern medicine and improved life expectancy. However, this tremendous achievement has also uncovered new sets of challenges that threaten to reverse many health gains made. Strategies to mitigate these challenges are now rightly the focus of intense research and policy initiatives.

1.1 A Brief History of Antibiotics

An antibiotic is a chemical substance produced by a microorganism which is capable of killing or inhibiting the growth of another microorganism.[1] This definition was fiercely protected in the 1940s by Selman A. Waksman who went to great lengths to differentiate antibiotics from other antimicrobials of non-microbial origin. Waksman also made sure the term was not extended to include microbial compounds with properties other than antibiosis.[1,2] The term was narrowed further to describe microbial compounds with antibacterial activity thereby differentiating it from antivirals, antifungals, etc. Contemporary use has however seen the term extended to include synthetic and semi-synthetic antibacterial agents - an extension that would have irritated Waksman!

Contrary to popular notion, the use of antibiotics predates the 'modern antibiotic era'. Ancient cultures used microbial products to treat infectious diseases as evidenced by traces of tetracycline found in human skeletal remains dating back to 350 AD,[3,4] and the use of mouldy bread or red soils to treat infections in some ancient civilisations.[5,6] The 'modern antibiotic era' began in the 1890s with the use of pyocyanase by two German physicians to treat infections in hospital settings.[7] The compound, which was isolated from extracts of *Pseudomonas aeruginosa* (formerly *Bacillus pycyaneus*), was later abandoned due to inconsistent efficacy and high toxicity. The next compound of therapeutic significance was salvarsan. Although not an antibiotic in the strictest sense of the word, salvarsan nonetheless played a major role in the treatment of bacterial infections in the early 1900s. It was discovered by Paul Ehrlich in 1909 after he observed that some dyes he had developed for histological examination of tissues also showed antibacterial activities.[8] Salvarsan was later found to be an effective treatment for syphilis, and its success initiated the systematic screening of various dyes for antibacterial

activity. Prontosil and the sulpha drugs were discovered through this initiative, offering physicians effective treatment options for prevailing infections at the time.[9,10]

The serendipitous discovery of penicillin by Alexander Fleming in 1928 is perhaps the most familiar account of the history of antibiotics. Even though there were earlier reports of antibacterial activities of extracts obtained from various *Penicillium* fungi,[5] it was the discovery by Fleming that would revolutionise antimicrobial chemotherapy. Penicillin was isolated by Fleming and co-workers from the extracts of *Penicillium notatum* (formerly *P. rubrum*, and latterly *P. chrysogeneum* or *P. rubens*).[11] It was deployed in clinical settings in the early 1940s with remarkable success.[9,11] Around the same time, after decades of research on soil microbes and effects of their metabolic products on the environment, Waksman discovered actinomycin from *Actinomyces antibioticus*.[12] The compound showed bacteriostatic and bactericidal activities against pathogenic microbes, but it was later found to be very toxic thus precluding its therapeutic use. Waksman however went on to discover other therapeutically useful compounds of microbial origin, including streptomycin[13] (the first antibiotic effective against tuberculosis) and neomycin.[14] He also set up a new screening platform for the discovery of therapeutically useful compounds from soil microbes. The Waksman Platform would earn him a Nobel Prize and usher in the 'golden era' of antibiotic discovery.

Many pharmaceutical companies adopted the Waksman Platform which involved screening of soil-borne actinomycetes against selected pathogens to assess antibacterial activity in the form of clear zones of inhibition on overlay agar plates.[15,16] The initiative led to the discovery of many classes of antibiotics (e.g. aminoglycosides, tetracyclines, macrolides, amphenicols, rifamycins, etc.) from the early 1940s to the early 1960s (Figure 1.1).[9,16] Medical practice was transformed by the deployment of these new therapeutic agents as infectious diseases were the leading cause of death at that time. The impressive rate at which new antibiotics were discovered back then meant that the period is referred to as the 'golden era' of antibiotic discovery. It is worth noting that the majority of antibiotics in clinical use today were developed from lead compounds discovered during this period.

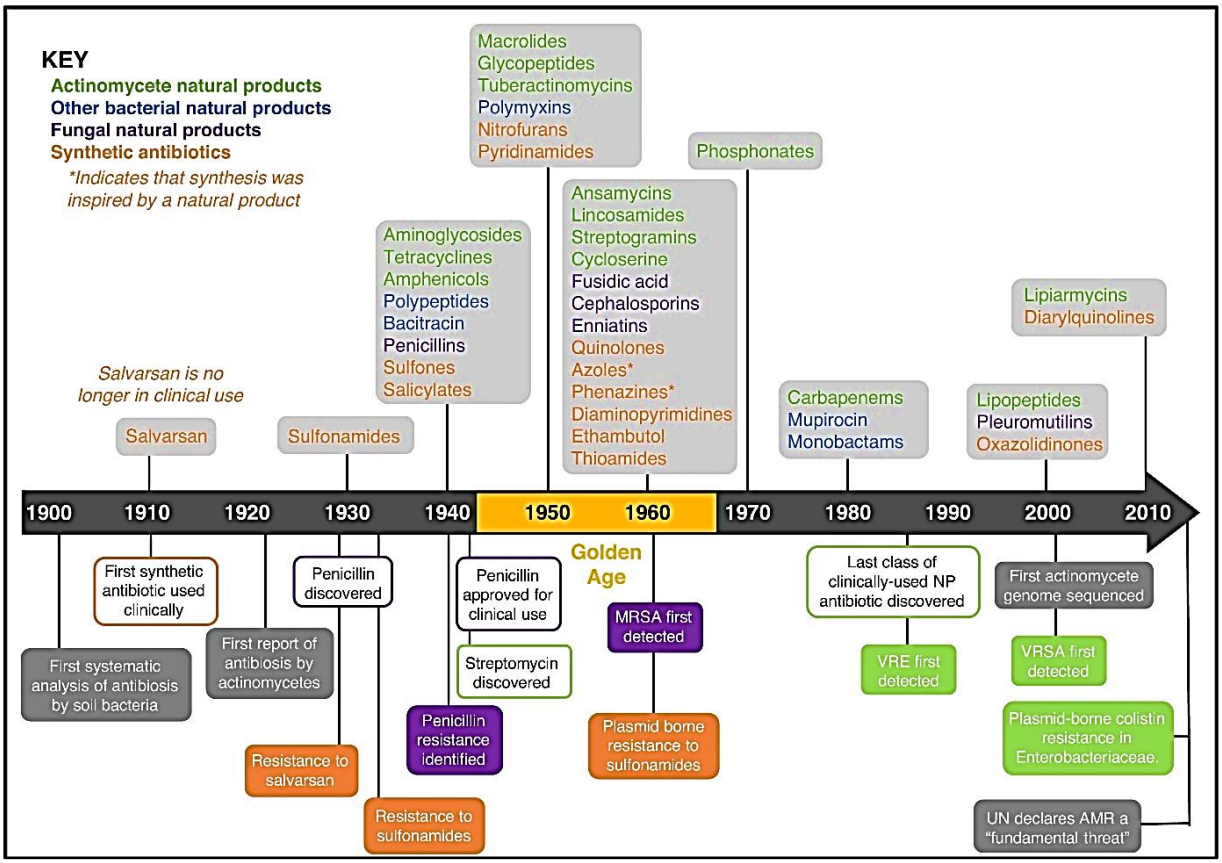


Figure 1.1: Timeline of antibiotic deployment and observation of antibiotic resistance. MRSA (methicillin resistant *S. aureus*), VRE (vancomycin resistant enterococci), VRSA (Vancomycin resistant *S. aureus*). Image adapted from Hutchings *et al*.[9]

Unfortunately, and to the disappointment of many researchers, the success rate of Waksman Platform in identifying new therapeutic leads started declining by the late 1960s. Rediscovery rate of known compounds was very high, to the extent that the platform was later completely abandoned. Several synthetic routes of antibiotic discovery were explored afterwards but with very limited success such that fluoroquinolones, oxazolidinones, and diarylquinoline are the only new classes of synthetic antibiotics developed in the last six decades.[9,16]

1.2 Antimicrobial Resistance (AMR)

Running in tandem with antibiotic discovery during the 'golden era' was the emergence of antimicrobial resistance (AMR). AMR is a natural phenomenon which reflects the innate or acquired abilities of microorganisms to withstand the effects of antimicrobial agents.[17-21] The term 'antimicrobial resistance' describes resistance to all types of antimicrobial agents i.e. resistance to antibiotics,

antifungals, antivirals, etc., but its use in this text will be limited to resistance to antibiotics. The challenge posed by AMR is illustrated in Figure 1.1 which shows the relative timelines between the deployment of a new class of antibiotics and the observation of antibiotic resistance. It is striking to observe how close these two timelines are, and even more striking are instances where resistance to antibiotics had emerged before clinical deployment.

An antibiotic is rendered ineffective when the target microorganism circumvents its main mode of action. Antibiotics exert their antibacterial effects in one of five ways: i) inhibition of bacterial cell wall synthesis, ii) inhibition of protein synthesis

Table 1.1: Mechanism of Action of Antibiotics

Mechanism of action	**Class of antibiotics**
Inhibition of bacterial cell wall synthesis	β-Lactams - penicillins cephalosporins carbapenems monobactams Glycopeptides - vancomycin teicoplanin
Inhibition of protein synthesis	Bind to 30S subunit - aminoglycosides tetracyclines Bind to 50S subunit - chloramphenicol macrolides oxazolidinones lincosamides
Inhibition of nucleic acid synthesis	Inhibit DNA synthesis - fluoroquinolones Inhibit RNA synthesis - rifamycins
Depolarization of bacterial cell membrane	Lipopeptides - daptomycin polymyxins
Disruption of metabolic process	Sulphonamides Trimethoprim

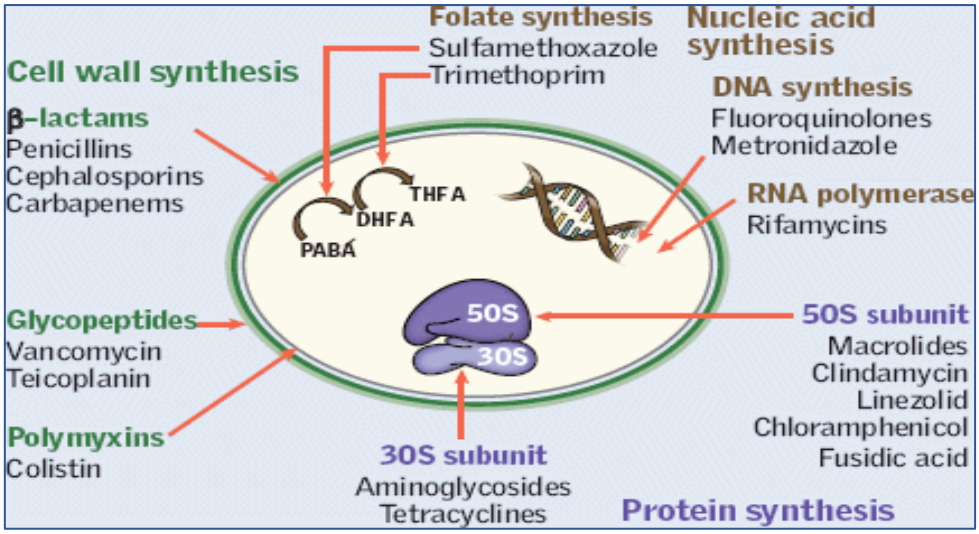


Figure 1.2: Classes of antibiotics and bacterial cellular targets.
Image source: Oxford Medicine Online (www.oxfordmedicine.com).

iii) inhibition of nucleic acid synthesis, iv) depolarisation of bacterial cell membrane, and v) disruption of vital metabolic processes.[17-19] Table 1.1 shows some examples of classes of antibiotics that employ each of these mechanisms, while Figure 1.2 illustrates the cellular targets disrupted by each class of antibiotics.

An antibiotic's main mode of action can be disrupted in an organism through various mechanisms mediated by resistance genes; and a resistance gene has been defined by Martínez *et al* as 'a gene in the presence of which an organism is able to withstand the effects of a given antibiotic and in the absence of which susceptibility is restored'.[22]

1.2.1 Mechanisms of Antibiotic Resistance

1.2.1.1 Innate Resistance

Innate resistance occurs naturally. This is expected given that most antibiotics are derived from metabolic products of microorganisms, including bacteria. For the event not to be suicide, the producing organism must possess some resistance determinants or resistance genes that protect it from the effects of its own metabolic product(s). Likewise, as bacterial cells do not exist in isolation, but in densely populated microbial communities, it is expected that some species in the same niche as antibiotic-producing strains would have developed some form of

natural resistance to the effects of those antibiotics in order to survive. Innate resistance, which is universal to all strains of a given species, may be intrinsic (i.e. always expressed regardless of antibiotic exposure) or induced (i.e. expressed after antibiotic exposure).[19-21] The most common forms of intrinsic innate resistance are seen in Gram-negative bacteria and are mediated by the reduced permeability of the outer membrane (OM), the presence of active multidrug (MD) efflux pumps, and the presence of restrictive porin channels in the OM.[19-21] These features reduce the entry of antibiotic molecules into bacterial cells thereby preventing the molecules from reaching their cellular targets or from reaching lethal intracellular concentrations.[19-21] Common forms of induced innate resistance seen in both Gram-negative and -positive bacteria are mediated by the activity of MD efflux pumps as well as by the production of inactivating enzymes that degrade or chemically alter antibiotic molecules.[19,20] *Pseudomonas aeruginosa* is an example of a multidrug resistant pathogen with intrinsic innate resistance mediated by reduced permeability of its OM and the presence of active efflux pumps. It also exhibits induced innate resistance mediated by the production of the inactivating enzyme, β-lactamase, which makes it resistant to β-lactam antibiotics.[19,21]

The reservoir of all naturally occurring innate resistance genes that mediate antibiotic resistance traits both in external environments (e.g. soil, water, air etc) and within living systems (e.g. gastrointestinal tracts of humans and animals) constitutes the 'antibiotic resistome'.[23] This reservoir has gained a lot of attention and study in recent years as it is believed to be the source of some clinically significant acquired resistance.[17,23]

1.2.1.2 Acquired Resistance

Acquired resistance, unlike innate resistance, is not native to the resistant strain but results from genetic mutation in chromosomal genes (vertical evolution) or acquisition of new genetic material from another microorganism (horizontal evolution).[17-20] Also unlike innate resistance, the genetic determinants of acquired resistance are not universally found in all strains of a given species. The clinical burden of antibiotic resistance stems mainly from acquired resistance where previously sensitive bacterial strains become resistant to treatment over time. A strain is said to have acquired resistance to an antibiotic if the minimum inhibitory

concentration (MIC) of the antibiotic against that strain is significantly higher than the corresponding MIC against the wildtype strain.[24] It is also worth noting that the term 'resistant strain' is subjective in clinical settings. Depending on the site of infection, a given strain may be resistant or susceptible to a particular antibiotic e.g. an otherwise gentamicin-resistant strain may become gentamicin-susceptible in the treatment of lower urinary tract infection where lethal concentration of the drug is easier to achieve than in other sites.[20]

Vertical evolution involving random genetic mutation in chromosomal genes brought about by errors in DNA replication is not uncommon in bacteria. Whilst many of these mutations are clinically insignificant, some indeed become significant if the mutations occur in genes encoding proteins associated with antibiotic targets or binding sites. The mutations may alter the architecture of the targets or binding sites such that the antibiotic is unable to induce its effects resulting in antibiotic resistance in the mutated strain.[17-20] A resistant strain may go on to become the dominant strain with continued antibiotic use as sensitive wildtype strains are eliminated by the drug while the mutant resistant strain persists. Replication of the resistant strain then proceeds with the resistance gene(s) passed down to its offspring. Genetic mutations can also mediate other forms of resistance traits including the production of inactivating enzymes, alteration of porin channels needed for drug entry into bacterial cells, or activation of MD efflux pumps.[18,20] Some clinically significant genetic mutations leading to antimicrobial resistance include penicillin resistance seen in *Streptococcus pneumoniae* mediated by a mutation in the gene encoding the penicillin-binding protein 2b (PBP2B);[18] methicillin resistance in methicillin-resistant *Staphylococcus aureus* (MRSA) mediated by the mutated *mecA* gene which encodes the penicillin-binding protein 2A (PBP2A) leading to decreased affinity for β-lactam antibiotics;[25] and resistance to daptomycin in *Staphylococcus aureus* mediated by mutation in *mprF* gene which changes cell membrane charge thereby inhibiting the binding of calcium and daptomycin.[26]

Acquired resistance can also result from the acquisition of new genetic material from another bacterium (horizontal gene transfer (HGT)). HGT is a common natural phenomenon in bacteria and can occur via conjugation, transduction, or transformation.[17-20] Conjugation is usually plasmid-mediated and involves direct exchange of genetic material between two cells through processes akin to mating.

Transduction involves DNA transfer through a vector, usually a bacteriophage. Transformation is acquisition of 'naked' genetic material released into the environment by a lysed bacterial cell. HGT-mediated transfer of antibiotic resistance genes usually occurs through conjugation or transformation, and very rarely through transduction.[19] HGT can occur between members of the same or different species, or even between members of different genera. HGT is believed to be the primary mechanism of transfer of resistance genes from innately resistant bacterial strains to otherwise sensitive strains as seen in the plasmid-mediated acquired resistance to quinolone observed in a clinical isolate of *Klebsiella pneumoniae*.[17,23] Likewise, many resistance genes observed in *Acinetobacter* spp. were shown to have been acquired through transformation.[17]

1.2.2 Factors Accelerating the Emergence and Spread of AMR

It is always the case that fewer antibiotics are effective against Gram-negative bacteria compared with Gram-positive strains. This is due to innate protection afforded by the relatively impermeable OM and active MD efflux pumps of the former. However as mentioned earlier, the burden of clinically significant AMR stems mainly from treatment failure encountered when previously sensitive bacterial strains (Gram-negative and Gram-positive strains) become resistant to treatment over time. In other words, acquired resistance remains the main challenge that must be overcome or mitigated.

Whilst random genetic mutations and acquisition of genetic materials via HGT are natural processes that may lead to emergence of AMR, the rates at which AMR emerges and spreads can be accelerated by both natural and man-made factors. Many authors have reviewed these factors extensively highlighting their complexity and interrelatedness.[17,22,23,27,28] A thorough review of the factors would be outside the scope of this thesis, but the salient points will be discussed briefly.

1.2.2.1 Antibiotic Tolerance

Antibiotic tolerance is another feature of certain bacterial cells that makes them insusceptible to antibiotics. Unlike antibiotic resistance that is mediated by genetic determinants, antibiotic tolerance is mediated by phenotypic attributes.[29-31] A tolerant cell (also referred to as a persister cell), does not possess genes that

encode resistance traits, but is a specialised survival cell that is able to spontaneously enter a dormant non-replicating state with little or no metabolic activity for a period of time.[30] Persister cells may account for up to 1% of any microbial population.[29] Given that most antibiotics exert their effects by disrupting vital cellular activities during replication and active metabolism (Figure 1.2), they are ineffective against persister cells in a dormant state as these cells are not actively dividing or carrying out metabolic activities. Persister cells remain dormant while the antibiotic eliminates actively dividing cells, and later come out of dormancy when antibiotic level drops off. They subsequently start replication thereby re-establishing the infection that was treated with the antibiotic in the first place. Many recurrent or chronic infections such as bacterial endocarditis and lung infections in cystic fibrosis are prolonged by the antibiotic tolerance of persister cells.[30]

Even though persister cells lack the genes that encode resistance traits, their ability to evade treatment results in prolonged colonisation and infection. This calls for prolonged antibiotic therapy that can inadvertently propagate resistance genes acquired either through genetic mutation or HGT.[31] As such, antibiotic tolerance potentially increases the emergence and spread of clinically significant antibiotic resistance.

1.2.2.2 Bacterial Biofilm Formation

Biofilm formation is a survival mechanism in which bacterial cells, belonging to the same or different species, attach to surfaces and/or each other thereby growing as multicellular aggregates.[32-34] Cells are encased in hydrated extracellular polymeric matrices comprised primarily of polysaccharides and proteins secreted by the cells (Figure 1.3).[32-34] Biofilms can form on biotic and abiotic surfaces including living or dead tissues, indwelling medical devices, surfaces of medical equipment, water pipes, etc. Their formation is facilitated by various processes including quorum sensing and the up- or downregulation of many genes involved in replication and metabolism.[33] Bacterial cells within biofilms can evade host defences and withstand antibiotic concentrations that are several-fold higher than concentrations that would kill planktonic (i.e. free floating) cells.[32-35] For instance, the MIC of ampicillin against the planktonic suspension of a β-lactamase-negative

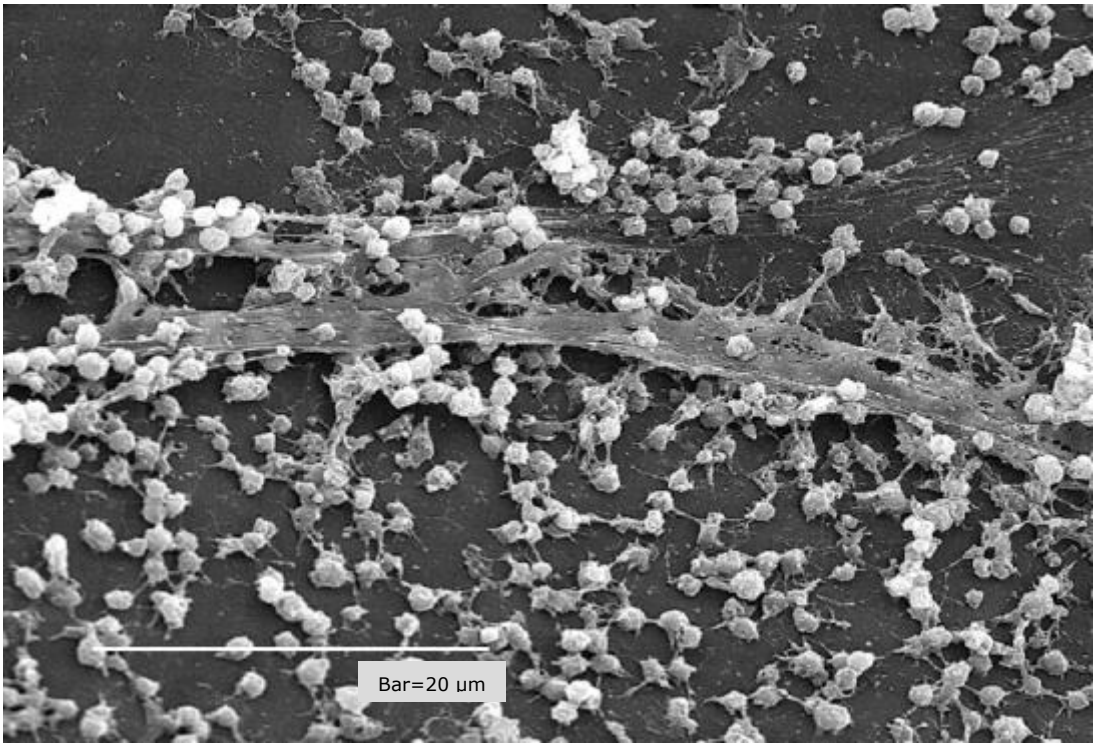


Figure 1.3: A Staphylococcal biofilm on an indwelling medical device.
Scanning electron micrograph of a Staphylococcal biofilm on the surface of an indwelling medical device. Image obtained from Donlan.[32]

strain of *Klebsiella pneumoniae* was 2 µg/mL whereas the same organism when grown in a biofilm withstood 5000 µg/mL ampicillin at 4 hours.[35] The antibiotic resistance observed within biofilms is not usually mediated by resistance genes but rather by the unique architecture and composition of extracellular matrices of biofilms, which in turn drive the phenotypic resistance traits displayed by resident cells. It has been shown that some antibiotics (e.g. aminoglycosides) penetrate biofilm matrices poorly or incompletely, while the altered chemical microenvironment within biofilms may inactivate other antibiotics or disrupt their cellular permeability through porin channels leading to loss of potency.[34] Also, the depletion of vital substrates within biofilms and the accumulation of metabolic waste products force some resident cells into dormant states with the consequence that antibiotics, which exert their effects on rapidly dividing cells, lose their potency on those cells. These dormant cells, along with persister cells which are known to be common features within biofilms, readily reseed biofilms after antibiotic therapy is withdrawn.[33,34] Although genetically mediated antibiotic resistance mechanisms are not the dominant resistance drivers within biofilms, they can act as contributory factors especially if some strains within the biofilms

possess innate or acquired resistance genes. The proximity of cells within biofilms also mean that these resistance genes are readily transferred to otherwise sensitive strains through HGT.

In living systems, biofilms serve as reservoirs of pathogenic organisms which are difficult to eliminate. While potent antibiotics readily kill planktonic pathogenic cells, cells embedded within biofilms are mostly spared and later dispersed to re-establish infections as seen in chronic otitis media, recurrent lung infections in cystic fibrosis, and chronic endocarditis.[32-34] This inevitably leads to prolonged proliferation of the pathogen within the host and prolonged antibiotic use with increased likelihood of mutant strains emerging and spreading as discussed below. Likewise, biofilms formed on abiotic surfaces may make disinfectants less effective against resident pathogens meaning these organisms can proliferate for longer and therefore constitute a steady source of contamination with the likelihood of resistant mutants emerging.

1.2.2.3 Selective Pressure Exerted by Antibiotic Chemotherapy

Effects of antibiotics on microbial populations may inadvertently drive the spread of antibiotic resistance. Acquisition of resistance genes through random genetic mutation usually poses a fitness cost to the mutant strain such that the higher the cost, the lower the chance the mutation is passed on through vertical transfer.[36,37] This is because wildtype strains are likely to readily outcompete any mutant strains for nutrients and other vital substrates. However, in the presence of antibiotics, sensitive wildtype strains are eliminated while resistant mutant strains remain unaffected. The latter may then go on to become dominant strains as they replicate without competition.[36,37]

Effects of antibiotics on microbial populations have also been shown to encourage horizontal transfer of innate or acquired resistance genes from one bacterium to another via conjugation or transformation.[38-40] This is particularly relevant in external environments (soil, water, etc) where innate antibiotic resistance is rife, and where resistance genes are readily acquired by otherwise sensitive strains through HGT upon exposure of the environment to antibiotics (e.g. exposure to antibiotics in wastewater).[21,40] Likewise, transfer of resistance genes can occur from non-pathogenic or commensal strains in living systems (e.g. in the

gastrointestinal tracts of humans and animals) to pathogenic strains via HGT following antibiotic therapy in the host.[38,39]

Arguably the most important driving factor for the emergence and spread of AMR is the selection for resistance traits driven by overuse and misuse of antibiotics.[41,42] Unnecessary human and veterinary prescriptions for antibiotics to treat non-bacterial infections, empirical treatment of bacterial infections with broad spectrum antibiotics, inappropriate dosing and duration of treatment, and the use of antibiotics for intense farming of food-producing animals all contribute to inappropriate use of antibiotics that can eventually accelerate the rate of emergence and spread of AMR.

1.2.2.4 Environmental Dissemination of Antibiotic Resistance Genes

As mentioned earlier, most antibiotics are derived from metabolic products of bacteria and other microorganisms. These metabolites serve various purposes in nature including use as warfare agents to gain competitive advantage. Given that there are $2x10^{30}$ bacteria cells on Earth according to an estimate by Flemming and Wuertz,[43] it is unsurprising that innate resistance genes abound in external environments where they protect antibiotic-producing strains from the lethal effects of their metabolic products. In addition to these pool of innate resistance genes are resistance genes acquired through random mutations or rearrangements in species occupying the same niche as antibiotic-producing strains.

In most cases, the fitness costs associated with the expression of these resistance genes prevent their expression in the absence of external selective pressures.[27,36,37] Furthermore, the presence of certain selective pressures initiate the transfer of chromosomal resistance genes onto mobile genetic elements like plasmids, integrons, and transposons from where they are more readily transferred to other bacterial cells through HGT.[27,44] Transfer of chromosomal genes in this way further increases the fitness costs to the organism because of the burden of keeping multiple copies of the same genes at different sites, and the difficulty in regulating expression independently from multiple sites. External selective pressures that are known to induce expression of these resistance genes and their transfer include antibiotics (present in wastewaters, sewage, and run-

off from agricultural farms), heavy metals, biocides, and other xenobiotic substances discarded into the environment.[27,44]

However, resistance genes in external environments are not clinically relevant until they are transferred into human commensals or pathogens. Since human microbiome is spatially removed from the microbiome in the environment, it is usually the case that a dispersal route is required for the transfer to occur.[27,44] Many events can facilitate this transfer, and to different extents, depending on the length of exposure. Examples of dispersal events include interactions between humans and wild animals (including birds and marine life), interactions with livestock, ingestion of raw contaminated food, and drinking contaminated water. These activities present potential routes for transferring environmental bacteria with innate or acquired resistance genes into human hosts. The resistance genes may be subsequently transferred to human pathogens or commensals in the presence of the right selective pressure e.g. antibiotic therapy in the host.

1.2.2.5 Poor Infection Prevention and Control

Poor sanitation (especially contaminated drinking water, poor hand hygiene and inadequate sewage disposal) increases the incidence and spread of infectious diseases within communities.[41,45] Not only does high incidence of bacterial infections result in rapid proliferation of pathogenic organisms and likelihood of resistant mutants emerging, it also calls for increased use of antibiotics which can exert selective pressure for resistant strains. Lack of adequate sanitation facilities is an ongoing problem in many low- and middle-income countries where there is evidence of increased incidence of bacterial infections caused by resistant *E. coli*, *Klebsiella* spp. and *S. aureus* strains.[45] Poor sanitation can also facilitate the dissemination of environmental resistance genes into human microbiome as highlighted above.

In a similar vein, poor infection control measures within healthcare settings have been associated with increased incidence of nosocomial infections and spread of resistant strains, notably methicillin-resistant *S. aureus* (MRSA) and vancomycin-resistant Enterococci (VRE).[39,42] Infrequent handwashing by healthcare staff and infrequent change of garments are significant contributory factors to the spread of infections and resistant strains within healthcare facilities.

Lastly, inadequate infection prevention through low vaccination coverage in vulnerable populations, especially in low- and medium-income countries, can also accelerate the rates at which antibiotic resistance emerges and spreads. Vaccines, when properly deployed in humans and animals, can drastically reduce infection rates and antibiotic use. Consequently, the emergence of resistant strains through random mutation during bacterial proliferation is curtailed, and the selective pressure exerted by antibiotic therapy for resistant strains is also minimised.[46,47] Notable examples of the effects of vaccination on AMR include the reduction in the incidence of infections caused by resistant pathogens after deployment of pneumococcal and *Haemophilus influenzae* vaccines in vulnerable populations in Finland and India, respectively.[48,49] Likewise, vaccines against viral pathogens can indirectly stem antibiotic resistance by preventing the occurrence and the need to treat secondary bacterial superinfections. For instance, influenza vaccination was shown to reduce antibiotic prescription by up to 50% in vaccinees in a study reported by Klugman and Black.[50] Low vaccine uptake and lack of investment and financial incentives remain major factors undermining the deployment of vaccines in many low- and middle-income countries.[46]

1.2.2.6 Global Travel

The world has become highly interconnected with very easy access to travel by air, land, or sea. Mass movement of people, livestock, and food products across international borders has meant that outbreaks of infectious diseases do not remain contained within geographic epicentres. The role of international travel in the spread of antibiotic resistance genes has been reviewed by many authors.[51,52] Even though AMR is observed everywhere, prevalence does vary across regions and rapid dissemination is facilitated by international travel. Travellers are likely to be colonised with resistant strains prevalent in their home countries which they may then go on to disseminate at their destinations given the right conditions. Likewise, returning travellers may be colonised or infected with antibiotic-resistant bacterial strains picked up from overseas destinations. This is particularly relevant to foodborne pathogens which are likely to colonise the gastrointestinal tract for a long period of time, therefore facilitating the transfer of resistant genes through HGT to the host microbiome.[51] Foodborne traveller's diarrhoea caused by resistant strains of *Campylobacte*r and *Salmonella* spp. are common among returning

international travellers who had visited South Asia and the Indian subcontinent such that the majority of cases seen in the USA, UK, Sweden, and Germany are linked to international travel.[51,53-56] The global spread of the multidrug resistant extended-spectrum-β-lactamase-producing *Enterobacteriaceae* (ESBL-E) strains has also been linked to international travel.[51,56] ESBL-E colonisation and infections are endemic in many popular tropical and sub-tropical tourist destinations with prevalence running as high as 50% in the population.[51] It is not uncommon to find returning travellers from these locations infected or colonised with ESBL-E strains.

Another area of great concern is medical tourism. People travel across borders for medical treatments or surgical procedures for various reasons, including shorter waiting times and significantly reduced medical costs. This practice exposes medical tourists to prevalent antibiotic-resistant nosocomial infections in overseas healthcare facilities.[51] The global spread of multidrug resistant MRSA and VRE has been linked to international travel, including medical tourism.[57,58] Also, the rapid global spread of the New Delhi Metallo-β-lactamase (NDM)-1 resistance gene seen in many Gram-negative bacteria has been linked to nosocomial infections contracted by returning medical tourists or travellers from the Indian subcontinent.[51,59] NDM-1 is a highly transmissible mobilised resistance gene that mediates resistance to penicillins, cephalosporins, and carbapenems; leaving very few treatment options.[60]

Other non-human vectors of resistant bacterial strains include wild animals, livestock and other food substances transported internationally. The plasmid-mediated colistin resistance gene, *mcr*-1, first detected in *E. coli* isolated from raw meat in China has spread across international borders through trade in livestock and meat products.[61] The gene is now detectable in strains that have colonised and infected humans, which is a very worrying trend as colistin is a last resort antibiotic.[51,61]

1.2.3 Clinical Implications of AMR

The emergence of resistance to a particular antibiotic in a previously susceptible pathogen increases the costs of treatment and the burden on patients and healthcare systems. Treatment failure with first line antibiotics require more expensive and perhaps more intensive treatments. The burden associated with treatment failure is further exacerbated if the pathogen is multidrug resistant. A multidrug resistant (MDR) strain is one that has acquired resistance to two or more antibiotics, and as such treatment options for infections caused by the organism become very limited. Although the extent to which a strain is resistant to existing antibiotics can be described by two other terms - extensively drug-resistant (XDR) strain i.e. resistant to most antibiotics, and pandrug resistant (PDR) strain i.e. resistant to all available antibiotics - for simplicity, MDR will be used in this text to describe all scenarios.

The incidence of infections caused by MDR pathogens is on the rise globally with dire consequences because clinicians are running out of treatment options. The commonly implicated organisms are resistant strains of the ESKAPE pathogens (***E****nterococcus faecium*, ***S****taphylococcus aureus*, ***K****lebsiella pneumoniae*, ***A****cinetobacter baumannii*, ***P****seudomonas aeruginosa*, and several species of ***E****nterobacter*).[62] These pathogens are resistant to virtually all first-line antibiotics (e.g. amoxicillin, ampicillin, erythromycin, trimethoprim, doxycycline) recommended for most infections and, more worryingly, resistant to many second- and third-line antibiotics like fluoroquinolones, vancomycin, cephalosporines, carbapenems and colistin.[62,63] There are high morbidity and mortality rates, and huge economic burden associated with MDR infections. Around 25,000 people lose their lives to MDR infections every year in the EU, 23,000 in the US and 700,000 globally.[64,65] The cost of treating a single MDR infection could be colossal, for instance a single carbapenem-resistant *Enterobacteriaceae* (CRE) infection is estimated to cost up to $66,031 to treat in the US.[66] The annual healthcare costs and lost productivity associated with MDR infections in the EU is estimated at around €1.5 billion, while the estimate is $4.6 billion in the US.[63,65] There are other significant indirect burdens on healthcare systems; several medical procedures (surgeries, organ transplants, cancer chemotherapies, etc.) require prophylactic antibiotic therapy to make them safe. Loss of antibiotic efficacy against MDR pathogens will increase the costs and

mortality rates associated with these procedures. Longer hospitalisation and more intensive care are often needed, putting more pressure on limited resources. Outbreaks of nosocomial MDR infections within hospitals may demand temporary shutdown of wards with disruption to vital services. Such outbreaks are also very expensive to eradicate as was the case in a large Australian hospital where an outbreak of VRE cost $1.5 million to eradicate.[67]

The clinical implications of AMR continue to be a matter of great concern, more so because the rate at which new antibiotics are discovered dramatically declined from the 1970s onwards (Figure 1.1). With some MDR pathogens resistant to virtually all antibiotics, treatment options available to clinicians are fast dwindling. The problem is further compounded by the fact that the rate at which AMR emerges and spreads far outweigh the rate at which new antibiotics are being developed.[68] If the current trajectory remains unchanged, AMR is predicted to be the leading cause of death by 2050, claiming 10 million lives globally every year with $100 trillion reduction in global economic output.[65] In 2017, The World Health Organisation (WHO) released the list of priority MDR pathogens against which new

Table 1.2: WHO Priority Pathogens Requiring New Antibiotics.

	Priority 1: Critical	**Priority 2: High**	**Priority 3: Medium**
Gram-positive	**None**	***Enterococcus faecium*****,** vancomycin-resistant ***Staphylococcus aureus*****,** methicillin-resistant, vancomycin-intermediate and resistant	***Streptococcus pneumoniae*****,** penicillin-non-susceptible
Gram-negative	***Acinetobacter baumannii***, carbapenem-resistant ***Pseudomonas aeruginosa*****,** carbapenem-resistant ***Enterobacteriaceae (including E. coli, Klebsiella pneumoniae)*****,** carbapenem-resistant, ESBL-producing	***Helicobacter pylori*****,** clarithromycin-resistant ***Campylobacter* spp.,** fluoroquinolone-resistant ***Salmonellae*****,** fluoroquinolone-resistant ***Neisseria gonorrhoeae*****,** cephalosporin-resistant, fluoroquinolone-resistant	***Haemophilus influenzae*****,** ampicillin-resistant ***Shigella* spp.,** fluoroquinolone-resistant

antibiotics are urgently needed (Table 1.2). The pathogens, the majority of which are Gram-negative bacteria, are grouped into three priority categories according to the urgency of need. The need for new antibiotics is greatest/critical for priority 1 MDR pathogens as they are already resistant to some last resort antibiotics including carbapenems and third generation cephalosporins.[69] These organisms also tend to cause very severe infections with higher mortality rates. The need for new antibiotics effective against priorities 2 and 3 MDR pathogens is not as critical, but nonetheless urgent as they cause more common infections like food poisoning and gonorrhoea.[69] The WHO priority list serves to strategically direct research into areas with the most critical needs to ensure finite resources are put to best use. In addition to developing new antibiotics, the judicious use of current antibiotics is paramount, as are many other policy measures aimed at tackling AMR.[41,69]

1.2.4 Policy Measures Aimed at Tackling AMR

Given the complexity and the multifaceted nature of the mechanisms that drive emergence and spread of AMR, the One Health approach to tackling AMR was adopted to ensure stakeholders and policymakers work collaboratively to design and implement mitigating measures. The approach encompasses strategies to raise public awareness of AMR, promote judicious use of current antimicrobials through antimicrobial stewardship, discourage misuse of antibiotics for intense animal farming, implement effective infection prevention and control measures in humans and animals, discourage indiscriminate disposal of environmental pollutants, improve global surveillance and reporting of AMR, and increase investment in research and development (R&D) of rapid diagnostics, vaccines, new antimicrobials, and alternative treatments.[41,70] The WHO, in conjunction with other international organisations, have launched various programmes and initiatives to achieve the objectives of the One Health Initiative.

In 2015, the **G**lobal **A**ction **P**lan on Antimicrobial Resistance (GAP) was developed and endorsed by member states of the WHO, Food and Agriculture Organisation (FAO), and World Organisation for Animal Health (OIE).[41] GAP sets out five strategic objectives, based on the One Health Initiative, that could form the framework of national and international AMR policies. The first objective of GAP is to raise public awareness of AMR. To that end, the **W**orld **A**ntimicrobial **A**wareness

Week (WAAW) was launched in 2015. It is an annual global campaign to raise awareness of AMR and to encourage best practices among the public, healthcare practitioners, and policymakers.[41] The campaign highlights the importance of adequate sanitation facilities and hygiene in preventing infectious diseases. It also emphasises the importance of antimicrobial stewardship, the judicious use of current antimicrobials, and the importance of access to good quality medicines. The Clean **Wa**ter **S**anitation and **H**ygiene (WASH) programme was subsequently launched in a bid to fulfil the objective to reduce the incidence of infectious diseases in humans and animals. Furthermore, the **Gl**obal **A**ntimicrobial Resistance and Use **S**urveillance **S**ystem (GLASS) was also launched by WHO in 2015 to fulfil another objective of GAP. GLASS collates and shares data on global surveillance of AMR, monitors the status of various national surveillance systems and facilitates enrolment of countries into GLASS.[41] In 2016, the United Nations General Assembly endorsed GAP as the global blueprint for tackling AMR and its principles have since been adopted by national governments and many international organisations in their efforts to tackle AMR.

Various global funding schemes and initiatives have also been launched to fund the implementation of GAP polices as well as fund R&D of new diagnostics, vaccines, and antibiotics.[41] These funding schemes include the **A**nti**m**icrobial **R**esistance **M**ulti **P**artner **T**rust **F**und (AMR MPTF) which committed to investing $70 million to implement various global and regional AMR action plans between 2019 and 2020; the **G**lobal **A**ntibiotic **R**esearch and **D**evelopment **P**artnership (GARDP), which is a joint initiative between WHO and the Drugs for Neglected Diseases *Initiative* (DND*i*), aims to develop and deliver five new treatments that target MDR pathogens by 2025; the AMR action fund is another funding initiative developed collaboratively by the WHO, European Investment Bank, and the Wellcome Trust and launched in 2020. It expects to invest more than $1 billion in R&D of new antimicrobials and aims to bring two to four new antibiotics to the market within the first decade. The **G**lobal **A**lliance for **V**accines and **I**mmunisations (GAVI, the Vaccine Alliance), which is a partnership between the WHO, the World Bank, UNICEF, the Bill & Melinda Gates Foundation, and other global partners, is also actively increasing access to immunisations in low- and middle-income countries - an initiative that serves to address the problem of low vaccination coverage in these communities.

On a national level, governments around the world are developing and implementing policies in line with the broad objectives of the One Health Initiative and GAP. The UK government published its first 5-year **N**ational **A**ction **P**lan (NAP) for AMR in 2013 setting out its strategy for tackling AMR over the following five years.[71] The strategies outlined in the 5-yearly NAPs will enable the government to achieve its broader long-term AMR objectives in line with the One Health Initiative.[71,72] With the full understanding that AMR is a global problem, the then UK prime minister commissioned 'The Review on Antimicrobial Resistance' in 2014, chaired by Jim O'Neill, to analyse the global impact of AMR and to propose concrete actions aimed at tackling AMR globally.[65] The review concluded in 2016 and its final report and recommendations have since informed the UK government's policies on AMR, including the current 5-year NAP and the UK's 20-year vision for AMR.[65,72] Amongst other recommendations, strong emphasis is placed on innovation and investment in basic and translational research. Another key recommendation is the need to address the economic and regulatory barriers to novel antibiotics R&D. Pharmaceutical companies are less likely to make good returns on investments in antibiotics R&D given the current antibiotic stewardship requirements where any novel antibiotic is recommended to be used very sparingly for the most serious MDR infections.[65,73] In recognition of these recommendations, the UK government has committed to funding various national and international research initiatives (in academia, industry, research councils, etc.) aimed at tackling AMR.[72] Pharmaceutical companies are also being incentivised to invest in the development of new antimicrobials through several initiatives including new reimbursement models for novel antibiotics and access to innovation funds.[72,74]

Similarly, there are global emphases on R&D of new antimicrobials in recognition of the fact that AMR is an inevitable natural phenomenon. While all other policy measures slow down the rates of emergence and spread of AMR, novel antibiotics (and antibiotic alternatives) have the potential to reverse the devastating effects of antibiotic resistance. More so if these new therapies belong to new drug classes or have novel mechanism of action. As the rate at which new antibiotics are discovered and developed post-1970 is worryingly low, tremendous efforts are now being made globally to increase the pipeline portfolio of novel antibiotics and antibiotic alternatives to meet current and future clinical needs.

1.3 Antibiotic Alternatives in the Pipeline (2019/2020)

Antibiotic alternatives are therapies that do not employ the traditional modes of action of conventional antibiotics to produce or enhance antibacterial effects.[75,76] Some of these therapies have direct effects on bacterial cells while others predominantly target host cells. There are increasing numbers of therapies categorised as antibiotic alternatives, many of which are currently in experimental stage of development. Therefore, the most promising or established alternative therapies will be discussed here briefly.

(i) Vaccines: vaccines are prophylactic antigenic therapies that significantly reduce the incidence of bacterial infections and the need for antibiotic therapy.[46,47] They are developed from antigenic materials extracted from pathogenic bacteria. Initial administration of a vaccine triggers an immune response in the vaccinee as antibodies and other immune system components against the target pathogen are produced. On subsequent exposure of the vaccinee to the pathogen through natural infection, the immune system promptly 'remembers' the pathogen and neutralises it. Unlike antibiotics that mainly attack single bacterial target sites, vaccines usually elicit polyclonal antibodies which target multiple bacterial epitopes therefore making bacterial resistance to vaccines relatively uncommon.[46] Deployment of vaccines have been shown to stem or eliminate AMR in many vaccinee populations .[46,47] The pipeline portfolio of vaccines appears promising with increasing focus on therapies that are effective against WHO MDR pathogens, many of which are still in the preclinical stage of development (Figure 1.4 and Table 1.3).

(ii) Antibodies: monoclonal antibodies (mABs) that bind to and inactivate proteins on the surface of pathogens, bacterial exotoxins, or virulent factors are becoming popular alternative therapies used in combination with standard antibiotics.[75,77] They are a viable option for immunocompromised patients who are unable to mount robust immune response to vaccines. mABs are highly specific to a given pathogen and bacterial resistance has been shown to be relatively uncommon.[77] Many of the mABs in the pipeline are being trialled against Gram-negative and Gram-positive organisms including *S. aureus*, *P. aeruginosa,* and *E. coli* MDR strains.[46,77,78]

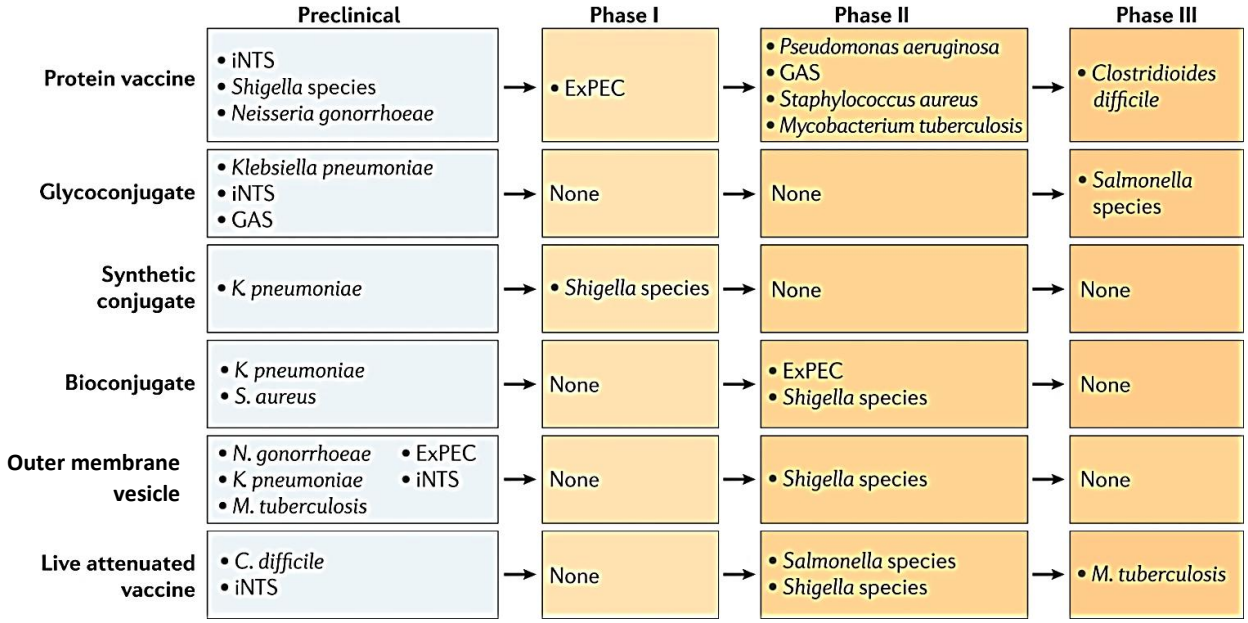


Figure 1.4: Vaccines in preclinical and clinical pipelines (2019/2020).
iNTS (invasive non-typhoidal *Salmonella*), GAS (group A *Streptococcus*), ExPEC (extraintestinal pathogenic *Escherichia coli*). Image adapted from Micoli *et al*.[46]

Table 1.3: Pipeline Portfolio of Selected Antibiotic Alternative Therapies

Therapy	Number in Preclinical Pipeline (year)	Number in Clinical Pipeline (year)	References
Vaccines	47 (2020)	38 (2020)	46,78
Antibodies	29 (2019)	9 (2020)	78,86
Phage + Phage enzymes	31 (2020)	4 (2020)	78,86
Antimicrobial peptides	29 (2020)	34 (2020)	78,86,90
Repurposed drugs	15 (2019)	3 (2019)	86,93,96

(iii) Bacteriophages: bacteriophages (or phages) are viruses that infect, replicate within, and eventually kill/lyse bacterial cells. As they are obligate parasites, they are found everywhere bacteria are found and it is estimated that there are up to 10^{32} bacteriophages on Earth - more than ten-fold the number of bacterial cells.[79,80.] Bacteriophages show high selectivity for their bacteria hosts compared with antibiotics, but some bacteriophages can infect multiple bacterial species. The narrow spectrum of action of bacteriophages is advantageous in that the normal human microbiome is not usually disrupted by phage therapy.[79,80] A lytic bacteriophage infects a bacterial cell by attaching itself irreversibly to receptors on the surface of the cell after which virion enzymes degrade the bacterial cell wall and the phage's genetic material is then injected into the bacterial cell. Within minutes of being inside the cell, the phage genome hijacks the host's DNA replication and metabolic machinery to produce viral nucleic acids and proteins, followed by assembling of daughter viral particles. Bacterial cell lysis is then initiated by virion endolysins and holins to release progeny virions which then go on to infect more bacterial cells.[79,80] A lysogenic (or temperate) bacteriophage on the other hand infects its bacterium host without inducing immediate self-replication and cell lysis. The phage genome (called a prophage) is incorporated into the bacterium's DNA where it usually remains dormant. For that reason, lytic bacteriophages are more suitable than lysogenic bacteriophages as alternative therapies to antibiotics.[79]

Bacteriophages were used for treating bacterial infections long before the discovery of antibiotics.[80] They are benign to human or animal cells and are degraded (without generating phage-derived toxic degradation products) by innate human or animal antibodies usually after they have exerted their antibacterial activities.[79] As bacterial resistance can develop to individual bacteriophage, a cocktail of different types of phages with different selectivity are usually administered to overcome resistance.[79,80] Interestingly, bacteriophages and phage cocktails are able to penetrate biofilms to eliminate susceptible cells, including persister cells, thereby eliminating resistance observed as a result of biofilm formation.[81,82]

The deployment of antibiotics at the beginning of the 20th century led to phage therapy being largely abandoned, but the dire impact of AMR has reinvigorated research in the field. Even though bacteriophages have the potential to be

highly effective antibacterial agents, their deployment is hampered by regulatory hurdles and insufficient clinical data to evaluate their safety and efficacy, as well as difficulties in conducting comparative clinical trials.[79,80] Bacteriophages exhibit complex pharmacokinetic and pharmacodynamic profiles which make standardised dosing difficult, and they are sometimes degraded by innate human antibodies faster than expected such that therapeutic levels are never achieved. There is also a tendency for phage therapy to induce toxic shock as endotoxins are released from lysed bacterial cells.[79,80] Nevertheless, phage therapy has been used successfully (mainly on compassionate grounds) to treat a handful of MDR infections including those caused by MDR *Acinetobacter baumannii*.[83] Many of the drawbacks associated with phage therapy are being addressed with genetically engineered bacteriophages and there are promising candidates in the pipeline that are effective against MDR pathogens (Table 1.3).[78,79]

(iv) Lysins: lysins (or bacteriophage endolysins) are enzymes produced by bacteriophages to degrade bacterial cell wall in order to release progeny virions after phage infection and replication. Lysins can also degrade bacterial cell wall when applied exogenously, although they are more effective against Gram-positive bacteria when used as such as the outer membrane of Gram-negative species offer extra protection against the enzymes.[84,85] However, the modular nature of lysins makes them amenable to bioengineering to produce enzymes with optimised properties including activity against some Gram-negative bacteria.[85] Lysins have very narrow spectrum of action and can selectively target specific pathogens without affecting human commensal microflora, although they can also be engineered to have a wider spectrum if desired.[84,85] In comparison with conventional antibiotics, bacterial resistance to lysins is harder to achieve as the enzymes target highly conserved bonds in the peptidoglycan cell wall.[85] Lysins are effective against both replicating and non-replicating bacterial cells, and are therefore able to eliminate planktonic persister cells as well as those embedded within biofilms.[82,85]

However, like all exogenous proteins administered systemically in animals or humans, lysins are rapidly degraded after systemic administration with a half-life of around 20 minutes necessitating frequent dosing or intravenous administration.[84] Lysins are also potentially immunogenic, which could reduce

bioavailability and *in vivo* efficacy. Compared with phage therapy, clinical trials involving lysins are more comparable to those of conventional antibiotics, and the pharmacokinetics of lysin-based therapies are less complicated. Lysins are also more amenable to established bioengineering techniques that could optimise their antibacterial effects and reduce or eliminate unwanted side effects. Research is ongoing to harness these qualities and there are currently a handful of phage-derived and recombinant lysins in both preclinical and clinical pipelines (Table 1.3).[78,86]

(v) Antimicrobial peptides (AMPs): AMPs or host defence peptides are highly diverse cationic short-chain peptides (usually comprised of 10-50 amino acids) produced by virtually all living organisms as the first line of defence against microorganisms.[87-90] AMPs are ubiquitous in nature where they act directly as antimicrobial agents or indirectly as host immune system modulators e.g. the human AMP cathelicidin LL-37 shows antibacterial, antifungal, antiviral, and immunomodulatory activities.[90] Antibacterial AMPs show rapid killing and broad-spectrum activity against Gram-positive and Gram-negative bacteria with many also showing antibiofilm properties.[87-90] Bacterial resistance to AMPs is slower to emerge compared with conventional antibiotics as the former usually target multiple sites including bacterial cell membrane.[87-90]

Despite their therapeutic potentials, only a handful of AMPs (daptomycin, polymyxins, gramicidin, nisin, and melittin) are currently licensed for clinical use.[90] However, the rise in AMR has refocused research on AMPs as alternative therapies, but not without significant challenges. AMPs acting on cell membranes are not always selective for microbial cells but can also target eukaryotic cells resulting in cytotoxicity even at therapeutic concentrations as seen with licensed polymyxin E (colistin).[90] AMR to AMPs, although slower to develop, is also a significant challenge. Development of resistance to licensed AMPs colistin and daptomycin exemplifies this challenge.[26,61,90] Furthermore, most AMPs that are currently licensed for use or in the pipeline are intravenous or topical preparations as they are susceptible to proteolytic degradation after systemic administration.[86-90] Consequently, it could be difficult to correlate *in vitro* antibacterial activity with *in vivo* efficacy. Low bioavailability after systemic administration is also a major problem. However, synthetic AMPs show better stability profiles, but they are quite expensive to design and

synthesise.[86,88-90] Overall, the development of AMPs (natural or synthetic) for clinical use is associated with higher production costs compared with conventional antibiotics.[89,90]

Research into innovative formulation strategies (e.g. production of ultra-short and/or truncated AMPs; and use of novel delivery systems such as nanoparticles) is ongoing to address some of the formulation challenges.[87-90] Likewise, Specifically Targeted AMPs (STAMPS) are being designed to improve the selectivity of AMPs for microbial cells thereby reducing cytotoxicity to mammalian cells. STAMPS are also designed to reduce the emergence of resistance.[90,91] These efforts and others have significantly increased the number of AMPs in the pipeline (Table 1.3).[78,86,90] Also, the Antimicrobial Peptide Database, which contains manually curated data of 3,273 AMPs (at the time of writing), is facilitating database-guided R&D of novel AMPs as alternative antibacterial agents.[92]

(vi) Repurposed drugs: repurposed or repositioned drugs are drugs being used for purposes other than their original licensed indications.[86,93-96] Drug repurposing is gaining traction in many areas of medicine and has become a credible research strategy.[95] Given that AMR needs to be tackled urgently, drug repurposing is all the more appealing as the costs and duration of drug development are massively reduced for drugs that have previously gone through preclinical and/or clinical trials with established safety, pharmacological, and formulation data, albeit for different indications. Repurposed drugs in the pipeline include anthelminthics (e.g. niclosamide and oxyclozanide), anticancer drugs (e.g. tamoxifen and raloxifene), anti-inflammatory drugs (e.g. celecoxib and diflunisal), and anti-cholesterol drugs (e.g. simvastatin and atorvastatin).[93,94] The mechanisms of action of these drugs are quite varied with some showing antibacterial activities in their own rights and others only effective as adjuncts to conventional antibiotics in the treatment of MDR infections.[93,94] While no repurposed drug has been licensed as an antibiotic alternative so far, researchers are quite optimistic for the prospects of the approach given recent advances in informatics and computational drug repositioning.[95,96] As of 2019 (latest data available at the time of writing), fifteen drug repurposing projects were in the preclinical pipeline and three were already being assessed in clinical trials.[86,93,96]

Other groups of antibiotic alternatives in the pipeline include antibiotic potentiators which primarily enhance or restore antibacterial activities of conventional antibiotics against antibiotic-resistant pathogens;[86,97,98] anti-virulence therapies that are designed as adjuncts to antibacterial agents where they disrupt bacterial processes like quorum sensing, biofilm formation and persistence;[99,100] immune system stimulators which enhance the production of host innate AMPs or other components of the immune system and are used as adjuncts to conventional antibiotics;[101] microbiota-modulating therapies (e.g. probiotics) which confer health benefits to the host when administered in adequate quantities;[76,102] and nanobiotics which are nanoparticles (e.g. colloidal forms of silver, zinc or copper) with direct antibacterial activities.[76,103]

It is encouraging to see increased research activities in the alternative therapy space as some of these therapies (especially those showing antibiofilm and anti-persistence activities) can restore lost efficacy of conventional antibiotics against MDR pathogens when used as adjuncts. However, antibiotics will remain the mainstay of antibacterial therapy, at least for the foreseeable future. Not least because majority of alternative therapies are only suitable as adjunctive or prophylactic treatments. It is also worth noting that conventional antibiotics, which are usually small molecules, are significantly cheaper to develop compared with many alternative therapies.[90] Furthermore, there are significant translational hurdles to overcome in developing alternative therapies for clinical use as many of these therapies (e.g. bacteriophages) are currently not suited to established clinical trials and regulatory protocols, and as such will require substantial funding and investments to create adapted or bespoke regulatory paths. Lastly, many alternative therapies are either pathogen-specific or show very narrow spectrum of activity compared with conventional antibiotics and as a result are likely to be licensed for limited or specific bacterial infections. This can disincentivise investments in alternative therapy projects unless policymakers take steps to guarantee good returns on investments. For all these reasons, it is unsurprising that R&D of novel antibiotics currently remains the most pragmatic strategy to adopt in order to urgently address AMR. Consequently, projects to increase the pipeline portfolio of conventional antibiotics dominate the research landscape as illustrated visually by Theuretzbacher *et al* (Figure 1.5).[86]

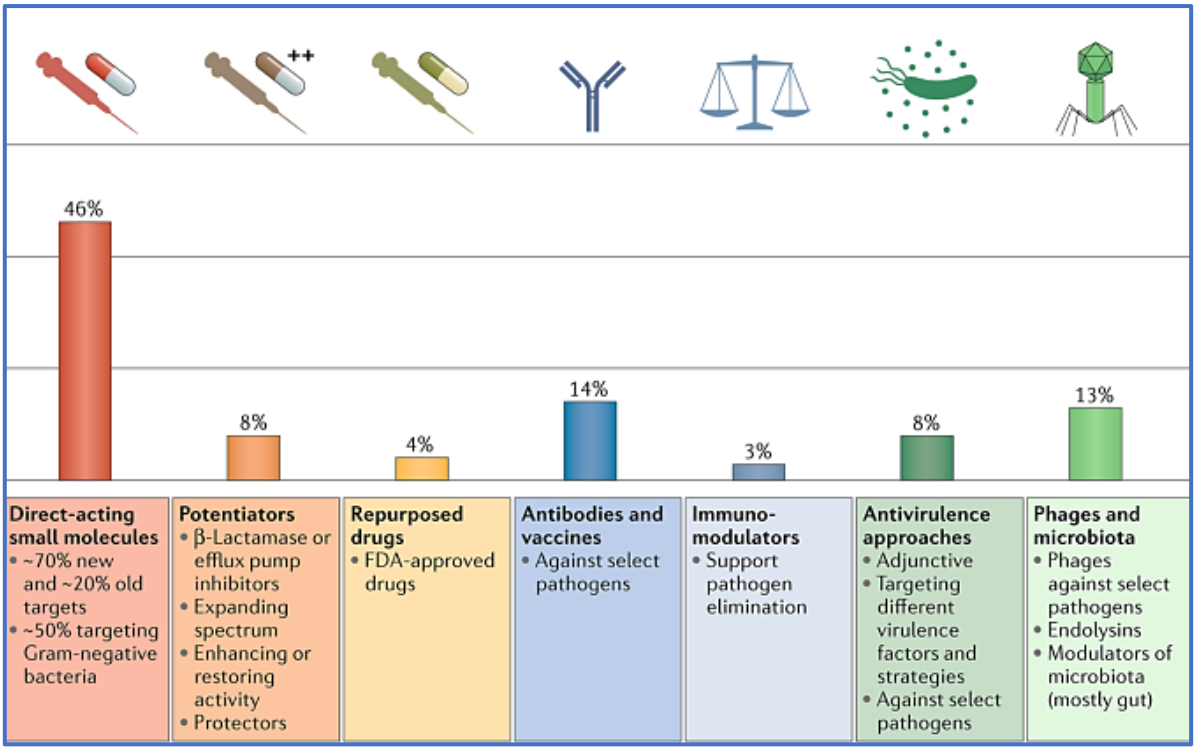


Figure 1.5: Preclinical pipeline portfolio of antibiotics and antibiotic alternatives (2019). Total number of projects = 407. Image adapted from Theuretzbacher *et al*.[86]

1.4 Antibiotics in the Pipeline (2020)

The WHO released its latest report and analysis of antibacterial agents in the global pipeline (at the time of writing) in April 2021. Preclinical and clinical data were collected from various sources including journal articles, the Pew Charitable Trusts' list of antibiotics in clinical development, pharmaceutical companies, and other relevant stakeholders.[78] The document reports on all therapies (i.e. antibiotics and antibiotic alternatives) in the pipeline as of 1st of September 2020.[78] The portfolio is restricted to therapies active against WHO priority pathogens (Table 1.2), *Mycobacterium tuberculosis* and *Clostridium difficile* to reflect treatment areas with the most critical needs.[78] Of the 292 therapies in the preclinical pipeline, 115 (39%) were conventional antibiotics of which 45 were pathogen-specific (mostly *M. tuberculosis*-specific). Nine of the antibiotic candidates have broad spectrum activity and many are active against at least two MDR pathogens with Figure 1.6 showing the number active against each MDR pathogen.[78]

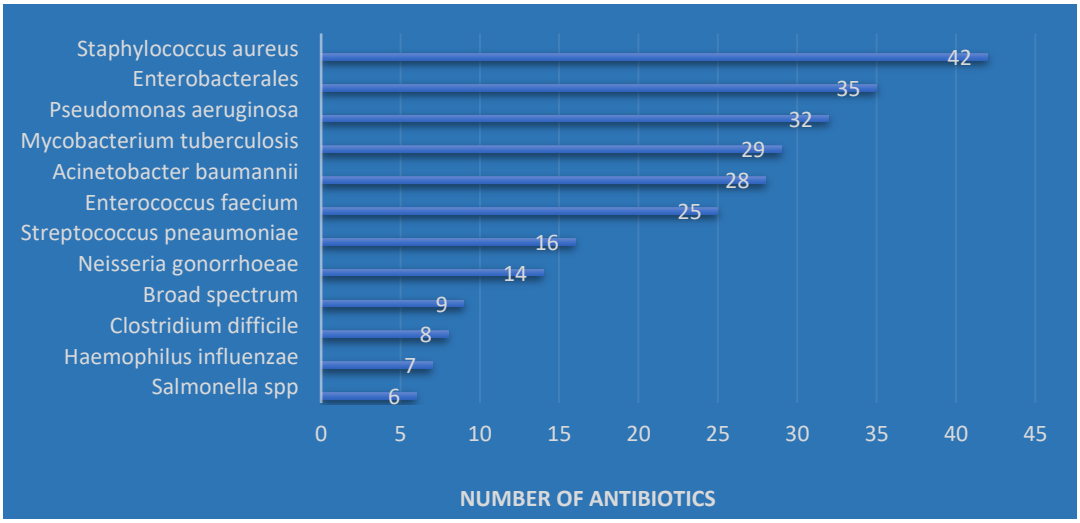


Figure 1.6: Number of antibiotics in the 2020 preclinical pipeline by MDR pathogen. Data extracted from the WHO report and analysis of antibacterial agents in the global pipeline.[78]

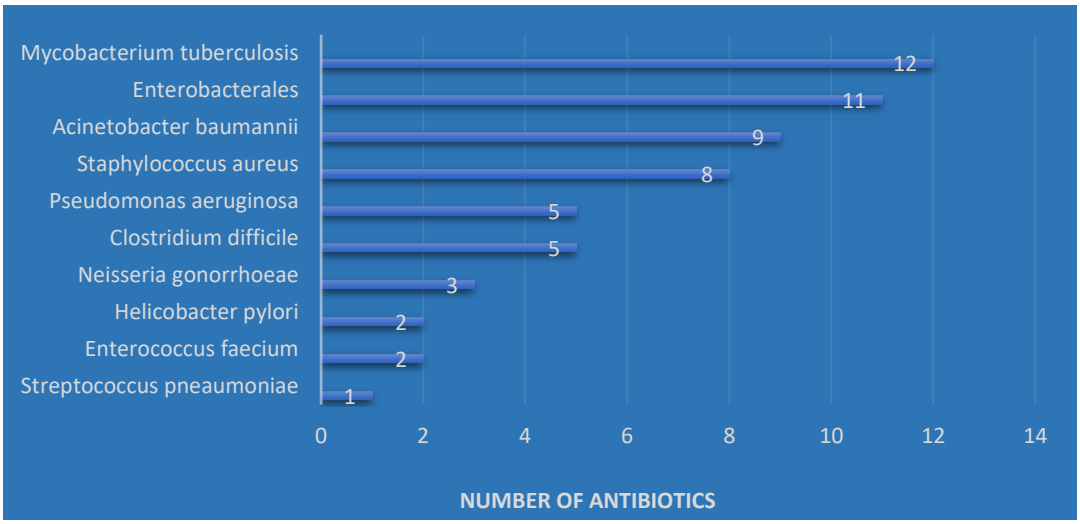


Figure 1.7: Number of antibiotics in the 2020 clinical pipeline by MDR pathogen. Data extracted from the WHO report and analysis of antibacterial agents in the global pipeline.[78]

In total, 68 therapies were in the clinical pipeline, 41 (60%) of which were conventional antibiotics. All the antibiotics are new chemical entities showing activity against at least one MDR pathogen, with none having broad spectrum activity. Figure 1.7 shows the number of antibiotics in the portfolio active against each MDR pathogen. Eighteen (44%) of these antibiotics were considered innovative i.e. fulfil at least one of the following criteria: new chemical class; new target or binding site; new mode of action; and/or no cross resistance to other antibiotic classes.[78]

According to WHO's assessment of the 2020 clinical pipeline portfolio of conventional antibiotics, progress is being made to address the urgent need for new antibiotics to treat the most critical MDR infections, but there are still significant gaps in the development landscape. For instance, MDR Gram-negative pathogens, especially carbapenem-resistant *A. baumannii* and *P. aeruginosa*, are still not sufficiently addressed in the current portfolio, and there are no candidates active against MDR *Salmonella* spp. (Figure 1.7). Also, more orally administered antibiotics active against extended-spectrum β-lactamase-producing-*Enterobacteriaceae* (ESBL-E) and carbapenem-resistant *Enterobacteriaceae* (CRE) strains would be needed to ensure treatments are accessible and available from outpatient care settings. The WHO hopes the current clinical pipeline could lead to the approval of eight new antibiotics in the next 5 years, but that is still insufficient to meet current and future needs according to their analysis.[78] Given that less than 15% of antibiotics in phase I clinical trials get approved for use,[78] and far higher attrition rates for preclinical drug candidates making it to phase I trials,[16] there is an imperative to substantially increase the preclinical pipeline portfolio of conventional antibiotics to meet critical current and future needs.

1.5 Strategies for the Discovery of New Antibiotic Lead Compounds

1.5.1 The Waksman Platform

The Waksman Platform remains the most successful discovery initiative to date. Many antibiotics in clinical use today were discovered through this initiative from the early 1940s to the early 1960s (Figure 1.1). Soil dwelling bacteria, mainly actinomycetes, were screened for their abilities to produce compounds with antibacterial activities.[15] The discovery platform was a huge success until the high rediscovery rate of known antibiotics rendered it obsolete. Consequently, the 1970s and beyond saw a sharp decline in discovery rates accompanied by a rise in the emergence of bacterial resistance to existing antibiotics. Various efforts were then made to restore the efficacy of these antibiotics and to explore other discovery initiatives.

1.5.2 Chemical Modification of Antibiotic Molecule, Combination Therapy, and Modification of Resistance Mechanisms

Modification of existing antibiotic molecule or scaffold, combination/hybrid therapies, and the use of resistance modifying agents were the most popular strategies adopted to restore lost efficacy of existing antibiotics. Semi-synthetic antibiotics, which are usually modifications or derivatives of existing antibiotics, were developed to restore or enhance efficacy. Excellent examples include carbapenems and cephalosporins which are β-lactam antibiotics that are relatively resistant to the effects of β-lactamases and are structural analogues of thienamycin and cephalosporin C, respectively;[9,104] azithromycin which was derived from the chemical modification of a macrolide ring;[105] and tigecycline, a derivative of minocycline, that is not affected by tetracycline resistance mechanisms.[106]

Combination therapies were also deployed to combat AMR and they involve the simultaneous use of two or more antibiotics to treat an infection, provided the mechanism of action of each component is not hindered or antagonised by the combination.[107] It is assumed that it is far less likely that a pathogen would develop resistance simultaneously to all components of a combination therapy than it would to each component used as monotherapy. Perhaps the most common combination therapy is the regimen comprising of three or four antibiotics (isoniazid, ethambutol, pyrazinamide, and rifampicin) used for treating *M. tuberculosis*.[107] Co-trimoxazole, which is a combination of trimethoprim and sulfamethoxazole, is another combination therapy that has seen antibacterial activity restored against some strains that are resistant to either component used as monotherapy.[107,108]

Several resistance modifying agents have been developed as adjuvants to restore the efficacy of existing antibiotics against resistant pathogens.[109,110] Clavulanic acid and tazobactam are examples of β-lactamase inhibitors that can restore efficacy of β-lactam antibiotics against β-lactamase-producing pathogens. Combination of clavulanic acid with amoxicillin in Augmentin® was a landmark achievement in the treatment of infections caused by amoxicillin-resistant pathogens. Efflux pump inhibitors (EPIs) e.g. phenothiazines and arylpiperazine, are another class of resistance modifying agents that block the extrusion of

antibiotics from bacterial cells thereby restoring efficacy.[110] Other resistance modifying agents include anti-virulence factors and membrane pemeabilisers.[109]

1.5.3 *De novo* Chemical Synthesis, High Throughput Screening, and Rational Drug Design

Although many of the aforementioned strategies restored or enhanced the efficacy of existing antibiotics, it is not uncommon for resistance to re-emerge after some time, highlighting the continuous need for novel antibiotics and novel resistance modifying agents. To that end, full chemical synthesis of antibiotic compounds, high throughput screening (HTS), and rational drug design (RDD) platforms were initiated. The chemical synthesis route to antibiotic discovery has so far had very limited success as only three classes of antibiotics - fluoroquinolones, oxazolidinones, and diarylquinoline - have emerged from the platform to date.[9,16]

Inspired by the availability of genomics data in the early 1990s, initiation of automated HTS platform appeared very promising.[111] Essential bacteria-specific receptor binding sites were screened against a library of synthetic compounds to identify lead compounds with affinity for those receptors. These compounds were then used in *in vitro* assays to assess their abilities to block vital bacterial processes and their effectiveness as antibacterial agents. Unfortunately, the platform has not yielded a single drug candidate despite best efforts.[16,111] The RDD route based on the analysis of 3D-structures of targets on bacterial receptors and designing of ligands that can interact with those targets has not been a successful platform either.[111]

De novo chemical synthesis, HTS, and RDD have been unsuccessful largely because synthetic compounds penetrate bacterial cells (especially cells of Gram-negative bacteria) very poorly or are extruded by MD efflux pumps and as such cannot access their target sites which are usually inside bacterial cells.[16,111] Also, most automated screening and chemical synthesis platforms adopted the Lipinski's rule of five to predict bacterial cell membrane penetration.[16,112] Whilst the rule works well for most molecules penetrating human cells, it appears to be unsuitable for predicting drug penetration into bacterial cells.[16,113] Consequently, based on the small library of compounds that tend to penetrate bacterial cells well and those that do not, the following physicochemical properties that influence

compound penetration into bacterial cells or extrusion from cells were compiled:[16,113]

(i) Small hydrophobic compounds tend to penetrate the outer membrane by diffusing through the lipid component of the membrane.
(ii) Hydrophobic cations and highly hydrophobic compounds are preferred substrates for MD efflux pumps and as such are readily extruded from bacterial cells after penetration.
(iii) Hydrophobic anions are not good substrates for MD efflux pumps.
(iv) Amphipathic compounds (which most antibiotics are) penetrate the outer membrane poorly and those that leak through are extruded by MD efflux pumps.
(v) Relatively hydrophilic compounds ≤600 Da in size tend to penetrate bacterial cells probably through porin channels.
(vi) Larger hydrophilic compounds are restricted by the inner membrane from entering the cell.
(vii) Incorporation of fluorine or boron atoms (or other rare atoms) into antibiotic molecules favours penetration as these atoms do not occur frequently in natural products and as such are not recognised or extruded by MD efflux pumps.

Adoption of these rules instead of the Lipinski's rule of five could improve the success rates of HTS and RDD and rejuvenate the platforms, as well as guide the development of new antibiotics generally.[16,113] For instance, hydrophobic lead compounds with intracellular targets could be formulated in combination therapies with EPIs. It is worth bearing in mind however that some antibiotics attack targets on the cell surface (e.g. daptomycin and vancomycin), or in the periplasm (e.g. penicillins) and in such cases the rules of compound penetration into bacterial cells become less relevant.

Another major shortcoming of HTS and RDD platforms as identified by many authors is the target-based screening adopted by these platforms as opposed to cell-based or *in situ* screening.[16,114] Whilst the platforms can identify compounds that block specific bacterial targets, antibacterial activities are often not achieved in *in vitro* assays because of the complexity of bacterial metabolism. Incorporating cell-based or *in situ* screening into HTS and RDD platforms might generate

successful lead compounds according to Lewis and Gajdács.[16,114] However, ethical concerns may arise with *in situ* screening as it requires excessive amounts of animal models. A viable solution as suggested by Lewis will be to use worms (e.g. *Caenorhabditis elegans*) as hosts instead of animals.[16]

HTS and RDD target-based screening will also miss prodrugs (e.g. metronidazole and isoniazid) which further reinforces the need to incorporate cell-based or *in situ* screening into these platforms.[16] Prodrug formulation strategy is becoming increasingly important when it comes to antibiotic therapies. It has been suggested that the 'ideal antibiotic' is preferably a prodrug that is converted into its active form inside bacterial cells by bacteria-specific enzymes (therefore not toxic to mammalian cells), with the active form binding covalently to many unrelated targets within bacterial cells.[16,114] Because the active form binds covalently to multiple targets within bacterial cells, it is unlikely to be extruded by MD efflux pumps thus creating an irreversible sink.[16,114] A broad-spectrum activity is also achievable because of its ability to bind to multiple unrelated targets and very importantly, it will eliminate persister cells as a result of targeting multiple sites.

1.5.4 Bioprospecting for Natural Products with Antibiotic Properties

Following the less encouraging experiences with chemical synthesis, HTS, and RDD, efforts are shifting towards further exploitation of natural products as sources of novel antibiotics. Natural products and their derivatives account for 60-80% of antibiotics in clinical use (Figure 1.8).[9,115,116] It is surprising there are no plant-based antibiotics in clinical use given the vast number of bioactive compounds that have been isolated from plant sources; this may therefore represent a credible research opportunity.[16] Currently, microorganisms (especially bacteria) are the main sources of these clinically useful natural products. Interestingly, less than 1% of all known bacterial species are the producers of all bacteria-derived therapeutically useful antibiotics. As chemical diversity follows biological diversity,[16] it is reasonable to assume that bacterial species represent an untapped resource in the search for novel antibiotic lead compounds. The assumption is further reinforced by the realisation that known antibiotic producers

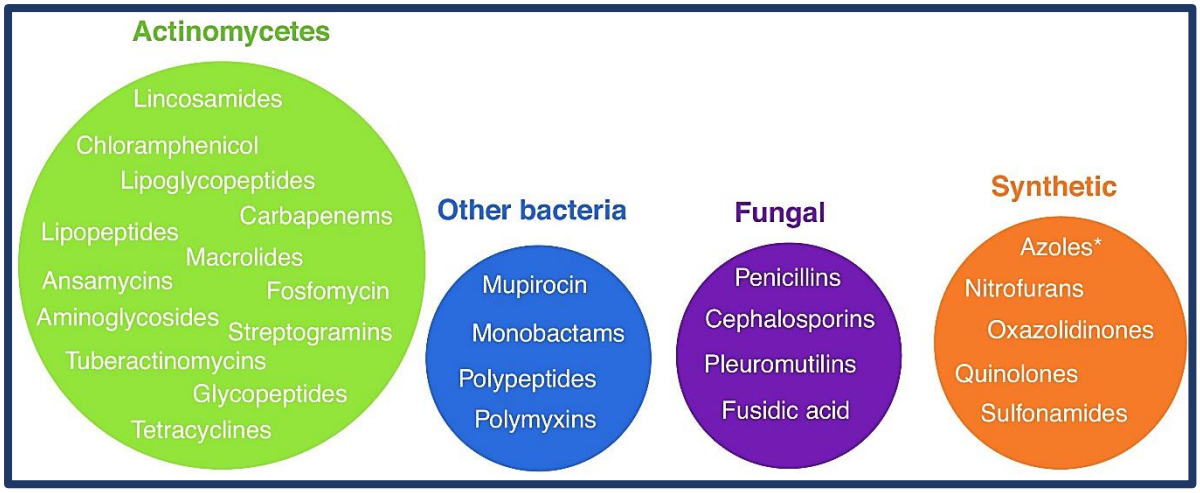


Figure 1.8: Classes of antibiotics and their sources.
Image adapted from Hutchings *et al*.[9]

can biosynthesise more bioactive compounds (or secondary metabolites) than initially thought. Recent developments in genomics and computational biology have given unprecedented insights into the metabolic and biosynthetic potentials of microorganisms, and there is abundant optimism that these potentials can be harnessed to develop new therapeutic agents, including novel antibiotics.

1.6 Microbial Secondary Metabolites as Sources of Bioactive Compounds

Microbial secondary metabolites, unlike primary metabolites, are low molecular weight (<3,000 Da) compounds produced by certain microbes which are not essential for normal growth processes but play some ecological roles in the organisms' microenvironment.[115-121] They are thought to confer competitive advantage on the producer and may also be produced in response to various stress situations such as nutrient or essential minerals deprivation - antibiotic production is thought to be linked to nutrient depletion or starvation stress.[115-120] Some secondary metabolites may function as signalling/communication molecules between microbes or between microbes and their environment.[115-120] The ability to biosynthesise secondary metabolites is not universal to all microorganisms, and some are better producers than others.[115,121] The most prolific producers of these compounds include *Streptomyces* species and other actinomycetes.[115,118] Other proficient producers include *Myxobacteria*, *Bacillus* and *Pseudomonas* species as well as some filamentous fungi.[115,119]

It is estimated that around 75,000-80,000 secondary metabolites have been isolated from various microbes, of which at least 38,000 are bioactive i.e. have a biological effect or trigger a response in a living organism, tissue, or cell.[119] Some of these bioactive metabolites have been exploited for therapeutic purposes including for antitumour, anticholesterol, antifungal, antiviral, and antibiotic purposes. Of the 28,500 metabolites known to show some antibiotic activity, less than 1% have been developed for clinical use as toxicity to mammalian cells remains a major barrier to clinical application.[115,119] Some metabolites are also very weak antibiotics requiring excessively high concentrations to produce desired antibacterial effects suggesting their primary role in nature may not be antibiosis.[120,122] However, the AMR crisis has meant that many of these supposedly toxic, weak, or even non-bioactive metabolites are being re-examined for clinical application given recent advances in chemical microbiology, combinatorial biosynthesis, and formulation technology.[119,123]

Despite the seemingly large number of microbial secondary metabolites that have been isolated and studied, the microbial natural product collection is dwarfed by the library of at least eight to ten million synthetic products that are being assessed for bioactivity through HTS or RDD platforms.[119] But as discussed earlier, combinatorial chemistry alone has been relatively unsuccessful in addressing the steep decline in the discovery of novel antibiotics. As succinctly stated by Bérdy, 'the best combinatorial chemists are microbes';[119] and by Lewis, 'prospecting for natural products is in essence a 'screen of screens': natural products produced by bacteria are the result of natural selection over billions of years;'[16] - microbial natural products appear to be the best match for pathogenic microbes. Therefore, there is an urgent need to substantially expand the chemical diversity of microbial natural products to create a larger library of natural compounds that can be screened for antibacterial activity. As fungal secondary metabolites tend to be more toxic to mammalian cells (hence only a few are therapeutically useful as antibiotics) compared with bacterial metabolites, efforts are being made to expand the chemical space of bacterial secondary metabolites by exploring diverse species from diverse niches, even the bottom of the ocean.[16,119]

Inability to culture over 90% of bacterial species under standard laboratory conditions has been a barrier to their exploitation for drug discovery purposes.[16,121,124] As such several unconventional methods have been developed

to cultivate previously unculturable bacterial species. These methods include cultivation in minimal substrate media, *in situ* cultivation, cultivation with specific growth factors, and co-cultivation.[121,124,125] Many previously unculturable species have been successfully cultured or isolated using these techniques. The isolation of *Eleftheria terrae* from a soil sample by *in situ* cultivation in multichannel iChip was reported by Ling *et al* in 2015.[126] A partially purified active fraction of extracts obtained from *E. terrae* contained teixobactin, a first in class antimicrobial agent.[126]

The sharp decline in the rate of discovery of new antibiotics from the 1970s onwards was also partly due to the limited insights into microbial secondary metabolism. Microbial secondary metabolites are products of biosynthetic pathways of which all enzymes, regulatory proteins, tailoring enzymes and transport proteins required to biosynthesise and export the products are encoded by adjacent genes on microbial chromosomes.[121,127-129] In other words, the genes encoding all components of a biosynthetic pathway are usually clustered together on chromosomes and are collectively referred to as the biosynthetic gene cluster (BGC) for that pathway. The first BGCs within partial genomes of *Streptomyces* spp. were discovered in the 1970s and 1980s, unveiling the connections between gene contents and secondary metabolism for the first time.[130-132] Following the availability of complete genomes of *Streptomyces coelicolor* and *Streptomyces avermitilis* in the early 2000s, it became apparent that the genomes of these prolific secondary metabolite producers contained up to thirty BGCs encoding natural products whereas only two to three of these products were known at the time.[121,133-135] It was quickly realised that microorganisms have much more potential and capacity to biosynthesise useful natural products than previously thought. This knowledge formed the basis of genome mining, where genomics data is used to assess the biosynthetic capacities of microorganisms thereby allowing predictions to be made about their potential to biosynthesise secondary metabolites that can be exploited for therapeutic purposes.[121,136,137]

Apart from the order of nucleotides/genes in microbial genomes, another crucial indicator of an organism's biosynthetic potential is the size of its genome.[127] The genomes of all microorganisms must contain genes encoding all necessary primary metabolites, only then can any spare capacity be dedicated to genes encoding secondary metabolites. The 3 Megabase (Mb) mark appears to be the threshold

above which increasing proportion of a genome is dedicated to secondary metabolism. Above 3 Mb, the larger the genome, the greater the coding capacity for secondary metabolites.[127] It is therefore not surprising that organisms like *Streptomyces* spp. with genomes as large as 8-9 Mb are prolific secondary metabolite producers.[121] Interestingly, the corelation between genome size and secondary metabolite biosynthesis has disproved the longstanding assumption that unculturable bacteria hold the key to natural products discovery renaissance. Several culture-independent sequencing projects have revealed that the majority of unculturable bacteria have genomes ranging from 0.6 to 3.1 Mb in size with an average of just 1.6 Mb.[138] Consequently, Baltz eloquently made the case for shifting research focus to 'gifted microbes' with genomes larger than 8 Mb, even though (by the author's own admission) that benchmark would exclude ≈85% of culturable bacteria.[138] To adopt the 8 Mb cut-off mark as suggested by Baltz, instead of 3 Mb, would mean to ignore the phylogenetic and biosynthetic diversity of the vast majority of culturable bacteria, a situation the research community can ill afford in the face of AMR.

1.7 Genome Mining for Secondary Metabolite Biosynthetic Gene Clusters

Availability of next-generation sequencing and a steep reduction in costs in recent years have made genome sequencing and gathering of genetic information common practice. Thousands of microbial genomes have been (and are still being) sequenced, generating vast amounts of data that can be mined for BGCs encoding secondary metabolites.[121] Known secondary metabolite families include polyketides (PKs) encoded by polyketide synthases (PKSs), non-ribosomal peptides (NRPs) encoded by non-ribosomal peptide synthetases (NRPSs), ribosomally synthesised and post-translationally modified peptides (RiPPs), terpenoids, bacteriocins, various PKs-NRPs hybrids, etc.[127,136] Although there are diverse chemistries and bioactivities within each group, the core enzymes involved in the biosynthetic machinery of respective groups are often conserved.[127,139] Therefore, secondary metabolite biosynthetic gene clusters (smBGCs) of these metabolic products can be recognised by looking for signature patterns within microbial genomes.[121] These patterns are usually identified by sequence comparison algorithms based on profile hidden Markov models (pHMMs) or Basic

Local Alignment Search Tool (BLAST) integrated into *in silico* bioinformatics tools used for genome mining.[136]

1.7.1 Bioinformatics Tools for smBGC Identification and Analysis

In silico identification of smBGCs within genomes can be performed with various robust and user-friendly bioinformatics tools such as antiSMASH,[140] BAGEL,[141] eSNaPD,[142] ClustScan,[143] NaPDos,[144] etc. (Table 1.4). Some tools only detect certain smBGC types e.g. RODEO[145] and NP.searcher[146] will only detect RiPPs and NRPS/PKS type smBGCs, respectively, while other tools like PRISM 4[147] and antiSMASH are comprehensive and will detect a wider range of smBGC types.

Table 1.4: Genome Mining Tools

Tool	Predicted smBGC	Structure Prediction	URL (Reference)
antiSMASH	Comprehensive	Yes	http://antismash.secondarymetabolites.org (140)
ClusterFinder	Comprehensive	No	http://github.com/petercim/ClusterFinder (128)
NP.searcher	NRPS, PKS	Yes	http://dna.sherman.Isi.umich.edu/ (146)
BAGEL	Bacteriocin, RIPP	No	http://bagel.molgenrug.ng/ (141)
ClustScan	NRPS, PKS	Yes	http://bioserv.pbf.hr/cms/ (143)
NaPDos	NRPS, PKS	No	http://napdos.ucsd.edu (144)
eSNaPD	Comprehensive	No	http://esnapd2.rockefeller.edu/ (142)
RODEO	RIPP	Yes	http://ripp.rodeo/ (145)
PRISM 4	Comprehensive	Yes	http://prism.adapsyn.com/ (147)
EvoMining	Comprehensive	No	https://github.com/nselem/evomining (150)

Many authors have reviewed these tools (and others not listed here), highlighting their merits and limitations.[129,139,148,149] Currently, antiSMASH remains the most comprehensive and user-friendly genome mining pipeline with high-confidence

and low false-positive ratings.[136] However, novel smBGC types are not detectable when using antiSMASH tool on the web interface as sequence homology to known smBGCs is required for detection.[140] ClusterFinder[128] and EvoMining[150] on the other hand are comprehensive tools with the added advantage of the ability to detect novel smBGCs, however their false-positive ratings are higher. The downloadable versions of antiSMASH have incorporated the ClusterFinder algorithm as an optional feature to enable mining for novel smBGCs, but with increased risk of false-positive results.[140] Some of these bioinformatics tools also have built-in functionalities to predict the structures of end products associated with biosynthetic pathways, especially PK and NRP pathways as they are the most studied. Furthermore, the novelty of detected smBGCs can be assessed with various incorporated modules. Figure 1.9 illustrates the role of bioinformatics tools in genome mining and natural product discovery.

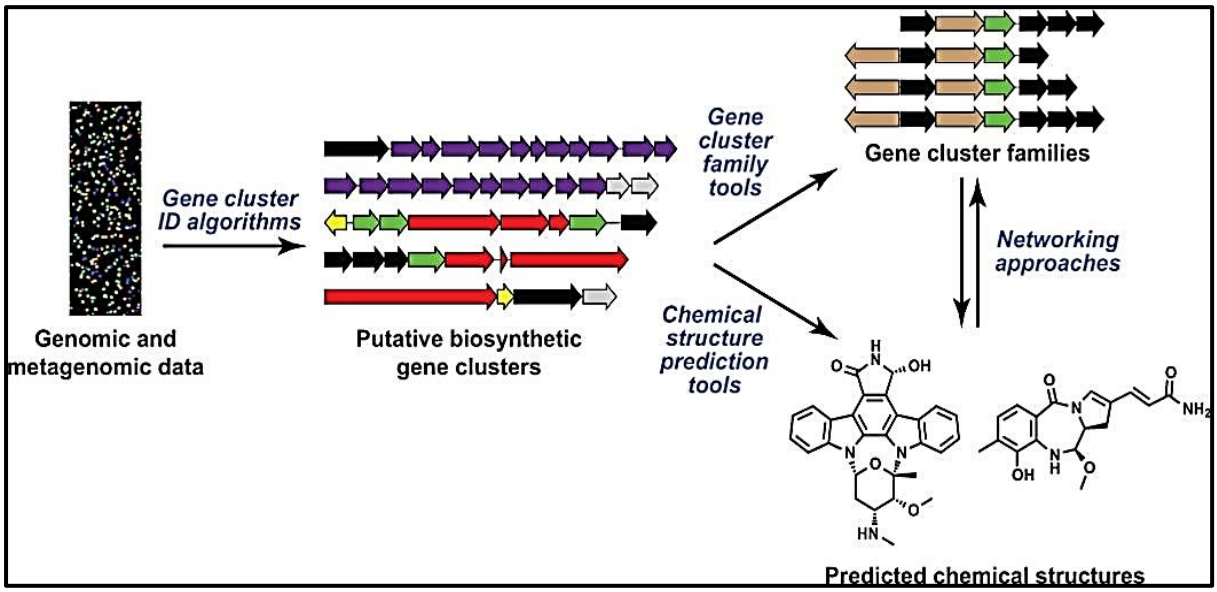


Figure 1.9: Illustration of the role of bioinformatics tools in natural product discovery.
Genomic and metagenomic data are mined using various bioinformatics tools to identify putative smBGCs and gene cluster families encoded in input data, with some tools able to predict chemical structures of encoded metabolites.
Image adapted from Medema and Fischbach.[129]

1.7.2 The Role of Genome Mining in the Discovery of Novel Natural Compounds

Genome mining can be applied to genomes of single cells or metagenomes (i.e. assembly of genomes obtained from complex environments). On scanning a genome or metagenome for smBGCs, priority is usually given to novel smBGCs, which are usually cryptic or silent, as their end products are more likely to be novel or unstudied.[121,151,152] There are no universal strategies for activating smBGCs of interest, but strategies are usually informed by detailed study of the gene clusters. Activation strategies can range from very simple fermentation experiments to those employing downright audacious genetic manipulation techniques.

Cryptic smBGCs are expressed in the native organism when the right standard fermentation conditions are used, but their metabolites are unknown and/or unstudied (i.e. not linked to their respective smBGCs). If accurate *in silico* end product structure prediction is possible, and metabolites are produced in sufficient amounts, the products of interest can be identified (based on predicted structures) and isolated from fermentation broths.[121] In many cases, the concept of OSMAC (**o**ne **s**train **ma**ny **c**ompounds) is adopted to access the full range of metabolites produced by an organism with changing fermentation parameters.[153] In the event of inaccurate structure prediction or where structure prediction is not possible, gene knockout techniques may be used to identify products encoded by cryptic smBGCs.[152,154] The gene(s) of interest is first deleted from the genome of the organism prior to fermentation experiments. Fermentation products obtained with the modified strain are then compared with those obtained with the wildtype strain. Missing metabolites in the former are attributed to the deleted gene(s) in the modified strain.[121] These products can then be isolated, characterised and screened for bioactivity. In cases where the product of interest cannot be expressed in sufficient amounts in the native organism and/or the organism grows extremely slowly, the smBGC can be transferred into a suitable host for heterologous expression; provided the end product is not toxic to the host, and gene activation in the host is successful.[152,154] Many novel compounds have been isolated from fermentation broths by activating cryptic smBGCs, these include coelichelin isolated from *Streptomyces coelicolor* by Lautru *et al*, and

aspoquinolones A-D isolated from *Aspergillus nidulans* by Scherlach and Hertweck.[155,156]

Silent smBGCs, unlike cryptic smBGCs, are not expressed in the native strain under normal fermentation conditions. Therefore their activation necessitates heterologous expression in a suitable host or genetic manipulation of the native strain to achieve gene expression.[151,154,157] Strategies to activate silent smBGCs in the native organism include transcription activation by an inducer, and removal or inactivation of a gene inhibitor.[151,154,157] If gene expression inducer or inhibitor is known, fermentation broths can be supplemented accordingly to facilitate expression of the silent smBGC. Silent smBGCs can also be activated by using different types of stress signals e.g. phosphate starvation stress, nitrogen source stress and metal toxicity stress.[121,158-160] Some researchers have been able to activate silent genes by co-cultivation of the organism of interest with other organisms,[161,162] while others have genetically modified the organism of interest to induce gene expression.[163,164] The most successful methods of activating silent smBGCs have been those requiring minimal manipulation of the native organism.[154] It is expected that techniques involving heterologous expression will be more successful as the technology develops. Production of many bioactive compounds have been induced by the activation of silent genes, these include stambomycins isolated from *Streptomyces ambofaciens* by Laureti *et al*; and holomycin, isolated from *Streptomyces clavuligerus* by Charusanti *et al*.[157,162,165.]

Genome mining has undoubtedly created new and exciting platforms for natural product discovery. The genomes of culturable microbes can now be mined to exploit their full biosynthetic potentials. Heterologous expression of environmental DNA in suitable hosts now enables culture-independent natural product discovery meaning the diverse chemistries encoded by unculturable or yet-to-be-cultured microbes in the environment can be accessed and exploited for drug discovery purposes.[166,167] Furthermore, comparative genome mining enables comparison of newly identified smBGC in an organism with curated smBGCs in various databases (e.g. the antiSMASH database[168] and the Minimum Information about a Biosynthetic Gene Clusters (MIBiG) database[169]) to aid quick characterisation and dereplication, meaning truly novel smBGCs are prioritised for further study. Comparative genomics also enables researchers to explore the evolutionary relatedness of species within phylogenetic groups in order to determine the roles

of vertical and horizontal evolution in inter- and intra-species biosynthetic diversity.[170] Lastly, genome mining offers a more rational approach to natural product discovery in contrast to the ad-hoc nature of traditional screening of microbial extracts for bioactivity.[121,129,136]

1.8 Project Aim and Objectives

Bioprospecting for novel bacterial secondary metabolites with antibiotic properties is crucial to the fight against AMR. There is a continuous need to expand the chemical diversity of bacteria-derived natural products that can be exploited as lead compounds for the development of therapeutically useful antibiotics. As chemical diversity follows phylogenetic diversity, an inexhaustible resource of under- or un-explored bacterial species is available to be leveraged for this purpose. Furthermore, the tasks of bioprospecting for novel natural compounds have never been easier given the availability of robust genome mining tools that can expedite the process, coupled with recent advances in computational and systems biology.

The overall aim of this research was to expand the chemical space of bacteria-derived natural products that can potentially serve as lead compounds in the development of novel antibiotics. This aim can be achieved here by:

(i) Selectively isolating potential antibiotic-producing bacterial strains from soil samples using ultra-minimal substrate medium that can induce starvation stress.
(ii) Preliminary identification of recovered isolates and assessment of their biosynthetic potentials based on initial taxonomic and genomic data; and afterwards prioritising candidate strains that are not readily associated with antibiotic production for whole-genome sequencing.
(iii) Obtaining whole-genome sequencing data and assembling high-quality complete genomes of candidate strains.
(iv) Performing detailed *in silico* analysis of assembled complete genomes to capture the full metabolic and biosynthetic capacities of candidate strains, including mining for putative smBGCs that may encode novel metabolites with antibiotic properties.

(v) Establishing fermentation conditions that may encourage activation of selected smBGCs of interest through in-depth *in silico* gene cluster/pathway analysis.

(vi) Performing wet lab fermentation experiments with conditions identified, followed by extract generation and preliminary targeted LC-MS analysis of crude extracts to determine if activation was successful (provided there is sufficient prior knowledge of the structural/chemical characteristics of secondary metabolite(s) of interest). Results of initial LC-MS analysis will inform future scale-up, isolation, characterisation, and bioassay experiments.

CHAPTER TWO

ISOLATION OF POTENTIAL ANTIBIOTIC-PRODUCING BACTERIAL STRAINS FROM SOIL SAMPLES.

2. Isolation of Potential Antibiotic-Producing Bacterial Strains from Soil Samples

Soil is an extensive ecosystem that supports microbial and plant lives and mediates many processes that make life conducive on Earth. Soil-dwelling microbes produce chemical substances that serve various ecological purposes in their microenvironments. Many of these chemical substances are bioactive and are exploited for various therapeutic purposes including use as antibiotics.

2.1 Introduction

Of all readily accessible external environments (i.e. soil, lakes, rivers, caves, air, etc.), soil remains the most biodiverse ecosystem with constant microbe-microbe and microbe-environment interactions.[43,121,171-174] It is estimated that 0.5 g of soil contains thousands of unique bacterial species and billions of individual bacterial cells.[43,171] Soil-dwelling bacteria produce certain chemical substances or secondary metabolites to gain competitive advantage for nutrient and vital substrates and/or to communicate with other microbes and other living components in the ecosystem.[115-117] Many of these secondary metabolites are bioactive and as such have been exploited for therapeutic purposes. The majority of antibiotics in clinical use today are derivatives of secondary metabolites of soil-dwelling bacteria, mainly actinomycetes.[115,119] However, less than 1% of all known bacterial species have had their metabolites exploited in this way.[119] Given the biodiversity present in the soil and its accessibility, it presents an expansive natural resource to explore for novel antibiotic-producing bacterial strains.

A cross-section of most soil types will reveal distinct layers with different physical, chemical, and microbial composition (Figure 2.1). Organic matter, water, and oxygen contents decrease down the horizons. As a result, microbial population or biomass also varies down the layers. The topsoil is the most nutritionally dense layer with good supply of organic matter, oxygen, and water. Consequently, the topsoil layer has the highest biomass while the bedrock layer has the lowest biomass.[171] Soils can also be differentiated based on their association with plant roots. Soil that is in direct contact with plant roots forms the rhizosphere, while bulk soil has no association with plant roots.[172-174] The plant rhizosphere has been

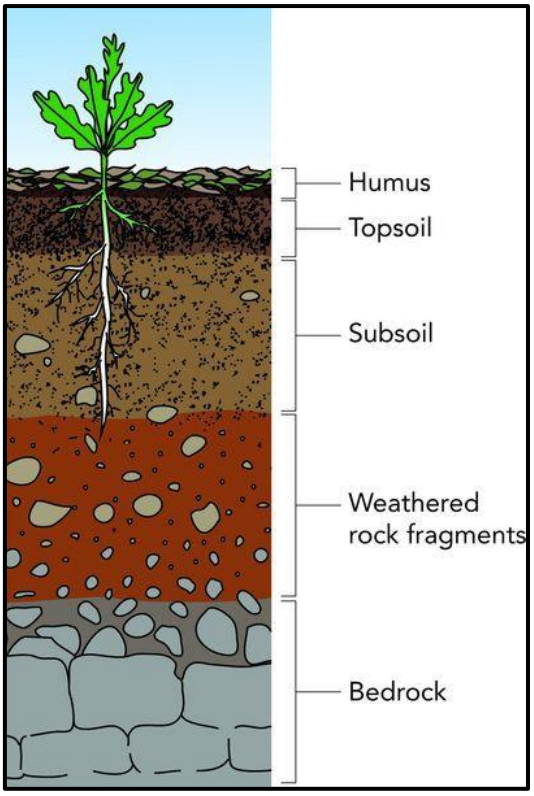


Figure 2.1: Cross-section of soil showing different layers or horizons.
Image source: Wikimedia (https://commons.wikimedia.org)

described as one of the most complex ecosystems on Earth with dynamic plant-microbe and microbe-microbe interactions.[172-174] Rhizobacteria (bacterial species found in the rhizosphere) have been shown to promote plant growth in the presence of biotic and abiotic stress.[172-174] They do so by enhancing nutrient and inorganic element uptake by root systems and by increasing plants' tolerance to environmental stressors including heavy metals, high salt concentrations, and phytopathogens.[173] Rhizobacteria that protect against phytopathogens do so through a variety of mechanisms including the ability to produce one or more secondary metabolites that have lethal effects on plant pathogens, and also afford the rhizobacteria some competitive advantage for vital nutrients and trace elements during starvation stress.[173] Hence, the rhizosphere represents a promising niche to explore for novel antibiotic-producing bacterial strains as antibiosis occurs naturally within it. However, as antibiotic production or secondary metabolism is not universal to all rhizobacteria, criteria that would improve the chances of recovering potential antibiotic producers are necessary. Therefore, isolates were recovered from the rhizosphere using starvation stress as antibiotic production is linked to survival traits expressed during nutrients

depletion.[120,175] Also, as a corelation exists between genome size and increasing capacity for secondary metabolism,[127] the 3 Mb size above which increasing proportion of a bacterial genome is devoted to secondary metabolism was adopted as a minimum threshold / genome size cut-off for this work.

2.2 Aims and Objectives

The aims of this section of the project were:

- To isolate nutritionally versatile bacterial strains from topsoil samples collected from the rhizosphere using starvation stress.
- To use preliminary taxonomic and genomic data to identify and prioritise potential novel antibiotic-producing isolates for further work.

These aims can be achieved by:

(i) Using ultra-minimal substrate medium to isolate bacterial strains from nutrient-dense topsoil samples collected from the rhizosphere.

(ii) Cultivating isolates in various media with varying nutrient concentration to confirm each isolate's ability to withstand starvation stress.

(iii) Performing preliminary identification of selected nutritionally versatile isolates using 16S rRNA gene sequence comparisons.

(iv) Prioritising isolates that are not readily associated with antibiotic production but have genomes typically larger than 3 Mb for whole-genome sequencing.

2.3 Methodology

2.3.1 Sampling

Soil samples collected from soil supporting vegetation at a few inches below the soil surface (i.e. from the topsoil layer) and close to plant roots (i.e. from the rhizosphere) are more likely to harbour antibiotic-producing bacterial strains than samples collected from bulk soil.[121,172,173] Hence, sampling was carried out here at a few inches below the soil surface close to plant roots.

2.3.2 Strain Isolation

Bacterial strains that can withstand starvation stress are likely to be isolated from a nutrient-dense topsoil sample by using an ultra-minimal substrate medium.

Therefore, 1:100 Ravan medium[125] was used for strain isolation here. There is a sharp contrast between the nutrient concentration in the nutrient-dense topsoil sample and the ultra-minimal substrate medium, therefore the medium should selectively isolate nutritionally versatile bacterial strains that are able to thrive in both nutritionally dense and sparse environments.

Nutritionally versatile bacterial strains isolated from the rhizosphere in this way are potential producers of secondary metabolites that can afford the producer some competitive advantage for scarce nutrients. Some of these metabolites may be bioactive and some may be suitable for exploitation as antibiotic lead compounds.[120,175]

2.3.3 Verification of Nutritional Versatility of Isolates

The nutritional versatility of isolates obtained as described above, and their ability to withstand starvation stress, can be verified by cultivating isolates in both liquid and solid media with varying nutrient concentration and incubation temperature. This would ensure isolates selected for further work can produce their own metabolites to withstand starvation stress and not just utilising communal metabolites secreted by other microbes in close proximity.

2.3.4 Preliminary Identification of Selected Isolates

Traditional means of identifying bacterial species based on comparisons of phenotypic attributes have been shown to be less accurate than identification based on comparisons of genetic attributes.[176,177] The genetic marker of choice in bacteria is the 16S rRNA gene, which is a housekeeping gene and therefore universally conserved.[121,176] Therefore, 16S rRNA gene sequence comparison was used here for preliminary identification of selected isolates.

2.3.4.1 The 16S rRNA Gene

The 16S rRNA gene encodes the 30S small subunit (SSU) of the bacterial ribosome. It is approximately 1500 base pairs (bp) long, consisting of regions of highly conserved nucleotide sequence interspersed with regions of hypervariable

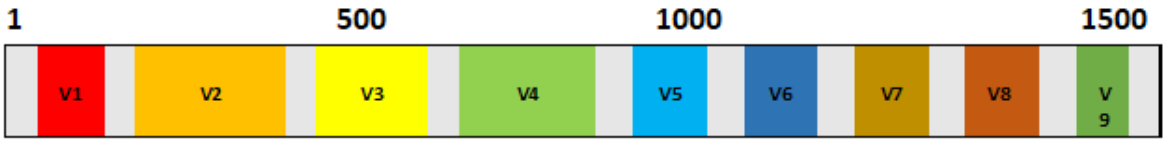


Figure 2.2: Illustration of the 16S rRNA gene.
Highly conserved regions are depicted by grey boxes that are interspersed by hypervariable regions V1-V9 that are depicted by coloured boxes.

nucleotide sequence (Figure 2.2). The nucleotide sequence in the highly conserved regions is quite similar in all bacterial species whereas sequences in the hypervariable regions V1-V9 are dissimilar in different species but quite similar between strains of the same species.[121] Therefore, comparisons of the 16S rRNA gene sequence of unknown isolates to those of known species contained in curated databases can aid in the identification of unknown isolates.[176-178]

2.3.4.2 16S rRNA Gene Nucleotide Sequence Determination

Prior to gene sequence comparison, the 16S rRNA gene is amplified in polymerase chain reactions (PCR) using genomic DNA (gDNA) extracted from lysed cells as template (Figure 2.3). A forward and a reverse primer pair complementary to two

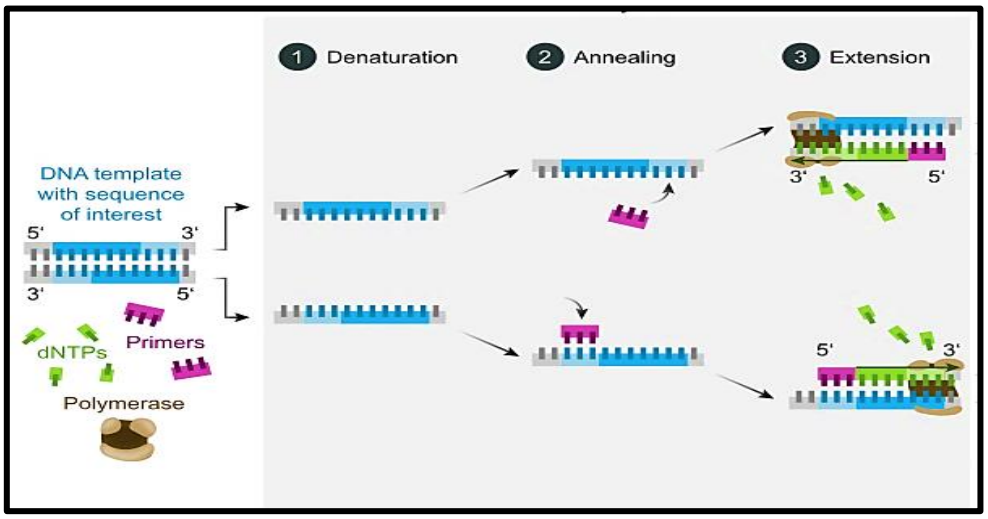


Figure 2.3: Stages of Polymerase Chain Reaction
PCR reactions involve three stages: (1) separation of double-stranded DNA into single strands at high temperature; (2) annealing of primers to the single strands at lower temperature; and (3) elongation of the primers to the full length of the single strands by insertion of complementary deoxynucleotides (dNTPs) in reactions catalysed by heat-resistant DNA polymerase.
Image source: Wikimedia (https://commons.wikimedia.org).

highly conserved regions on the gene can be used to copy and amplify intervening hypervariable regions. Appendix 1 shows the colour-coded 16S rRNA gene sequence of *E. coli,* highlighting highly conserved regions that can serve as primer binding sites in PCR reactions.[179]

The exact nucleotide sequence of amplified 16S rRNA gene can be determined by Sanger sequencing reactions similar to PCR but utilising a single primer. A chromatogram is produced at the end of reactions with each peak on the chromatogram corresponding to a nucleotide on the template (Figure 2.4). A good sequencing reaction will generate few or no indeterminate nucleotides (i.e. N, Y, W, S, etc) with well separated peaks.[121,178] Crucially, the quality of the chromatogram validates or invalidates the nucleotide sequence data. Therefore, sequencing data must be viewed and edited by cutting off regions of poor peak resolution from trace and text data, if necessary. This can be performed with many user-friendly software e.g. Chromas v 2.6.5 (Technelysium Pty Ltd, Australia).[180]

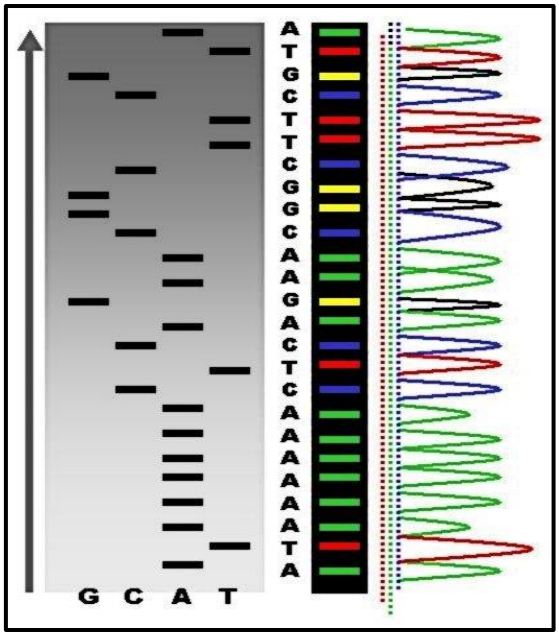


Figure 2.4: A visualised chromatogram and associated nucleotide sequence. A good sequencing reaction will generate well resolved peaks with few or no indeterminate nucleotides. Image source: Wikimedia (https://commons.wikimedia.org).

2.3.4.3 16S rRNA Gene Sequence Repositories/Databases

16S rRNA gene sequence data have been obtained for thousands of bacterial species and are deposited in various public repositories, with the NCBI's (National

Centre for Biotechnology Information) database, GenBank, being the largest.[181] GenBank accepts sequence data from laboratories around the world. A number of earlier deposits were not peer-reviewed and as such might have been of poor quality.[176,178] As a result, RefSeq (the Reference Sequence database) which contains manually curated and peer-reviewed sequence data was created by NCBI.[182] The database entries include the 16S RefSeq nucleotide sequence record.[182] Other databases with curated 16S rRNA nucleotide sequences include SILVA,[183] EzTaxon,[184] and Greengenes;[185] but RefSeq remains the largest, regularly updated curated repository for prokaryotic 16S rRNA gene sequences, and as such was used here.[182] Identification of unknown isolates is carried out by comparison of the 16S rRNA gene nucleotide sequence of unknown isolates to those of organisms contained within databases or repositories.

2.3.4.4 Limitations of 16S rRNA Gene Sequence Comparisons

Species identification with 16S rRNA gene sequence comparisons is not without its limitations:

(i) The hypervariable regions V1-V9 on the gene are not evenly distributed across its full length (Figure 2.2). Therefore, comparisons of full gene sequences (≈1500 bp) are likely to result in more accurate identification than comparisons of partial sequences (<1500 bp).[176,186,187] However, sequencing by Sanger method is only efficient up to 1000 bp, meaning the method will generally produce partial sequences of the gene.[178,186,187] Different segments of the gene can be sequenced and overlapped to produce the full sequence, but the process can be time consuming. Many authors have evaluated the accuracy of taxonomic placements based on partial 16S rRNA sequences, and the recommendation is to sequence the first 500 to 750 bp (spanning V1-V3 or V1-V4 hypervariable regions) as a minimum as these areas have more variable nucleotides and therefore provide more reliable preliminary identification.[178,186,187] Confirmation of identity will however necessitate comparison of full gene sequences.[186,187]

(ii) There is no consensus regarding the degree of sequence similarity required for taxonomic placements. Some researchers will assign isolates with ≥98% sequence similarity scores to the same species, while others adopt the higher benchmark of ≥99% similarity score.[176,178] Similarly, isolates can be

assigned to the same genus with sequence similarity scores ranging from 95% to 99%.[176,178]

(iii) Whilst correct genus identification is achievable in most cases (i.e. more than 90% of the time), species resolution is not always achievable with 16S rRNA gene sequence comparison as some species within the same genus have very similar 16S rRNA gene sequences. For instance, in the genus *Bacillus*, *B. globisporus* and *B. psychrophilus* share 99.5% similarity in their 16S rRNA gene sequences.[176]

Because of these limitations, 16S rRNA gene sequence comparisons result alone is not adequate for identity confirmation, but it is robust enough for preliminary identification of unknown bacterial isolates as most isolates will be correctly identified to the genus level.[176,178] It can however be used in combination with other molecular markers for identity confirmation.[176,178]

2.3.5 Selection of Potential Antibiotic-Producing Isolates

After preliminary identification of isolates, a search on NCBI's GenBank (microbial genome) database should give an indication of the typical genome size expected for each isolate as a 3 Mb genome size threshold was set for this work. GenBank database was chosen as it remains the largest and most widely used biological sequence database.[181,188] Thereafter, antiSMASH database[168] search by phylogeny should indicate if isolates are strains of known secondary metabolite producers or belong to genera of known secondary metabolite producers, and as such may have some capacity for secondary metabolism. The database contains precomputed BGCs of all complete and annotated bacterial genomes in the GenBank bacterial genome database.[168] It therefore facilitates comparative genomics and enables BGC distribution within taxonomic groups to be evaluated quickly.

2.4 Materials and Methods

2.4.1 Sample Collection

A topsoil sample was collected in a sterile 50 mL collection bottle at 9:30 a.m. on the 28th of June 2017, from a private garden in Aberdeen, Scotland (57.101 N 2.078 W). Sample was collected from the rhizosphere approximately 5 cm beneath

the garden surface. Air temperature at the time of collection was 14.5°C. Soil temperature was not taken.

2.4.2 Isolation of Nutritionally Versatile Bacterial Strains from Soil Sample

At approximately 5 hours after collection, the sample was weighed (8.2 g) and agitated gently in 40 mL of sterile distilled water for 1 minute in the collection bottle. Soil particles were left to settle for 5 minutes after which 100 µL aliquots of supernatant were added to three sterile universal bottles each containing 10 mL of 1:100 Ravan medium (prepared by adding 1 mL of sterile Ravan medium (Table 2.1) to 99 mL of sterile distilled water).[125] The contents of each bottle were thoroughly mixed after inoculation and the bottles were designated A1, A2, and A3. A1 was later incubated at 6°C, A2 at 17°C, and A3 at 37°C for 24 hours.

Table 2.1: Composition of Ravan Medium (pH 7.2).

Component	Quantity
Glucose (Fisher Scientific, UK)	5g
Peptone (Oxoid, UK)	5g
Yeast extract (Sigma-Aldrich, USA)	5g
Sodium citrate (Sigma-Aldrich, USA)	5g
Sodium acetate (Sigma-Aldrich, USA)	5g
Pyruvic acid (Sigma-Aldrich, USA)	2g
Distilled water	1000ml

At 24 hours, the change in turbidity of contents of all bottles were observed. To prevent overgrowth, A3 was transferred to the refrigerator after 24 hours as monitoring for growth at 37°C every 24 hours was not feasible at the time the experiments were carried out. A1 and A2 were kept at 6°C and 17°C, respectively. All bottles were then incubated for 6 more days. The change in turbidity of the contents of each bottle was observed at 7 days. On day 7, the contents of each bottle were thoroughly mixed and 100 µL aliquots of suspension from each bottle were aseptically spread over 1:100 Ravan medium plates with 1% agarose

(Bioline, UK) as gelling agent.[125] The plate corresponding to bottle A1 was incubated at 6°C, the one corresponding to A2 at 17°C, and the one corresponding to A3 at 37°C; plates were designated B1, B2, and B3, respectively. All plates were incubated for 7 days (plate B3 was transferred to the refrigerator after 24 hours and kept at 6°C for the rest of the incubation period). Plates were inspected for discrete colonies at 24 hours and at 7 days.

On day 7, four colonies with different morphology (i.e. different pigmentation, shape, or texture) were randomly selected from the plate with the most discrete and visible colonies after 7-day incubation (Plate B1). These colonies were aseptically streaked onto separate 1:100 Ravan medium plates for colony purification. All four plates were incubated at 6°C for 7 days. On day 7, a single colony from each plate was aseptically streaked onto fresh 1:100 Ravan medium plates and incubated at 6°C for 7 days to further purify the colonies. This was repeated one more time. The four pure isolates obtained therefrom were designated A23AA, A23BA, A13BB and A13AA.

2.4.3 Verification of Nutritional Versatility of Isolates

2.4.3.1 Cultivation in Liquid Media

For each isolate, single pure colonies were aseptically transferred from agarose plate into two sterile universal bottles each containing 10 mL 1:100 Ravan medium. Inoculated bottles were agitated gently after which one was incubated at 6°C and the other at 37°C for 24 hours. After 24 hours, the bottles initially incubated at 37°C were transferred to the refrigerator and all bottles were kept at 6°C for further 6 days. The turbidity of the contents of the bottles were observed at 24 hours and at 7 days. The same procedure was repeated using 10 mL 1:10 Ravan medium and 10 mL nutrient broth (Oxoid, UK).

2.4.3.2 Cultivation on Solid Media

For each isolate, single pure colonies were aseptically streaked onto two 1:10 Ravan medium plates. Thereafter one plate was incubated at 6°C and the other at 37°C, for 7 days (the plates initially incubated at 37°C were transferred to the refrigerator after 24 hours and kept at 6°C for the rest of the incubation period).

Plates were inspected for growth at 24 hours and at 7 days. The procedure was repeated with nutrient agar (Oxoid, UK) plates.

2.4.4 Preliminary Identification of Isolates by 16S rRNA Gene Sequence Comparisons

2.4.4.1 Extraction of Genomic DNA (gDNA) from Isolates

The DNeasy® Ultraclean® Microbial Kit for DNA isolation (Qiagen, UK) was used with modified protocol to lyse and extract gDNA from isolates. gDNA was also extracted from *E. coli* NCTC 4171 and *S. aureus* NCTC 6571 (positive controls for Gram-negative and Gram-positive bacteria, respectively). Isolates were cultured in nutrient broth and incubated at 6°C for 7 days, while control strains were cultured in nutrient broth and incubated at 37°C for 24 hours before cell lysis. All isolates and control strains were cultured in duplicate.

Three millilitre aliquots of each 7-day isolate culture were transferred into sterile 15 mL centrifuge tubes and centrifuged at 4,200 rpm for 5 minutes. Thereafter, 1.2 mL of supernatant was removed from each tube and discarded, and pellet re-suspended by gently vortexing the tube. The contents of each tube (1.8 mL) were then transferred into a 2 mL collection tube (Qiagen, UK). Likewise, 1.8 mL of overnight culture of control strains were transferred into 2 mL collection tubes. These tubes, along with those containing isolates were centrifuged at 10,000 rpm for 5 minutes. Supernatants were completely removed from all tubes, after which pellets were re-suspended in 300 µL of PowerBead solution (salts and buffer solution that stabilises and homogeneously disperses cells before lysis). The contents of the collection tubes were gently vortexed and transferred to separate PowerBead tubes (contain beads to mechanically lyse cells). Fifty microlitre of solution SL (solution containing sodium dodecyl sulphate and other disruption agents to breakdown cell wall and denature proteins) was added to each PowerBead tube before incubating the tubes at 70°C for 10 minutes in a water bath. The contents of the PowerBead tubes were later vortexed vigorously for 2 minutes before centrifuging at 10,000 rpm for 3 minutes. Supernatant was removed from each PowerBead tube and transferred into a clean 2 mL collection tube before discarding PowerBead tube. Thereafter, 100 µL of solution IRS (reagent to precipitate non-DNA materials, proteins, and cell debris) was added to each collection tube before vortexing the contents for 5 seconds and incubating

at 6°C for 5 minutes. The contents of the tubes were later centrifuged at 10,000 rpm for 2 minutes and the supernatants transferred into clean 2 mL collection tubes. Nine hundred microlitre of solution SB (highly concentrated salt solution to bind DNA to spin column) was added to each collection tube and the entire contents of each tube were loaded onto MB spin columns in two equal portions, centrifuging at 10,000 rpm for 30 seconds after each portion. The flowthroughs were discarded (DNA is now bound to spin columns). Thereafter, 300 µL of solution CB (ethanol wash solution to clean bound DNA by removing residues of salts and other contaminants) was added to each spin column before centrifuging at 10,000 rpm for 1 minute. The flowthroughs were discarded. The spin columns were centrifuged again at 10,000 rpm for 1 minute to remove all traces of ethanol as ethanol can interfere with downstream DNA applications.

Finally, 50 µL of solution EB (eluting buffer- 10 mM Tris pH 8; unbinds DNA from the columns) was added to the centre of each spin column placed in a clean 2 mL collection tube and centrifuged at 10,000 rpm for 1 minute to elute extracted gDNA. Eluted gDNA from each isolate and control strains were stored in the freezer at -20°C in 5 µL aliquots for further analysis and downstream applications.

2.4.4.2 Confirmation of the Presence and Quality of gDNA in Eluted Extracts from Isolates and Control Strains

The presence and quality of gDNA in extracts obtained as per Section 2.4.4.1. (referred to hereafter as samples) were confirmed by band visualisation after electrophoresis on a 1% (w/v) agarose gel as detailed below:

A 1% (w/v) agarose gel was prepared by adding 2 g of agarose powder (Bioline, UK) to 200 mL 1 X TBE buffer (Sigma-Aldrich, UK), and heating up in a Toshiba microwave oven (EC042A5C-SS) on 800 watts for 3 minutes to dissolve the powder. The gel solution was left to cool before adding 20 µL of GelRed® (Biotium, USA) to aid visualisation of DNA under UV light. The solution was then poured into a sealed gel cast containing a 16-well comb. The gel was left to set for approximately 30 minutes before removing the seals. The electrophoresis tank was filled with 1 X TBE buffer to the fill mark after which the comb was removed. For each sample, 5 µL of extracted gDNA was mixed with 2 µL of 5 X loading buffer (Bioline, UK) before loading into a well in the gel. Five microlitre of 1kb HyperLadder™ (Bioline, UK) was also loaded into a separate well as reference

standard. Electrophoresis was carried out at 110 V for 90 minutes, after which the gel was visualised and photographed under UV light using the Biomolecular Imaging System (Uvitec, UK) and the Digital Graphic Printer (Sony, UK).

2.4.4.3 Amplification of the 16S rRNA Gene Segment of Samples by Polymerase Chain Reaction

The entire length of the 16S rRNA gene was amplified with the 27F and U1510R universal primers (primer characteristics as outlined Appendix 2). PCR reaction mixtures were prepared in sterile PCR tubes on ice according to the formula in Table 2.2. PCR tubes were loaded onto the GeneAmp® PCR System 9700 (Applied Biosystems, USA) and PCR was undertaken according to the following protocol: Initial denaturation at 95°C for 2 minutes followed by 30 cycles of further denaturation at 95°C for 30 seconds, primer annealing at 45°C for 30 seconds and elongation at 72°C for 105 seconds. A final elongation was carried out at 70°C for 5 minutes.

Table 2.2: Components of PCR Reaction Mixtures.

Component	Volume (µL)
GoTaq® G2 Colourless Master Mix (Promega, USA)	25
27F primer (10 µM)	5
U1510R primer (10 µM)	5
Template (extracted gDNA sample; ≈20 ng/µL)	5
Nuclease free water (Promega, USA)	10

2.4.4.4 Purification of Products of Polymerase Chain Reaction

Excess primers and nucleotides were removed from PCR products with the Wizard® SV Gel and PCR Clean-Up System (Promega, USA) as detailed below:

Forty microlitre of each PCR product was added to 40 µL of membrane binding solution in a sterile 1.5 mL microcentrifuge tube and thoroughly mixed. The mixture was transferred onto an SV mini column inserted into a collection tube and incubated for 1 minute at room temperature before centrifuging at 13,500

rpm for a further 1 minute. The flowthrough was discarded while the mini column was re-inserted into the collection tube and washed with 700 µL of membrane wash solution (diluted with 95% ethanol), then centrifuging at 13,500 rpm for 1 minute. The flowthrough was discarded, and the process was repeated with 500 µL of membrane wash solution and centrifuging at 13,500 rpm for 5 minutes. The collection tube was emptied, and the mini column re-centrifuged for a further 5 minutes to remove residual ethanol. The column was inserted into a sterile 1.5 mL microcentrifuge tube and bound PCR product was eluted by adding 50 µL of nuclease free water onto the column and incubating for 1 minute at room temperature before centrifuging at 13,500 rpm for 1 minute. Cleaned PCR products were stored at -20°C in aliquots of 5 µL.

2.4.4.5 Preparation of Products of Polymerase Chain Reaction for Sequencing

The concentration of DNA in each cleaned PCR product was estimated after electrophoresis on a 1% (w/v) agarose gel (as described in Section 2.4.4.2) by comparing band intensity of PCR product to those of the Hyperladder™ reference standard (Figure 2.5). DNA concentration of each cleaned product was approximately 16 ng/µL. This was adjusted by diluting 10 µL of cleaned PCR product with 20 µL of sterile nuclease free water to give approximately 160ng of

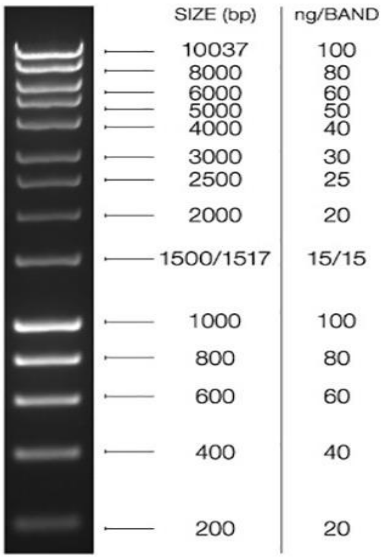


Figure 2.5: 1kb Hyperladder™.
Image source: Bioline, UK.

DNA in 30 µL of solution prior to shipping to sequencing facility. 27F primer was used for sequencing reactions and was diluted with sterile nuclease free water to give a working concentration of 3.2 µM before shipping. DNA sequencing was performed by DNA Sequencing & Services (School of Life Sciences, University of Dundee, Scotland, www.dnaseq.co.uk) using Applied Biosystems Big-Dye v3.1 chemistry on an Applied Biosystems 3730 capillary DNA sequencer.

2.4.4.6 Analysis of Sequencing Data

Text and trace data were viewed and edited with Chromas v2.6.5 before database search. Database was searched in April 2018 with the following criteria:

URL: http://www.ncbi.nlm.nih.gov/blast/Blast.cgi

Search type: Nucleotide vs nucleotide database (blastN)

Database: 16S Reference RNA sequences

Database ID: refseq_rna

Limited by Entrez search: Bacteria (ORGN)

The start point of analysis was set at position 70 on the chromatogram given that 27F primer anneals to the template from positions 8 to 27 (Appendix 1), and nucleotide sequences immediately adjacent to primer binding sites were unreliable as evidenced on trace data. Also, as sequencing reactions utilise a single primer, elongation efficiency diminishes with sequence length. Hence, the endpoint of analysis was set at position 650 which is upstream of the point of peak deterioration as observed on trace data. The first three species with highest identity scores returned by database search were included in the BLAST search results (Section 2.5.6).

2.4.5 Selection of Potential Antibiotic-Producing Isolates

After preliminary identification, the average genome size of strains of identified species were obtained from NCBI's GenBank microbial genome database. Database was accessed in July 2018.

Thereafter, the antiSMASH database was browsed by phylogeny down to the genus level to determine if isolates belong to genera of secondary metabolite

producers. The average number of BGCs within the genomes of database entries were noted, as were the average number of BGCs associated with antibiotic production (if any). Database was accessed in October 2018.

2.5 RESULTS

2.5.1 Isolation of Bacterial Strains from Topsoil Sample

2.5.1.1 Change in Turbidity of the Contents of Bottles A1, A2 and A3 at 24 hours and at 7 Days

The ultra-minimal substrate medium (1:100 Ravan medium) used for strain isolation supported the growth of nutritionally versatile bacterial strains as shown in Table 2.3, with increase in turbidity of contents of bottles A1 and A3 (incubated at 6°C and 37°C, respectively) at the end of the 7-day incubation period. However, contents of A2 that was incubated at 17°C throughout the incubation period did not show a noticeable change in turbidity, perhaps because incubating at 17°C did not support optimal growth of recovered strains.

Table 2.3: Change in Turbidity of Culture Media at Different Time Intervals.

	Turbidity		
Time	**A1**	**A2**	**A3**
0 hours	**+**	**+**	**+**
24 hours	**+**	**+**	**++**
7 Days	**++**	**+**	**+++**

Growth Key: - = no turbidity
+ = slight turbidity
++ = moderate turbidity
+++= considerable turbidity
++++ = extensive turbidity

2.5.1.2 Appearance of Plates B1, B2, and B3 at 24 Hours and at 7 Days.

On cultivation of 100 µL of contents of A1, A2, and A3 at 7 days on solid media (1:100 Ravan medium plates) and incubation for a maximum of 7 days, plate B1 (corresponding to bottle A1) which was incubated at 6°C for 7 days produced the most discrete colonies (Figure 2.7).

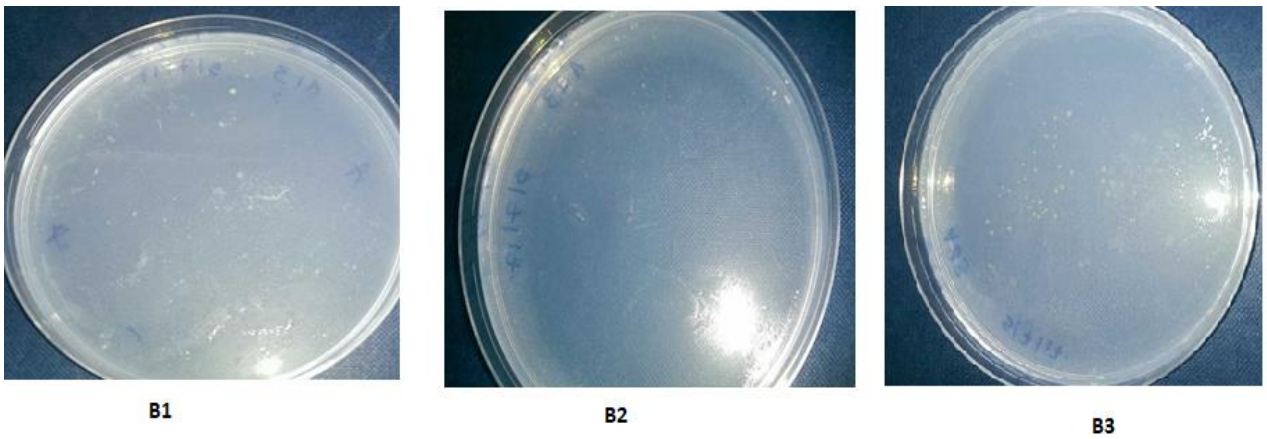


Figure 2.6: Plates B1, B2, and B3 at 24 hours.
At 24 hours, there were no visible colonies on plates B1 and B2 after incubation at 6°C and 17°C, respectively. Few pinpoint colonies were visible on plate B3 that was incubated at 37°C.

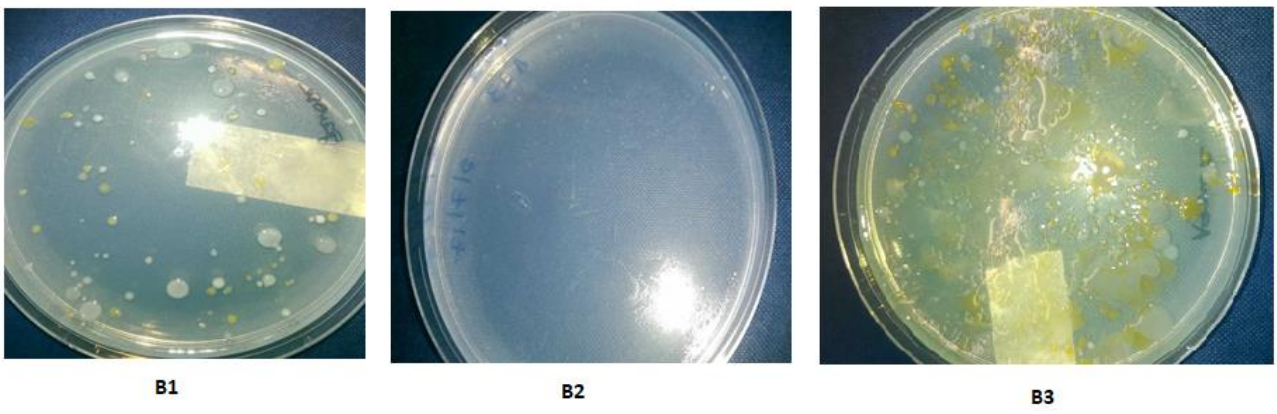


Figure 2.7: Plates B1, B2, and B3 at 7 days
At 7 days, there were no visible colonies on plate B2 incubated at 17°C. Also, there were no discrete colonies on plate B3 incubated at 37°C for 24 hours, and then kept at 6°C for 6 more days. Plate B1 incubated at 6°C for the entire incubation period produced the most discrete colonies.

2.5.2 Verification of Nutritional Versatility of Isolated Strains

Tables 2.4 and 2.5 show all four selected isolates to be nutritionally versatile as growth occurred in/on complex media with varying nutrient concentration and at different incubation temperatures. Isolates A23AA and A23BA appeared to have similar growth patterns, while A13BB and A13AA exhibited different growth patterns.

Table 2.4: Growth Pattern of Isolates at 24 hours

	1:100 Ravan medium broth		1:10 Ravan medium broth		Nutrient broth		1:10 Ravan medium plate		Nutrient agar plate	
Isolate	**6°C**	**37°C**	**6°C**	**37°C**	**6°C**	**37°C**	**6°C**	**37°C**	**6°C**	**37°C**
A23AA	-	+	-	++	-	+++	-	++	-	+++
A23BA	-	+	-	++	-	+++	-	++	-	+++
A13BB	-	+	+	+	+	++	+	+	+	++
A13AA	-	+	-	+	-	++	-	+	-	++

Table 2.5: Growth Pattern of Isolates at 7 days

	1:100 Ravan medium broth		1:10 Ravan medium broth		Nutrient broth		1:10 Ravan medium plate		Nutrient agar plate	
Isolate	**6°C**	**37°C**	**6°C**	**37°C**	**6°C**	**37°C**	**6°C**	**37°C**	**6°C**	**37°C**
A23AA	+	+	+	++	+++	+++	+	++	+++	+++
A23BA	+	+	+	++	+++	+++	+	++	+++	+++
A13BB	++	++	++	+++	++++	++++	++	+++	++++	++++
A13AA	+	+	+	+	++	++	+	+	++	++

Growth Key: - = No growth
+ = slight growth
+ + = moderate growth
+ + + = considerable growth
++++ = extensive growth

2.5.3 Confirmation of the Presence and Quality of gDNA in Extracts from Isolates and Control Strains

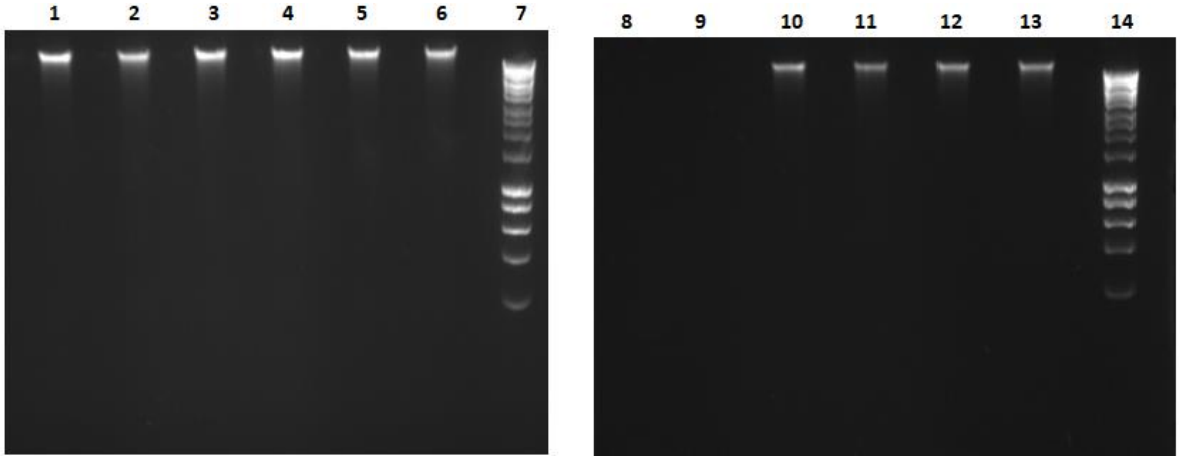


Figure 2.8: gDNA in extracts as visualised on agarose gel after electrophoresis.
E. coli (lanes 1&2), A13BB (lanes 3&4), A23BA (lanes 5&6), *S. aureus* (lanes 8&9), A13AA (lanes 10&11), A23AA (lanes 12&13), 1kb Ladder (lanes 7&14).

Extraction of high-quality gDNA from isolates and *E. coli* was confirmed after gel electrophoresis by the presence of fluorescent bands in respective wells (Figure 2.8), with no smearing. However, extraction was not successful in the case of *S. aureus*, hence the absence of fluorescent bands in wells 8 and 9.

2.5.4 Estimation of Purity and DNA Concentration of PCR Products

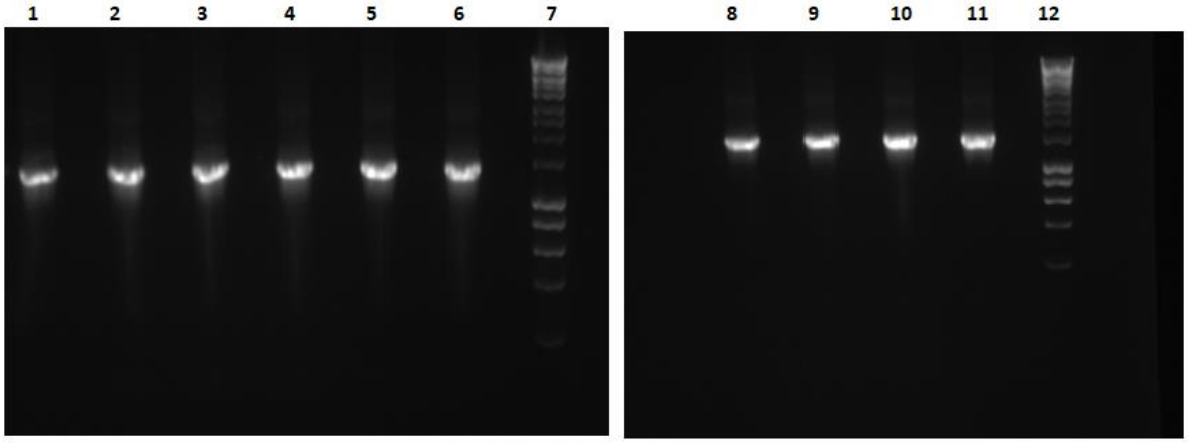


Figure 2.9: PCR products of extracts as visualised on agarose gel after electrophoresis.
E. coli (lanes 1&2), A13BB (lanes 3&4), A23BA (lanes 5&6), A13AA (lanes 8&9), A23AA (lanes 10&11), 1kb Ladder (lanes 7&12).

PCR products were of good quality as respective wells contained single fluorescent bands of expected size i.e. 1500 bp as compared with 1Kb Ladder (Figure 2.9). DNA concentration in each PCR product was approximately 16 ng/µL as estimated from band intensity comparisons between PCR products and 1Kb Ladder.

2.5.5 Edited Sequencing Data

PCR reactions copied entire length (i.e. ≈1500 bp) of the 16S rRNA gene in control strain and isolates. Entire 1500 bp lengths were sequenced but sequencing data were edited to remove regions of poor peak resolution leaving regions of well resolved peaks which were ≈580 bp in length, with sections shown in Figure 2.10.

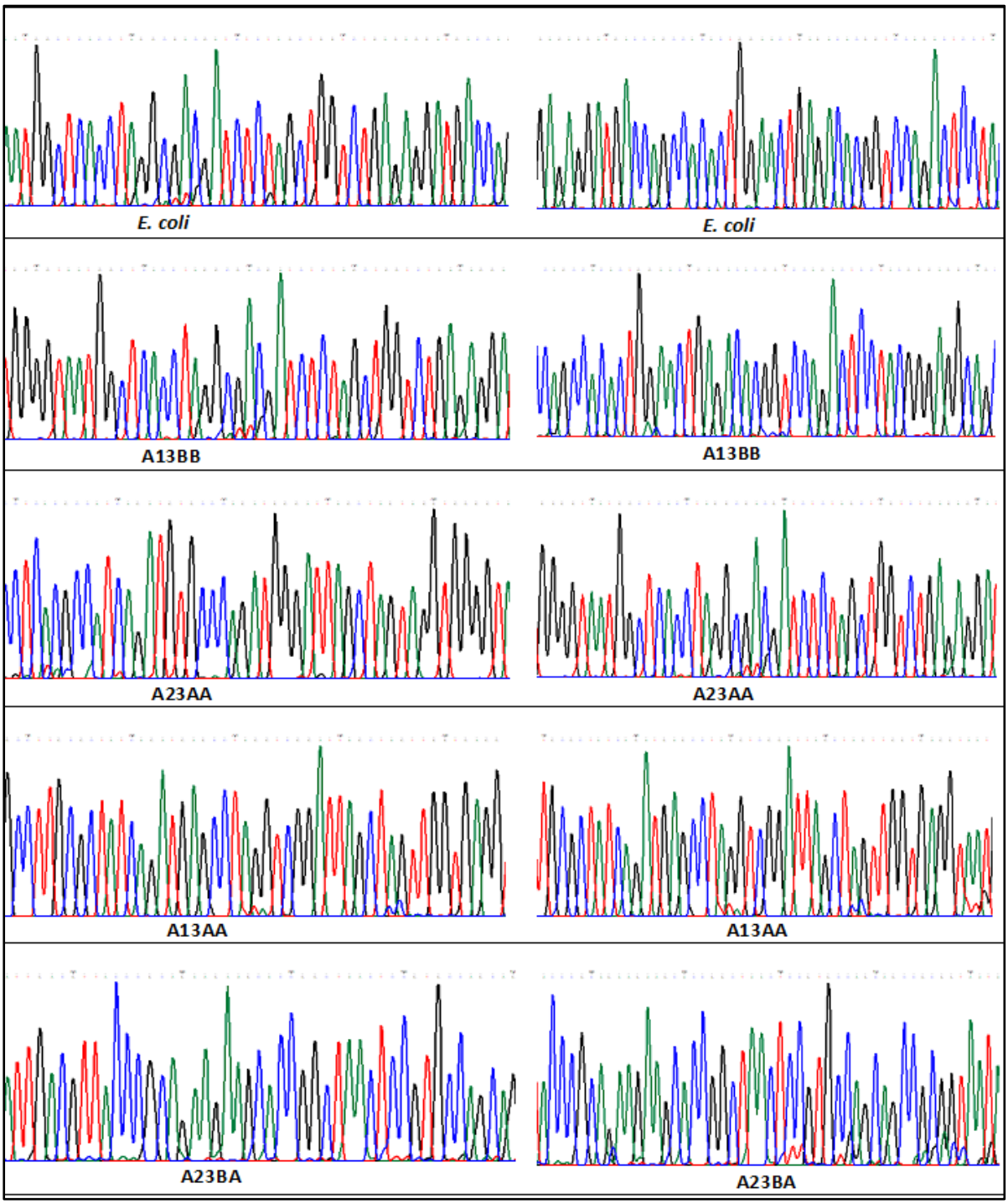


Figure 2. 10: Sections of edited chromatograms obtained for isolates and control strain. Colour key: **G**, **A**, **T**, **C**

2.5.6 BLAST Search Results

Table 2.6: BLAST Search Results.

Isolate/ Control	Description[1]	Identity Score[2]	Query Coverage[3]	E-value[4]
E. coli	*Escherichia fergusonii* strain	99%	100%	0.0
	Escherichia coli strain	99%	100%	0.0
	Escherichia coli strain	99%	100%	0.0
E. coli	*Escherichia fergusonii* strain	99%	100%	0.0
	Escherichia coli strain	99%	100%	0.0
	Escherichia coli strain	99%	100%	0.0
A13BB	*Pseudomonas fragi* strain	99%	100%	0.0
	Pseudomonas fragi strain	99%	100%	0.0
	Pseudomonas psychrophila strain	99%	100%	0.0
A13BB	*Pseudomonas fragi* strain	99%	100%	0.0
	Pseudomonas fragi strain	99%	100%	0.0
	Pseudomonas psychrophila strain	99%	100%	0.0
A23AA	*Hafnia alvei* strain	99%	100%	0.0
	Obesumbacterium proteus strain	99%	100%	0.0
	Hafnia alvei strain	99%	100%	0.0
A23AA	*Hafnia alvei* strain	99%	100%	0.0
	Obesumbacterium proteus strain	99%	100%	0.0
	Hafnia alvei strain	99%	100%	0.0
A13AA	*Hafnia alvei* strain	99%	100%	0.0
	Obesumbacterium proteus strain	99%	100%	0.0
	Hafnia alvei strain	99%	100%	0.0
A13AA	*Hafnia alvei* strain	99%	100%	0.0
	Obesumbacterium proteus strain	99%	100%	0.0
	Hafnia alvei strain	99%	100%	0.0
A23BA	*Hafnia alvei* strain	99%	100%	0.0
	Obesumbacterium proteus strain	99%	100%	0.0
	Hafnia alvei strain	99%	100%	0.0
A23BA	*Hafnia alvei* strain	99%	100%	0.0
	Obesumbacterium proteus strain	99%	100%	0.0
	Hafnia alvei strain	99%	100%	0.0

Key: 1- identifiers for sequences similar to the query sequence in GenBank database.
2- the extent to which query sequence and database sequence have the same nucleotide at the same positions in the alignment generated by BLAST.
3- the percentage of query sequence aligned to database sequence.
4- describes the number of matches expected to happen purely by chance.

There were no species resolution for the control strain and isolates with the partial length 16S rRNA gene (≈580 bp) used for BLAST analysis as more than one species in the database showed ≥99% identity to query sequence in all instances as shown in Table 2.6 above.

2.5.7 Overview of Biosynthetic Capacities of Identified Species

The GenBank database search gave an indication of the typical genome size expected for strains of all four species identified by BLAST analysis. All strains of these species in the GenBank database have genomes larger than the 3 Mb minimum genome size set for this bioprospecting work (Table 2.7).

Table 2.7: Results of GenBank Microbial Genome Database Search

Identified species	Number of strains in database	Genome size range (Mb)	Average genome size (Mb)	Number of complete genomes
Pseudomonas fragi	19	4.97-5.22	5.09	2
Pseudomonas psychrophila	5	5.15-5.33	5.26	0
Hafnia alvei	13	4.47-4.97	4.75	3
Obesumbacterium proteus	2	4.67-5.01	4.84	1

Of all three genera identified by BLAST analysis, species belonging to the *Pseudomonas* genus appear to encode the most BGCs in the antiSMASH database. Consequently, *Pseudomonas* species also appear to encode the most BGCs associated with antibiotic production (Table 2.8).

Table 2.8: Results of antiSMASH Database Search by Genera

	Pseudomonas	***Hafnia***	***Obesumbacterium***
Number of entries in database	1,236	14	1
Average number of BGCs per strain	11	3	2
Average number of BGCs associated with antibiotic production	7	2	0

2.6. Discussion

Bacterial secondary metabolites constitute a relatively untapped natural resource in the development of novel antibiotics. The ability to biosynthesise these metabolites is not universal to all bacterial species, therefore isolation strategies that are selective for potential secondary metabolite producers are desirable. A notable attribute of some antibiotic-producing bacteria is the ability to withstand starvation stress;[120,175] therefore, an ultra-minimal substrate medium was used to isolate nutritionally versatile bacterial strains from a nutrient-dense topsoil sample. Furthermore, sample was collected from the rhizosphere to increase the likelihood of isolating antibiotic-producing bacterial species as antibiosis occurs naturally within the niche.[173]

Nutritionally versatile bacterial strains were isolated as described above using 1:100 Ravan media. Incubation of seeded plates at 6°C for up to 7 days produced at least 65 discrete colonies (Plate B1; Figure 2.7) out of which four colonies with different morphology were randomly selected and purified, and later designated as isolates A23AA, A23BA, A13BB, and A13AA. All four isolates were shown to be nutritionally versatile with the ability to grow in/on complex culture media with different nutrient concentrations (Tables 2.4 and 2.5).

16S rRNA gene sequence comparison was used to preliminarily identify isolates. This was preceded by gDNA extraction and purification. The extraction protocol used is more suited to Gram-negative bacteria, while further modification will be required for Gram-positive strains as extraction was not successful with *S. aureus* i.e. no visible bands corresponding to extracted gDNA on agarose gel after electrophoresis (lanes 8 and 9; Figure 2.8). Perhaps the mechanical lysis stage can be extended from 2 minutes to 10 minutes for Gram-positive bacteria, but care should be taken to avoid excessive DNA shearing. Where gDNA extraction was successful, DNA extracts obtained were of good quality and contained no other cellular macromolecules as single distinct bands were obtained for each extract under UV light (Figure 2.8). Extracts were above 10,000 bp in size which showed gDNA were not excessively sheared.

27F and U1510R universal primers were used to amplify the 16S rRNA gene segment of the gDNA extracts in PCR reactions. The entire length of the gene was

amplified in all samples with amplicon size of ≈1500 bp (Figure 2.9). The single bands visualised under UV light showed that each reaction produced a single PCR product which is crucial to the success of subsequent gene sequencing. The DNA concentration of PCR products was estimated to be 16 ng/µL. 27F primer was used for sequencing reactions, copying the entire length of the 16S rRNA gene. The trace and text data obtained from sequencing reactions were edited to remove regions of poor peak resolution leaving well separated peaks with little or no background interference, and no indeterminate nucleotides (Figure 2.10). These regions spanned positions 70 and 650 (≈580 bp in length) on the gene covering the V1-V4 hypervariable regions (Figure 2.2). Edited sequence data were later used for BLAST analysis with isolates and control strain (*E. coli*) preliminarily identified as per Table 2.6. The first three organisms with identity scores of ≥99% in the BLAST reports were included in the description column.

Control strain and all unknown isolates were identified to the genus level (identity scores >95%). Although it appeared from the BLAST results that isolates A23AA, A13AA and A23BA could be species from two different genera - *Hafnia* or *Obesumbacterium.* In fact, *Obesumbacterium proteus* is the only species in its genus but strains are grouped into bio-group 1 and bio-group 2.[189] Strains in *Obesumbacterium* bio-group 1 have been shown to be genetically more closely related to *Hafnia* species than to strains in bio-group 2.[189] Therefore, these two genera may represent an example of the few instances where 16S rRNA gene comparisons fail to correctly identify an organism to the genus level.[176,178] This may arise if sequence homology between the genomes of species in the genera in question are extremely high.[178] Furthermore, there were no resolutions between species (more than one species in the database showing identity scores of ≥99% to queries) in the BLAST results. Using full sequence of the 16S rRNA gene of each isolate for comparison may result in better resolution between species. This can be achieved by sequencing partial segments of the gene by Sanger method and overlapping all segments with bioinformatics tools to give the full-length sequence. Alternatively, full gene sequence can be obtained with next-generation sequencing techniques that can generate longer sequencing reads, but these are usually more expensive. Species resolution at this stage of the project is however not compulsory provided all matches to queries have genome sizes typically greater than 3 Mb. A13BB was preliminarily identified as a *Pseudomonas* species

whereas A23AA, A23BA and A13AA could be either *Hafnia* or *Obesumbacterium* species. Whilst A23AA and A23BA may be strains of the same species as they both have the same growth characteristics in different culture media (Tables 2.4 and 2.5), A13AA is most likely a different but closely related species as its growth characteristics are quite different. In essence at least three nutritionally versatile bacterial species were selected.

Strains of all identified species typically have genomes larger than 3 Mb (Table 2.7), hence isolates may have increased coding capacity for secondary metabolites and are therefore potential secondary metabolite producers.[127] Isolate A13BB more so than others as it may have the largest genome being a *Pseudomonas* species. As of October 2018, there were 1,236 entries for members of the *Pseudomonas* genus in the antiSMASH database (Table 2.8). Each entry had an average of 11 BGCs of which an average of 7 were potentially antibiotic-encoding. These included PKs, bacteriocins, and RiPPs. It was interesting to note that *Pseudomonas aeruginosa* accounted for 26% (n=320) of all entries, followed by *Pseudomonas syringae* which accounted for 7% (n=87). Both species are common pathogens, supporting the argument that the set of sequenced bacterial genomes available in various databases is biased towards pathogenic organisms, the greater proportion of which are not secondary metabolite producers.[127] There is clearly a need to address this bias. *Pseudomonas* species however are good secondary metabolite producers with the most popular species used for biocontrol purposes because of their ability to biosynthesise antimicrobials being *P. fluorescens* and *P. chlororaphis.*[190,191] Isolate A13BB is preliminarily identified as *P. fragi* or *P. psychrophlia* (Table 2.6) both of which are not readily associated with antibiotic production but are genetically very similar to *P. chlororaphis.*[192]
There were very few entries for members of the *Hafnia* and *Obesumbaterium* genera in the antiSMASH database as of October 2018 (Table 2.8). This may reflect the fact that these genera are relatively unresearched. Entries for both genera had fewer or no antibiotic-encoding BGCs within their genomes compared with *Pseudomonas*. Again, this may be because these genera are unresearched. Moreover, the typical genome size of species within these genera (Table 2.7) would suggest a greater capacity to encode secondary metabolites than reported in the antiSMASH database. Members of *Hafnia* and *Obesumbacterium* genera are

also not readily associated with antibiotic production as evidence suggesting otherwise was not found in literature.

In summary, at least three nutritionally versatile bacterial species were isolated from the topsoil sample. Isolate A13BB could be a strain of *P. fragi* or *P. psychrophila*, while isolates A23AA, A13AA, and A23BA could be strains of *H. alvei* or *O. proteus*. With all isolates likely to have genomes larger than 3 Mb, they may have the potential to biosynthesise secondary metabolites; isolate A13BB more so than others. The fact that these isolates were recovered from the rhizosphere also makes them potential antibiotic producers. Another advantage of a genome larger than 3 Mb is the increased capacity to accommodate genes acquired by HGT into non-core or accessory regions of the genome as the size of the non-core region increases with genome size.[170] Consequently, these isolates may possess unique (i.e. strain-specific) antibiotic-encoding BGCs which were acquired by HGT from other antibiotic producers within the rhizosphere.

Isolates A23AA, A13AA, and A23BA appeared to show less potential as antibiotic-producers compared to A13BB, but their biosynthetic potentials may be understated as they belong to relatively unresearched genera. Consequently, all isolates were considered good candidates for genome mining for potential antibiotic-encoding BGCs given their expected genome size and BGC distribution, ability to withstand starvation stress, and ecological origin. Furthermore, as none of the isolates are readily associated with antibiotic production, mining their genomes for antibiotic-encoding BGCs may unveil unique chemistries and novel natural products that may serve as lead compounds for developing new antibiotics.

2.7. Conclusions

All selected isolates recovered using an ultra-minimal substrate medium are nutritionally versatile and were able to withstand starvation stress, which is a key attribute of some antibiotic producers. All isolates were genuine candidates for genome mining for potential antibiotic-encoding BGCs. A13BB, a *Pseudomonas* species, may have the greatest potential to biosynthesise antibiotics as species within its genus have on average 7 antibiotic-encoding BGCs compared to 2 or 0 for strains belonging to *Hafnia* or *Obescumbacterium* genus, respectively. All

isolates may possess BGCs acquired through VGT and HGT within their genomes, some of which may encode novel metabolites that can be exploited as antibiotics. A23AA, A13AA and A23BA, which may be *Hafnia* and/or *Obescumbacterium* species, hold additional potential to encode novel bioactive compounds because the biosynthetic capacities of members of their genera are relatively unresearched. Although all isolates were suitable candidates for genome mining, it is highly likely that the BGC distribution in isolates A23AA, A13AA, and A23BA will be very similar being members of the same genus (or highly genetically similar genera), and from the same ecological origin. Therefore, whole-genome sequence data of isolates A13BB and A23BA were obtained for genome mining for potential antibiotic-encoding BGCs in the first instance.

In addition to expediting this bioprospecting work, obtaining the complete genome sequence data of these isolates, and depositing the data in repositories of the International Nucleotide Sequence Database Collaboration (comprising the European Nucleotide Archive, the DNA Data Bank of Japan, and GenBank) will serve to address the need to archive sequence data of bacterial genomes that can be mined for novel bioactive natural compounds in publicly accessible databases.

CHAPTER THREE

WHOLE-GENOME SEQUENCING AND ASSEMBLY OF ISOLATES A23BA AND A13BB.

3. Whole-genome Sequencing and Assembly of Isolates A23BA and A13BB

Next-generation sequencing techniques have revolutionised genome research by providing affordable and rapid means of generating vast amounts of sequencing data. Accurate and contiguous bacterial genomes that are essential for effective genome mining can be assembled *de novo* from these data, using freely accessible robust and validated tools and pipelines.

3.1 Introduction

An organism's genome is its complete set of genetic information stored as nucleotide sequences in RNA or DNA molecules. The exact order of nucleotides in single-stranded RNA molecules were the first sequencing projects to be undertaken in the 1960s.[193,194] First-generation sequencing techniques involved cleavage of RNA strands with RNase enzymes into fragments that were separated by chromatography and electrophoresis. Nucleotide sequence in the fragments were then deduced by sequential exonuclease digestion. The first RNA sequencing project comprising just 76 nucleotides took three years to complete reflecting the complexities of earlier techniques.[193]

DNA sequencing started in the late 1960s to early 1970s when some researchers copied short DNA molecules into RNA and then sequenced the latter to deduce the nucleotide sequence in the original DNA strands.[195-197] Others used primer extension methods involving DNA polymerases and labelled oligonucleotides to decipher nucleotide sequences of short DNA strands.[195-197] All these methods were still very laborious, and it was not until 1977 when Frederick Sanger and co-workers developed the chain termination or dideoxy method that DNA sequencing was truly transformed.[198] As the method was suited to automation, it became possible, for the first time, to sequence DNA fragments up to 1000 bp long in a matter of hours. Furthermore, shotgun Sanger sequencing enabled the nucleotide sequences of many microbial genomes to be determined in a reasonable length of time.[195,199-201] However, there is a ceiling to sequencing speed with increasing genome size, which meant that the Human Genome Project undertaken with so called hierarchical shotgun Sanger sequencing technique took thirteen years to complete at a cost of ≈ $3 billion.[202]

Limitations of shotgun Sanger sequencing were overcome by next-generation (i.e. second- and third-generation) sequencing techniques. The initial second-generation technique was pyrosequencing developed by 454 Life Sciences. It involves DNA amplification and elongation reactions using unlabelled nucleotides, with base calling inferred from lights emitted during pyrophosphates formation.[195,203] The genius of the method is its enablement of high throughput 'massively parallel' sequencing reactions that generate vast amounts of 400-500 bp reads that are assembled *in silico* in a relatively short time, such that Wheeler *et al* used pyrosequencing to re-sequence the human genome in two months at a cost of $1,500,000.[195,204,205] Other second-generation sequencing techniques (e.g. SOLiD and Ion Torrent) were introduced, but with only marginal benefits over pyrosequencing.[195-197] However, Illumina (previously Solexa) sequencing platform would take second-generation sequencing to the next level. It utilises DNA polymerase and fluorescent nucleotides for polymerisation reactions preceded by rapid *in situ* 'bridge amplification' of template DNA fragments ligated to oligonucleotide adapters.[195,205,206] As sequencing also proceeds in massively parallel fashion, high throughput, highly accurate paired-end reads ranging from 35 bp to 300 bp are generated at massively reduced costs and time.[195,205,206]

Assembling short Illumina reads into contiguous genome sequences can however be challenging.[207,208] The challenge is addressed by third-generation single molecule sequencing (SMS) techniques that generate long reads in excess of 10 kb.[195-197] Single molecule real time (SMRT) sequencing was the first SMS technique introduced and marketed by Pacific Biosciences. It utilises a single molecule of DNA polymerase and nucleotides with different fluorophores to copy native DNA fragments inside arrays of nanoholes in massively parallel reactions with base calling achieved in real time.[195,209] Oxford nanopore sequencing (ONS) technique marketed by Oxford Nanopore Technologies (ONT) was later introduced.[210,211] Unlike all other platforms, ONS is not a sequence by synthesis (SBS) technique involving polymerisation reactions. Instead, it relies on the change in ionic current observed as single molecules of native single-stranded DNA or RNA pass through biological or solid-state nanopores.[195,210,211] Sequencing proceeds in massively parallel fashion in thousands of nanopores generating vast amounts of ultra-long reads faster than ever.[210,212] Hybrid assembly of Illumina and SMS reads have been shown to produce the most accurate and contiguous bacterial genomes.[208]

3.2: Aim and Objectives

The aim of this section of the project was to assemble accurate and contiguous genomes of isolates A23BA and A13BB. This can be achieved by:

(i) Outsourcing DNA sequencing reactions to commercial sequencing facilities to obtain short and long sequencing reads for each isolate.
(ii) Performing quality control checks and trimming of reads with state-of-the-art bioinformatics tools to ensure accurate and high-quality reads are assembled.
(iii) Performing *de novo* hybrid genome assembly using high-quality short and long sequencing reads.
(iv) Assessing the quality and completeness of assemblies generated using state-of-the-art tools and pipelines.

3.3 Methodology

3.3.1 Selection of Sequencing Platforms/Techniques

BGCs can span thousands of adjacent nucleotides in microbial genomes. Therefore, an effective genome mining exercise would be contingent on the presence of unfragmented BGCs in accurately assembled contiguous genomes. Illumina remains the most cost effective and accurate short read sequencing platform, especially with the ability to generate paired-end short reads which further increases accuracy to >99%.[205,208,213] Although highly accurate Illumina short reads can be assembled *de novo* and mapped to high-quality reference genomes, it is often the case that repeat-rich regions (common feature within BGCs), homopolymers, and regions of high GC contents are poorly resolved. Therefore, generating a contiguous genome sequence using short reads alone is challenging, time consuming and costly.[208,213] Conversely, long reads generated by SMS platforms have overcome many of these limitations and are more likely to generate contiguous genomes when used alone, however their base call accuracy is usually lower compared with Illumina reads.[208,212] Whilst base call error rates of SMRT and ONS platforms were historically greater than 10%, many improvements have been made in recent years to achieve <1% and <5% error rates, respectively.[208,212,214] Although the latest SMRT High Fidelity (HiFi) reads generated by Circular Consensus Sequencing (CCS) appear to match Illumina in base calling accuracy, i.e. >99%, the costs of HiFi sequencing are currently

prohibitive.[212,214] Therefore, hybrid assembly of less expensive long SMS reads and Illumina reads, which harnesses advantages of both techniques, is currently the most cost-effective means of generating high-quality contiguous microbial genomes. As Illumina reads are often relied upon for hybrid assembly accuracy, it is crucial to ensure that Illumina sequencing depth or coverage generates a high proportion of error free reads to cover the entire genome.[213] Optimal coverage is determined by the organism, the genome size, the GC content, and the proportion of the genome made up of repeat or homopolymer sequences.[213] For most organisms, 30x coverage is sufficient for genome assemblies but more challenging microbial genomes will require a minimum of 50x coverage.[213]

The choice between the two SMS platforms for microbial genome sequencing is largely driven by costs and research objectives. ONS is currently significantly less expensive than SMRT sequencing. In addition, ONS can generate ultra-long reads greater than 2 Mb in length, as opposed to an average of 30 kb for SMRT as read length is restricted by polymerase longevity in the latter.[208,212,214] This means ONS is more likely to generate fewer contigs while challenging regions are also better resolved.[208,212,214] Furthermore, ONS is polymerisation-free which means polymerisation-associated biases and errors are not carried through.[195] However, as base calling is inherently more complex in ONS, its base call accuracy is usually lower compared with SMRT sequencing.[212,214] But subsequent polishing of hybrid assemblies with highly accurate Illumina reads compensates for this flaw.[208,212]
In essence, long read ONS and paired-end short read Illumina sequencing with a minimum coverage of 30x were the most cost effective and accurate sequencing techniques to adopt here.

3.3.2 Sequencing Reads Quality Assessment and Filtering

As no sequencing platform is 100% accurate, base calling errors do occur during reactions. Therefore, it is important to perform initial quality assessment and filtering (if needed) to ensure reads are of sufficient qualities for downstream applications. There are publicly available robust and validated tools that can be used for this purpose, many of which are more suited to analyse data generated by specific sequencing platforms or data presented in specific formats.

Illumina output files are presented in FASTQ format which comprises the sequence identifier, the nucleotide sequence, and associated base quality scores encoded by ASCII (American Standard Code for Information Interexchange) printable characters 33 to 126 (decimal numeral system).[215] Each ASCII code assigned to each base corresponds to its Phred or Quality (Q) score, which is the probability of that base call being incorrect (Table 3.1).

Table 3.1: Phred Quality Score Table

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10000	99.99%
50	1 in 100000	99.999%
60	1 in 1000000	99.9999%

Source: Biostars (www.biostars.org).

Quality assessment of raw Illumina sequencing data can be performed with several tools that take input files in FASTQ format, but the most popular tool is FastQC which generates quality reports in the form of summary graphs and tables.[216] Amongst other reports, per base sequence quality report and per sequence quality scores can give a quick assessment of the overall quality of input data. After quality assessment with FastQC, raw reads will usually need trimming to remove adapter sequence and other Illumina-associated sequences (if not already removed by sequencing facility) or to remove low-quality bases from the start or end of sequences. Sequences with low mean quality scores may be filtered from the data before downstream use, but as there is a trade-off between sequencing coverage and Phred Quality Scores, sequences with mean quality scores above Q15 are usually retained while those with lower scores are generally filtered.[213] The trimming and filtering exercise can be performed with various tools, but the most popular for Illumina data is Trimmomatic.[217] As Illumina sequencing data usually consist of two files, i.e. forward and reverse reads, both datasets can be trimmed by Trimmomatic simultaneously and paired to further increase accuracy as unpaired reads are filtered out. The quality of trimmed reads is usually re-

assessed with FastQC after which quality reports of individual datasets can be aggregated into a single report file with MultiQC to make thorough quality assessment and comparisons easier to perform.[218]

ONS output files are presented in Fast5 format which are usually converted to FASTQ files by sequencing facilities so that service users receive ONS data in FASTQ file format. However, FastQC is not the best tool for assessing the quality of ONS data as quality thresholds for Illumina and ONS data are quite dissimilar.[219] Various long read quality assessment tools are more suitable, with the NanoPack suite of tools being one of the most efficient.[219] NanoStat identifies the encoding system in the FASTQ input file and generates corresponding Phred scores which is presented in a comprehensive statistical data summary along with other data attributes.[219] In addition, various quality plots depicting read length distribution and Phred scores can be generated with NanoPlot and NanoComp to aid quality assessment.[219] As ONS data (especially those generated with MinION and GridION sequencers) are more error prone than Illumina data, the majority of reads in the former will have mean Phred scores lower than Q15 but higher than Q7. Reads with very low mean Phred scores can be filtered from the data with NanoFilt.[219] NanoFilt can also be used to trim low quality bases at either end of sequences.[219]

3.3.3 *De novo* Hybrid Genome Assembly

Hybrid assembly of bacterial genomes using long and short sequencing reads have been shown to generate the most accurate assemblies.[208] There are several hybrid assemblers for bacterial genomes, but Unicycler and Trycycler currently (at the time of assembling and writing) generate the most accurate and contiguous hybrid assemblies.[220,221] Assembly accuracy and contiguity with Unicycler relies heavily on the quality and depth of Illumina reads, while these parameters are less important when using Trycycler. Also, Unicycler works well with long reads of any depth and accuracy, but a depth above 25x or ideally around 100x is needed for Trycylcer hybrid assemblies.[221] This very high long read depth is achievable with the recently introduced ultra-high throughput PromethION sequencers, but not readily, if at all, with MinION or GridION sequencers that are currently in use in most labs and sequencing facilities. For that reason, Unicycler is currently the hybrid assembler of choice for bacterial genomes, and as such was used here.

Unicycler accepts input files in FASTA or FASTQ formats and performs hybrid assembly in multiple steps using various in-built tools. The first step is autocorrection of Illumina reads with SPAdes.[222] Thereafter, De Bruijn assembly graphs are constructed with Illumina reads using optimised k-mers that yield the fewest number of contigs and the fewest number of dead ends.[220] Assembly graphs are then bridged or simplified with a combination of short reads contigs and long reads to ensure that challenging regions are fully resolved. Replicons generated are circularised, and the final assembly is polished several times with Pilon[223] (using Illumina reads) until assembly improvements are no longer observed. This final stage guarantees the accuracy of the assembly. Provided high-quality Illumina reads with sufficient depths are used in the hybrid assembly, Unicycler will almost certainly assemble complete and contiguous bacterial genomes irrespective of the depth and quality of long reads.[220]

3.3.4 Assembly Quality Assessment

The quality of assembled genomes can be assessed with a number of robust tools, many of which are complementary. QUAST is the most widely used quality assessment tool because a reference genome is not necessary for analysis.[224] It generates useful quality metrics such as the number of contigs in the assembly, the number of contigs longer than certain lengths, the length of the largest contig, and the total number of bases in the assembly (i.e. the genome size). Researchers can easily estimate the quality or contiguity of final assemblies from these metrics. Contiguity is also assessed with the N50 and N75 values generated by QUAST. Where N50 is the length of the shortest contig in the set of largest contigs that together make up at least 50% of the entire assembly; with N75 being the same metric but generated for 75% of the total length of the assembly. Furthermore, QUAST generates interactive plots to enable further exploration of assembly quality.

Bandage is a complementary quality assessment tool that enables visualisation of final assembly graphs.[225] An example of Bandage output is presented in Figure 3.1 showing a contiguous assembly (Figure 3.1a) and a highly fragmented assembly

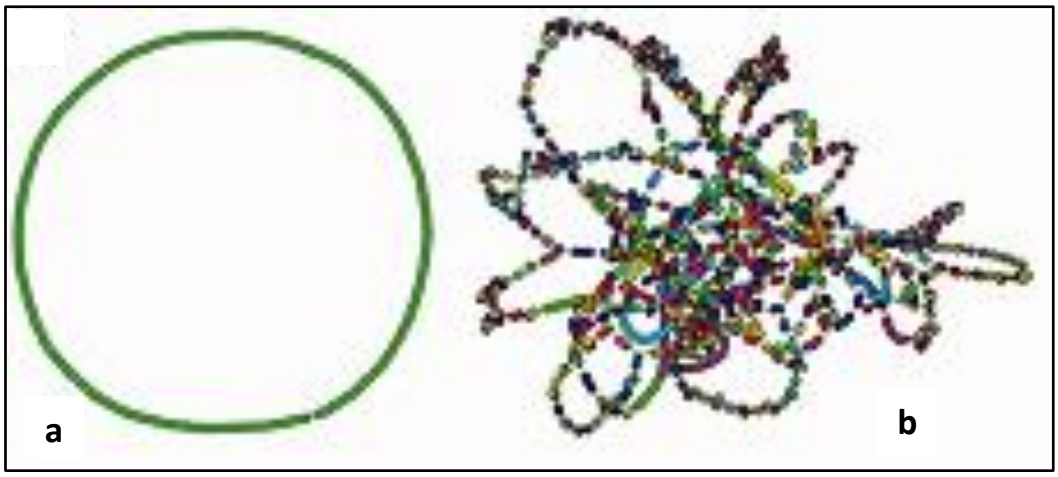


Figure 3.1: An example of a Bandage visualisation output.
Bandage output showing **a)** a contiguous assembly with no dead ends, **b)** a fragmented assembly with many contigs and dead ends. Image adapted from Wick *et al*.[225]

with multiple contigs and many dead ends (Figure 3.1b). Bandage can be useful for performing assembly improvement tasks as it provides an interactive interface where users can identify and investigate poorly resolved regions in final assemblies.[225]

3.3.5 Confirmation of Identities of Isolates Using Whole-genome Sequence Data

Isolates were preliminarily identified in Chapter 2 using partial 16S rRNA gene sequences for gene sequence comparisons. Confirmation of identity requires more robust analyses which can be carried out in two stages. The first stage utilises the typing tool, SpeciesFinder [226] to predict isolates' identities based on full-length 16S rRNA gene sequence comparisons. The tool extracts the full-length sequence of the 16S rRNA gene from the query and compares it to 16S rRNA gene sequences in the SpeciesFinder database, which is extracted from the Small Subunit rRNA SILVA database.[226,227] The analysis is rapid, and an output is generated where the 'best match' to the query in the SpeciesFinder database is displayed with a PASS or FAIL result. A PASS result indicates that the query has an identity score of >98% when mapped to the 'best match' in the database, while a FAIL result indicates that the query has an identity score ≤98% when mapped to the 'best match'. An identity score >98%, i.e. a PASS, is required for correct identity prediction.

As even full-length 16S rRNA gene comparison does not always resolve different species, the results obtained with SpeciesFinder can be validated using FastANI tool. FastANI uses a rapid alignment-free whole-genome mapping algorithm to compute the benchmark typing metric, ANI (**A**verage **N**ucleotide **I**dentity of all orthologous genes present in a pair of genomes).[228,229] FastANI estimates the ANI value between two genomes, usually a reference genome and a query. The analysis has robust and high-resolution power because nucleotide sequences of genes covering entire genomes are compared.[228] Typically, ANI value between two genomes of strains of the same species is above 95%.[228,229]

3.3.6 Assessment of Assembly Completeness

Although hybrid assembly of high-quality long and short reads should theoretically generate contiguous and complete genomes, a validation step is nevertheless necessary to confirm completeness. Many validation tools have been proposed including Assembly Likelihood Evaluation (ALE),[230] REAPR,[231] and QUAST,[224] but these tools do not assess genomes in terms of expected gene content which determines assembly completeness.[232] To address this gap, Simão *et al* introduced the BUSCO (**B**enchmarking **U**niversal **S**ingle-**C**opy **O**rthologs) tool which compares gene contents of an assembly against the database of conserved single-copy orthologous genes (termed 'BUSCOs') present in more than 90% of species belonging to its taxonomic group.[233] After BUSCO analysis, orthologous genes found intact (i.e. ≈95% of expected length) and in single copy in the assembly are classified as 'Complete'; partially recovered orthologous genes are classified as 'Fragmented'; missing orthologs are reported as 'Missing'; and expected orthologs that are present in more than one copy are classified as 'Duplicated'.[233]

Although BUSCO 'Complete' score is currently the widely accepted metric for assessing microbial genome completeness, there is no objective cut-off for a good or bad 'Complete' score.[234] The consensus for prokaryotic genomes is that a BUSCO 'Complete' score >95% would describe a complete assembly.[233]

3.4 Materials and Methods

3.4.1 Whole-genome Sequencing

3.4.1.1 Culture/Sample Preparation Prior to Shipping to Sequencing Facilities

Single pure colonies of isolates A23BA and A13BB were aseptically transferred into separate sterile microcentrifuge tubes each containing 100 µL sterile distilled water. Contents of each tube were gently vortexed and loopful quantities of each culture were aseptically transferred onto separate nutrient agar (Oxoid, UK) plates with a third of the surface of each plate covered in a lawn and cultures streaked out to cover remaining segments. Plates were incubated at 37°C and 28°C for 24 hours for isolates A23BA and A13BB, respectively. After 24 hours, plates were inspected to confirm culture purity after which entire bacterial cultures were scraped off the plates and aseptically transferred into separate Microbank™ cryovials (Pro-Lab Diagnostics, UK) before shipping at room temperature to MicrobesNG (Birmingham, UK) for Illumina (30x coverage) and Oxford nanopore sequencing.

On a separate occasion, as 30x Illumina sequencing depth was inadequate, a single pure colony of isolate A13BB was aseptically transferred into 10 mL sterile nutrient broth (Oxoid, UK) contained in a sterile universal bottle and incubated at 28°C for 24 hours, after which gDNA was extracted from pelleted cells as described in Section 2.4.4.1. Quality of gDNA was assessed as described in Section 2.4.4.2, then 2 x 50 µL solution of gDNA in eluting buffer (Qiagen, UK) was shipped at room temperature to Glasgow Polyomics (Glasgow, UK) for Illumina sequencing (100x coverage).

3.4.1.2 Generation of Sequencing Data

On receipt of cultures at MicrobesNG, gDNA was extracted from each isolate and libraries were prepared for Illumina sequencing according to in-house protocol supplied by MicrobesNG: 20 µL of culture suspension were lysed with 120 µL of TE buffer containing lysozyme (final concentration 0.1 mg/mL) and RNase A (ITW Reagents, Spain) (final concentration 0.1 mg/mL), incubated for 25 min at 37°C. Proteinase K (VWR Chemicals, USA) (final concentration 0.1 mg/mL) and SDS (Sigma-Aldrich, USA) (final concentration 0.5% v/v) were added and incubated for 5 min at 65°C. gDNA was purified using an equal volume of SPRI beads and

resuspended in eluting buffer (Qiagen, Germany). DNA was quantified with the Quant-iT dsDNA HS kit (Thermo Fisher Scientific, UK) assay in an Eppendorf AF2200 plate reader (Eppendorf UK Ltd, UK). gDNA libraries were prepared using the Nextera XT Library Prep Kit (Illumina, USA) following the manufacturer's protocol with the following modifications: 2 ng of DNA were used as input, and PCR elongation time was increased to 1 min from 30 sec. DNA quantification and library preparation were carried out on a Hamilton Microlab STAR automated liquid handling system (Hamilton Bonaduz AG, Switzerland). Pooled libraries were quantified using the Kapa Biosystems Library Quantification Kit for Illumina on a Roche light cycler 96 qPCR machine. Libraries were sequenced with Illumina HiSeq using a 250 bp paired-end protocol.

For ONS, gDNA was extracted from each isolate and libraries were prepared for sequencing according to in-house protocol supplied by MicrobesNG: broth cultures of each isolate were pelleted out and the pellets were resuspended in the cryopreservative of a Microbank™ (Pro-Lab Diagnostics, UK) tube. Approximately $2x10^9$ cells were used for high molecular weight DNA extraction using Nanobind CCB Big DNA Kit (Circulomics, USA). DNA was quantified with the Qubit dsDNA HS assay (Invitrogen, UK). Long read gDNA libraries were prepared with Oxford nanopore SQK-RBK004 kit and/or SQK-LSK109 kit with Native Barcoding EXP-NBD104/114 (ONT, UK) using 400-500 ng of high molecular weight DNA. Twelve to twenty-four barcoded samples were pooled together into a single sequencing library and loaded in a FLO-MIN106 (R.9.4 or R.9.4.1) flow cell in a GridION (ONT, UK).

On receipt of the gDNA solution at Glasgow Polyomics, DNA was quantified, and libraries were prepared for Illumina sequencing according to in-house protocol supplied by Glasgow Polyomics: gDNA was quantified with Qubit Fluorometer (ThermoFisher Scientific, USA) and libraries were prepared for Illumina sequencing using the Nextera XT DNA Library Preparation Kit (Illumina, USA) following manufacturer's protocol. Libraries were sequenced with Illumina MiSeq sequencer using a 300 bp paired-end protocol.

3.4.2 Quality Control of Sequencing Data

Raw and trimmed Illumina data were received from MicrobesNG for both isolates. Trimmed data were processed by the facility using Trimmomatic v0.30 with a sliding window quality cut-off of Q15. Therefore, no further trimming or filtering was done. Quality reports for trimmed forward and reverse reads were generated with FastQC v0.11.8 and results were aggregated into single report files with MultiQC v1.8. ONS long reads data for both isolates were assessed with NanoStat v1.28.2, while quality plots were generated with NanoPlot v1.28.2.

Raw forward and reverse Illumina reads data for isolate A13BB received from Glasgow Polyomics were quality assessed with FastQC v0.11.8. Initial quality reports of raw data showed poor basic statistics, therefore trimming and filtering were needed. Datasets were trimmed and filtered with Trimmomatic v0.36 with a sliding window quality cut-off of Q25 and minimum read length of 100. Adapter sequences (TruSeq3) and other Illumina-specific sequences were also removed. However, sequences generated from the genome of bacteriophage φX174 which acts as a control spike in Illumina sequencing reactions were not filtered from the raw data at this stage. The quality of trimmed forward and reverse reads were re-assessed with FastQC and quality reports were aggregated into a single report file with MultiQC v1.8. All tools were operated with default settings or as otherwise stated.

3.4.3 *De novo* Genome Assembly

For isolate A23BA, trimmed short reads and long reads received from MicrobesNG were assembled *de novo* with Unicycler (Galaxy v0.4.8.0).

For isolate A13BB, two assemblies - assembly 1 and assembly 2 - were generated with Unicycler (Galaxy v0.4.8.0). Assembly 1 was generated with trimmed short reads and long reads received from MicrobesNG, while assembly 2 was generated with trimmed short reads received from Glasgow polyomics and long reads received from MicrobesNG. As assembly 2 included the complete genome sequence of bacteriophage φX174 acting as internal standard, the 5,386 bp bacteriophage genome, which was assembled as a separate replicon, was extracted from the data before further analyses. Unicycler was operated with default settings for all assemblies.

3.4.4 Assembly Quality Assessment

Quality metrics for all three assemblies were generated with QUAST v5.0.2 using default settings. Assembly graphs were visualised with Bandage v0.8.1 (operated with default settings) to further assess assembly quality and contiguity. For isolate A13BB, the contiguous assembly (assembly 2) based on QUAST report and Bandage visualisation was selected for further analyses.

3.4.5 Confirmation of Identities of Isolates Using Whole-genome Sequence Data

For isolate A23BA, whole-genome sequence data was uploaded onto SpeciesFinder v1.2 in FASTA format to predict identity. SpeciesFinder prediction was verified by using FastANI v1.3 to estimate ANI values between the genome of isolate A23BA and the *Hafnia alvei* reference genome in the NCBI microbial genome database. Analyses were carried out in October 2020.

For isolate A13BB, whole-genome sequence data of assembly 2 was uploaded onto SpeciesFinder v1.2 in FASTA format to predict identity. SpeciesFinder prediction was verified by using FastANI tool v1.3 to estimate ANI values between the genomes of isolate A13BB and the *Pseudomonas psychrophila* reference genome in the NCBI microbial genome database, in the first instance. As this analysis was less straight forward, ANI values were later computed between the genomes of isolate A13BB and the *P. fragi* reference genome (i.e. *P. fragi* strain P121) in the NCBI microbial genome database, followed by further analyses using the complete genomes of *P. fragi* NL20W and *P. fragi* NMC25 downloaded from the NCBI database. Analyses were carried out in December 2020. SpeciesFinder and FastANI tools were operated with default settings for all analyses.

3.4.6 Confirmation of Assembly Completeness and Genome Annotation

The completeness of both assemblies was assessed with BUSCO using the enterobacteriales_odb9 lineage dataset (BUSCO v3.0.2) and the pseudomonadales_odb10 lineage dataset (BUSCO v4.1.2) for *Hafnia alvei* strain A23BA and *Pseudomonas fragi* strain A13BB, respectively. Both assemblies were later annotated with PGAP v4.11 before genome maps were created with CGView.[235,236]

3.5 Results

3.5.1 Quality Assessment of Trimmed Short Reads and Long Reads

After trimming and filtering as stated in Section 3.4.2, all Illumina datasets (forward and reverse reads) used for assembling were of high quality with mean quality scores (Phred scores) across base positions in all datasets being well above Q20 as shown in respective MultiQC output files (Figures 3.2, 3.4, and 3.6). The reverse reads dataset of the 100x sequencing reactions for isolate A13BB appeared to be the most challenging to generate with fluctuating mean quality scores across base positions (Figure 3.6). However, at no point did mean quality score across base positions dip below Q30, meaning these reads were of high quality regardless.

The ONS datasets received were not trimmed or filtered as 100% of reads had quality scores > Q7 and N50 values ≥ 9 kbp (Figures 3.3 and 3.5).

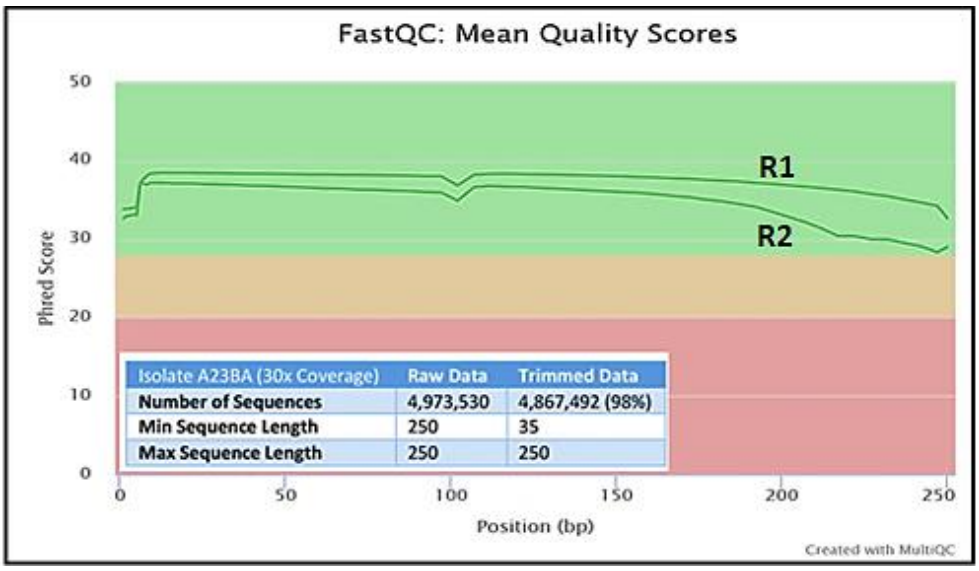


Figure 3.2: Mean quality scores across trimmed short reads (isolate A23BA).
After trimming, 98% of reads were retained with mean quality scores across reads of R1 dataset (forward reads) and R2 dataset (reverse reads) being >Q30 and ≥Q28, respectively, i.e. mean base call accuracy across reads was >99.9% and >99% for R1 and R2 datasets, respectively.

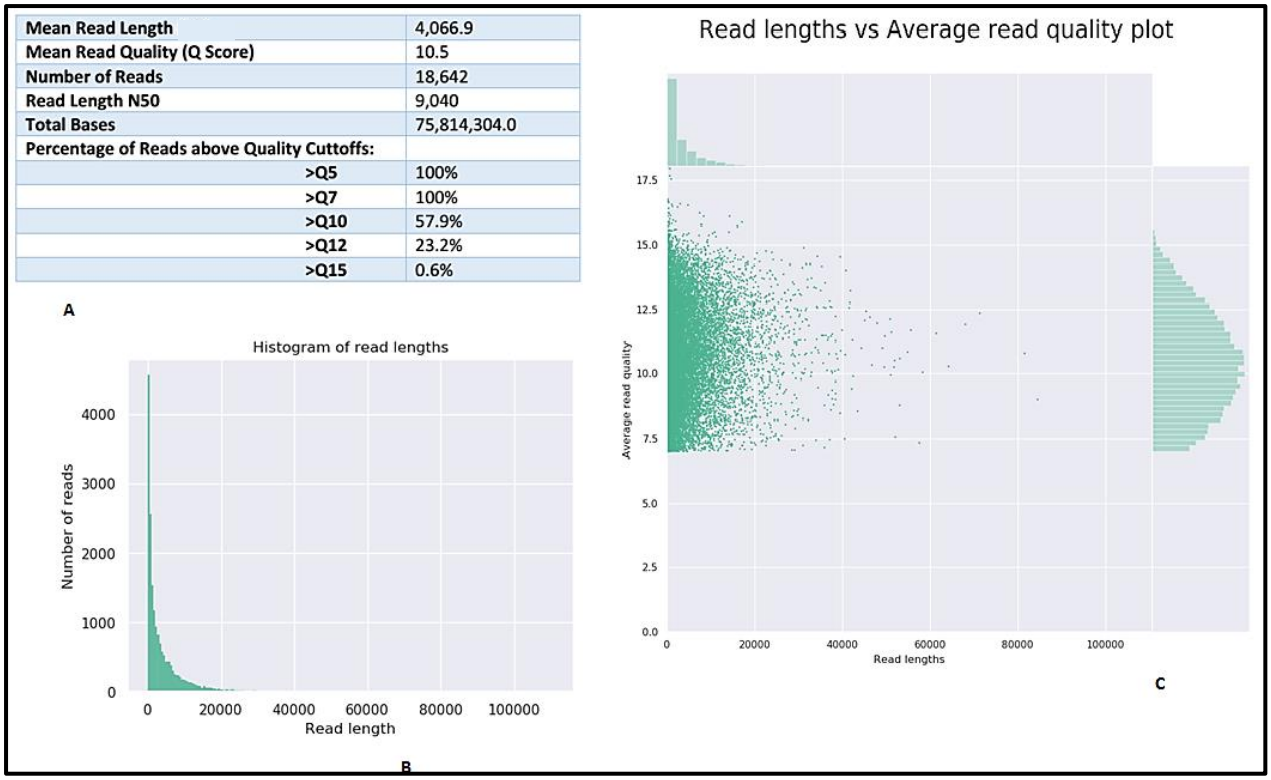


Figure 3.3: Quality assessment report of long reads (isolate A23BA).
A) basic statistics, **B)** read lengths histogram, **C)** read lengths vs average read quality plot. The majority of long reads generated were >4 kbp in length with mean quality score across reads of Q10.5 i.e. mean base call accuracy across reads was >90%.

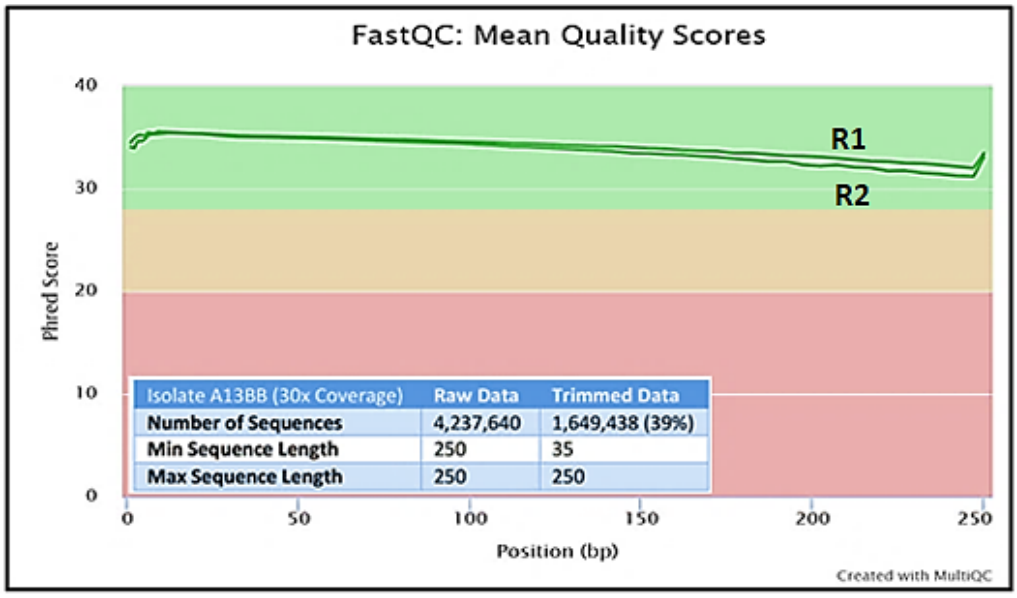


Figure 3.4: Mean quality scores across trimmed short reads (1) (isolate A13BB).
After trimming, only 39% of reads were retained with mean quality scores across reads of both R1 dataset (forward reads) and R2 dataset (reverse reads) being >Q30 i.e. mean base call accuracy across reads was >99.9% for both datasets.

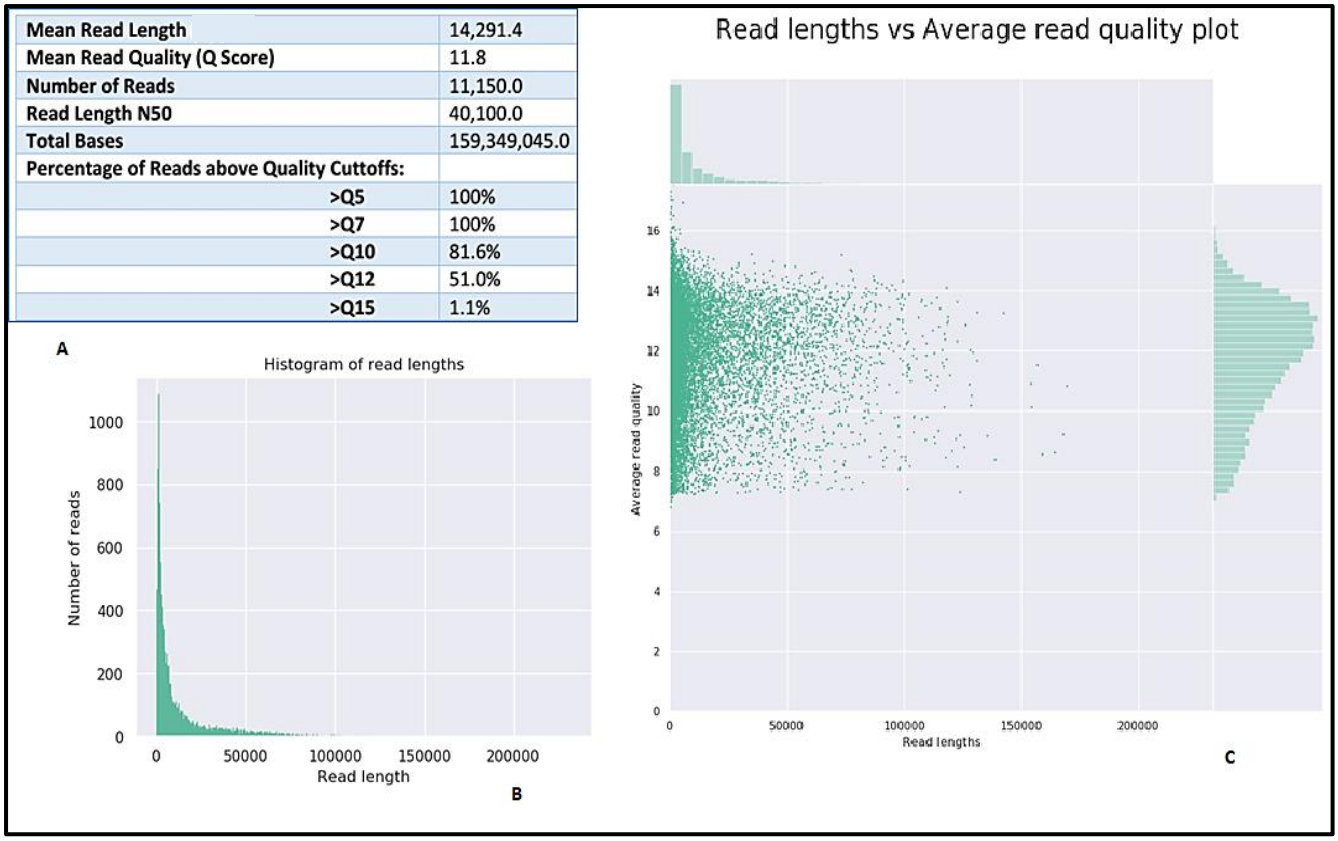


Figure 3.5: Quality assessment report of long reads (isolate A13BB).
A) basic statistics, **B)** read lengths histogram, **C)** read lengths vs average read quality plot. The majority of long reads generated were >14 kbp in length with mean quality score across reads of Q11.8 i.e. mean base call accuracy across reads was >90%.

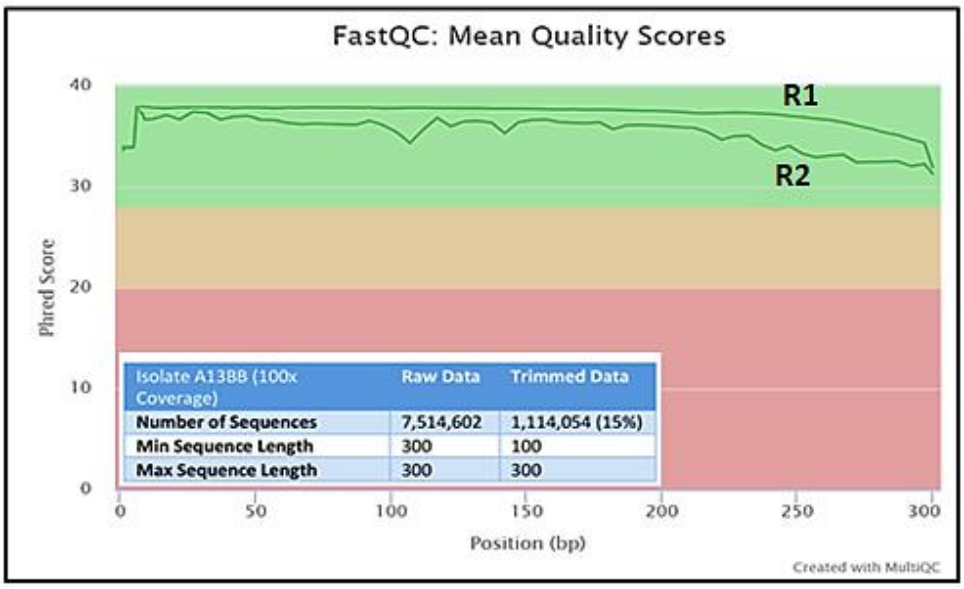


Figure 3.6: Mean quality scores across trimmed short reads (2) (isolate A13BB).
After trimming, only 15% of reads were retained with mean quality scores across reads of both R1 dataset (forward reads) and R2 dataset (reverse reads) being >Q30 i.e. mean base call accuracy across reads was >99.9% for both datasets.

3.5.2 Assembly Quality Assessment

The quality metrics obtained for isolate A23BA assembly (Figure 3.7A) showed an assembly that is 4,442,047 bp in length comprising 2 contigs, with the larger contig accounting for 98.2% of the assembly length. Both N50 and N75 values are the same, also representing 98.2% of the assembly. Bandage visualisation (Figure 3.7B) showed the assembly graph with two contigs that are most likely two separate replicons, comprising a chromosome and a plasmid.

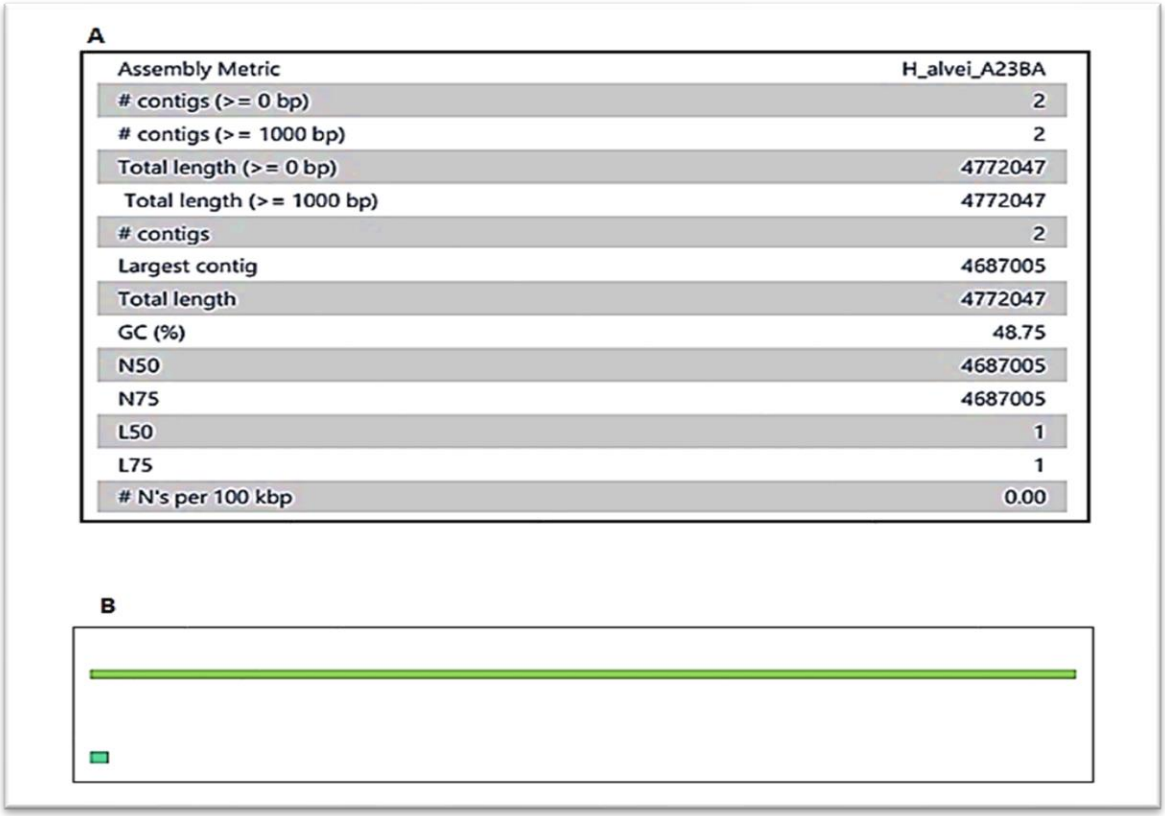


Figure 3.7: Isolate A23BA assembly quality report.
A) QUAST report, **B)** Bandage visualisation of assembly graph. Both QUAST report and Bandage visualisation showed an assembly comprising 2 contigs with the larger contig constituting over 98% of the assembly and representing the N50 and N75 values. These metrics are suggestive of a contiguous assembly comprising a chromosome and a plasmid.

The quality metrics obtained for isolate A13BB assembly 1 (Figure 3.8A) showed an assembly that is 4,941,220 bp in length comprising 4 contigs, with the largest contig accounting for just 53% of the assembly. And with N50 and N75 values accounting for 53% and 30% of the assembly, respectively. These metrics are suggestive of a fragmented assembly as also depicted by the Bandage assembly graph (Figure 3.8B).

A

Assembly Metric	P_fragi_A13BB
# contigs (>= 0 bp)	4
# contigs (>= 1000 bp)	4
Total length (>= 0 bp)	4941220
Total length (>= 1000 bp)	4941220
# contigs	4
Largest contig	2623109
Total length	4941220
GC (%)	59.44
N50	2623109
N75	1481853
L50	1
L75	2
# N's per 100 kbp	0.00

B

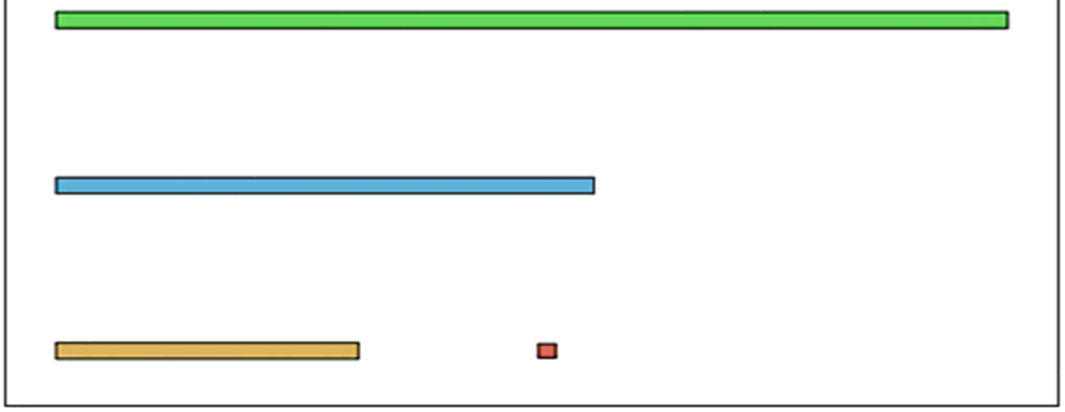
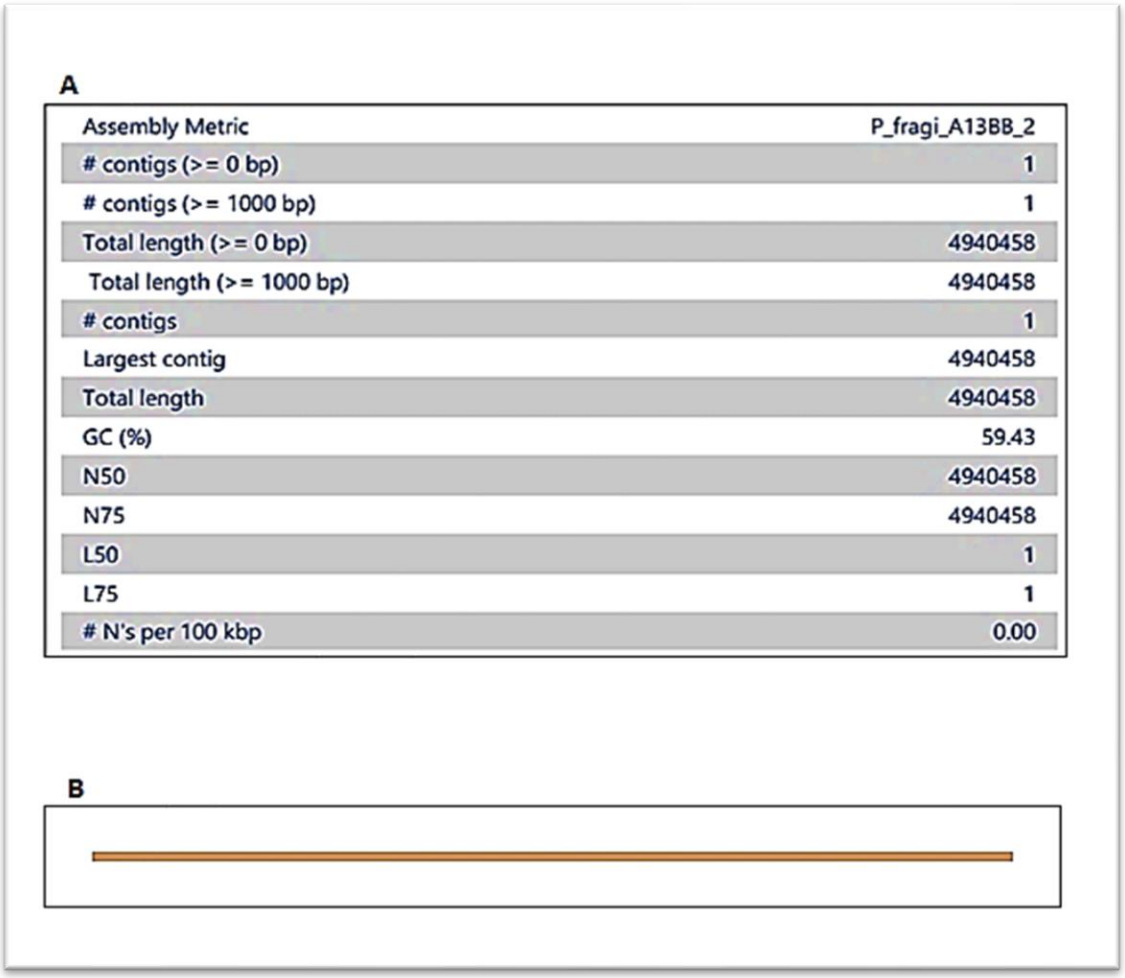


Figure 3.8: Isolate A13BB assembly 1 quality report.
A) QUAST report, **B)** Bandage visualisation of assembly graph. QUAST report and Bandage visualisation showed an assembly comprising 4 contigs with the largest contig constituting 53% of the assembly; and N50 and N75 values representing 53% and 30% of assembly length, respectively. These metrics are suggestive of a fragmented assembly.

The quality metrics obtained for isolate A13BB assembly 2 (Figure 3.9A) showed an assembly that is 4,940,458 bp in length with only 1 contig, and with both N50 and N75 values being the same value as the total assembly length. Bandage visualisation (Figure 3.9B) also shows the assembly graph of a contiguous assembly with just 1 contig.



A

Assembly Metric	P_fragi_A13BB_2
# contigs (>= 0 bp)	1
# contigs (>= 1000 bp)	1
Total length (>= 0 bp)	4940458
Total length (>= 1000 bp)	4940458
# contigs	1
Largest contig	4940458
Total length	4940458
GC (%)	59.43
N50	4940458
N75	4940458
L50	1
L75	1
# N's per 100 kbp	0.00

Figure 3.9: Isolate A13BB assembly 2 quality report.
A) QUAST report, **B)** Bandage visualisation of assembly graph. QUAST report and Bandage visualisation showed a contiguous assembly comprising one contig and with N50 and N75 values being the same value as total assembly length.

3.5.3 Confirmation of Identities of Isolates A23BA and A13BB

Isolate A23BA was predicted to be a strain of *Hafnia alvei* by SpeciesFinder and this was confirmed as such with FastANI estimate between the genome of isolate A23BA and the *H. alvei* reference genome being greater than 95% (Table 3.2).

The 'best match' for isolate A13BB in the SpeciesFinder database was *Pseudomonas psychrophila,* but with a 'FAIL' result. This meant the identity of the isolate could not be correctly predicted by the tool at the time of analysis. FastANI analysis of isolate A13BB mapped to the reference genomes of *P. psychrophila* and *P. fragi* suggested the isolate was not a strain of either species as ANI estimates were less than 95% (Table 3.2). However, mapping the genome of isolate A13BB to the genomes of two other strains of *P. fragi* (strains NL20W and NMC25) confirmed its identity as a strain of *P. fragi* with ANI estimates greater than 95%.

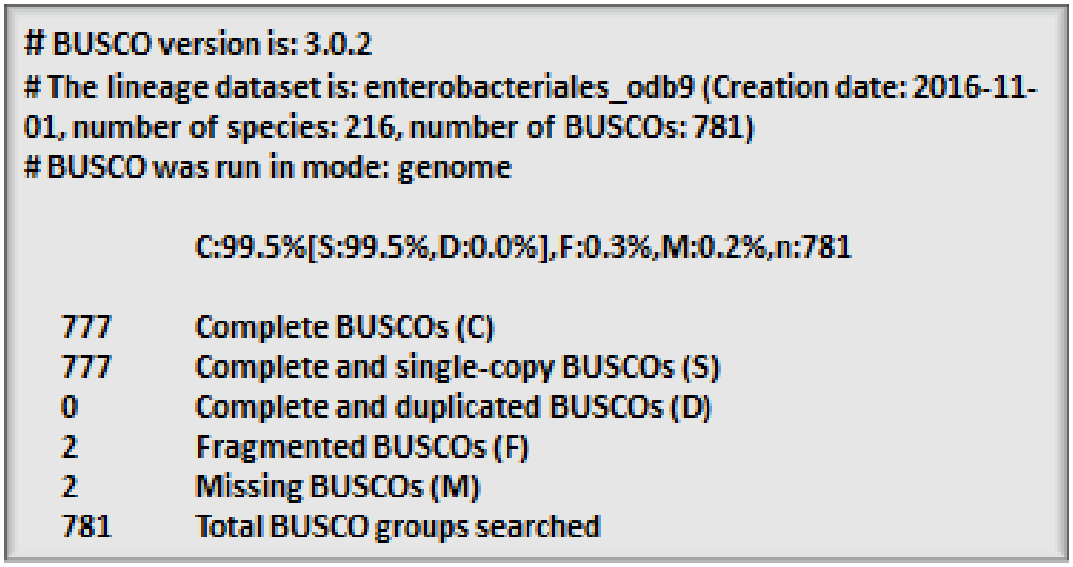
Table 3.2: Confirmation of Identities of Isolates

	SpeciesFinder Prediction	**FastANI Estimate**
Isolate A23BA	*Hafnia alvei* (PASS)	97.82 (mapped to *H. alvei* RG)
Isolate A13BB	*Pseudomonas psychrophila* (FAIL)	87.72 (mapped to *P. psychrophila* RG)
		85.85 (mapped to *P. fragi* P121, RG)
		98.91 (mapped to *P. fragi* NL20W)
		98.85 (mapped to *P. fragi* NMC25)

RG=Reference Genome

3.5.4 Assessment of Assembly Completeness with BUSCO

BUSCO 'Complete' score for *H. alvei* A23BA assembly was 99.5% (Figure 3.10), and 99.2% for *P. fragi* A13BB assembly (Figure 3.11). With the value of the metric greater than 95% for both assemblies, both were deemed complete.

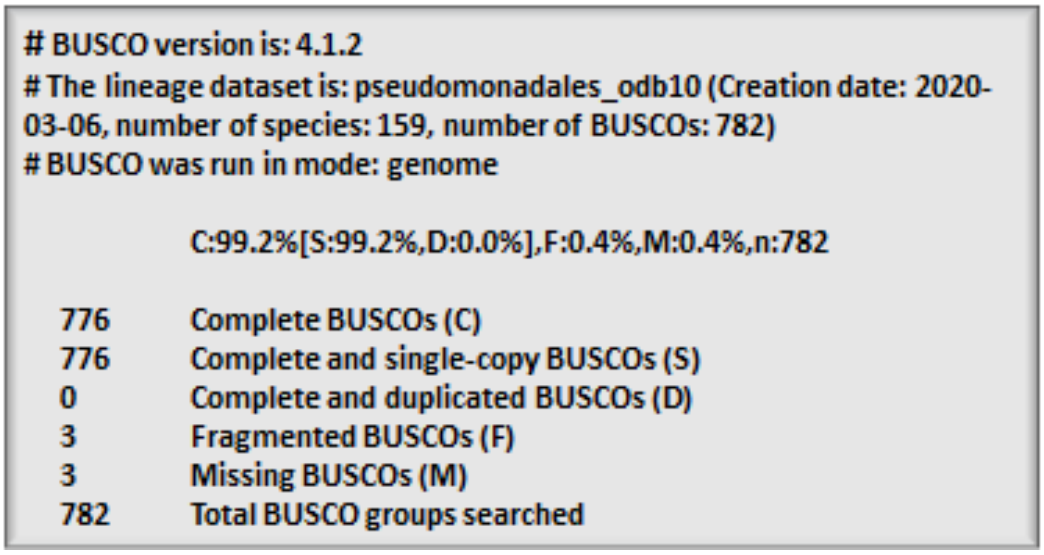


```
# BUSCO version is: 3.0.2
# The lineage dataset is: enterobacteriales_odb9 (Creation date: 2016-11-
01, number of species: 216, number of BUSCOs: 781)
# BUSCO was run in mode: genome

        C:99.5%[S:99.5%,D:0.0%],F:0.3%,M:0.2%,n:781

  777   Complete BUSCOs (C)
  777   Complete and single-copy BUSCOs (S)
  0     Complete and duplicated BUSCOs (D)
  2     Fragmented BUSCOs (F)
  2     Missing BUSCOs (M)
  781   Total BUSCO groups searched
```

Figure 3.10: BUSCO report for *Hafnia alvei* A23BA assembly.
Of the 781 orthologous genes (BUSCOs) expected for species in its taxonomic group, the genome of *H. alvei* A23BA contains 777 intact BUSCOs, representing 95% of expected BUSCOs. Therefore, the assembly was deemed 99.5% complete. Two expected BUSCOs are missing in the assembly and a further 2 are fragmented.



```
# BUSCO version is: 4.1.2
# The lineage dataset is: pseudomonadales_odb10 (Creation date: 2020-
03-06, number of species: 159, number of BUSCOs: 782)
# BUSCO was run in mode: genome

        C:99.2%[S:99.2%,D:0.0%],F:0.4%,M:0.4%,n:782

  776   Complete BUSCOs (C)
  776   Complete and single-copy BUSCOs (S)
  0     Complete and duplicated BUSCOs (D)
  3     Fragmented BUSCOs (F)
  3     Missing BUSCOs (M)
  782   Total BUSCO groups searched
```

Figure 3.11: BUSCO report for *Pseudomonas fragi* A13BB assembly.
Of the 782 orthologous genes (BUSCOs) expected for species in its taxonomic group, the genome of *P. fragi* A13BB contains 776 intact BUSCOs, representing 99.2% of expected BUSCOs. Therefore, the assembly was deemed 99.2% complete. Three expected BUSCOs are missing in the assembly and a further 3 are fragmented.

3.5.5 Genome Maps

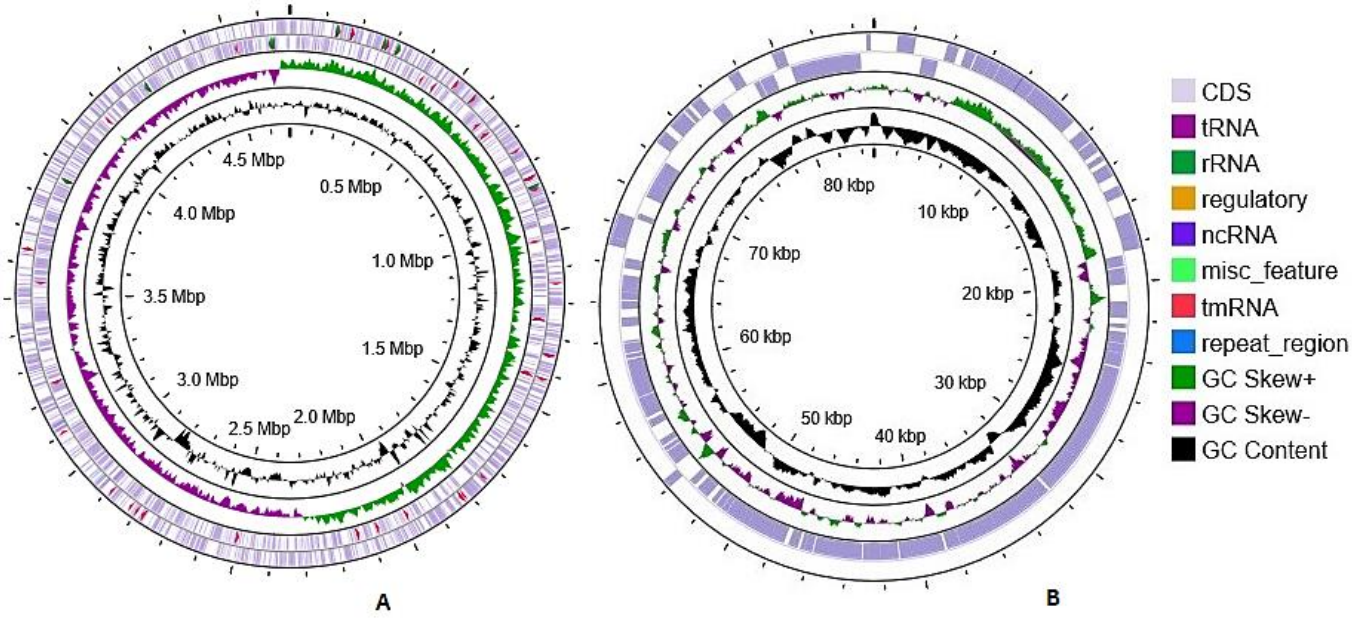


Figure 3.12: Circular representation of *H. alvei* A23BA genome.
A) chromosome and **B)** plasmid; showing basic genomic features.
Genome map was created with CGView.[236]

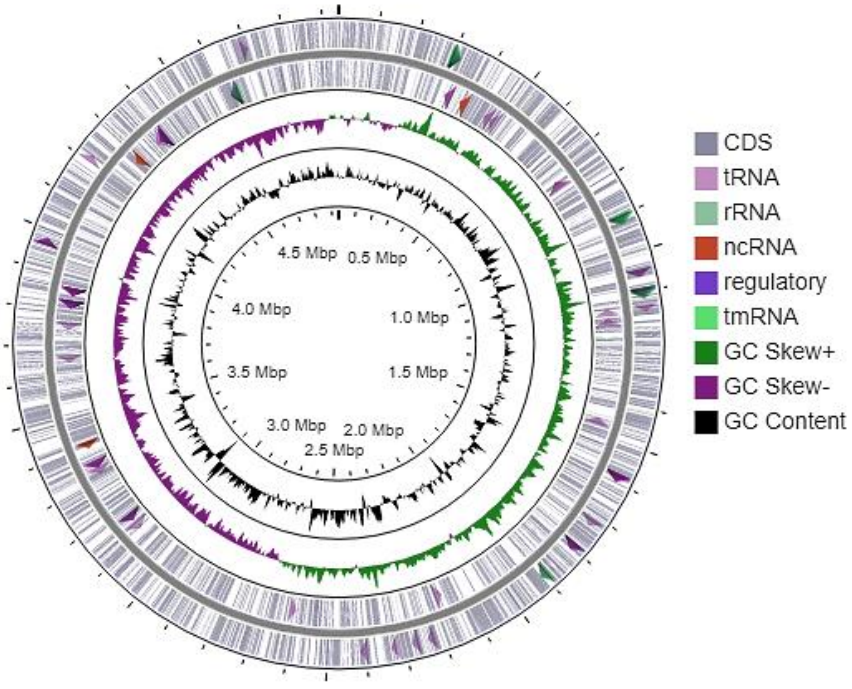


Figure 3.13: Circular Representation of *P. fragi* A13BB genome.
Circular chromosome showing basic genomic features.
Genome map was created with CGView.[236]

3.6 Discussion

The complete genomes of isolates A23BA and A13BB were assembled *de novo* with high-quality short Illumina reads and long Oxford nanopore reads. For isolate A23BA, Illumina sequencing (30x coverage) runs generated 4,973,530 short reads that were trimmed and paired with Trimmomatic with a quality cut-off of Q15 to further improve accuracy.[237] Ninety eight percent of reads were retained with a mean quality score of ≥Q28 across base positions (Figure 3.2).[237] Oxford nanopore sequencing generated 18,642 long reads with a mean read quality score of Q10.5 (Figure 3.3A). Figures 3.3B and 3.3C provide snapshots of read length distribution and read quality distribution, respectively. Trimming of long reads was not necessary as 100% of reads had quality scores greater than Q7 (Figures 3.3A and 3.3C). These high-quality short and long reads obtained for isolate A23BA were used in its genome assembly generating 2 contigs with N50 and N75 values of 4,687,005 (98.2% of total assembly length) and zero indeterminate bases (Figure 3.7A). As one of the contigs represents 98.2% of the genome, it was speculated that the larger contig is a chromosome while the smaller contig, making up just 1.2% of the assembly, is a plasmid. The assembly was uploaded onto PlasmidFinder[238] v2.0 (using the Enterobacteriaceae database) which confirmed the smaller contig to be a plasmid. Therefore, the two contigs in QUAST report (Figure 3.7A) are two separate contiguous replicons as opposed to a single fragmented replicon. Bandage visualisation showed the replicons are linear with 4 dead ends and not automatically circularised as expected (Figure 3.7B). This might suggest that replicon circularisation is not performed with a default command in Unicycler as the tool was operated with default parameters. Nonetheless, the assembly generated is of high quality (i.e. not fragmented) and suitable for downstream analyses.

Using full-length 16S rRNA gene sequence, isolate A23BA was predicted (with a PASS result) to be a strain of *Hafnia alvei* with SpeciesFinder. Identity was confirmed as such by FastANI whole-genome comparison when the genome of isolate A23BA was mapped to the reference genome for *H. alvei* species in the NCBI microbial genome database. ANI value was found to be 97.8% (Table 3.2), which is above the 95% threshold for identity confirmation.[228,229] BUSCO analysis of assembled genome of *H. alvei* A23BA found 777 BUSCOs in 'Complete' and single copy out of the expected 781 orthologous genes for its taxonomic group.

Two BUSCOs were found 'fragmented' and a further 2 were reported as 'missing' (Figure 3.10). The assembly was deemed complete with the BUSCO 'Complete' score of 99.5%, which is above the 95% consensus for completeness for prokaryotic genomes.[233] The complete genome of *H. alvei* A23BA is 4,772,047 bp in size with the overall GC content of 48.77%. It comprises one circular chromosome (4,687,005 bp; GC content 48.8%) and one circular plasmid (85,042 bp; GC content 47.2%). Genomic features include 4,217 CDSs, 25 rRNA, 92 tRNA, 8 ncRNA, and 30 pseudogenes (Figure 3.12).[237]

For isolate A13BB, Illumina sequencing (30x coverage) runs generated 4,237,640 short reads that were trimmed and paired with Trimmomatic with a quality cut-off of Q15. Thirty-nine percent of reads were retained with a mean quality score of ≥Q31 across base positions (Figure 3.4). A high percentage of reads (69%) were discarded after trimming, perhaps indicating a heavily biased Illumina data. ONS generated 11,150 long reads with a mean read quality score of Q11.8 (Figure 3.5A). Figures 3.5B and 3.5C provide snapshots of read length distribution and read quality distribution, respectively. As with isolate A23BA, trimming of long reads was not necessary as quality scores were greater than Q7 (Figures 3.5A and 3.5C). Trimmed short reads and long reads obtained were used for genome assembly 1 which generated 4 contigs with N50 and N75 values of 2,623,109 and 1,481,853, representing 53% and 30% of total assembly length, respectively (Figure 3.8A). These N50 and N75 values are suggestive of a non-contiguous assembly confirmed by the Bandage output showing 4 contigs with 8 dead ends (Figure 3.8B). It was evident from the quality metrics of assembly 1 for isolate A13BB that Unicycler failed to generate a contiguous assembly. This was not surprising given the high GC content (59.44%) of the assembly as opposed to the low to medium GC contents to which 30x coverage Illumina sequencing is better suited.[213,239] The quality and accuracy of Illumina reads are compromised by PCR-related biases encountered within regions of high GC contents which means these regions are poorly amplified or not amplified at all, generating data that leave gaps in assembly graphs.[208,213,239] Furthermore, gaps also arise due to repeat or homopolymer regions that are common features of BGCs and microbial genomes in general.[208,213,239] A high sequencing depth or coverage is usually needed to compensate for these biases.[213,239] As hybrid assembly accuracy and contiguity

when using Unicycler rely heavily on the accuracy and depth of Illumina reads, it was not surprising that assembly 1 was fragmented.

In order to improve assembly quality, Illumina reads were regenerated with a sequencing depth of 100x producing 7,514,602 reads. The high sequencing depth and the volume of data generated made a more aggressive quality trimming possible such that trimming threshold was set to Q25 with a minimum sequence length of 100 to ensure highly accurate Illumina data was used for subsequent hybrid assembly.[240] Fifteen percent of reads were retained after trimming as described with a mean quality score of ≥Q31 across base positions (Figure 3.6).[240] The new set of trimmed short reads and ONS reads previously obtained were used in generating assembly 2 for isolate A13BB. Assembly 2 consists of 1 contiguous replicon with N50 and N75 metrics being the same value as the total assembly length i.e. 4,940,458, and with zero indeterminate bases (Figure 3.9A). Bandage visualisation also showed a contiguous linear replicon (Figure 3.9B). The high-quality assembly (i.e. assembly 2) now obtained for isolate A13BB is suitable for downstream analyses.

The 'best match' for isolate A13BB in the SpeciesFinder database was *Pseudomonas psychrophila,* but with a FAIL result (Table 3.2); meaning the identity score was ≤ 98% and an accurate identity prediction could not be made with the tool at the time of analysis. This may suggest that full-length 16S rRNA gene comparison alone cannot provide accurate taxonomic placement of the species or the SpeciesFinder database, at the time of analysis, was not comprehensive enough to provide resolution of the species as the accuracy of the results of 16S rRNA gene comparisons can be database dependent. The results of the preliminary identification of isolate A13BB using partial 16S rRNA gene comparison and RefSeq database (Section 2.5.6) predicted the isolate could be a strain of *P. fragi* or *P. psychrophila* with ≥ 99% identity scores. Therefore, ANI values were computed between the genome of isolate A13BB and the reference genomes of *P. psychrophila* and *P. fragi* in the NCBI microbial genome database. The ANI value between the genomes of isolate A13BB and the reference genome of *P. psychrophila*, which also happens to be the only complete genome of the species in NCBI database at the time of analysis, was 87.72% (Table 3.2). This ANI value being well below the 95% benchmark for identity confirmation suggested isolate A13BB was highly unlikely to be a strain of *P. psychrophila*.

Likewise, the ANI value between the genomes of isolate A13BB and the reference genome for *P. fragi* species (i.e. *P. fragi* P121) in the database at the time of analysis was 85.85%, also suggesting that isolate A13BB was highly unlikely to be a strain of *P. fragi*. However, as there were two other complete genomes of *P. fragi* strains (strains NL20W and NMC25) in the NCBI database at the time of analysis, ANI values were obtained between the genomes of isolate A13BB and the genomes of these two strains, and were found to be 98.91% and 98.85%, respectively (Table 3.2). The ANI values obtained with strain NL20W and strain NMC25 suggested that isolate A13BB is indeed a strain of *P. fragi.* This raised suspicions about the accuracy of the taxonomic placement of *P. fragi* strain P121 that was the reference genome of the species at the time of analysis, more so because strain P121 also had ANI values below 86% when mapped to the genomes of strain NL20W and strain NMC25, as well as to the refence genome for *P. psychrophila*. Interestingly, Stanborough *et al* have also raised suspicions about the accuracy of the taxonomic placement of *P. fragi* P121 as the strain was clearly distinct from all other *P. fragi* strains analysed in their work in which they used various genetic and metabolic markers, including ANI values.[241] They suggested that *P. fragi* P121 may need taxonomic reclassification - a notion that is reinforced by the findings of the work carried out here. The findings here also emphasise the importance of using accurately classified reference genomes for ANI analysis, or for any other typing analysis.

BUSCO analysis of the assembled genome of *P. fragi* A13BB found 776 BUSCOs 'Complete' and in single copy out of the expected 782 orthologous genes for its taxonomic group. Three orthologs were found 'fragmented' and a further 3 were reported as 'missing' (Figure 3.11). Therefore, the assembly was deemed complete with a 'Complete' score of 99.2%. The complete genome sequence of *P. fragi* A13BB comprises a single circular chromosome 4,940,458 bp in size with a GC content of 59.40%. Genomic features include 4,410 CDSs, 25 rRNA, 73 tRNA, 4 ncRNA, and 41 pseudogenes (Figure 3.13).[240]

3.7 Conclusions

The complete genome sequences of isolates A23BA and A13BB were generated *de novo* by hybrid assembly of high-quality Illumina and Oxford nanopore reads. Identity of isolate A23BA was confirmed through whole-genome ANI analysis as a

strain of *Hafnia alvei*. Isolate A13BB was preliminarily identified by partial 16S rRNA gene sequence comparison as a strain of *Pseudomonas fragi* or *Pseudomonas psychrophila.* It appears that these two species cannot be resolved by full-length 16S rRNA gene sequence comparison either. However whole-genome ANI analysis confirmed the identity of isolate A13BB as a strain of *P. fragi,* although the accuracy of the taxonomic classification of the reference genome for *P. fragi* species in the NCBI microbial genome database at the time of analysis was doubtful.

The 4.77 Mb genome of *H. alvei* A23BA consists of a circular chromosome (4,687,005 bp) and a circular plasmid (85,042 bp), while the 4.94 Mb genome of *P. fragi* A13BB consists of a single circular chromosome. As both strains have genome sizes greater than the 3 Mb threshold set for this bioprospecting work, they were both suitable candidates for genome mining for secondary metabolite biosynthetic gene clusters. They are also potential producers of metabolites with antibiotic properties given their ecological origin. The contiguous and accurate genomes assembled here will facilitate and expedite exploration of the metabolic and biosynthetic potentials of both strains. In addition to expediting this bioprospecting work, it is believed that the complete genome sequences generated here will serve as valuable publicly accessible resources. Therefore, the datasets have been deposited in the repositories of the International Nucleotide Sequence Database Collaboration.

CHAPTER FOUR

EXPLORATION OF THE METABOLIC AND BIOSYNTHETIC POTENTIALS OF *HAFNIA ALVEI* A23BA AND *PSEUDOMONAS FRAGI* A13BB

4. Exploration of the Metabolic and Biosynthetic Potentials of *Hafnia alvei* A23BA and *Pseudomonas fragi* A13BB

Analysis of genome sequences of bacterial species associated with beneficial metabolic and biosynthetic activities has uncovered molecular mechanisms that underpin these useful attributes. Likewise, analysis of genome sequences of species that are not readily associated with such attributes can uncover hidden potentials or identify novel metabolic pathways.

4.1 Introduction

The immense metabolic and biosynthetic capacities of bacteria have been the subject of numerous studies. These attributes support (or antagonise) other life forms, maintain niche integrity, and drive many biogeochemical processes occurring within ecosystems. Exploitation of these natural processes for beneficial purposes in pharmacotherapy, bioremediation, sustainable agriculture, biotechnology, etc. is expedited by whole-genome analysis to mine for key genetic determinants in order to identify and prioritise beneficial strains. *H. alvei* and *P. fragi* represent species that are currently relatively under- or unexplored for beneficial metabolic and biosynthetic activities, therefore whole-genome analysis of *H. alvei* A23BA and *P. fragi* A13BB may uncover hidden potentials.

H. alvei is a Gram-negative bacterium commonly isolated from clinical materials, gastrointestinal tract of animals, plant surfaces, soil, and water. Its general and biochemical characteristics are outlined in Appendix 3.[237,242] Some strains are commensals of the gastrointestinal tract while others are opportunistic pathogens implicated in both nosocomial and community-acquired infections.[237,243,244] It is almost never associated with secondary metabolite biosynthesis or other beneficial metabolic activities except for the antimicrobial activities reported for a strain isolated from the gut of honeybees.[245] Phylogenetic studies of this little-known species involving 26 strains isolated from different sources showed it has an open pan-genome with each strain possessing sets of unique genes which are acquired mainly by horizontal gene transfer (HGT) from other species within their natural habitats.[237,246] As a result of the open pan-genome, *H. alvei* strains show considerable metabolic pathway diversity and varied biosynthetic potentials.[237,246] Appendices 4 and 5 show the evolutionary relationship between *Hafnia alvei*

strains in the NCBI genome database (accessed in October 2021), and the genomic islands found in the genome of *H. alvei* A23BA, respectively.

P. fragi is a Gram-negative bacterium that is widely distributed in nature and commonly associated with meat and dairy spoilage.[241,247] The general and biochemical characteristics of the species are outlined in Appendix 6. It is rarely reported for its beneficial metabolic activities, except for its plant-growth promotion reported by Selvakumar *et al* and Farh *et al,* and its bioremediation potentials reported by Singha *et al*.[248-250] However, a thorough literature search has not uncovered a previous report of its potential as a beneficial secondary metabolite producer. Phylogenetic studies of 14 strains of *P. fragi* isolated from meat, milk, and soil samples showed a core/pan-genome ratio of 45%, which indicates a high degree of genetic diversity within the species.[241] As the accessory genome of *P. fragi* is quite substantial, it is expected that unique genes acquired through HGT would be common features within it.[170,241] Appendices 7 and 8 show the evolutionary relationship between *P. fragi* strains in the NCBI genome database (accessed in October 2021), and the genomic islands detected in the genome of *P. fragi* A13BB, respectively.

Given the ecological origin of *H. alvei* A23BA and *P. fragi* A13BB, and the fact that both species have dynamic accessory genomes that can accommodate foreign genes acquired through HGT, it is expected that both strains will harbour metabolic and biosynthetic gene clusters that are not readily associated with their taxonomic groups. More importantly, and given the context of this project, both strains are also potential producers of metabolites that can be exploited as antibiotics. Therefore, whole-genome analysis to identify metabolic and biosynthetic gene clusters of interest was the first crucial step taken in assessing the potentials of the strains.

Whole-genome analysis can be performed with various freely available tools and pipelines, with the choice of tools usually determined by research objectives. To meet the objectives of this part of the project, two robust and comprehensive pipelines for genome analysis, i.e. the Rapid Annotations using Subsystems Technology tool kit (RASTtk)[251] and the Antibiotics and Secondary Metabolites Analysis Shell (antiSMASH (bacterial version)),[140] were used to mine for genes or gene clusters that encode enzymes associated with various metabolic and

biosynthetic pathways of interest. RASTtk can also be used to search for other genomic features such as CRISPRs, insertion sequences (IS), and repeat regions, all of which are important indicators of HGT and may also have useful applications in biotechnology.[251]

4.2 Aim and Objectives

The aim of this section of the project was to assess the metabolic and biosynthetic potentials of *H. alvei* A23BA and *P. fragi* A13BB, based on the repertoire of genes in their genomes. This can be achieved by:

(i) Performing whole-genome analysis with RASTtk to identify genes or gene clusters that encode enzymes required for key reactions in metabolic pathways of interest.
(ii) Performing whole-genome analysis using antiSMASH to mine for secondary metabolite biosynthetic gene clusters (smBGCs) that may encode enzymes necessary for the biosynthesis of useful bacterial natural compounds.
(iii) Using integrated tools in antiSMASH for further analysis of selected smBGCs that may encode metabolites with antibiotic properties.

4.3 Methodology

4.3.1 Assessment of Metabolic Potentials of Bacterial Species with RASTtk

RASTtk is a comprehensive genome annotation and analysis pipeline that was introduced as an improvement to the RAST pipeline.[251,252] Upon uploading genomes in FASTA format onto the RAST server, default mode of RASTtk is automatically enabled for annotations and analyses (with the option to use 'Classic RAST' or 'Customised RASTtk' if desired). Firstly, tRNA- and rRNA-encoding genes are called using tRNAscan-SE and the 'rRNA finder' tool, respectively.[251,252,253] Thereafter, Glimmer3 and Prodigal are used to call putative coding sequences (CDSs) i.e. genes that are likely to encode proteins.[254,255] In order to 'determine' putative genes and assign functions to them, RASTtk compares and matches their nucleotide sequences to those of known genes that encode FIGfams; where FIGfams are Fellowship for Interpretation of Genomes' families of proteins derived

from the manually curated library of subsystems in the SEED database. [252,256,257,258] And where subsystems are biological processes or structural complexes implemented by a group of proteins (i.e. enzymes) with each protein performing a defined functional role.[257] Putative genes that cannot be 'determined' and assigned functional roles with FIGfams are then compared and matched to genes encoding proteins found in other curated protein family databases, or compared and matched to genes encoding proteins that contain certain signature protein k-mers.[251,252] After annotations and gene function assignment, two classes of 'asserted gene functions' are automatically created by RASTtk i.e. 'subsystem-based assertions' and 'nonsubsystem-based assertions', with the former being gene function assertions based on subsystems-derived FIGfams.[252] As subsystems are predominantly experimentally characterised and expert-curated and therefore highly reliable, RASTtk uses the 'subsystem-based assertions' as the basis for metabolic reconstruction in which, for each genome analysed, all functional variants of subsystems present in the genome are identified along with all genes encoding the enzymes that implement the functional variants in the organism.[252,257,258]

The very high degree of confidence in analysis results and in the metabolic reconstruction generated by RASTtk as described above stems from the rigour and robustness built into each modular analysis, especially the use of subsystems technology. Therefore, RASTtk represents a public resource that enables rapid and validated preliminary assessments of metabolic potentials of bacterial and archaeal species. However, as metabolic reconstruction with RASTtk is heavily reliant on subsystems technology, any genome that is significantly divergent (at the time of analysis) from the genomes in the SEED database may be under-characterized as a result. In other words, the metabolic processes identified in a genome by RASTtk at a point in time may not necessarily reflect the full metabolic capacity of the species, but rather the metabolic potentials that can be validated at the time of analysis. Whilst this may appear to be a flaw, it is indeed a strength as assertion of metabolic potential without evidence from wet lab experiments should only be based on robust and validated *in silico* methods. This 'flaw' is being addressed as the SEED database is constantly updated to include functional variants of subsystems curated from genomes of divergent strains as these become available.[252,258]

4.3.2 Genome Mining with antiSMASH to Assess the Potentials of Bacterial Species to Biosynthesise Secondary Metabolites

The genes encoding enzymes required for the biosynthesis of a specific secondary metabolite are usually co-located in a region within the bacterial genome, constituting the secondary metabolite biosynthetic gene cluster (smBGC) for that biosynthetic pathway. Furthermore, each biosynthetic pathway has unique core or signature enzymes that catalyse key reactions of the pathway. These important features form the basis of genome mining where entire bacterial genomes are scanned to detect regions of the genomes harbouring co-located genes that encode proteins involved in secondary metabolite biosynthesis.

antiSMASH remains the most comprehensive genome mining pipeline to date. The latest version at the time of analysis (v5.0) can detect 52 types of smBGCs in bacterial genomes.[140] Upon loading a bacterial genome onto the antiSMASH webserver, the amino acid sequence translation of all CDSs detected in the genome is searched against a library of manually curated profile hidden Markov models (pHMMs) that are constructed based on multiple sequence alignments of experimentally characterised signature proteins (i.e. conserved core enzymes) unique to each of the 52 smBGC types detectable with antiSMASH.[140,259] A smBGC is predicted when a stretch of nucleotides contains clusters of genes that encode signature proteins found in a particular smBGC type all spaced within <10 kb mutual distance.[259] The predicted smBGC region will usually include accessory genes adjacent (usually up to 5-20 kb distance depending on the type of smBGC) to the last signature gene on either side of the stretch of nucleotides.[259] This rule-based smBGC detection means that antiSMASH predictions have high confidence and low false-positive ratings, but the pipeline (webserver version) is unable to detect novel smBGCs as sequence homology to known and experimentally curated core proteins is needed for detection.[140] However, downloadable standalone versions of antiSMASH include scripts for the probabilistic ClusterFinder algorithm that uses statistical methods for BGC detection and therefore can detect novel smBGCs, albeit with high false-positive results which often include some primary metabolite BGCs.[128,140,260]

After smBGC detection, antiSMASH incorporates various modules for further downstream analyses. The KnownClusterBlast module searches the MIBiG database for experimentally characterised and manually curated smBGCs that

show similarities to detected smBGCs in antiSMASH to aid dereplication.[169,261] smCOG analysis is enabled in antiSMASH, where each gene in a smBGC is compared to those in a database of manually curated clusters of orthologous groups (COG) of proteins that are involved in secondary metabolism. This analysis seeks to predict the functional roles of proteins encoded within detected smBGCs.[259] A module is also included in the pipeline which uses pHMMs to detect antibiotic resistance genes (ARGs) in smBGCs as these may indicate that the smBGC may be antibiotic-encoding.[140] However, inability to detect ARGs in a smBGC will not necessarily imply the smBGC is not antibiotic-encoding as sequence homology to known ARGs is a prerequisite for ARG detection.

Although antiSMASH v5.0 can detect 52 smBGC types, not all smBGC types are drug-like.[260] Drug-like smBGCs are likely to encode metabolites that have properties that can be exploited for medicinal purposes.[260] Example of smBGCs that are not generally considered to be drug-like include those encoding melanin, indole, ectoine, butyrolactones, etc.[260] Examples of drug-like smBGCs on the other hand will include those encoding polyketides (PKs), non-ribosomal peptides (NRPs), various PK-NRP hybrids, ribosomally synthesised and post translationally modified peptides (RiPPs), aminoglycosides, etc.[260] Furthermore, Baltz observed that drug-like smBGCs are generally large (average size 54 kb) and are typically PKs, NRPs, or PK-NRP hybrid smBGC types.[260] The author's observation was drawn from a selection of 50 drug-like smBGCs with sizes ranging from 10.5 to 143.9 kb. However, it is worth noting that all 50 smBGCs selected in the study are encoded in actinomycetes which tend to have large (> 7 Mb) genomes and therefore can devote up to 3 Mb genome capacity to secondary metabolism.[138] As such these strains tend to harbour large smBGCs.[260] It will be reasonable however to expect other species with small- to medium-sized genomes to harbour drug-like smBGCs that are generally < 54 kb in size. Examples of which include spectinomycin, chloramphenicol, and gentamicin encoded by smBGCs which are 18.4, 22, and 32.6 kb in size, respectively.[169,262,263] A comprehensive genome mining effort would therefore seek to prioritise drug-like smBGCs (regardless of size) for activation, especially if the end products are also likely to be novel.

4.4 Methods

4.4.1 Whole-genome Analysis to Identify Metabolic Pathways of Interest Encoded in the Genomes of *H. alvei* A23BA and *P. fragi* A13BB

The complete genome sequences of *H. alvei* A23BA and *P. fragi* A13BB were uploaded onto the RAST server (v2.0) in FASTA formats. Annotations and metabolic reconstructions were performed with RASTtk using default parameters and with the options 'automatically fix errors' and 'build metabolic models' checked. After metabolic reconstructions, each subsystem category and subcategory identified in the genomes were expanded down to individual subsystem levels. The spreadsheets of subsystems implementing metabolic pathways of interest (i.e. metabolism of aromatic compounds, stress response, iron acquisition and metabolism, etc) were viewed to determine the functional variants present in the genomes and to collate all associated genes/gene clusters.

Analysis of *H. alvei* A23BA genome was carried out in June 2020, and *P. fragi* A13BB genome in December 2020.

4.4.2 Genome Mining to Identify smBGCs of Interest Encoded in the Genomes of *H. alvei* A23BA and *P. fragi* A13BB

The complete genome sequences of *H. alvei* A23BA and *P. fragi* A13BB were uploaded onto the antiSMASH bacterial version webserver (v5.0) in FASTA formats. Genome mining was carried out in 'relaxed mode' with default parameters and with the following modules enabled: KnownClusterBlast analysis, SubClusterBlast analysis, smCOG analysis, and Active site finder. After smBGCs detection, selected drug-like smBGCs were further analysed to predict end products of genes contained within the clusters and their functions.

Genome mining of *H. alvei* A23BA was carried out in June 2020, and *P. fragi* A13BB was carried out in December 2020.

4.5 Results

4.5.1 Assessment of Metabolic Potentials of *H. alvei* A23BA and *P. fragi* A13BB

The closest species to *H. alvei* in the SEED database at the time of analysis was *Serratia proteamaculans*. This significant divergence may account for the low

percentage of genes associated with subsystems in *H. alvei* A23BA genome (Table 4.1 and Figure 4.1). Gene/ gene clusters of subsystem variants potentially associated with bioremediation, biocontrol, plant-growth promotion, and environmental adaptation were collated and presented in Table 4.2.

Table 4.1: Genomic Features of *H. alvei* A23BA Identified by RASTtk

Feature	Count
Coding sequences (genes)	4444
Subsystems	364
Genes connected to subsystems	1432 (33%)
Genes not connected to subsystems	3012 (67%)
tRNA	92
rRNA	25
CRISPRs	2
Repeat regions	82

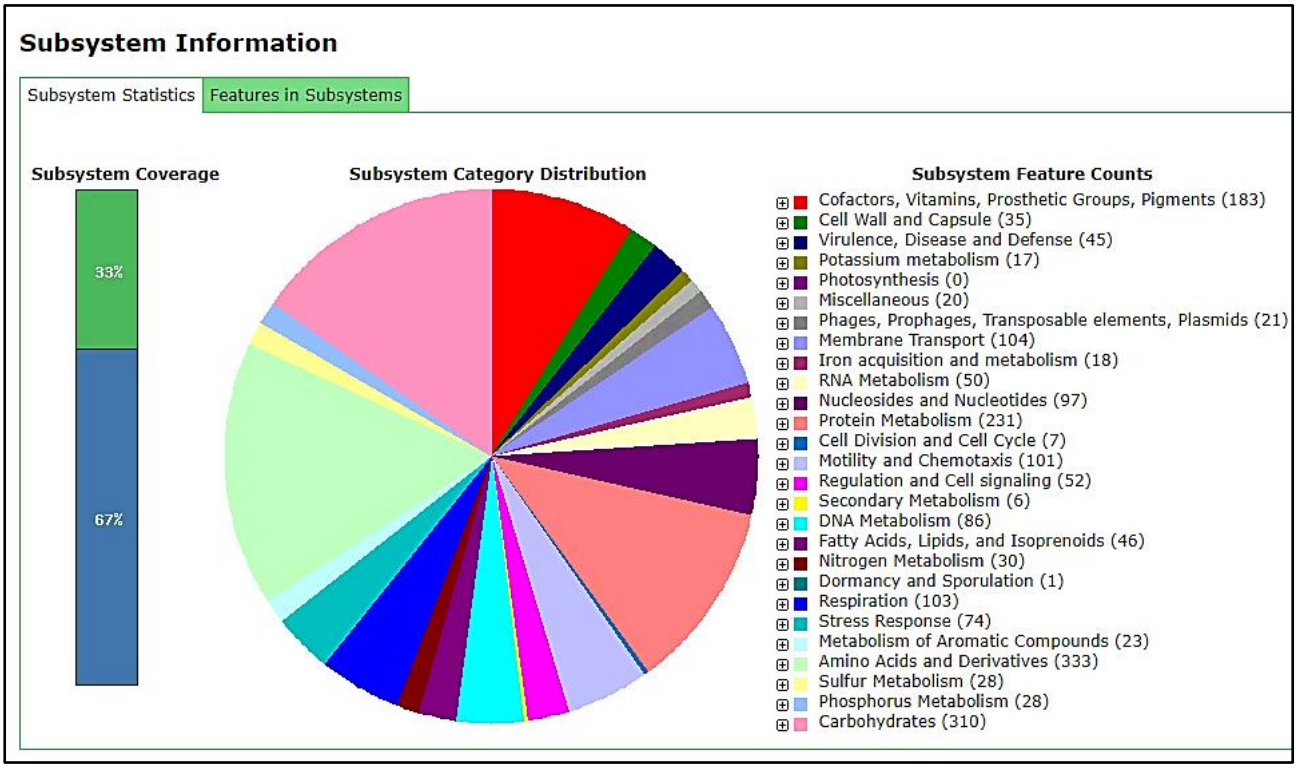


Figure 4.1: Subsystem categories encoded in *H. alvei* A23BA.
The significant divergence of *H. alvei* from species in the SEED database meant that only 33% of genes found in the genome of *H. alvei* A23BA are associated with subsystems. Potentially beneficial metabolic processes/subsystems identified include those associated with stress response, metabolism of aromatic compounds, and disease and defence.

Table 4.2: Genes / Metabolic Pathways of Interest in *H. alvei* A23BA

Genes potentially associated with environmental adaptation, bioremediation, biocontrol, and plant-growth promotion in *H. alvei* A23BA		
Pathway	**Gene**	**Gene Product**
copper resistance	*cop*C	copper resistance protein
	*cop*D	copper resistance protein
	bco	blue copper oxidase *cue*O precursor
	*cue*A	copper-translocating P-type ATPase
	*scs*A	suppressor for copper sensitivity; copper binding protein
	*scs*B	suppressor for copper sensitivity; membrane binding protein
	*scs*C	suppressor for copper sensitivity; secreted protein
	*scs*D	suppressor for copper sensitivity; membrane protein
	*cut*A	periplasmic divalent cation tolerance protein
	*cut*C	cytoplasmic copper homeostasis protein
	*cut*F	copper homeostasis protein
	*cut*E	copper homeostasis protein
cobalt-zinc-cadmium resistance	*tr*Mer	transcriptional regulator; MerR family
quinate degradation	*qui*B2	3-dehydroquinate dehydratase II
biphenyl degradation	*bph*C	biphenyl-2,3-diol 1,2-dioxygenase III-related protein
	*bph*E1	acetaldehyde dehydrogenase
4-hydroxyphenylacetic acid catabolism	*hpa*X	4-hydroxyphenylacetate symporter
	*hpa*A	transcriptional activator of 4-hydroxyphenylacetate 3-monooxygenase operon
	*hpa*C	4-hydroxyphenylacetate 3-monooxygenase, reductase component
	*hpa*H	2-oxo-hepta-3-ene-1,7-dioic acid hydratase
	*hpa*F	5-carboxymethyl-2-hydroxymuconate delta-isomerase
	*hpa*D	3,4-dihydroxyphenylacetate 2,3-dioxygenase
	*hpa*E	5-carboxymethyl-2-hydroxymuconate semialdehyde dehydrogenase
	*hpa*G1	2-hydroxyhepta-2,4-diene-1,7-dioate isomerase
	*hpa*G2	5-carboxymethyl-2-oxo-hex-3- ene-1,7-dioate decarboxylase
	*hpa*R	homoprotocatechuate degradative operon repressor
central meta-cleavage pathway of aromatic compound degradation	*hpc*B	3,4-dihydroxyphenylacetate 2,3-dioxygenase
	*hpc*H	2-oxo-hepta-3-ene-1,7-dioic acid hydratase
	*hpc*C	5-carboxymethyl-2-hydroxymuconate semialdehyde dehydrogenase
	*hpc*D	5-carboxymethyl-2-hydroxymuconate delta-isomerase
aromatic amine catabolism	*mao*A	monoamine oxidase
	*hpc*B	3,4-dihydroxyphenylacetate 2,3-dioxygenase
	*fea*B	phenylacetaldehyde dehydrogenase
	nacb	nitrilotriacetate monooxygenase component B
	*hpa*C	4-hydroxyphenylacetate 3-monooxygenase, reductase component
glutathione-dependent pathway of formaldehyde detoxification	*frm*A	S-(hydroxymethyl)glutathione dehydrogenase
	*frm*B	S-formylglutathione hydrolase
glutathione-dependent xenobiotic degradation	*gst*O	glutathione S-transferase, omega
	gst	glutathione S-transferase
	*glo*A	lactoylglutathione lyase

Table 4.2 continued

Pathway	Gene	Gene Product
	*glo*B	hydroxyacylglutathione hydrolase
	*sam*1	SAM-dependent methyltransferase
	gsr	glutathione reductase
nitrogen metabolism	*nir*1a	nitrite reductase [NAD(P)H] large subunit
	*nir*1b	nitrite reductase [NAD(P)H] small subunit
	*nir*C	nitrite transporter
	*nar*G	respiratory nitrate reductase alpha chain
	*nar*H	respiratory nitrate reductase beta chain
	*nar*I	respiratory nitrate reductase gamma chain
	*nar*J	respiratory nitrate reductase delta chain
	*nap*B	nitrate reductase cytochrome c550-type subunit
	*nap*C	cytochrome c-type protein NapC
	*nap*F	ferredoxin-type protein NapF
	*nap*G	ferredoxin-type protein NapG (periplasmic nitrate reductase)
	*nap*H	polyferredoxin NapH (periplasmic nitrate reductase)
ferric ion uptake	*Fhu*A	ferric hydroxamate outer membrane receptor
	*Fhu*D	ferric hydroxamate ABC transporter, periplasmic substrate binding protein
	*Fhu*C	ferric hydroxamate ABC transporter, ATP-binding protein
	*Fhu*B	ferric hydroxamate ABC transporter, permease component
plant alkaloid biosynthesis from L-lysine	*oxc*	oxalyl-CoA decarboxylase
alkanesulfonate assimilation	*ssu*A	alkanesulfonates-binding protein
	sox	dibenzothiophene desulfurization enzyme
	ars	arylsulfatase
phosphate metabolism	*pho*U	phosphate transport system regulatory protein
	*pho*R	phosphate regulon sensor protein
	*pho*B	phosphate regulon transcriptional regulatory protein
	ppk	polyphosphate kinase
	ppx	exopolyphosphatase
	gdp	guanosine-5'-triphosphate,3'-diphosphate pyrophosphatase
	*pho*A	alkaline phosphatase
	ppa	inorganic pyrophosphatase
	*pho*H	phosphate starvation-inducible protein

RASTtk analysis of *P. fragi* A13BB genome also reflected its divergence from the entries in the SEED database at the time of analysis, with the closest species to *P. fragi* being *P. fluorescens.* Hence the low percentage of detected genes associated with subsystems (Table 4.3 and Figure 4.2) in the genome of *P. fragi* A13BB. Gene/ gene clusters of subsystem variants potentially associated with bioremediation, biocontrol, plant-growth promotion, and environmental adaptation were also collated and presented in Table 4.4.

Table 4.3: Genomic Features of *P. fragi* A13BB Identified by RASTtk

Feature	Count
Coding sequences (genes)	4528
Subsystems	353
Genes connected to subsystems	1401 (31%)
Genes not connected to subsystems	3127 (69%)
tRNA	73
rRNA	25
CRISPRs	0
Repeat regions	32

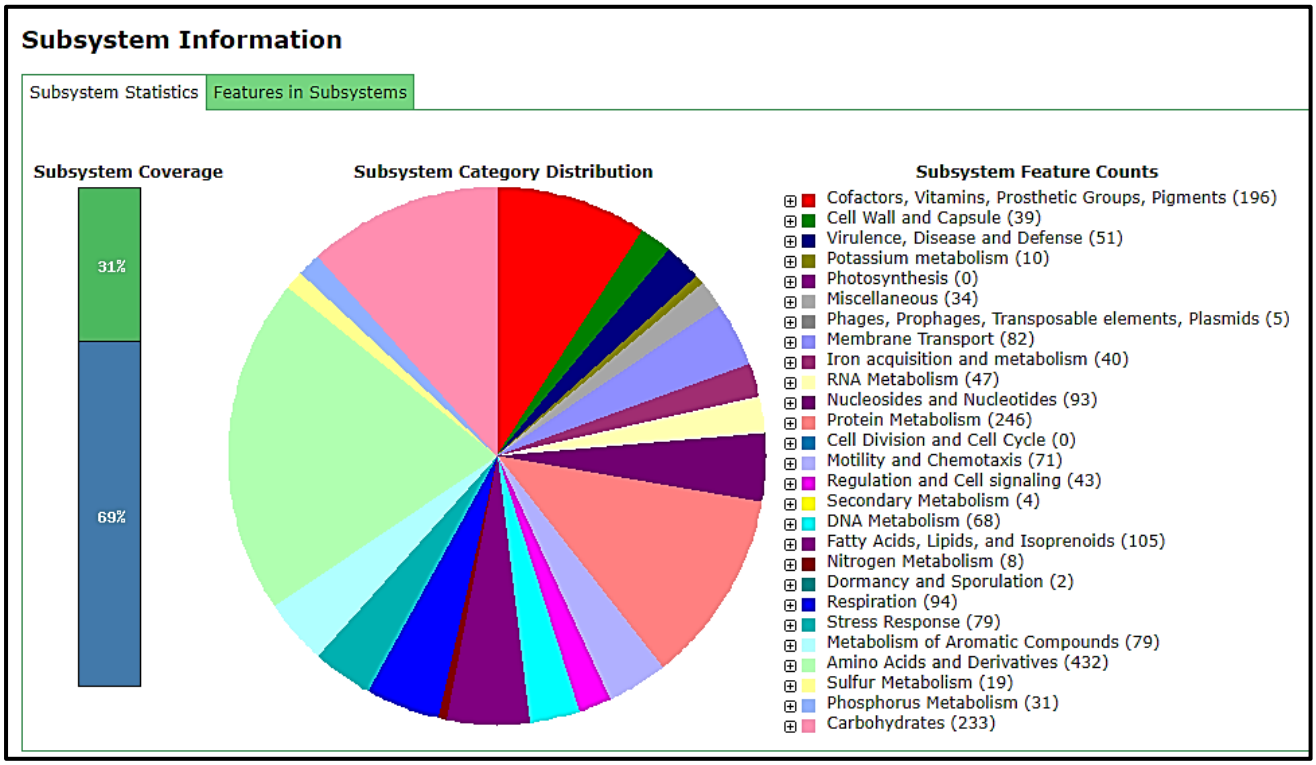


Figure 4.2: Subsystem categories encoded in *P. fragi* A13BB.
The significant divergence of *P. fragi* from species in the SEED database meant that only 31% of genes found in the genome of *P. fragi* A13BB are associated with subsystems. Potentially beneficial metabolic processes/subsystems identified include those associated with stress response, metabolism of aromatic compounds, and disease and defence.

Table 4.4: Genes / Metabolic Pathways of Interest in *P. fragi* A13BB

Genes potentially associated with environmental adaptation, bioremediation, biocontrol, and plant-growth promotion in *P. fragi* A13BB		
Pathway	**Gene**	**Gene Product**
copper resistance	*copG*	copG protein
	*cop*D	copper resistance protein
	*cop*Z	copper chaperone
	*cue*A	copper-translocating P-type ATPase
	mo	multicopper oxidase
	ct	copper tolerance protein
	*clf*A	multidrug resistance transporter, Bcr/CflA family
	*ccm*H	cytochrome c heme lyase subunit CcmH
	*ccm*F	cytochrome c heme lyase subunit CcmF
	*cus*S	copper sensory histidine kinase CusS
	*cus*R	copper-sensing two-component system response regulator CusR
	*cut*E	copper homeostasis protein
	*cor*C	magnesium and cobalt efflux protein CorC
cobalt-zinc-cadmium resistance	*tr*Mer	transcriptional regulator; MerR family
	*czc*R	cobalt-zinc-cadmium resistance protein
	fp	probable Co/Zn/Cd efflux system membrane fusion protein
	*czr*R	DNA-binding heavy metal response regulator
	hmhk	heavy metal sensor histidine kinase
	*cad*R	Cd(II)/Pb(II)-responsive transcriptional regulator
	hmrR	heavy metal resistance transcriptional regulator HmrR
quinate degradation	*qui*B2	3-dehydroquinate dehydratase II
biphenyl degradation	*bph*C	2,3-dihydroxybiphenyl 1,2-dioxygenase
	*bph*F	4-hydroxy-2-oxovalerate aldolase
n-phenylalkanoic acid degradation	*fad*A	3-ketoacyl-CoA thiolase
	*fad*B	enoyl-CoA hydratase
	*fad*D	long-chain-fatty-acid--CoA ligase
	*pha*J1	enoyl-CoA hydratase, R-specific
	*fad*B2	delta(3)-cis-delta(2)-trans-enoyl-CoA isomerase
	*fad*B3	3-hydroxyacyl-CoA dehydrogenase
	*fad*B4	3-hydroxybutyryl-CoA epimerase
benzoate degradation	*ben*B	benzoate 1,2-dioxygenase beta subunit
	*ben*A	benzoate 1,2-dioxygenase alpha subunit
	*ben*C	benzoate 1,2-dioxygenase, ferredoxin reductase component
	*ben*D	1,2-dihydroxycyclohexa-3,5-diene-1-carboxylate dehydrogenase
	*ben*E2	benzoate transport protein
	*ben*R	benABC operon transcriptional activator BenR
p-hydroxybenzoate degradation	*pob*A	p-hydroxybenzoate hydroxylase
	*pca*K	4-hydroxybenzoate transporter
beta-ketoadipate pathway of aromatic compound degradation	*cat*A	catechol 1,2-dioxygenase
	*cat*B	muconate cycloisomerase
	*cat*C	muconolactone isomerase
	*cat*D	beta-ketoadipate enol-lactone hydrolase
	*cat*EA	3-oxoadipate CoA-transferase subunit A
	*cat*EB	3-oxoadipate CoA-transferase subunit B
	*cat*A2	catechol 1,2-dioxygenase 1
	*pca*R	Pca regulon regulatory protein PcaR

Table 4.4 continued

Pathway	Gene	Gene Product
	*pca*H	protocatechuate 3,4-dioxygenase beta chain
	*pca*G	protocatechuate 3,4-dioxygenase alpha chain
	*pca*B	3-carboxy-cis,cis-muconate cycloisomerase
	*pca*C	4-carboxymuconolactone decarboxylase
	*pca*I	succinyl-CoA:3-ketoacid-coenzyme A transferase subunit A
	*pca*J	succinyl-CoA:3-ketoacid-coenzyme A transferase subunit B
	*pca*T	dicarboxylic acid transporter PcaT
homogentisate pathway of aromatic compound degradation	*hmg*A	homogentisate 1,2-dioxygenase
	hppd	4-hydroxyphenylpyruvate dioxygenase
	mai	maleylacetoacetate isomerase
	faa	fumarylacetoacetase
	aaa	aromatic-amino-acid aminotransferase
	*hmg*R	transcriptional regulator, IclR family
gentisate degradation	*fhf*	fumarylacetoacetate hydrolase family protein
	mai	maleylacetoacetate isomerase
	pca	4-hydroxybenzoate transporter
glutathione-dependent pathway of formaldehyde detoxification	*frm*A	S-(hydroxymethyl)glutathione dehydrogenase
	*frm*B	S-formylglutathione hydrolase
	*frm*F	transcriptional regulator, LysR family, in formaldehyde detoxification operon
glutathione-dependent xenobiotic degradation	*gst*Z	glutathione S-transferase, zeta
	gst	glutathione S-transferase
	*gst*U	glutathione S-transferase, unnamed subgroup
	*glo*A	lactoylglutathione lyase
	*glo*B	hydroxyacylglutathione hydrolase
	*sam*1	SAM-dependent methyltransferase
	gsr	glutathione reductase
	gpx	glutathione peroxidase
	grx3	glutaredoxin 3
	grx	glutaredoxin
	*gsh*A	glutamate--cysteine ligase
	*gsh*B	glutathione synthetase
	glt	gamma-glutamyltranspeptidase
choline and betaine uptake, and betaine biosynthesis	*bet*A	choline dehydrogenase
	*bet*B	betaine aldehyde dehydrogenase
	*bet*T	high-affinity choline uptake protein BetT
	*sox*A	sarcosine oxidase alpha subunit
	*sox*B	sarcosine oxidase beta subunit
	*sox*D	sarcosine oxidase delta subunit
	*sox*G	sarcosine oxidase gamma subunit
	*gbc*A	GbcA Glycine betaine demethylase subunit A
	*gbc*B	GbcB Glycine betaine demethylase subunit B
nitrogen metabolism	*nor*R	anaerobic nitric oxide reductase transcription regulator NorR
	*gln*E	glutamate-ammonia-ligase adenylyltransferase

Table 4.4 continued

Pathway	Gene	Gene Product
	amt	ammonium transporter
	*gs*1	glutamine synthetase type I
	*gln*D	[Protein-PII] uridylyltransferase
	*glt*B	glutamate synthase [NADPH] large chain
	*glt*D	glutamate synthase [NADPH] small chain
	gogat	ferredoxin-dependent glutamate synthase
ferric ion uptake	*pvs*A	vibrioferrin ligase/carboxylase protein PvsA
	*pvs*B	vibrioferrin amide bond forming protein PvsB
	*pvs*C	vibrioferrin membrane-spanning transport protein PvsC
	*pvs*D	vibrioferrin amide bond forming protein PvsD
	feat	iron-chelator utilization protein
	*iut*A	aerobactin siderophore receptor IutA
sulphur metabolism	*ssu*A	alkanesulfonates-binding protein
	*ssu*F	organosulfonate utilization protein SsuF
	*ssu*C	alkanesulfonates transport system permease protein
	dde	probable dibenzothiophene desulfurization enzyme
	pc	ABC-type nitrate/sulfonate/bicarbonate transport systems, periplasmic components
	td	alpha-ketoglutarate-dependent taurine dioxygenase
	ars	arylsulfatase
	*trx*R	thioredoxin reductase
	tpx	thiol peroxidase, Tpx-type
	bcp	thiol peroxidase, Bcp-type
	*ahp*C	alkyl hydroperoxide reductase subunit C-like protein
	as	arylsulfatase
phosphate metabolism	*pho*U	phosphate transport system regulatory protein
	*pho*R	phosphate regulon sensor protein
	*pho*B	phosphate regulon transcriptional regulatory protein
	ppk	polyphosphate kinase
	ppx	exopolyphosphatase
	phn-reg	phosphonate uptake and metabolism regulator, LysR-family
	*pho*A	alkaline phosphatase
	*phn*W	2-aminoethylphosphonate:pyruvate aminotransferase
	*phn*X	phosphonoacetaldehyde hydrolase
	*pho*H	phosphate starvation-inducible protein
	*th*B	NAD(P) transhydrogenase subunit beta
	*pho*Q	response regulator in two-component regulatory system with PhoQ
	*th*2	soluble pyridine nucleotide transhydrogenase
	lat	low-affinity inorganic phosphate transporter
	napi	sodium-dependent phosphate transporter

4.5.2 Assessment of Potentials of *H. alvei* A23BA and *P. fragi* A13BB to Biosynthesise Secondary Metabolites

4.5.2.1 antiSMASH Outputs

antiSMASH analysis of the genomes of *H. alvei* A23BA and *P. fragi* A13BB detected 4 and 5 smBGCs, respectively (Figures 4.3 and 4.4). Thiopeptide, siderophore, and β-lactone smBGCs are drug-like and suitable for further exploration. However, as β-lactone smBGCs are common to both genomes, with both showing little or no homology to known smBGCs (Figures 4.5, 4.6, and 4.7), they were selected for further analysis in the first instance.

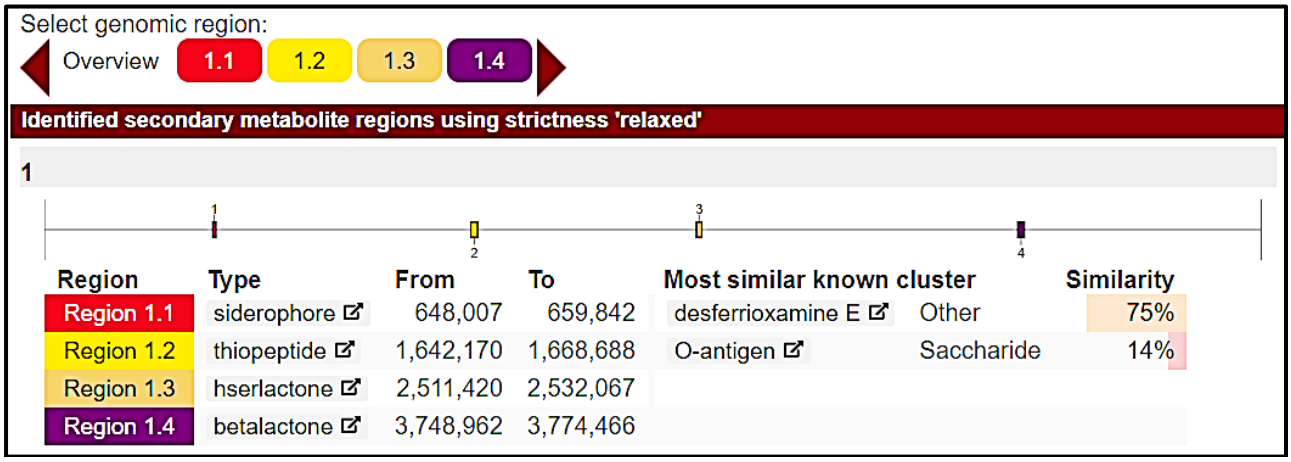


Figure 4.3: Overview of all detected smBGCs in the genome of *H. alvei* A23BA.
Overview of the four smBGC types detected in the genome of *H. alvei* A23BA and the degree of similarity to known smBGCs.

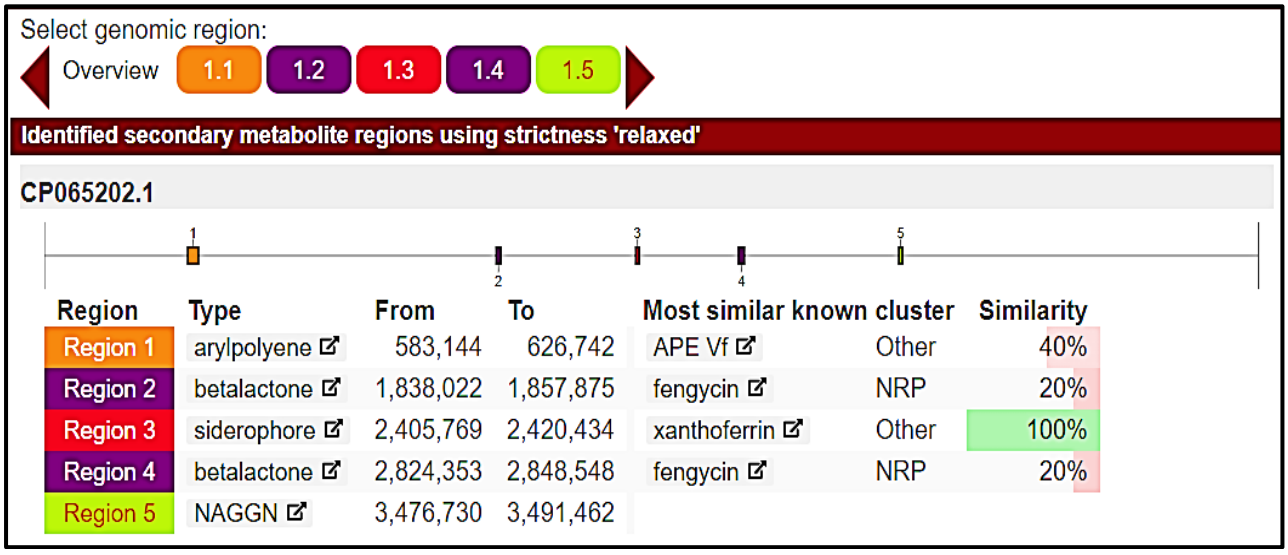


Figure 4.4: Overview of all detected smBGCs in the genome of *P. fragi* A13BB.
Overview of the five smBGC types detected in the genome of *P. fragi* A13BB and the degree of similarity to known smBGCs.

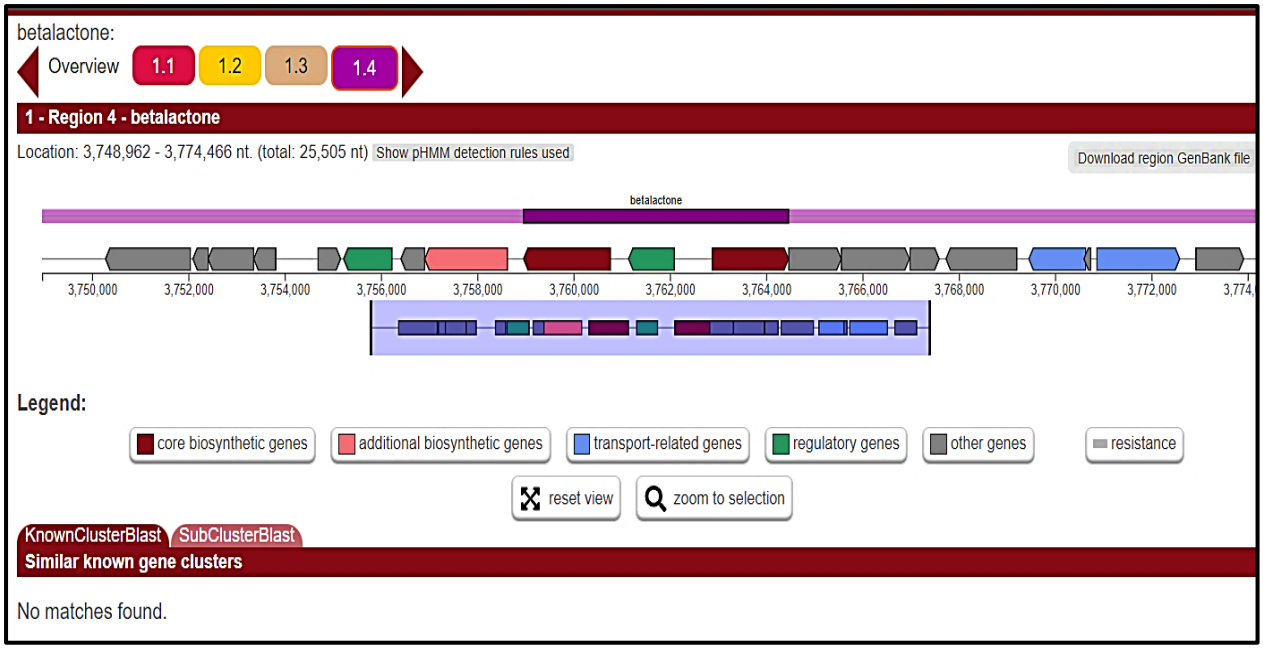


Figure 4.5: Detailed view of the β-lactone smBGC detected in *H. alvei* A23BA.
Colour-coded illustration of the β-lactone smBGC in *H. alvei* A23BA. The 25,505 bp long cluster contains 2 core biosynthetic genes, 1 additional biosynthetic gene, 2 transport genes, 2 regulatory genes, and 12 other genes. The cluster shows no similarity to known smBGCs.

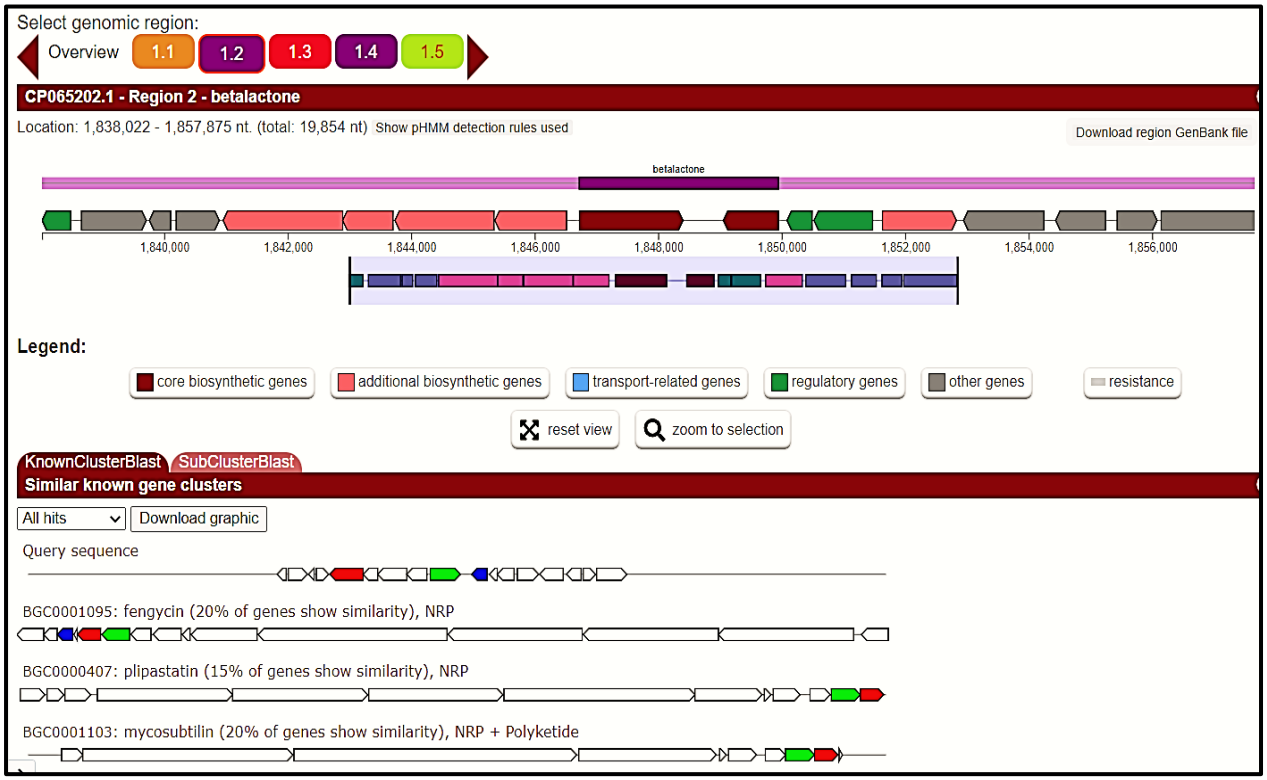


Figure 4.6: Detailed view of the β-lactone smBGC (1) detected in *P. fragi* A13BB.
Colour-coded illustration of β-lactone smBGC (1) in *P. fragi* A13BB. The 19,854 bp long cluster contains 2 core biosynthetic genes, 5 additional biosynthetic gene, 3 regulatory genes, and 7 other genes. The cluster shows low similarity (15-20%) to known smBGCs.

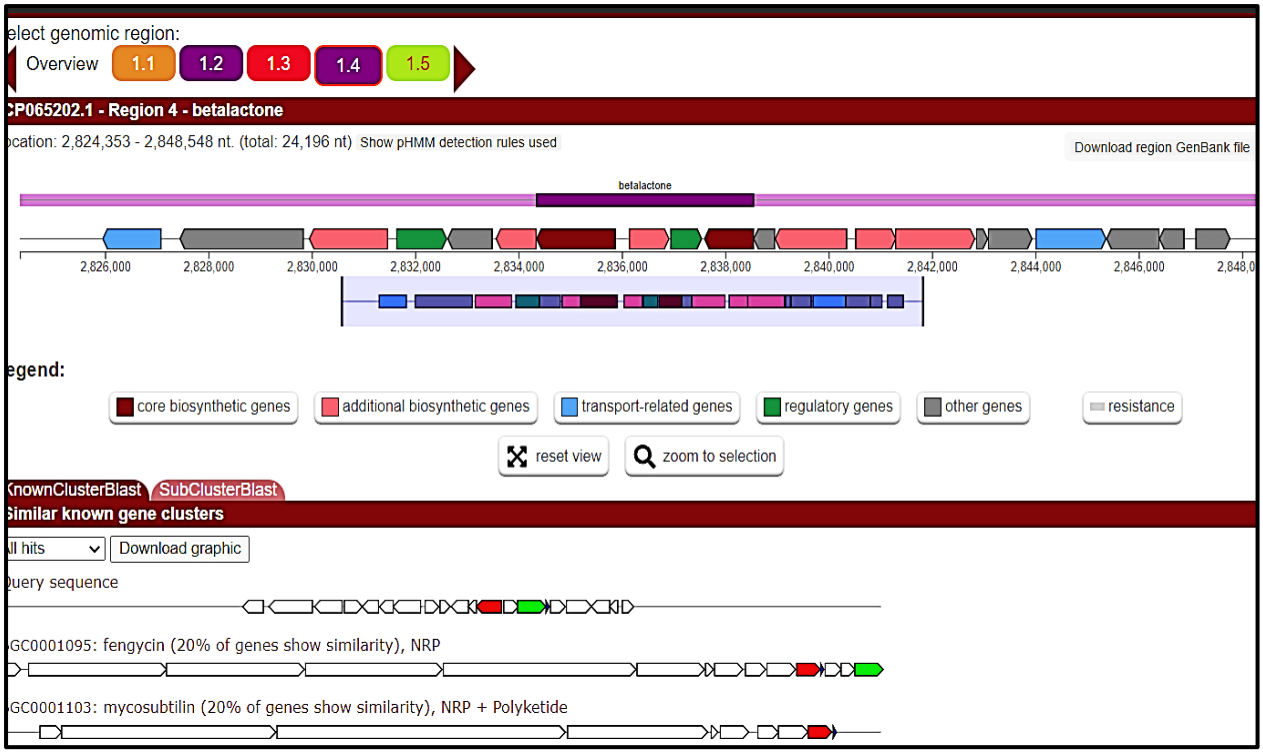


Figure 4.7: Detailed view of the β-lactone smBGC (2) detected in *P. fragi* A13BB.
Colour-coded illustration of β-lactone smBGC (2) in *P. fragi* A13BB. The 24,196 bp long cluster contains 2 core biosynthetic genes, 6 additional biosynthetic gene, 2 transport genes, 2 regulatory genes, and 8 other genes. The cluster shows low similarity (20%) to known smBGCs.

4.5.3 smCOG and BlastP Analyses of the β-Lactone smBGCs in *H. alvei* A23BA and *P. fragi* A13BB

Table 4.5: Analysis of the β-Lactone smBGC in *H. alvei* A23BA

smBGC Features: 25,505 nucleotides; 19 genes.		
Gene	**smCOG description of gene product**	**BlastP description of gene product (% Identity)**
A	unknown	peptidoglycan glycosyltransferase (100)
B	unknown	cell division protein (100)
C	unknown	SAM-dependent methyltransferase (100)
D	unknown	transcriptional repressor MraZ (100)
E	unknown	L-alanine exporter (100)
F	transcriptional regulator LacI family	catabolite repressor/activator (99.7)
G	unknown	acetolactate synthase SS (100)
H	pyruvate oxidase/decarboxylase	acetolactate synthase LS (100)
I	AMP-dependent synthetase & ligase (core enzyme)	acyl CoA ligase/synthetase (100)
J	transcriptional regulator LysR family	transcriptional regulator LeuO (100)

Table 4.5 continued		
Gene	**smCOG description of gene product**	**BlastP description of gene product (% Identity)**
K	2-isopropylmalate synthase (core enzyme)	2-isopropylmalate synthase (100)
L	unknown	3-isopropylmalate dehydrogenase (100)
M	unknown	3-isopropylmalate dehydratase LS (100)
N	unknown	3-isopropylmalate dehydratase SS (100)
O	unknown	DASS family anion symporter (100)
P	major facilitator transporter	efflux transporter (99.74)
Q	unknown	unknown
R	transport protein	transcriptional regulator SgrR (100)
S	unknown	Thiamine ABC transporter (99.70)

Table 4.6: Analysis of the β-Lactone smBGC (1) in *P. fragi* A13BB

smBGC Features: 19,854 nucleotides; 17 genes.		
Gene	**smCOG description of gene product**	**BlastP description of gene product (% Identity)**
A	transcriptional regulator AsnC family	transcriptional regulator AsnC family (100)
B	unknown	glu/leu/phe/val dehydrogenase (100)
C	unknown	duf3077 domain-containing protein (100)
D	unknown	transcriptional repressor LexA (100)
E	acetyl-CoA carboxylase (biotin carboxylase)	acetyl-CoA carboxylase subunit alpha (99.85)
F	enoyl-CoA hydratase	enoyl-CoA hydratase (100)
G	acetyl-CoA carboxylase (carboxyl transferase)	methylcrotonyl-CoA carboxylase (100)
H	acyl-CoA dehydrogenase	isovaleryl-CoA dehydrogenase (100)
I	AMP-dependent synthetase and ligase (core enzyme)	acyl CoA ligase/synthetase (100)
J	2-isopropylmalate synthase (core enzyme)	isopropylmalate synthase (100)
K	transcriptional regulator MerR family	transcriptional regulator MerR family (100)
L	transcriptional regulator LysR family	transcriptional regulator LysR family (100)
M	acyl-CoA dehydrogenase	acyl-CoA dehydrogenase (100)
N	unknown	phosphate ABC transporter (100)
O	unknown	exodeoxyribonuclease III (100)
P	unknown	acetyl transferase GNAT family (99.54)
Q	unknown	autotransporter protein TamA (99.83)

Table 4.7: Analysis of the β-Lactone smBGC (2) in *P. fragi* A13BB

smBGC Features: 24,196 nucleotides; 20 genes.		
Gene	**smCOG description of gene product**	**BlastP description of gene product (% Identity)**
A	RND family efflux transporter MFP subunit	efflux RND transporter (100)
B	unknown	MB PQQ-dependent dehydrogenase (99.62)
C	aldehyde dehydrogenase	aldehyde dehydrogenase (100)
D	transcriptional regulator LysR family	transcriptional regulator LysR family (100)
E	unknown	acyl-CoA thioesterase II (100)
F	enoyl-CoA hydratase	enoyl-CoA hydratase (100)
G	AMP-dependent synthetase and ligase (core enzyme)	acyl CoA synthetase/ligase (100)
H	short-chain dehydrogenase/reductase	glucose 1-dehydrogenase (99.61)
I	transcriptional regulator TetR family	transcriptional regulator TetR family (99.47)
J	2-isopropylmalate synthase (core enzyme)	isopropylmalate synthase (100)
K	unknown	cupin domain-containing protein (100)
L	acetyl-CoA carboxylase (biotin carboxylase)	acetyl-CoA carboxylase (biotin carboxylase) (99.78)
M	enoyl-CoA hydratase	SDR family oxidoreductase (99.61)
N	acetyl-CoA carboxylase (carboxyl transferase)	acyl-CoA carboxylase subunit beta (100)
O	unknown	acetyl-CoA carboxylase biotin carrier protein (100)
P	unknown	fumarylacetoacetate hydrolase protein (100)
Q	major facilitator transporter	MFS transporter (100)
R	unknown	flavin-dependent oxidoreductase (99.70)
S	unknown	transport protein (100)
T	unknown	hypothetical protein (100)

4.6 Discussion

Assessment of the metabolic potentials of *H. alvei* A23BA with RASTtk identified a total of 4, 444 coding sequences of which 1, 432 (33%) encode proteins associated with 364 subsystems encoded in its genome (Table 4.1). As expected, the total number of subsystems and associated genes in *H. alvei* A23BA predictable with RASTtk were relatively low given that the three closest neighbours to *H. alvei* in the SEED database at the time of analysis were *Serratia proteamaculans*, *Erwinia*

carotovora, and *Edwardsiella tarda*. Nonetheless, a metabolic reconstruction with RASTtk provided a robust (albeit incomplete) picture of the organism's metabolic profile. There are 27 subsystems categories in the RASTtk metabolic reconstruction output file which cover all metabolic processes and structural complexes encoded in an organism (Figure 4.1). Some of these categories (e.g. virulence, disease, and defence; iron acquisition and metabolism; nitrogen metabolism; metabolism of aromatic compounds, etc.) include metabolic processes that confer fitness, environmental adaptation, and competitive advantage on an organism. These natural processes can be exploited for useful purposes such as bioremediation, biocontrol, and plant-growth promotion.[172,173,264] Genes encoded in the genome of *H. alvei* A23BA that are associated with metabolic processes that can be exploited for useful purposes are listed in Table 4.2. Of note are genes involved in the degradation of heavy metals, xenobiotics, and other aromatic compounds. These genes are usually present in the genomes of bacterial species commonly employed in bioremediation, which is a cheap and environment-friendly detoxification process that is gaining increasing importance given the negative impacts of environmental pollutants.[264] Additionally, genes that mediate plant-growth promotion and biocontrol (i.e. genes mediating heavy metal resistance, ferric ion uptake, nitrogen and phosphate metabolism, etc.) were also found in the genome and listed in Table 4.2. All these genetic attributes are not associated with *H. alvei* species and as such were reported for the first time for strain A23BA.[237] Other genomic features identified by RASTtk include two CRISPRs and 82 repeat regions (Table 4.1). CRISPRs and associated proteins encoded in some bacterial genomes mediate adaptive immunity against invading viral genetic elements.[265,266] Furthermore, endogenous CRISPR/Cas systems can also be exploited for various microbial engineering purposes such as strain typing, and engineering of biosynthetic/metabolic pathways of interest in the native host.[265,266]

Assessment of the metabolic potentials of *P. fragi* A13BB with RASTtk identified a total of 4, 528 coding sequences of which 1, 401 (31%) encode proteins associated with 353 subsystems encoded in its genome (Table 4.3). The three closest neighbours to *P. fragi* A13BB in the SEED database at the time of analysis were *P. fluorescens*, *P. putida*, and *P. syringae*, therefore the rather low percentage of genes associated with subsystems was not surprising. As with *H. alvei* A23BA, *P. fragi* A13BB encodes genes that mediate various metabolic processes that can be

exploited for bioremediation, biocontrol, and plant growth promotion (Table 4.4). Interestingly, *P. fragi* A13BB appears to encode genes involved in the biosynthesis of a vibrioferrin-like siderophore that mediates ferric ion uptake, even though *P. fragi* species are classified as non-siderophore producing pseudomonad.[267] As some of the metabolic attributes uncovered here are not readily associated with *P. fragi* species, they were also reported for *P. fragi* A13BB.[240]

Genome mining of *H. alvei* A23BA with antiSMASH identified four smBGCs types - siderophore, thiopeptide, hserlactone (homoserine lactone), and β-lactone smBGCs (Figure 4.3). Thiopeptide and β-lactone smBGCs are known to encode metabolites that can be exploited for medicinal purposes, and both smBGCs types found in *H. alvei* A23BA show little or no similarities to known smBGCs.

Thiopeptides are a diverse group of structurally complex RiPPs that are potent antibiotics effective in nanomolar concentrations against Gram-positive organisms.[268,269] They inhibit protein synthesis by binding to highly conserved sites on bacterial ribosome and as such are able to overcome established drug resistance seen in MDR strains.[268,269] They are generally ineffective against Gram-negative species. However, Ranieri *et al* recently reported potent antibacterial activity of thiostrepton against MDR *P. aeruginosa* and *A. baumannii* in which the drug utilised membrane pyoverdine receptors to cross the outer membrane.[270] Also, Imai *et al* recently reported the activity of darobactin against Gram-negative pathogens.[271] Darobactin is intriguing in that it is encoded by RiPPs-type smBGC, but its mode of action is not inhibition of protein synthesis. Its target is a vital outer membrane protein, hence its inactivity against Gram-positive strains.[271] Despite the potency of thiopeptides, their poor solubility in water and low bioavailability have been barriers to successful application in human medicine.[268,269] However, recent advances in thiopeptide bioengineering have yielded analogues with improved pharmacological and stability profiles.[268] Examples of these analogues include LFF571 and BIK0379, both of which are in clinical trials.[268,272] Various genome mining projects have uncovered many medium-sized smBGCs (i.e. < 40 kb) that are likely to encode novel thiopeptides.[260] The fact that these metabolites are encoded by medium-sized clusters means they are widely distributed in bacteria irrespective of genome size.[260] Thiopeptides therefore represent an underexploited class of bacterial secondary metabolites that hold huge potentials as antibiotic lead compounds.

Likewise, β-lactone natural products are chemically diverse compounds that have been isolated from bacteria, fungi, insects, and plants.[273,274] The strained β-lactone ring in the core structures (Figure 4.8) makes these compounds highly reactive. The strain is released when the electrophilic centre in the lactone ring reacts with nucleophilic species to form stable covalent adducts.[273-276] In various *in vitro* studies, β-lactones were shown to react with highly nucleophilic catalytic amino acid residues (i.e. serine, cysteine, or threonine) in enzyme active sites, in effect inhibiting the normal functions of these enzymes.[273-276] In theory, any enzyme with these residues in its catalytic site can be inhibited. Demonstration of some of these

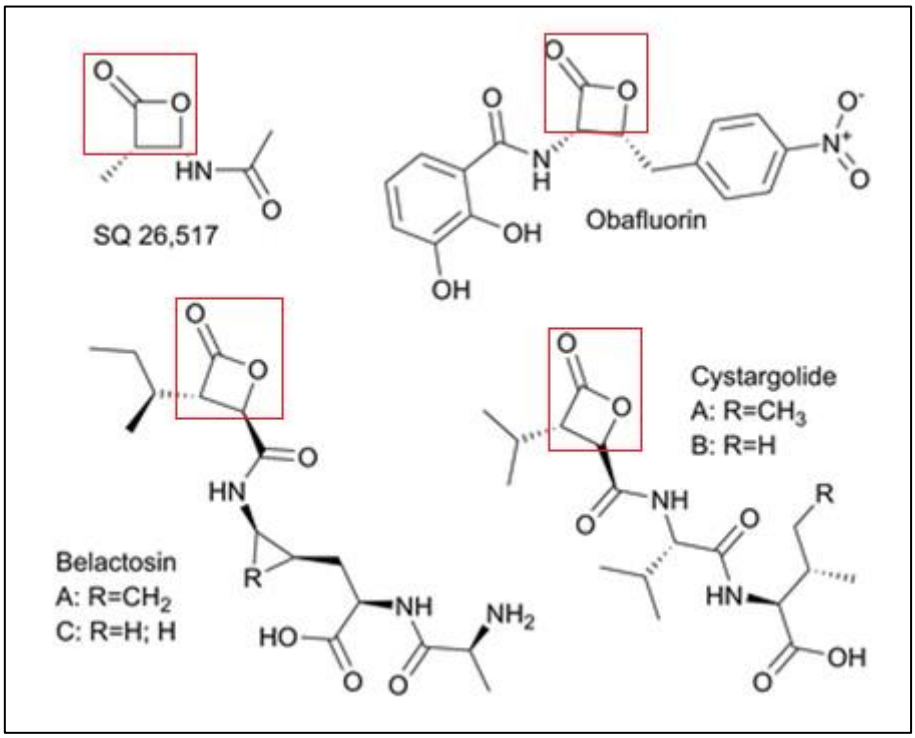


Figure 4.8: Examples of β-lactone natural compounds.
Strained lactone rings highlighted in red boxes. Image adapted from Robinson *et al*.[273]

inhibitory effects *in vivo* underpin the antibacterial, antifungal, anticancer, anticholesterol, and antiobesity properties of known β-lactone compounds.[273] Their antibacterial activity is fascinating as De Pascale *et al* demonstrated that some β-lactone compounds (i.e. ebelactone A, hymeglusin, and two synthetic analogues of ebelactone A) showed inactivating effects on homoserine transacetylase (HTA) that catalyses the transfer of acetyl group from acetyl CoA to hydroxyl group of homoserine in the biosynthesis of the amino acid methionine, in microorganisms.[275] The β-lactones used in the study showed some antibacterial

and antifungal activities by forming covalent adducts with the serine residue in the catalytic site of HTA thereby blocking the biosynthesis of methionine, which is required for protein synthesis and other critical metabolic functions.[275] As higher eukaryotes including humans do not synthesis methionine or have HTA homologues, De Pascale *et al* reasoned that β-lactones are potential lead compounds for novel antimicrobials.[275] However, some of the results obtained by De Pascale *et al* suggested that inhibition of methionine biosynthesis may not be the only mechanism at play in the antimicrobial effects of the β-lactone compounds investigated, as methionine supplementation of culture broth did not restore cell viability in all instances.[275] In fact, Böttcher and Sieber had previously demonstrated that β-lactone compounds can inhibit multiple bacterial enzymes (ligases, oxidoreductases, hydrolases, and transferases) through their *in vitro* study of the effects of these compounds on the proteomes of Gram-positive and -negative bacteria.[276] In total, 23 different bacterial enzymes involved in vital cellular processes were inhibited, with the homologue of only one of these enzymes (i.e. S-formylglutathione hydrolase) found in the mouse liver cytosol included in the study as eukaryotic reference proteome.[276] Therefore, β-lactone compounds potentially have a broad target range, meaning antimicrobial resistance will be less of an issue. Despite these potentials however, β-lactone compounds (e.g. obafluorin, SQ-26517, ebelactone A, hymeglusin, etc) appear to only show weak broad-spectrum antibacterial activity *in vivo,* if at all. Böttcher and Sieber speculated that cellular uptake of β-lactone compounds might be limited in intact bacterial cells as evidenced by an *in vivo* study in which cellular uptake of a β-lactone probe was limited such that full inhibition of a target bacterial enzyme was not achieved.[276] This may suggest that the ability to limit cellular uptake of β-lactone compounds is a common resistance trait in bacteria given their potentially lethal and far-reaching effects. Interestingly, β-lactone compounds have the hallmarks of 'ideal antibiotics' which, according to both Lewis and Gajdács, are capable of disrupting many vital bacterial cellular processes simultaneously, and which should ideally be formulated as prodrugs that are converted into active drug forms inside bacterial cells by bacteria-specific enzymes.[16,114] It is also worth pointing out that a compound that can disrupt many bacterial cellular functions simultaneously will have the potential of eliminating persister cells and cells embedded within biofilms.[16,114] Therefore, these compounds hold huge potentials as novel antibiotic lead compounds.

H. alvei A23BA also encodes a siderophore smBGC (Figure 4.3). Although siderophore smBGCs are not generally regarded as drug-like, the potential role of microbial siderophores in combating antibiotic resistance mediated by restricted cellular uptake of antibiotic drug is gaining a lot of attention. Siderophores are structurally diverse small molecular weight (150-2000 Da) iron-chelating organic compounds that are biosynthesised and secreted by many microbial species in response to extremely low concentration of soluble ferrous ion (Fe^{2+}) in extracellular microenvironments.[277-279] Iron is required by all living organisms for vital metabolic processes; even though it is one of the most abundant metals on Earth, it predominantly exists in the insoluble ferric ion (Fe^{3+}) form at physiological pH.[277] Siderophores secreted by producing species into extracellular spaces have very high affinity for ferric ions with which they form stable complexes (i.e. ferrisiderophores) that are transported into bacterial cells through active transport mediated by high affinity receptor proteins present in bacterial cell wall or membrane.[277-279] Inside bacterial cells, the ferric ion is reduced to ferrous ion, through the actions of bacterial ferric reductases, and decoupled from the siderophore.[279] Aside from being a vital metabolite to producing species, other non-producing species (or indeed species that would rather not expend the energy required for siderophore biosynthesis) in the same niche as the producer benefit from xenosiderophores (i.e. siderophores produced by other organisms) by assimilating ferrisiderophores present in the microenvironment in the same way the producing species would.[278,280] As well as serving a communal good, siderophores can also be used as weapons of warfare where the producing strain secretes a siderophore with a functional group not recognised by competing cells thereby depriving those cells of vital iron, or the producer secretes a siderophore conjugated to a lethal antibiotic. In the latter scenario, the microbial siderophore-antibiotic complex (or sideromycin) is transported into competing Gram-negative and -positive bacterial cells by active transport mediated by membrane proteins that bind ferrisiderophores.[278,280,281] Once inside the cells, the antibiotic is decoupled from the siderophore where it is free to effect its lethal actions.[278,280,281] Very few sideromycins have been discovered compared to natural antibiotics, with albomycin (Figure 4.9), isolated from several *Streptomyces* species, being the most studied.[278,280,281] It is potent against several Gram-positive and -negative pathogens including some MDR species.[278,281]

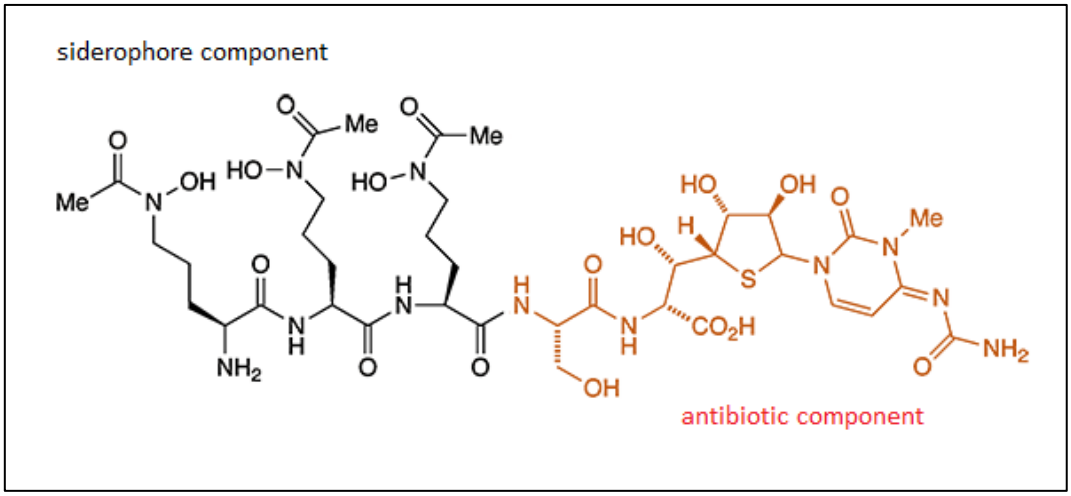


Figure 4.9: Structure of albomycin showing siderophore and antibiotic components.

Even though albomycin is not used widely because its recovery from microbial sources is very minimal, its discovery nonetheless gave researchers the inspiration to synthesise antibiotics conjugated to natural or synthetic siderophores resulting in complexes with increased potency and broader spectrum of activity (with natural siderophores yielding better results).[278,280] In some instances, potency of siderophore-conjugated antibiotics were increased up to a thousand-fold compared to unconjugated drugs as cellular uptake was enhanced by the complex formation.[282] In other cases, antibiotics that were previously only effective against Gram-positive pathogens showed potent activity against Gram-negative pathogens when conjugated with siderophores, as the complexes formed were actively transported through the restrictive outer membrane via ferrisiderophore uptake pathways into the cytoplasm.[283] This 'Trojan horse' strategy has seen further success in the recent FDA-approved cefiderocol, which is a catechol-cephalosporin conjugate antibiotic (Figure 4.10) effective against many MDR Gram-negative pathogens.[284,285] It is currently licensed for use in chronic urinary tract infections caused by MDR pathogens in patients with limited or no alternative treatment options.[284,285]

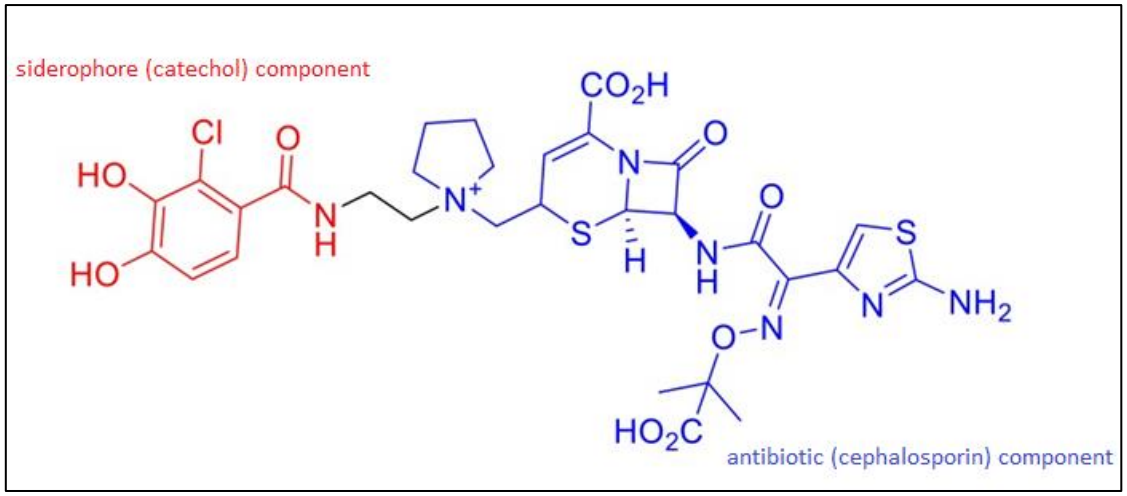


Figure 4.10: Structure of cefiderocol showing siderophore and antibiotic components.

In essence, three smBGCs of interest (i.e. thiopeptide, β-lactone, and siderophore smBGCs) were detected in the genome of *H. alvei* A23BA (Figure 4.3). Likewise, *P. fragi* A13BB encodes two β-lactone smBGCs and one siderophore smBGC that can be exploited for therapeutic purposes (Figure 4.4). Detection of a siderophore smBGC in the genome of *P. fragi* A13BB further strengthens the notion that the classification of *P. fragi* as non-siderophore producing pseudomonad is erroneous. This was also recently pointed out by Stanborough *et al* who identified conserved vibrioferrin-like gene clusters in 13 strains of *P. fragi* in which siderophore production was observed.[286]

The β-lactone smBGCs in *H. alvei* A23BA and *P. fragi* A13BB were selected for further analysis given the potentials of β-lactone natural compounds as 'ideal antibiotics' and coupled with the fact the three smBGCs show little or no homology to known smBGCs (Figures 4.5, 4.6, and 4.7). To that end, smCOG analysis of each cluster was carried out to predict the end products of genes found in the clusters. This analysis was complemented with BlastP analysis that is hyphenated to antiSMASH. The β-lactone smBGC in *H. alvei* A23BA spans a genomic region that is 25,505 bp long and contains 19 genes of which only 7 were predicted to encode homologues of enzymes found in the smCOG database (Table 4.5). Two genes were predicted to encode the core enzymes i.e. genes encoding AMP-dependent synthetase and ligase, and 2-isopropylmalate synthase. In addition, there are 2 putative transport-related genes, 2 putative regulatory genes, and 1 additional biosynthetic gene predicted to encode pyruvate oxidase/decarboxylase.

Genes that encode enzymes that are not homologous to any entries in the smCOG database are designated as having "unknown" end products. Almost all the genes with "unknown" end products had their end products predicted with BlastP analysis.

The first β-lactone smBGC in *P. fragi* A13BB spans a genomic region that is 19,854 bp long and contains 17 genes of which 10 are predicted to encode enzymes homologous to proteins found in the smCOG database (Table 4.6). Two of these genes are predicted to encode the core enzymes, 3 are predicted to encode regulatory enzymes, and 5 to encode additional biosynthetic enzymes. Genes assigned "unknown" end products had their end products predicted by BlastP analysis. The second β-lactone smBGC in *P. fragi* A13BB spans a genomic region that is 24,196 bp long and contains 20 genes of which 12 were predicted to encode enzymes that are homologous to proteins in the smCOG database (Table 4.7). Of these, 2 genes were predicted to encode the core enzymes, and 6 to encode additional biosynthetic enzymes. Two genes were predicted to encode transport proteins and a further 2 to encode regulatory enzymes. Genes assigned "unknown" end products also had their end products predicted by BlastP analysis.

With the necessary caveat in mind, i.e. *in silico* prediction of gene cluster boundaries may be imprecise, it can be reasonably assumed that this sub-group of amino acid β-lactone smBGCs recently detectable with antiSMASH are medium-sized, i.e. < 40 kb, given the size of the clusters detected here. As such, this smBGC type is likely to be widely distributed in bacterial genomes irrespective of genome size. Although some of the genes in the three clusters analysed here are predicted to encode enzymes with unknown (or more precisely unverified) functions, a detailed study of the roles of encoded enzymes with known functions, especially the core enzymes, can form the basis of subsequent gene cluster activation.

4.7 Conclusions

Whole-genome analysis of *H. alvei* A23BA and *P. fragi* A13BB with RASTtk gave indications of the metabolic potentials of both strains and uncovered metabolic potentials that are not readily associated with the species. The bioremediation, biocontrol, and plant-growth promotion potentials of *H. alvei* A23BA were the first

to be reported for *H. alvei* species. Likewise, the same attributes were reported for *P. fragi* A13BB to add to the very limited evidence currently found in literature, and more importantly to refute the longstanding classification of *P. fragi* as non-siderophore producing pseudomonad.

Genome mining of both strains with antiSMASH detected drug-like smBGCs that may encode metabolites with antibiotic properties. *H. alvei* A23BA encodes a thiopeptide, a β-lactone, and a siderophore smBGC, while *P. fragi* A13BB encodes two β-lactone smBGCs and a siderophore smBGC. Thiopeptides are potent antibiotics effective against Gram-positive bacteria, while β-lactone compounds show weak broad-spectrum antibacterial activities. Siderophores can be used as 'Trojan horse' to enhance cellular uptake of antibiotics resulting in improved potency and/or broader spectrum of activity. The weak antibacterial activities of β-lactone compounds are probably due to genetic or phenotypic resistance traits seen in many bacterial species that limit their cellular uptake and intracellular accumulation, which is not surprising given the lethal consequences of high intracellular concentrations of these compounds. Nonetheless, β-lactones have the hallmarks of 'ideal antibiotics' that can disrupt multiple bacterial processes simultaneously. Enhancement of cellular uptake, perhaps by conjugating the compounds to natural siderophores, may drastically improve their potency. Therefore, the three β-lactone smBGCs detected here were prioritised for activation. Furthermore, as these β-lactone smBGCs show little or no homology to known smBGCs, they may also encode novel metabolites.

CHAPTER FIVE

ACTIVATION OF THE β-LACTONE smBGCs IN *HAFNIA ALVEI* A23BA AND *PSEUDOMONAS FRAGI* A13BB

5. Activation of the β-Lactone smBGCs in *Hafnia alvei* A23BA and *Pseudomonas fragi* A13BB

β-Lactone natural compounds have highly reactive electrophilic centres that react covalently with nucleophilic amino acid residues in enzyme active sites resulting in inactivation of many bacteria-specific enzymes. Hence these compounds are potential leads in the development of 'ideal antibiotics'. Detailed study of the functional roles of enzymes encoded in the β-lactone smBGCs detected in *H. alvei* A23BA and *P. fragi* A13BB genomes can inform activation strategies.

5.1 Introduction

As highlighted in section 4.6, β-lactone natural compounds have strained four-membered lactone rings with highly reactive electrophilic centres. These electrophilic centres form stable covalent adducts with nucleophilic catalytic serine, cysteine, or threonine residues in enzyme active sites.[273-276] More than twenty bacteria-specific enzymes belonging to four enzyme classes (i.e. hydrolases, ligases, transferases, and oxidoreductases) are susceptible to inhibition by β-lactone compounds.[276] In comparison with β-lactam antibiotics that predominantly inhibit PBPs, β-lactone compounds can potentially inhibit many vital bacterial enzymes simultaneously. Harnessing this potential may produce novel antibiotic compounds that are more successful than β-lactam antibiotics. However, the success of β-lactam antibiotics stems from the fact that PBPs have no homologues in humans, whereas β-lactone compounds can potentially inhibit enzymes with serine, cysteine, and threonine catalytic sites. Therefore, whilst the broad target range of β-lactone compounds is desirable for optimum antibiotic activity, unwanted off-target reactions in the host must be guarded against. A prudent drug design consideration would be to formulate β-lactone antibiotic lead compounds as prodrugs that are perhaps conjugated to natural siderophores that enhance bacterial cellular uptake. This will also ensure that the active forms of the drugs are predominantly released inside bacterial cells thereby averting off-target effects in the host.

Of the twenty-nine distinct scaffolds of natural β-lactone compounds known to date, sixteen are of bacterial origin from a few taxonomic groups.[273] But it is believed that these compounds are more widely distributed in bacteria, and as

such many are waiting to be discovered. It is also believed that many β-lactone compounds are missed during rapid screening of culture broths because of their instability as they readily undergo thermal degradation and acid/base hydrolysis.[273] To aid rapid screening of crude extracts for β-lactone compounds, Castro-Falcón *et al* designed the cysteine thiol probe that tags β-lactone compounds in crude extracts making them more stable and detectable.[287]

Genome mining for β-lactone compounds is not entirely straight forward either. This is because the compounds are structurally diverse with diverse biosynthetic logic.[273] The recently incorporated pHMMs for detection of β-lactone compounds in antiSMASH v5.0 are based on the signature/core enzymes found in the amide ligase-type cluster.[140] The discovery of this unusual cluster type was made by Wolf *et al* who were the first researchers to identify the smBGCs of belactosins and cystargolides.[288] They were surprised that the formation of peptide bonds in these compounds was not catalysed by enzymes associated with NRPS or ribosomal systems, but by rare single-enzyme amino acid ligases.[288] They also observed that the formation of the β-lactone warhead in these metabolites did not follow the characteristic pattern for most β-lactone compounds, but it instead followed an unprecedented enzymatic activation and lactonization of an intermediate metabolite, 3-isopropylmalate (3-IPM).

Wolf *et al* identified genes encoding homologues of 2-isopropylmalate synthase (2-IPMS), *cys*A and *bel*J, in the smBGCs of cystargolides and belactosins, respectively, that initiate the biosynthesis of 3-IPM in the same way that 3-IPM is formed in the biosynthesis of leucine - a primary metabolic reaction that is catalysed by 2-IPMS.[288] *cys*A is responsible for the Claisen condensation of acetyl CoA and α-ketoisovalerate (a degradation product of valine) to yield 2-isopropylmalate (2-IPM), while *bel*J is responsible for the Claisen condensation of acetyl CoA and 2-keto-3-methylvalerate (a degradation product of isoleucine) to also yield 2-IPM.[288] The reaction catalysed by *bel*J is unprecedented in bacteria as 2-IPMS has always been thought to have very narrow substrate specificity for α-ketoisovalerate.[288] In both biosynthetic pathways, 2-IPM is isomerised to 3-IPM. This step is catalysed in the biosynthesis of belactosins by *bel*F and *bel*G i.e. homologues of 3-isopropylmalate dehydratase small subunit and 3-isopropylmalate dehydratase large subunit, respectively.[288] As these enzymes are not encoded in the smBGC of cystargolides, they reasoned that the isomerisation

steps in the biosynthesis of cystargolides were probably catalysed by enzymes from primary metabolism.[288] Therefore, Wolf and colleagues speculated that the core enzymes of this unusual amide ligase cluster-type are likely to include 2-IPMS homologues, AMP-dependent synthetase and ligase-type enzymes, ATP-grasp enzymes, methylesterases, and/or SAM-dependent methyltransferases; all of which are encoded in the smBGCs of cystargolides and belactosins.[288] They reasoned that 2-IPMS is necessary for generating 3-IPM; AMP-dependent synthetase and ligase-type enzymes possibly needed for amide bond formation and/or β-lactone warhead formation; ATP-grasp enzymes possibly for amide bond formation; and lastly, methylesterases and SAM-dependent methyltransferases possibly required for β-lactone warhead formation.[288] However, the detection rules for β-lactone compounds in antiSMASH v5.0 are based on genes encoding 2-IPMS homologues and AMP-dependent synthetase and ligase-type enzymes, suggesting that the latter can perform the dual role of catalysing amide bond formation and β-lactone warhead formation.

These two core enzymes are encoded in the β-lactone smBGCs found in *H. alvei* A23BA and *P. fragi* A13BB, along with other biosynthetic enzymes (Tables 4.5, 4.6, and 4.7). An understanding of the functional roles of enzymes encoded in these clusters and the metabolic reactions they catalyse can provide some insights into standard fermentation conditions that may encourage gene cluster activation, provided the clusters are not silent in these strains. Efforts to activate a novel smBGC or identify its metabolic product should start with the assumption that the cluster is cryptic, rather than silent, in the native strain. Only after a range of standard fermentation parameters have been exhausted without successful activation should the cluster be assumed to be silent. Therefore, the initial assumption here was that the β-lactone smBGCs in *H. alvei* A23BA and *P. fragi* A13BB are cryptic, and as such are likely to be expressed in fermentation experiments conducted with the right standard fermentation conditions.

5.2 Aim and Objectives

The aim of this section of the project was to activate the β-lactone smBGCs in *H. alvei* A23BA and *P. fragi* A13BB by employing fermentation conditions that will facilitate gene cluster expression. This can be achieved by:

(i) Establishing the canonical functions of key enzymes encoded in the clusters and their respective substrates, from literature and the MetaCyc database.
(ii) Identifying physiological/metabolic conditions that may cause genes encoding many of these enzymes to be simultaneously upregulated.
(iii) Performing fermentation experiments with parameters that are likely to induce the physiological/metabolic conditions identified.
(iv) Performing reactivity-guided screening of crude extracts obtained therefrom for β-lactone compounds to determine if gene cluster activation was successful.

5.3 Methodology

5.3.1 Gene Cluster Analysis

In order to analyse the three β-lactone smBGCs and determine the functions of enzymes encoded in the clusters, a review of relevant literature was carried out using primary sources where possible. Literature review was complemented with MetaCyc database search. The MetaCyc database of metabolic pathways and enzymes was chosen for this piece of work because it is the most comprehensive reference database of its kind and the largest curated collection of metabolic pathways to date.[289] Also, unlike other databases that contain a hybrid of experimentally elucidated metabolic pathways and those predicted by *in silico* methods, MetaCyc only features curated and experimentally validated enzymatic metabolic pathways.[289]

After gene cluster analysis, the genes common to all three clusters studied here are those encoding the two core enzymes i.e. genes encoding homologues of 2-IPMS and AMP-dependent synthetase and ligase (Tables 4.5, 4.6, and 4.7).

5.3.1.1 2-Isopropylmalate Synthase (2-IPMS)

The canonical function of homologues of 2-IPMS is catalysis of the condensation of acetyl CoA and 3-methyl-2-oxobutanoate (α-ketoisovalerate) to form 2-IPM i.e.

Acetyl CoA + 3-methyl-2-oxobutanoate + H2O ⇌ 2-isopropylmalate + CoA

The substrates of this reaction are acetyl CoA and 3-methyl-2-oxobutanoate.[289,290] The first substrate, acetyl CoA, is an indispensable primary metabolite from which energy (in the form of ATP) is produced in reactions of the tricarboxylic acid (TCA)

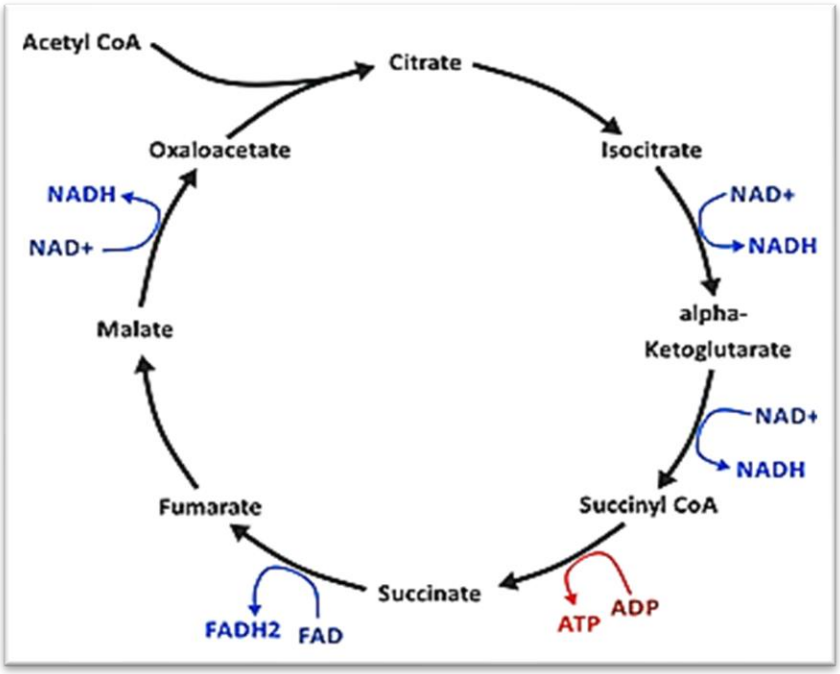


Figure 5.1: The TCA cycle.
Acetyl CoA is an indispensable primary metabolite from which ATP is produced in reactions of the TCA cycle.
Image source: Wikimedia (https://commons.wikimedia.org).

cycle (Figure 5.1).[289,291,292] It is biosynthesised preferably from glucose through the glycolytic intermediate, pyruvate. During rapid bacterial growth on glucose, excess acetyl CoA not required for the TCA cycle is used for the biosynthesis of fatty acids and amino acids, or is converted to acetate in a reversible reaction with the acetate ultimately excreted from the cell.[291-293] As such, in the absence of the preferred acetyl CoA precursor, i.e. glucose, TCA cycle-bound acetyl CoA is derived from catabolism or β-oxidation of fatty acids, degradation of amino acids, or from acetate; with the pathway from acetate to acetyl CoA being the shortest.[289,291-293] Extracellular acetate is assimilated into bacterial cells and converted to acetyl CoA via two pathways - the reversible low affinity Pta-AckA pathway favoured for high extracellular acetate concentrations (≥25 mM) or the unidirectional high affinity AMP-forming acetyl CoA synthetase (Acs) pathway favoured for low extracellular acetate concentrations (≤10 mM).[291,293]

The second substrate of 2-IPMS, i.e. 3-methyl-2-oxobutanoate or α-ketoisovalerate, is the last intermediate product in the biosynthesis of valine, which also serves as a precursor in the biosynthesis of leucine (Figure 5.2).[289,290] Conversely, α-ketoisovalerate is also the first product in the degradation pathway

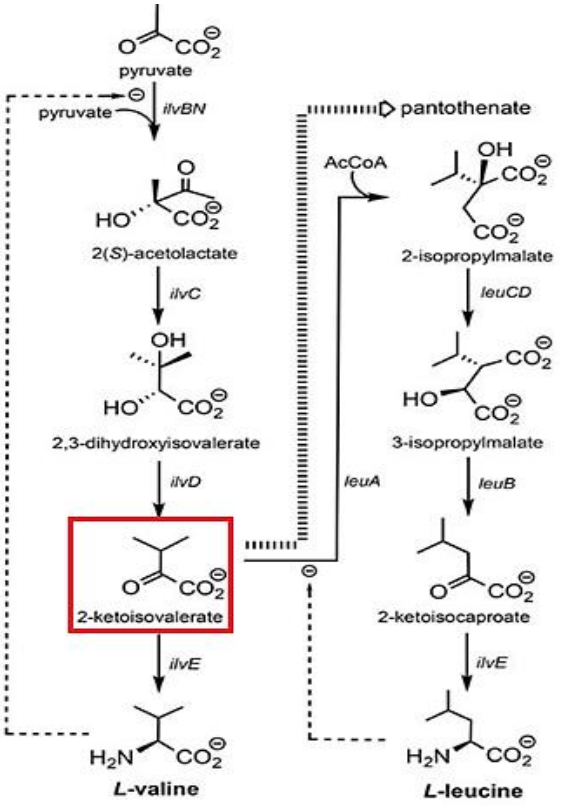


Figure 5.2: Biosynthetic pathways of valine and leucine.
α-ketoisovalerate, a substrate of 2-IPMS, is the last intermediate product in the biosynthesis of valine and the precursor in the biosynthesis of leucine. Image adapted from Franco *et al*.[290]

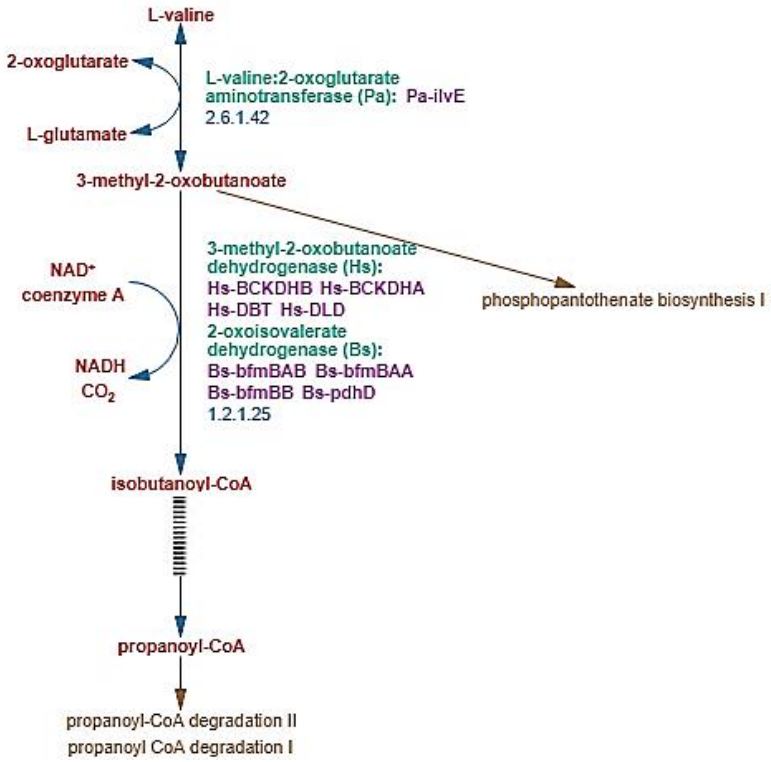


Figure 5.3: Truncated valine degradation pathway.
α-ketoisovalerate, a substrate of 2-IPMS, is the first intermediate product in the degradation pathway of valine. Image adapted from MetaCyc.org.[289]

of valine, which is not a reversal of the anabolic pathway (Figure 5.3).[289] Amino acids are degraded for various reasons, including utilisation as both nitrogen and carbon sources in the absence of preferred substrates i.e. ammonia and glucose, respectively. They are also degraded because excessive amounts of certain amino acids (e.g. valine, cysteine, and leucine) not required for protein synthesis are known to inhibit growth in some organisms.[294,295.]

5.3.1.2 AMP-dependent Synthetase and Ligase

The second core enzyme encoded in all three β-lactone smBGCs is the AMP-dependent synthetase and ligase-type enzyme, or acyl CoA ligase and synthetase to be specific (Tables 4.5, 4.6, and 4.7). Acyl CoA ligase homologues are found in both prokaryotes and eukaryotes where they perform the canonical function of activating fatty acids to initiate β-oxidation of fatty acids to produce acetyl CoA.[289,296,297] However, according to Wolf *et al*, homologues of this enzyme have also been shown elsewhere to activate amino acids and to catalyse the formation of amide bonds.[288] This led to their speculation that the acyl CoA ligase homologues encoded in the smBGCs of cystargolides and belactosins probably activated 3-IPM and catalysed the spontaneous intramolecular cyclisation of the activated product to form the β-lactone warhead.[288] They also speculated that these enzymes catalysed the formation of amide bonds in the cystargolides and belactosins peptide backbone.[288] Therefore, it can be inferred that the substrates for the acyl CoA ligases encoded in the β-lactone smBGCs are 3-IPM and amino acids that form the β-lactone warhead and the peptide backbone, respectively.

5.3.1.3 Additional Biosynthetic Enzymes

In addition to genes encoding core enzymes found in all three smBGCs, the two clusters in *P. fragi* A13BB contain other homologous gene pairs that are likely to encode similar enzymes with similar biosynthetic functions (Tables 4.6 and 4.7). Both clusters encode homologues of acetyl CoA carboxylase that catalyses the carboxylation of acetyl CoA to produce malonyl CoA in the first step of fatty acid biosynthesis;[289,298] enoyl CoA hydratase homologues that are involved in β-oxidation of fatty acids to produce acetyl CoA and ATP;[289,296,299] and flavin-dependent oxidoreductases (i.e. isovaleryl CoA dehydrogenase) involved in the

degradation of leucine;[289] with methylcrotonyl CoA carboxylase which also features in leucine degradation pathway being encoded in the first *P. fragi* β-lactone smBGC only (Table 4.6).[289] Other enzymes involved in fatty acid metabolism encoded in either of the two clusters include homologues of acyl CoA thioesterase and acyl CoA dehydrogenase both of which are involved in β-oxidation of fatty acids.[289,296] Whilst the β-lactone smBGC in *H. alvei* A23BA does not appear to encode homologues of these additional biosynthetic enzymes, enzymes from primary metabolism can be recruited to perform similar functions, if needed. The β-lactone smBGC in *H. alvei* A23BA does however encode three additional biosynthetic enzymes involved in leucine biosynthesis i.e. homologues of 3-isopropylmalate dehydrogenase, 3-isopropylmalate dehydratase large subunit, and 3-isopropylmalate dehydratase small subunit (Table 4.5).[289,290]

In summary, and in speculation about the biosynthetic logic, the smBGCs studied here appear to encode secondary metabolites that are biosynthesised in a similar fashion to leucine. The secondary metabolic pathway is probably initiated to efficiently utilise less preferred carbon and nitrogen sources during nutrient-limited conditions without having to accumulate excessive amounts of an intermediate metabolite (i.e. 3-methyl-2-oxobutanoate) that is energy-intensive to degrade (Figure 5.3). The enzymes involved in the biosynthesis of β-lactone secondary metabolites are recruited or repurposed from primary metabolism. Whilst many of these enzymes are likely to perform the same functions as in primary metabolism, neofunctionalization cannot be ruled out entirely. The initial substrates for the secondary biosynthetic reactions are the same as those involved in the biosynthesis of leucine, i.e. acetyl CoA and 3-methyl-2-oxobutanoate. Acetyl CoA can be derived from glucose, fatty acids, amino acids, or acetate. As several genes in the smBGCs of β-lactone compounds encode enzymes involved in the metabolism of fatty acids to produce acetyl CoA and ATP, these genes are likely to be upregulated in many organisms cultivated in growth media with poor carbon sources (e.g. acetate and amino acids) and downregulated in the presence of glucose.[300] 3-methyl-2-oxobutanoate can be diverted from the biosynthetic pathway of valine (Figure 5.2) or from the degradation pathway of valine (Figure 5.3), with the latter scenario more likely in the presence of a poor nitrogen source.[301]

5.3.2 Fermentation Conditions

The choice of carbon and nitrogen sources in a fermentation broth will influence the types of genes that are upregulated or downregulated in an organism during growth. The goal of fermentation experiments here was to encourage upregulation of the majority of genes found in the three β-lactone smBGCs by creating the right metabolic/physiological conditions and providing the right substrates.

5.3.2.1 Choice of Carbon Source

A poor carbon source (e.g. acetate or amino acids) should induce upregulation of the majority of additional biosynthetic genes found in the clusters. Using acetate as the carbon source during bacterial growth has been shown to lead to increase in intracellular cAMP, catabolite derepression, and upregulation of metabolic genes including *acs* that encodes acetyl CoA synthetase (Acs), and genes involved in fatty acid metabolism.[300,302] *acs* was significantly upregulated (more than 8-fold) in one study during bacterial growth on acetate.[300] The unidirectional Acs pathway has high affinity for extracellular acetate and is responsible for acetate assimilation at low extracellular concentrations (≤10mM).[293,302] It is interesting to note that the Acs pathway is also responsible for assimilation of previously excreted acetate (excreted during exponential growth on glucose) that supports microbial metabolism during late exponential and stationary phases of growth, and its activities have been linked to the biosynthesis of many secondary metabolites, including antibiotics.[293,300,302-304] Inside the cell, Acs catalyses the synthesis of acetyl CoA from acetate i.e.

Acetate + CoA + ATP → Acetyl CoA + AMP + PP_i.

Acetyl CoA is then fed into the TCA cycle to generate ATP and other intermediate metabolites, or is fed into the glyoxylate cycle for gluconeogenesis, or is utilised for fatty acid biosynthesis and other acetylating metabolic reactions.[293,300] As acetyl CoA is the precursor for fatty acid biosynthesis, genes involved in fatty acid catabolism are upregulated when organisms grow on poor carbon sources like acetate in readiness for utilization of fatty acids as sources of acetyl CoA if required.[300]

Therefore, using acetate at a low concentration (i.e. ≈10 mM) as the carbon source in growth media will satisfy provision of acetyl CoA substrate for 2-IPMS

and also lead to upregulation of many genes found in the clusters, as well as other genes associated with secondary metabolism e.g. *acs.* As cultivation on acetate leads to upregulation of these genes earlier on in the growth cycle rather than during stationary phase, increased amounts of secondary metabolites may be recovered from batch cultures as demonstrated by Leone *et al*.[305] However, bacterial growth in media with acetate as sole carbon source proceeds with a long lag phase, perhaps because many genes regulating biosynthesis and metabolism of primary metabolites (e.g. amino acids and nucleotides) are downregulated when microorganisms utilise acetate as the sole carbon source.[300,305] Therefore, supplementation of growth media with yeast extract that supplies essential growth factors, amino acids, and other micronutrients is necessary to shorten the lag phase.[305] The consequence of yeast extract supplementation however is an increase in extracellular acetate concentration as cells assimilate more preferable carbon sources (e.g. gluconeogenic amino acids) present in the yeast extract first to support exponential growth, with acetate being excreted as a waste product.[305] Initial increased extracellular acetate concentration can decrease the pH of growth medium, leading to increased concentration of undissociated acetic acid in the medium. Undissociated acetic acid can diffuse freely across cell membranes resulting in growth inhibition.[305,306] To avoid this scenario, a buffered fermentation broth in the pH range of 7.5-8.5 has been shown to be desirable.[305,306] Potassium phosphate buffer will be suitable for this purpose as it has an excellent buffering capacity between pH 5.8 to 8.0.[307]

5.3.2.2 Choice of Nitrogen Source

Excluding ammonium salts/nitrates from fermentation broths can encourage degradation of amino acids to obtain nitrogen.[295,301] Therefore, supplementing ammonium salts- and nitrates-free growth media with valine should encourage valine assimilation and degradation here for two reasons: **a)** the first step in the degradation pathway of valine (Figure 5.3) yields L-glutamate which is one of the two preferred amino acids utilised as nitrogen source (the other being L-glutamine which is first converted to L-glutamate).[289,295,301] L-glutamate is subsequently degraded in one simple enzymatic step to give 2-oxoglutarate (synonym α-ketoglutarate) which is fed into the TCA cycle whilst also releasing much needed

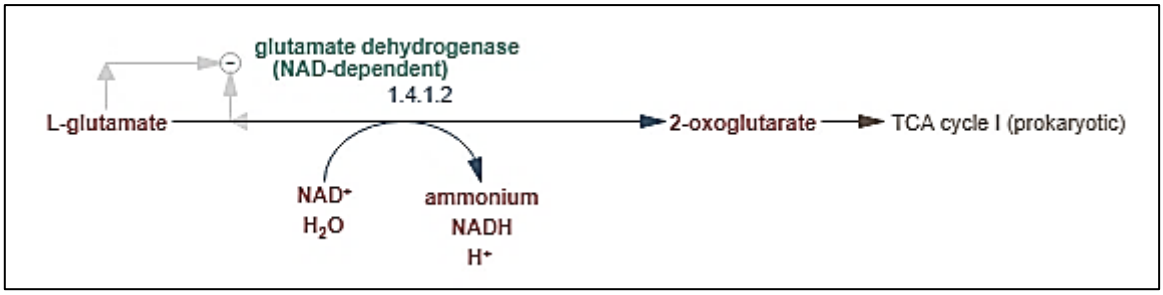


Figure 5.4: L-glutamate degradation pathway.
Image adapted from MetaCyc.org.[289]

ammonia (Figure 5.4);[289] **b)** the full valine degradation pathway is not necessarily invoked, thereby conserving much needed energy. This is because the first intermediate in the degradation pathway (i.e. 3-methyl-2-oxobutanoate) can be diverted to synthesise the β-lactone warhead in a similar fashion to leucine biosynthesis. Biosynthesis of the β-lactone warhead and subsequent ligation to peptide backbone to form β-lactone compounds occur in few enzymatic steps compared with the much longer and energy intensive enzymatic steps involved in valine degradation to yield propanoyl CoA (Figure 5.3).[288,289] Also, diversion from leucine biosynthesis to β-lactone warhead biosynthesis prevents accumulation of unneeded leucine which is also energy intensive to degrade.[289]

Therefore, ammonium salts and nitrates were excluded from media composition to encourage utilisation of amino acids as nitrogen source. It was expected that glutamate and glutamine contents of yeast extract would be utilised first, followed by other amino acids that are easily degraded to obtain glutamate. To ensure valine was degraded for this purpose, growth media were supplemented with valine to create a steady supply. Degradation of valine will produce the second substrate of 2-IPMS (i.e. 3-methyl-2-oxobutanoate). It is reasonable to assume that genes encoding homologues of 2-IPMS in the smBGCs will be upregulated in this scenario, but no reference to that effect was found in literature; therefore the hypothesis will need to be experimentally tested. Furthermore, as Wolf *et al* uncovered neofunctionalization of a 2-IPMS homologue in catalysing the condensation of acetyl CoA with a degradation product of isoleucine (i.e. 2-keto-3-methylvalerate) to produce 2-IPM in belactosins biosynthesis,[288] some fermentation experiments were carried out here with valine component of growth media replaced with isoleucine. It is also worth noting that 2-keto-3-

methylvalerate is the first intermediate product in isoleucine degradation pathway, with the step also yielding L-glutamate.[289]

Other necessary components of growth media were sources of ferrous ion (Fe^{2+}) and magnesium ion (Mg^{2+}), both of which participate in many critical metabolic and biosynthetic reactions.[308,309] Ferrous ion is a micronutrient required in micromolar concentrations usually 0.3 to 1.8 μM, while magnesium ion is usually required in concentrations between 1 to 22 mM.[308-310]

5.3.3 Reactivity-guided Screening of Crude Extracts for β-Lactone Compounds

Given that most therapeutically useful microbial natural products are present in crude extracts in minute quantities amongst a myriad of other components, and the propensity of β-lactone compounds to undergo thermal degradation or hydrolysis, screening crude extracts for these compounds can be challenging. To make the tasks of identifying natural compounds with electrophilic moieties in crude extracts less onerous, Castro-Falcón *et al* designed the cysteine thiol probe (Figure 5.5 (a)) that reacts covalently and selectively with β-lactone, β-lactam, and enone compounds in crude extracts to form adducts that are highly UV- and MS-visible.[287] The probe is furnished with a thiol group with which it mimics *in vivo* reactions between the thiol group on catalytic cysteine residues of several enzymes (Figure 5.5 (b)) and the electrophilic centres in β-lactone, β-lactam, and

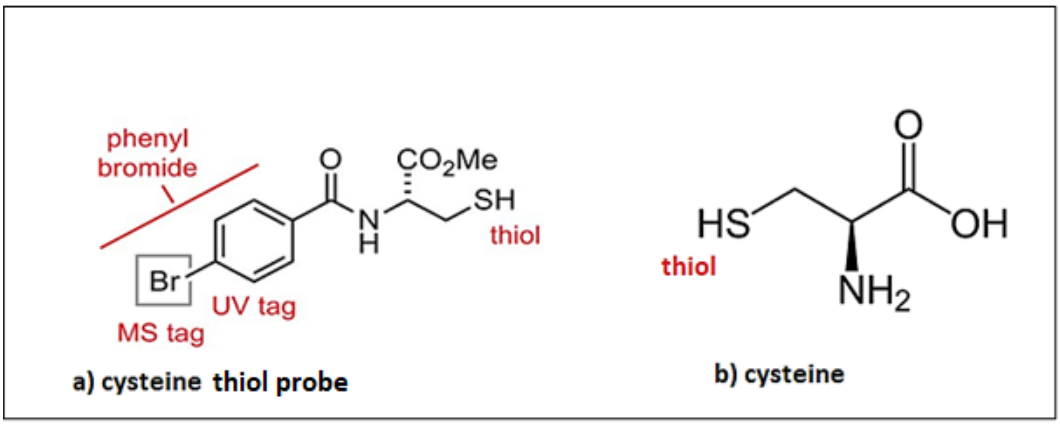


Figure 5.5: Nucleophilic compounds with thiol functionalities.
Cysteine thiol probe is furnished with a thiol group with which it mimics *in vivo* reactions of cysteine with electrophilic compounds. Image adapted from Castro-Falcón *et al*.[287]

enone natural compounds. The benzene ring of the probe's phenyl bromide group ensures adducts formed are highly UV visible, while the bromine atom affords conspicuous MS isotopic patterns: ^{81}Br:^{79}Br = 1 (i.e. two peaks with the same signal intensity and a mass difference of two). The phenyl bromide group also ensures crystalline adducts are formed, meaning chemical structures of reacted natural compounds can be deduced readily with X-ray crystallographic methods.[287]

To demonstrate proof of concept, Castro-Falcón *et al* carried out reactivity-guided isolation of salinosporamide A (a β-lactone natural compound) from the crude extract of *Salinispora tropica* strain CNB-392.[287] Despite the extract containing numerous components, salinosporamide A was the only metabolite labelled to produce two adducts that were both retrosynthetically traced back to salinosporamide A.[287] LC-MS analysis of reaction mixture showed that adducts began to form from 2 hours, with maximum amounts detected at 24 hours. However, unwanted probe dimerization in the presence of atmospheric oxygen can also occur during reactions which can reduce the amount of free probe available for adduct formation. This unwanted reaction can be minimised by adding tris (2-chloroethyl) phosphine hydrochloride salt to reaction mixtures if necessary.[287] Here, cysteine thiol probe was reacted with commercially obtained salinosporamide A to observe its reactivity with a known β-lactone compound before reacting the probe with crude extracts.

5.4 Materials and Methods

5.4.1 Fermentation and Extract Production

Single pure colonies of *H. alvei* A23BA and *P. fragi* A13BB were aseptically transferred from agar plates stored at 6°C into separate universal bottles each containing 15 mL sterile nutrient broth (Oxoid, UK). The seed broths were incubated at 28°C for 24 hours at 150 rpm. After 24 hours, for each strain, 5 mL aliquots of seed broth were aseptically transferred into two separate 250 mL Erlenmeyer flasks, each flask containing 45 mL sterile fermentation broth (0.1% anhydrous sodium acetate, 0.5% yeast extract, 0.3% $MgSO_4.7H_2O$, 0.0002% $FeSO_4.7H_2O$, and 5 mM L-valine) and 5 mL sterile 0.7 M potassium phosphate buffer; each component was prepared and sterilised separately as detailed in Appendix 9, and aseptically mixed just before inoculation. Uninoculated medium

of 45 mL fermentation broth and 5 mL buffer solution contained in a separate 250 mL Erlenmeyer flask acted as control. The two flasks inoculated with *P. fragi* A13BB were designated A and B, while those inoculated with *H. alvei* A23BA were designated C and D. Uninoculated control flask was designated E. Contents of all flasks were incubated at 28°C for 6 days at 150 rpm.

After 6 days incubation, the contents of flasks A to D (whole cultures) and the contents of flask E (control) were extracted with two volumes (2 X 50 mL) of ethyl acetate (laboratory reagent grade; Fisher Scientific, UK). The two organic layers corresponding to each flask were combined and centrifuged at 4000 rpm at 10°C for 30 minutes to remove trapped water and cell debris before evaporation to dryness at 45°C under reduced pressure. Extracts obtained therefrom were stored at room temperature until further analysis.

Fermentation and extraction procedures were repeated as described above but with 5 mM L-valine in the fermentation broths replaced with 5 mM L-isoleucine. Flasks inoculated with *P. fragi* A13BB were designated F and G, and those inoculated with *H. alvei* A23BA were designated H and I. Uninoculated control flask was designated J.

5.4.2 Reaction of Cysteine Thiol Probe with Salinosporamide A and LC-MS Analysis of Reaction Mixture

Cysteine thiol probe (molar mass 316.9 g) and salinosporamide A (molar mass 313.1 g) were obtained from Hughes Research Group, University of Tübingen, Germany. A 10 mM solution of probe was prepared by dissolving 6.3 mg of probe in 2 mL acetonitrile (LC-MS grade, Chromasolv™, Honeywell, Germany) after which 0.5 mL of probe solution was transferred into a Chromacol HPLC glass vial (ThermoFisher Scientific, UK) for LC-MS analysis of probe solution only. A 1 mM solution of salinosporamide A was prepared by dissolving 2.5 mg of salinosporamide A in 8 mL acetonitrile. Afterwards, 100 µL of probe solution was added to 500 µL of salinosporamide A solution followed by 5 µL of triethylamine (Sigma Aldrich, USA) which acted as a catalyst. The reaction was carried out at room temperature and allowed to proceed for 3 hours before LC-MS analysis of reaction mixture.

All LC-MS analyses performed were carried out with a Hypersil GOLD™ C18 3 µm, 100 X 2.1 mm column (ThermoFisher Scientific, UK) fitted with a 4 X 2 mm SecurityGuard™ standard analytical guard cartridge system (Phenomenex, USA) and maintained at room temperature on an Agilent 1200 series HPLC system (Agilent Technologies, USA) coupled to an Agilent 6130 series single quadrupole mass analyser (Agilent Technologies, USA) with an electrospray ionisation (ESI) interface. Diode array detector (DAD) data acquisition was carried out at two wavelengths - 200 and 254 nm, with bandwidths set at 4 nm at both wavelengths. Reference wavelengths and bandwidths were set at 360 and 100 nm, respectively. Mass spectrum data acquisition was carried out in dual polarity in scan mode (mass range of *m/z* 100-1000), with adduct ions *m/z* $[M+H]^+$ presented in the results. Fragmentor voltage was set at 70 V, capillary voltage at 4 kV and drying gas was kept at 350°C with the flow rate of 10 L/min. Nebuliser pressure was set at 30 psig.

For LC-MS analysis of probe solution, 2 µL of solution was injected onto the HPLC column at a flow rate of 0.2 mL/min and a linear mobile phase gradient of t_o=0% solvent B to t_{30}=100% solvent B; where solvent A is water (LC-MS grade, Chromasolv™, Honeywell, Germany) with 0.1% formic acid (LC-MS grade, ThermoFisher Scientific, UK) and solvent B is acetonitrile (LC-MS grade, Chromasolv™, Honeywell, Germany) with 0.1% formic acid.
For LC-MS analysis of probe and salinosporamide A reaction mixture, 2 µL of reaction mixture was injected onto the column at a flow rate of 0.4 mL/min with the mobile phase composition and gradient kept as described above.

5.4.3 Sample Preparation and LC-MS Analysis of Crude Extracts

Crude extracts obtained above (Section 5.4.1) were designated correspondingly as samples A to J and analysed in batches. Each sample was dissolved in 1.5 mL acetonitrile, after which 3 µL of sample solution was injected onto the HPLC column at a flow rate of 0.4 mL/min with the following mobile phase gradient: t_o=0% solvent B, t_{10}=60% solvent B, to t_{25}=100% solvent B; where solvent A is water (LC-MS grade, Chromasolv™, Honeywell, Germany) with 0.1% formic acid (LC-MS grade, ThermoFisher Scientific, UK) and solvent B is acetonitrile (LC-MS grade, Chromasolv™, Honeywell, Germany) with 0.1% formic acid. Blank runs were carried out in between samples.

5.4.4 Reaction of Cysteine Thiol Probe with Crude Extracts and LC-MS Analysis of Reaction Mixtures

Reaction of probe solution with extract solutions were also carried out in batches. For each extract, 500 µL of sample solution (as obtained above in Section 5.4.3) was transferred into an HPLC glass vial after which 100 µL of 10 mM probe solution was added, followed by 5 µL triethylamine. Reactions were carried out at room temperature and proceeded for 3-5 hours after which 3 µL of reaction mixtures were injected onto the HPLC column at a flow rate of 0.4 mL/min with mobile phase gradient and composition as detailed above in Section 5.4.3.

5.4.5 Screening of Crude Extracts for Antibacterial Activity

While it was likely that any bioactive compound(s) in crude extracts obtained in Section 5.4.1 were present in very minute quantities and therefore unlikely to exhibit antibacterial activities without scale-up and extract concentration, the crude extracts were nonetheless screened for activity. For each extract (A-J), approximately 1 mL of crude extract solution in acetonitrile obtained above (Section 5.4.3) was evaporated to dryness at 45°C under reduced pressure and redissolved in 500 µL of 50:50 dimethyl sulfoxide (Sigma Aldrich, USA) solution in water. Afterwards, for each extract obtained from inoculated broth (i.e. A-D and F-I), 20 µL aliquots of extract solution were spotted onto two blank Oxoid™ antimicrobial susceptibility testing discs (ThermoFisher Scientific, UK) and allowed to air-dry. One disc was later placed in a quadrant of an agar plate overlaid with a lawn of *E. coli* NCTC 4174 and the other disc was placed in a quadrant of another agar plate overlaid with a lawn of *S. aureus* NCTC 6571. The remaining three quadrants on the two agar plates each had a blank disc impregnated with 20 µL of 50:50 dimethyl sulfoxide solution in water, or an Oxoid™ gentamicin 10 µg antimicrobial susceptibility disc (ThermoFisher Scientific, UK), or a blank disc impregnated with 20 µL of a solution of extract E or J in 50:50 dimethyl sulfoxide in water as appropriate. All agar plates were incubated at 37°C for 24 hours after which plates were inspected for clear zones of inhibition.

5.5 Results and Discussion

5.5.1 Strain Fermentation

Both strains utilised acetate as a carbon source in the presence of 0.5% yeast extract during fermentation. Initial concentration of acetate in fermentation broths was approximately 10.8 mM, which would have risen slightly at the start of bacterial growth as more preferred carbon sources (e.g. gluconeogenic amino acids) in the yeast extract would have been assimilated first and metabolised to pyruvate to produce ATP, accompanied with the release of acetate into extracellular fluids. Acetate assimilation via the Acs pathway and subsequent metabolism into acetyl CoA would have sustained growth and cellular metabolism after the pool of more preferred carbon sources was depleted. Both strains were also able to utilise amino acids as nitrogen source. Interestingly, *H. alvei* A23BA appeared to have accumulated white insoluble plastic-like residues in fermentation broths supplemented with L-isoleucine (Figure 5.6). This is reminiscent of bacterial

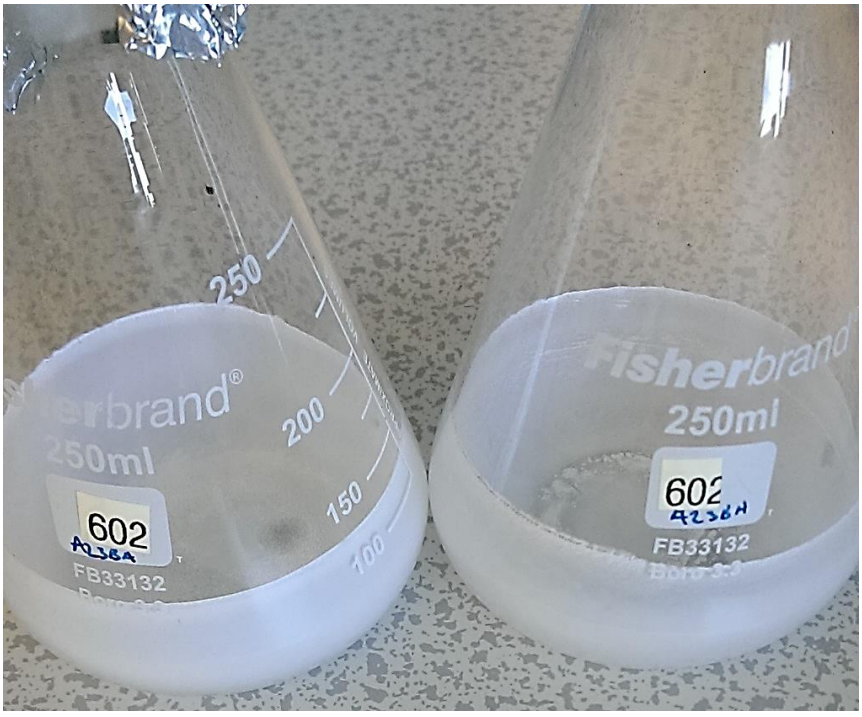


Figure 5.6: White insoluble plastic-like residues produced by *H. alvei* A23BA.

biosynthesis of biodegradable polyhydroxyalkanoates (PHAs). PHAs are biosynthesised when producing strains grow in nutrient-limited (especially nitrogen-limited) environments in the presence of excess carbon source.[311-313] This observation prompted the re-examination of the metabolic reconstruction generated by RASTtk for *H. alvei* A23BA (Section 4.5.1), but no variants of the PHA subsystem was identified. Perhaps because a novel biosynthetic pathway or novel variant of the PHA subsystem exists in *H. alvei* A23BA. Although not the main aim of this bioprospecting work, possible production of PHAs by *H. alvei* A23BA through a novel pathway or subsystem variant is noteworthy as it is unprecedented in the species. This attribute is worth investigating further given increasing industrial and environmental relevance of biodegradable polyesters.

5.5.2 LC-MS Analysis of Cysteine Thiol Probe Solution in Acetonitrile

Probe only solution was analysed within 30 minutes of preparing the solution. Figure 5.7 shows a section of the LC-MS chromatogram obtained, while the full chromatogram is presented in Appendix 10. LC chromatogram at 254 nm (Figure 5.7) shows three peaks corresponding to: **a)** an unreactive impurity (molar mass 285.0 g), which is a side-product of probe synthesis, eluting at t_R=18.5 minutes; **b)** unreacted probe eluting at t_R=20 minutes; and **c)** probe dimer (molar mass 631.9 g) eluting at t_R=24.5 minutes. The characteristic isotopic pattern resulting from the presence of a single bromine atom in a molecule is observed in the mass spectrum patterns of peaks a and b, while the mass spectrum pattern of peak c is consistent with the unique isotopic pattern obtained when two bromine atoms are present in a molecule (Figure 5.8).

Performing LC-MS analysis of probe solution in acetonitrile gave indications of retention times of analytes with the mobile phase flow rate and gradient employed. As retention times were rather long, mobile phase flow rate was increased from 0.2 mL/min to 0.4 mL/min for subsequent analyses. LC-MS analysis of probe solution in acetonitrile also provided reference MS patterns for the probe, impurity, and probe dimer.

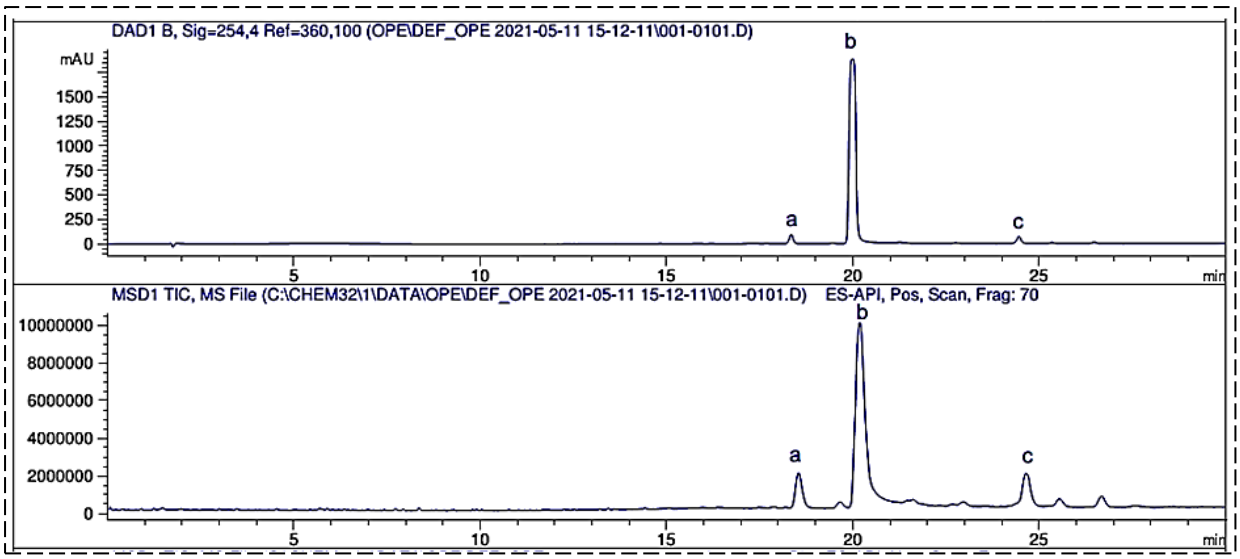


Figure 5.7: A section of the LC-MS chromatogram of probe solution in acetonitrile. Probe solution was analysed within 30 minutes of preparation. Chromatogram shows 3 peaks corresponding to unreactive impurity (peak a), unreacted probe (peak b), and probe dimer (peak c).

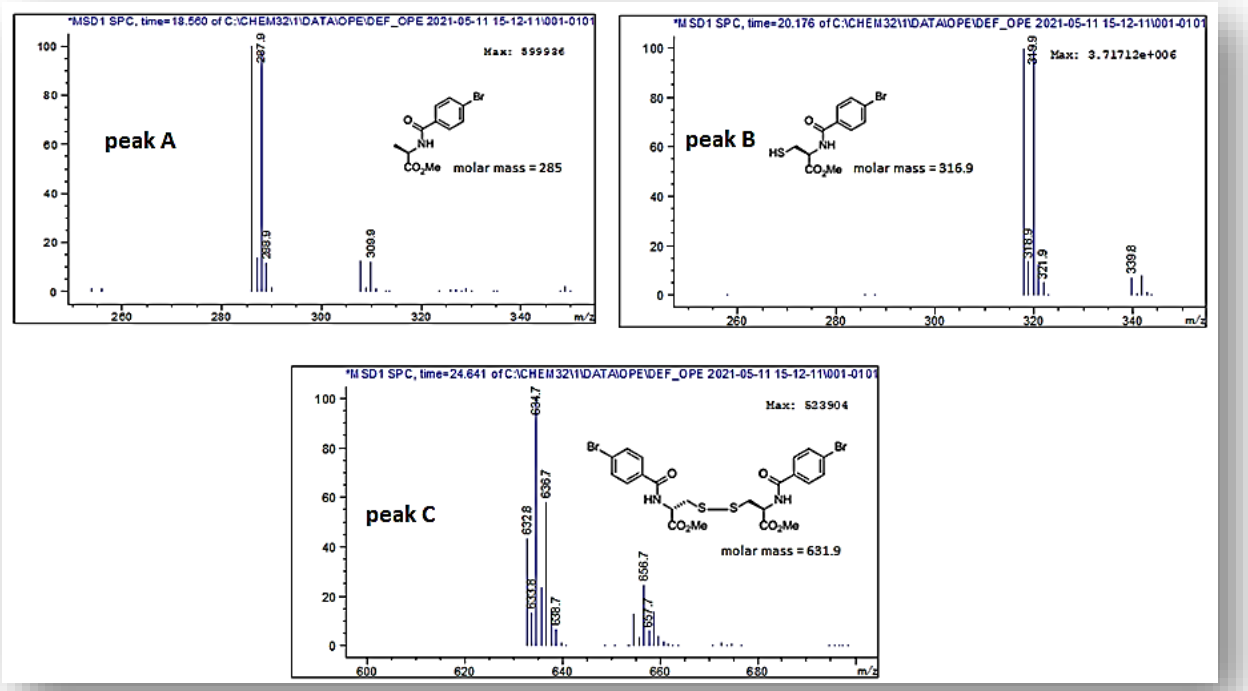


Figure 5.8: Mass spectrum patterns of peaks a, b, and c. peak A (unreactive impurity; mass=285.0), peak B (unreacted probe; mass=316.9), peak C (probe dimer; mass=631.9). The presence of halogen atoms in these molecules resulted in characteristic MS patterns that were easy to identify.

5.5.3 LC-MS Analysis of the Reaction Mixture of Cysteine Thiol Probe and Salinosporamide A

LC-MS analysis of reaction mixture was carried out at 3 hours after initiation of reaction. The probe reacts with salinosporamide A to form two products (product 1: molar mass 594.0 g, and product 2: molar mass 630.0 g) that can be retrosynthetically traced back to salinosporamide A.[287] Figure 5.9 shows a section of the LC-MS chromatogram obtained, while the full chromatogram is presented in Appendix 11. LC chromatogram at 254 nm (Figure 5.9) shows four peaks corresponding to the impurity (peak a) at t_R=13.25 minutes, unreacted probe (peak c) at t_R=14.9 minutes, product 2 (peak e) at t_R=17.9 minutes, and probe dimer (peak f) at t_R=19.6 minutes. The MS total ion chromatogram (TIC) however showed additional peaks: peak b corresponding to salinosporamide A (not fully resolved from probe) at t_R=14.7 minutes, and peak d corresponding to product 1 at t_R=17.2 minutes. Salinosporamide A has no UV absorption at 254 nm, hence the absence of a corresponding peak on the LC chromatogram. Likewise, product 1 was present in the reaction mixture at 3 hours at a concentration below the detection limit of the LC detector hence the absence of a corresponding peak on the LC chromatogram, but a corresponding peak was captured on the MS TIC, demonstrating the robustness of hyphenated systems. It should be noted that even though salinosporamide A is not UV visible at 254 nm, it however formed products with the probe that are UV visible, albeit product 1 was present at undetectable concentration at 3 hours. It also appears that probe dimerization is the thermodynamically preferred reaction as peak height for the probe dimer (peak f) on both LC and MS chromatograms is significantly higher than the peak height for products 1 and 2 (i.e. peaks d and e).

The presence of halogen atoms in the molecules of salinosporamide A, unreacted probe, product 1, and product 2 made their corresponding peaks on the MS TIC easy to identify by looking for characteristic patterns attributable to chlorine or bromine atoms (Figure 5.10). Salinosporamide A showed MS pattern characteristic of a molecule containing a chlorine atom due to the relative abundance of the two most abundant chlorine isotopes i.e. ^{35}Cl:^{37}Cl=3:1. Likewise product 1 and probe showed MS pattern characteristic of a molecule containing one bromine atom, while product 2 was easily identified because of the characteristic pattern attributable to both chlorine and bromine atoms being present in a molecule.

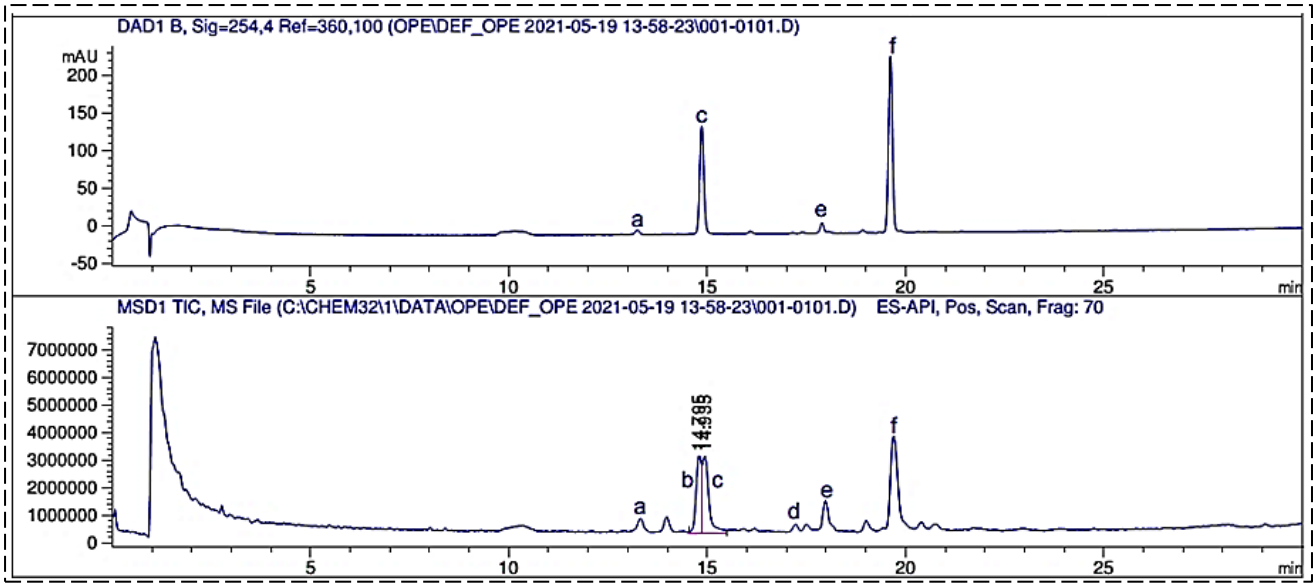


Figure 5. 9: A section of the LC-MS chromatogram of reaction mixture of probe and salinosporamide A.
Reaction mixture was analysed at 3 hrs after initiation of reaction. Chromatogram shows peaks corresponding to unreactive impurity (peak a), salinosporamide A (peak b), unreacted probe (peak c), product 1 (peak d), product 2 (peak e), and probe dimer (peak f).

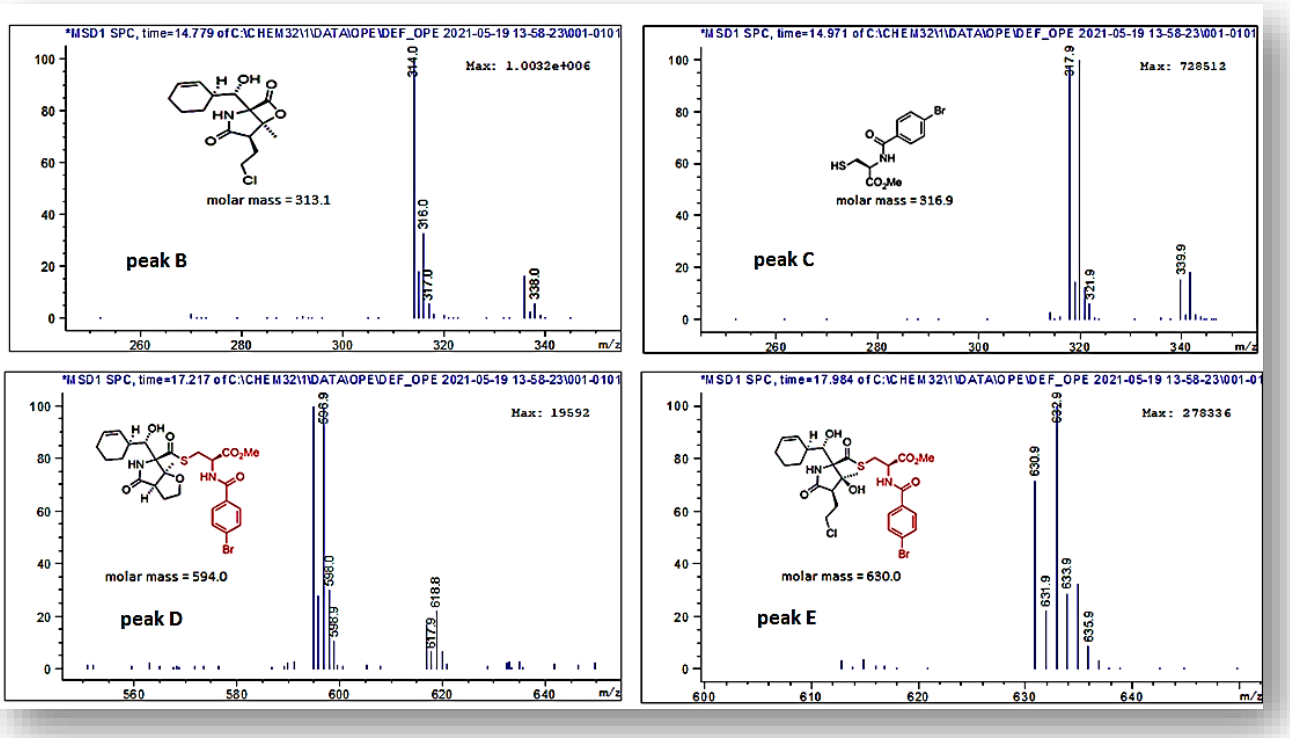


Figure 5. 10: Mass spectrum patterns of peaks b, c, d, and e.
peak B (salinosporamide A; mass=313.1), peak C (unreacted probe; mass=316.9), peak D (product 1; mass=594.0), peak E (product 2; mass=630.0). The presence of halogen atoms in these molecules resulted in characteristic MS patterns that were easy to identify. Probe component of products 1 and 2 is shown in red ink.

The peaks corresponding to unreactive impurity (peak a) and probe dimer (peak f) were also easy to identify on the MS TIC because of their characteristic patterns and a foreknowledge of their respective molar masses (Figure 5.11).

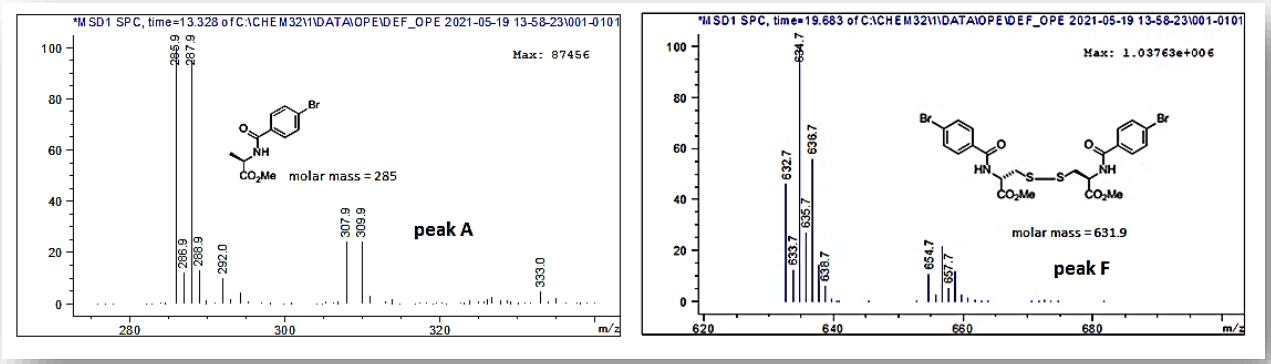


Figure 5.11: Mass spectrum patterns of peaks a and f.
peak A (unreactive impurity; mass=285.0), peak F (probe dimer; mass=631.9).

The reaction of probe with salinosporamide A to form stable adducts (products 1 and 2) that were both UV-visible and conspicuous on MS TIC has demonstrated the important role the probe can play in reactivity-guided screening of crude extracts with complex matrices for electrophilic natural compounds, including β-lactone compounds.

5.5.4 LC-MS Analysis of Solutions of Crude Extracts in Acetonitrile

Solutions of crude extracts in acetonitrile were analysed prior to reaction with probe to ensure extracts did not already contain brominated metabolites, and to obtain reference LC-MS chromatogram for each extract. Mobile phase gradient for these runs needed further alteration to obtain better chromatographic separation. Therefore, a non-linear gradient was used as stated in Section 5.4.3. Injection volume was also increased from 2 µL to 3 µL to increase sample load. All other LC-MS parameters were left unchanged.

LC-MS chromatograms of control samples E and J (Figures 5.12 and 5.13) show large void peaks as expected for unretained relatively polar broth components that were extracted into ethyl acetate. However, analysis of inoculated samples A-D and F-I showed narrower void peaks for unretained broth components compared

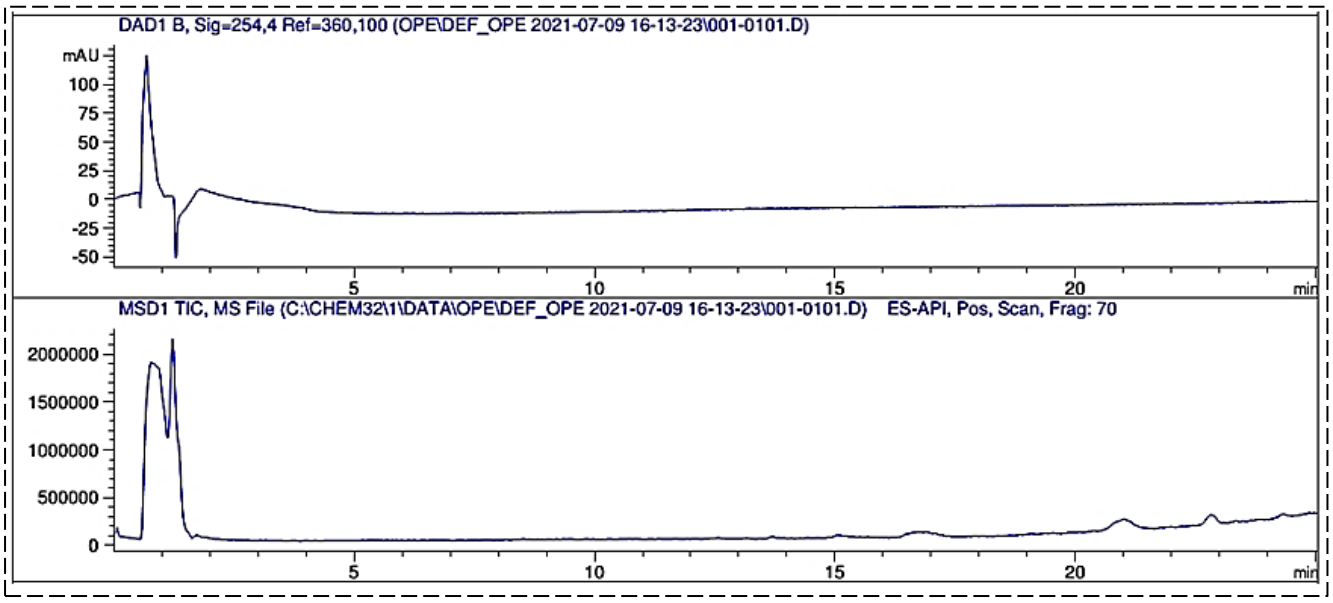


Figure 5.12: A section of the LC-MS chromatogram of sample E solution in acetonitrile.
(Extract of uninoculated broth with L-valine supplementation).
Chromatogram shows large void peak as expected for unretained relatively polar broth components that were extracted into ethyl acetate.

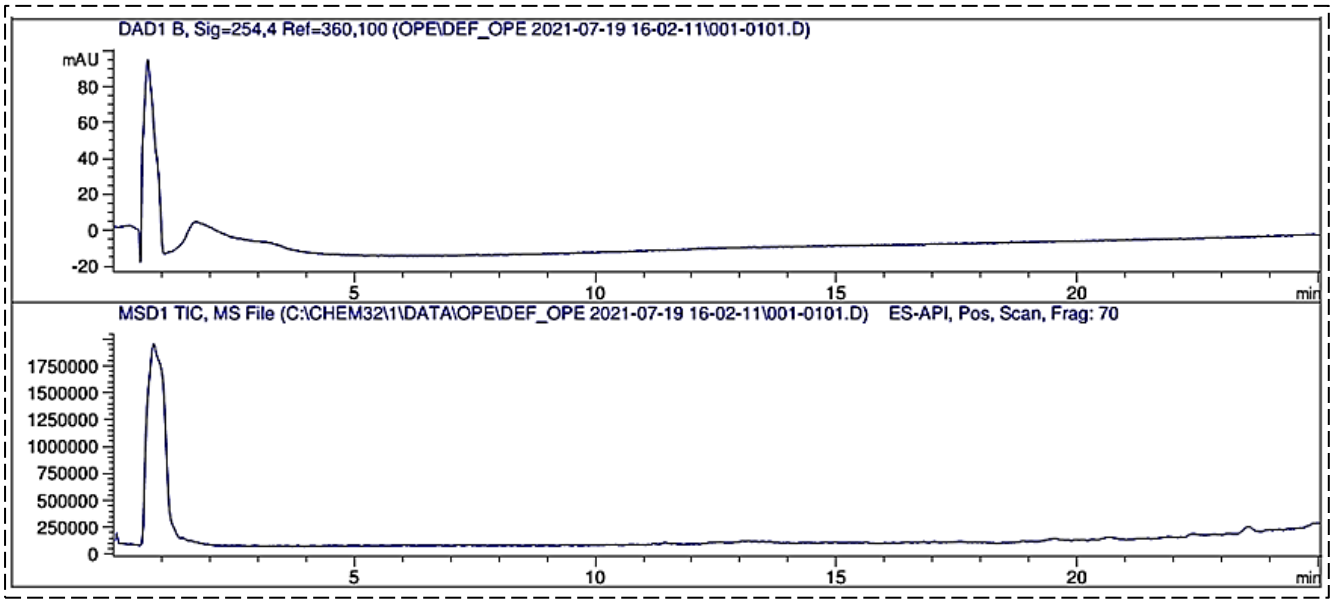


Figure 5.13: A section of the LC-MS chromatogram of sample J in acetonitrile.
(Extract of uninoculated broth with L-isoleucine supplementation).
Chromatogram shows large void peak as expected for unretained relatively polar broth components that were extracted into ethyl acetate.

to the large void peaks in Figures 5.12 and 5.13. As well as the narrower void peaks, analysis of crude extracts obtained from inoculated broths showed various retained peaks of detectable analytes in the extracts, most of which either had no

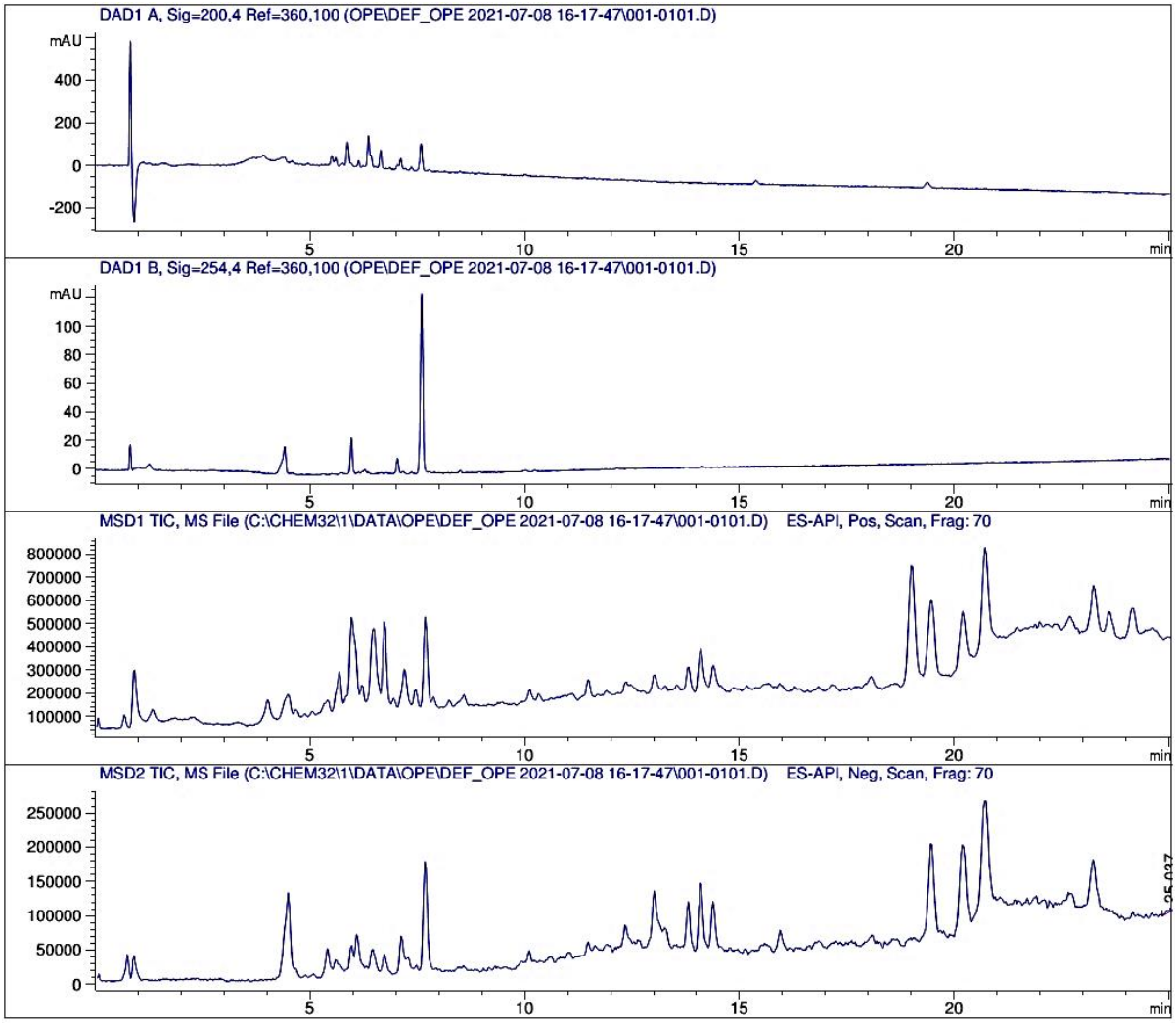


Figure 5.14: LC-MS chromatogram of sample A in acetonitrile.
(Extract of *P. fragi* A13BB inoculated broth with L-valine supplementation).
Chromatogram shows various retained peaks corresponding to detectable analytes in the crude extract.

UV absorption at 200 or 254 nm or were present in extract solutions at extremely low concentrations below the detection limit of the DAD. Consequently, only a handful of analyte peaks were readily visible on LC chromatograms compared to MS TICs as shown in Figures 5.14 and 5.15 corresponding to samples A and H, respectively. LC-MS chromatograms obtained for samples A-D and F-I show some similar features including similar analyte retention times and MS patterns. This probably reflects the likelihood of some metabolites being common to all extracts obtained from inoculated broths. This was expected as both strains were recovered from the same ecological niche and therefore may share some ubiquitous metabolites encoded by HGT-acquired BGCs.[170,189] There is also the possibility that some analytes common to all extracts from inoculated broths were retained broth

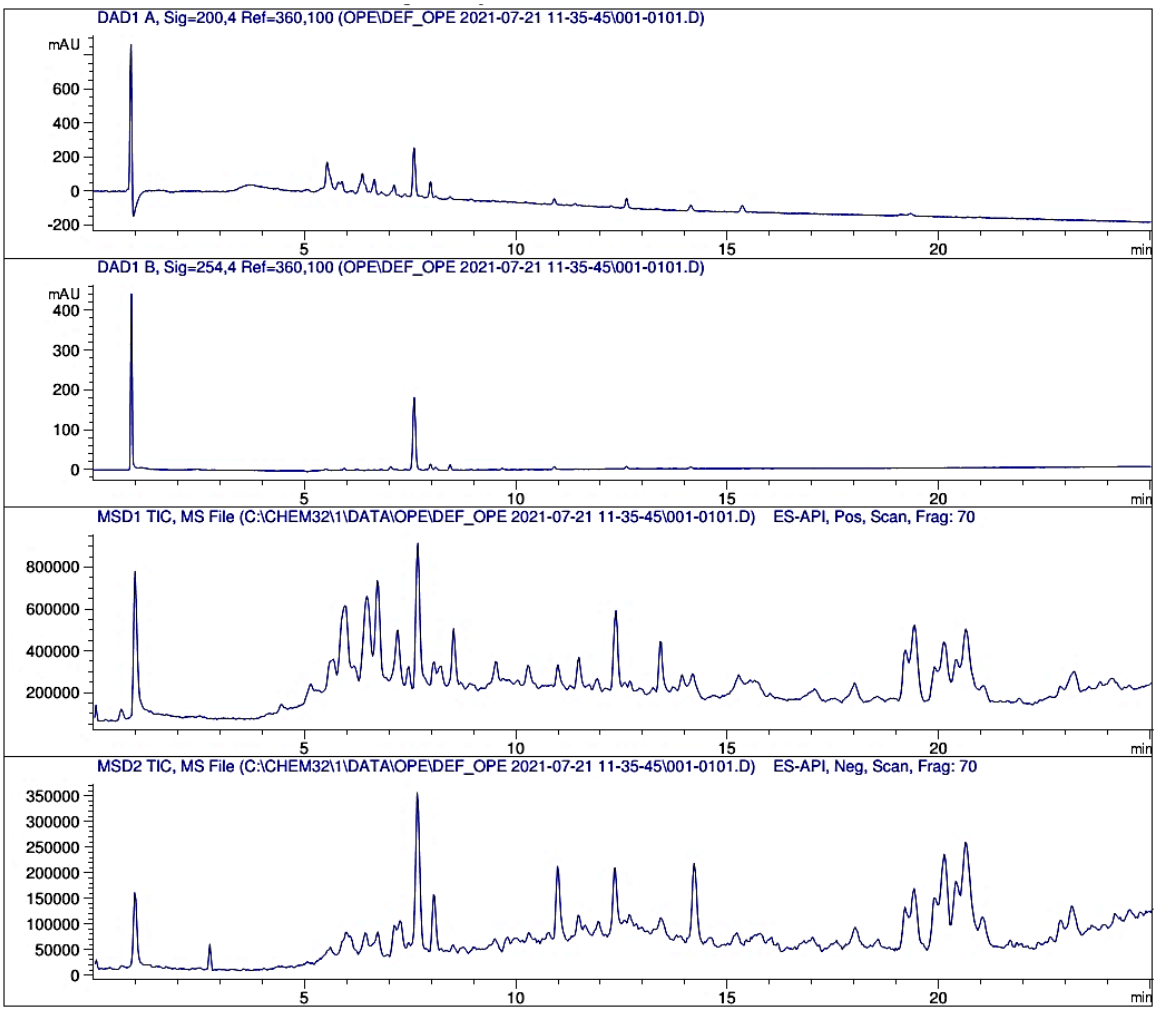


Figure 5.15: LC-MS chromatogram of sample H in acetonitrile.
(Extract of *H. alvei* A23BA inoculated broth with L-isoleucine supplementation). Chromatogram shows various retained peaks corresponding to detectable analytes in the crude extract.

components, but these were expected to be very few indeed if at all present, as no such retained peaks were detected on the LC-MS chromatograms of samples E and J i.e. extracts of uninoculated broths (Figures 5.12 and 5.13). At least 37 metabolite ions were detectable across all inoculated extracts with at least 12 ions common to all, and at least 2 metabolite ions were unique to extracts obtained from inoculated broths supplemented with isoleucine. Chromatograms corresponding to samples A and H (Figures 5.14 and 5.15) are presented here as representative chromatograms, while those corresponding to other inoculated samples are presented in Appendix 12.

The majority of metabolite ions detected in inoculated extracts have masses in the range *m/z* 250-730. If it is assumed that each ion bears a single charge and fragmentation was minimal as a 'soft' ionisation method was used, then a

molecular ion bearing a single charge would have been retained for each metabolite. As such many analytes in inoculated extracts could be assumed to have masses in the range 250-730 Da, which is ideal for antibiotic compounds as compounds ≤ 600 Da tend to penetrate bacterial cell wall better.[16,113,114]

Examination of MS pattern of metabolite ions on TICs showed none were halogenated as none showed the characteristic MS patterns associated with halogen atoms. MS pattern of some ions common to all inoculated extracts are presented in Appendix 13, while those of ions unique to extracts obtained from inoculated broths supplemented with isoleucine are presented in Appendix 14.

5.5.5 LC-MS Analysis of Reaction Mixtures of Cysteine Thiol Probe with Crude Extracts

Reactions between samples A-E (i.e. crude extracts obtained from fermentation broths supplemented with L-valine) and probe were conducted in two batches and reactions were allowed to proceed for 3-5 hours before reaction mixtures were analysed (each LC-MS sample run followed by blank run took approximately one hour, hence reaction mixtures were analysed in sequence, one hour apart). The reaction mixture of sample E (extract of uninoculated broth) with probe was first to be analysed at 3 hours after initiation of reaction. A section of the LC-MS chromatogram obtained is presented here (Figure 5.16). It is similar to the LC-MS

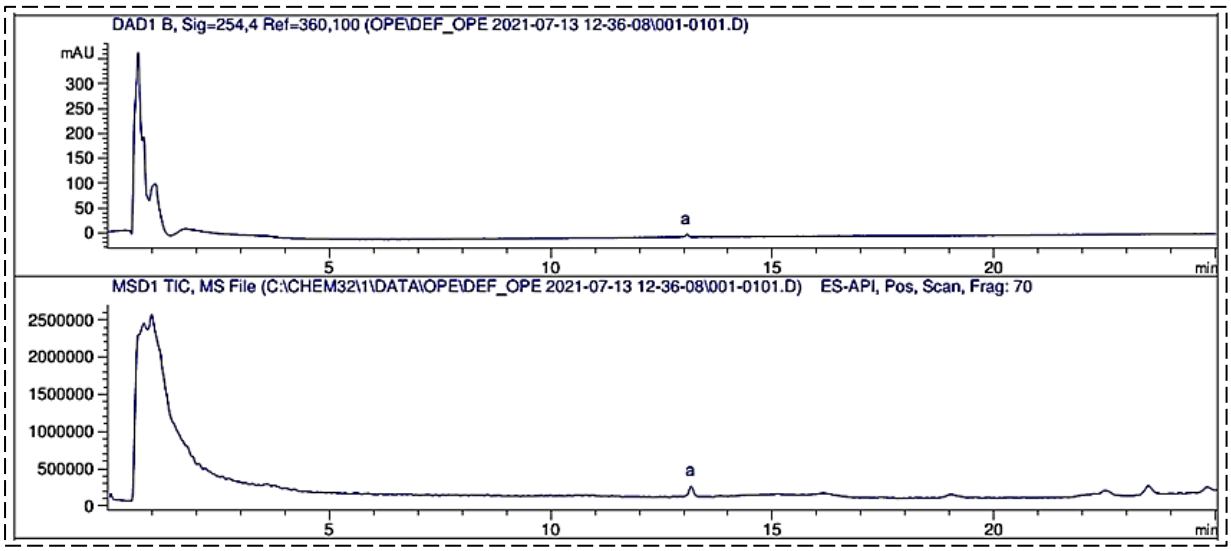


Figure 5.16: LC-MS chromatogram of reaction mixture of sample E with probe.
(Reaction mixture of uninoculated broth with L-valine supplementation and probe).
Chromatogram shows void peak and a retained peak (peak a) corresponding to probe dimer.

chromatogram obtained from the analysis of solution of sample E in acetonitrile (Figure 5.12) with a large void peak for unretained broth components. However, Figure 5.16 also shows a retained peak at t_R=13.1 minutes (peak a). Examination of the MS pattern of peak a confirmed the identity of the analyte ion as the probe dimer. There were no peaks corresponding to unreacted probe on the MS TIC implying that the probe had either completely dimerised at 3 hours (which would have proceeded at a significantly faster rate than was the case in the reaction of probe with salinosporamide A) or any unreacted probe was present at a concentration below the detection limits of the DAD and the mass detector. There was also significant peak height suppression of the probe dimer (peak a) due to matrix effects i.e. ionisation suppression by broth components present in sample E. The full LC-MS chromatogram of the reaction mixture of sample E with probe and the MS pattern of peak A are presented in Appendix 15.

The reaction mixture of sample A with probe was next to be analysed at 4 hours, followed by reaction mixture of sample B at 5 hours. Samples C and D were reacted with the probe the following day and reaction mixtures were analysed at 3 hours and 4 hours, respectively. LC-MS chromatograms obtained from analysis of reaction mixtures of samples A-D showed some similar patterns, with the chromatogram corresponding to sample B presented here as the representative chromatogram (Figure 5.17) while the rest are presented in Appendix 16. Unlike in Figure 5.16, peak a (probe dimer) at t_R=13 minutes was quite prominent on the chromatogram in Figure 5.17 in the absence of the suppressive effects of the broth components in the latter. The two brominated ions in the TIC in Figure 5.17 were those corresponding to probe dimer and unreactive impurity (Figure 5.18), i.e. peaks a and b. There were no other detectable brominated ions. This was also the case with the TICs obtained for the reaction mixtures of samples A, C, and D with probe (Appendix 16). Absence of a peak corresponding to unreacted probe in any of the TICs further reinforced the suspicion that the probe had dimerised completely at a much faster rate than expected and therefore was not available for labelling reactions, more so as unreactive impurity (peak b) was visible at t_R=9.68 minutes (Figures 5.17) meaning unreacted probe, if present, would have been detected too. That being the case, it was necessary to design measures that would minimise or eliminate probe dimerization as it appeared to be proceeding at a faster-than-expected rate in crude extract matrices with current experimental

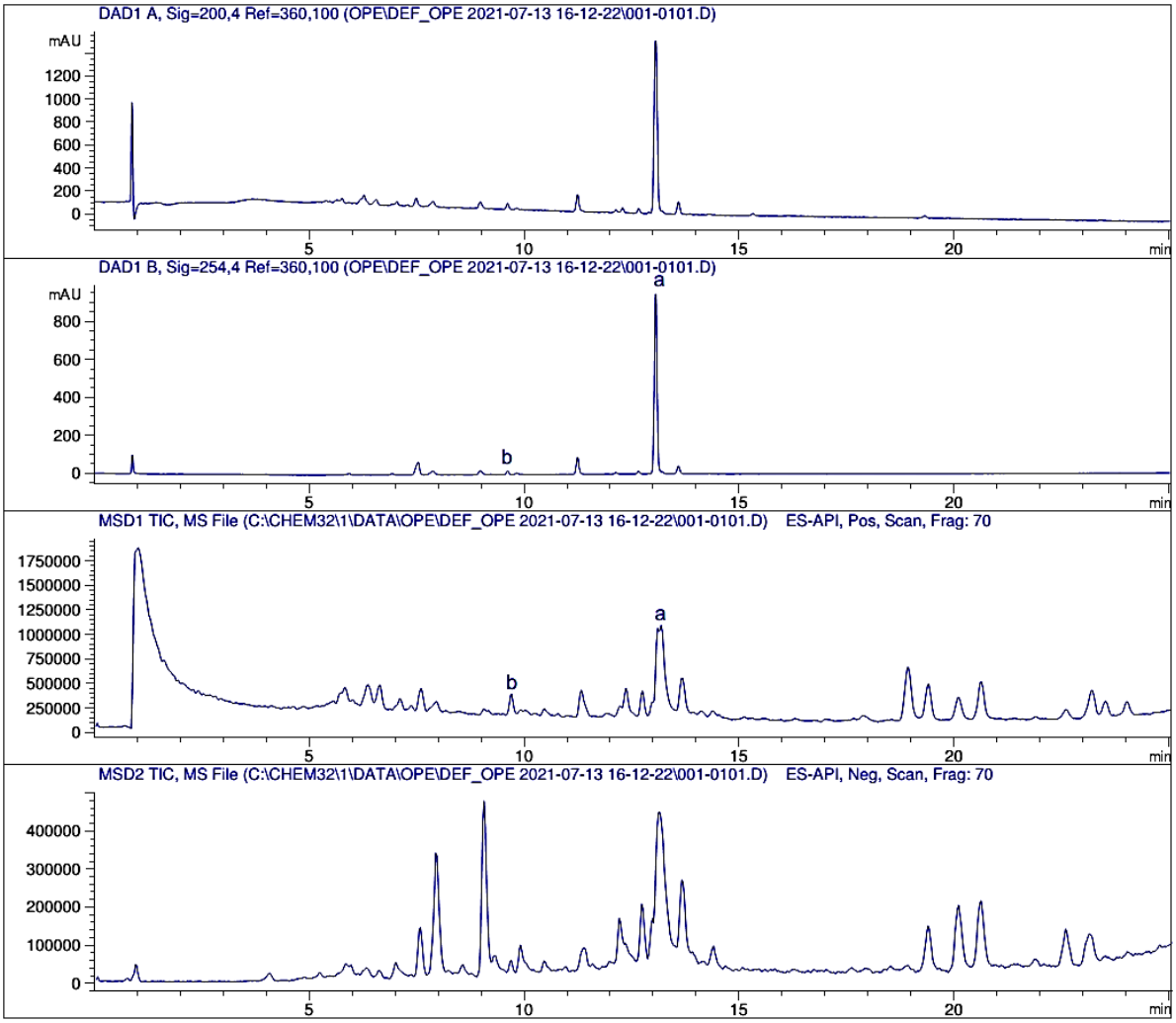


Figure 5.17: LC-MS chromatogram of reaction mixture of sample B with probe.
(Reaction mixture of *P. fragi* A13BB inoculated broth with L-valine supplementation and probe). Chromatogram shows peaks for probe dimer (peak a) and unreactive impurity (peak b), and peaks for other unbrominated analytes. Absence of a peak corresponding to unreacted probe showed probe had completely dimerised, and therefore unavailable for labelling reactions.

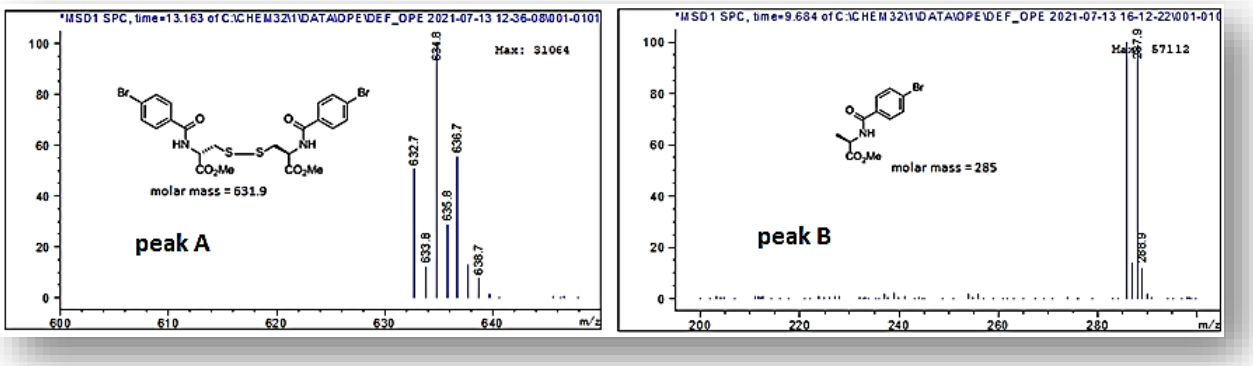


Figure 5.18: Mass spectrum patterns of peaks a and b.
peak A (probe dimer; mass=631.9), peak B (unreactive impurity; mass=285.0).

protocol. Such measures will include degassing solvents used for dissolving extracts and probe, performing all reactions under inert atmosphere using nitrogen gas, and adding tris (2-chloroethyl) phosphine hydrochloride salt to reaction mixtures. Therefore, reactions of probe with samples F-J (i.e. crude extracts obtained from fermentation broths supplemented with L-isoleucine) were not carried out at this stage until these measures are evaluated and implemented if found to be effective in minimising or eliminating probe dimerization.

5.5.6 Screening of Crude Extracts for Antibacterial Activity

Test strains were sensitive to 10 µg gentamicin as clear zones of inhibition were observed in respective quadrants on all plates. No zones of inhibition were observed for solutions of extracts A-D or F-I on all plates. Likewise, no zones of inhibition were observed for solution of dimethyl sulfoxide in water, or for solutions of extract E or J. Essentially all plates were identical after 24 hours incubation, therefore only plate A (Figure 5.19) overlaid with *E. coli* NCTC 4174 with a blank disc impregnated with 20 µL of solution of extract A is presented here. It was already envisaged that any bioactive metabolites in extracts of inoculated broths

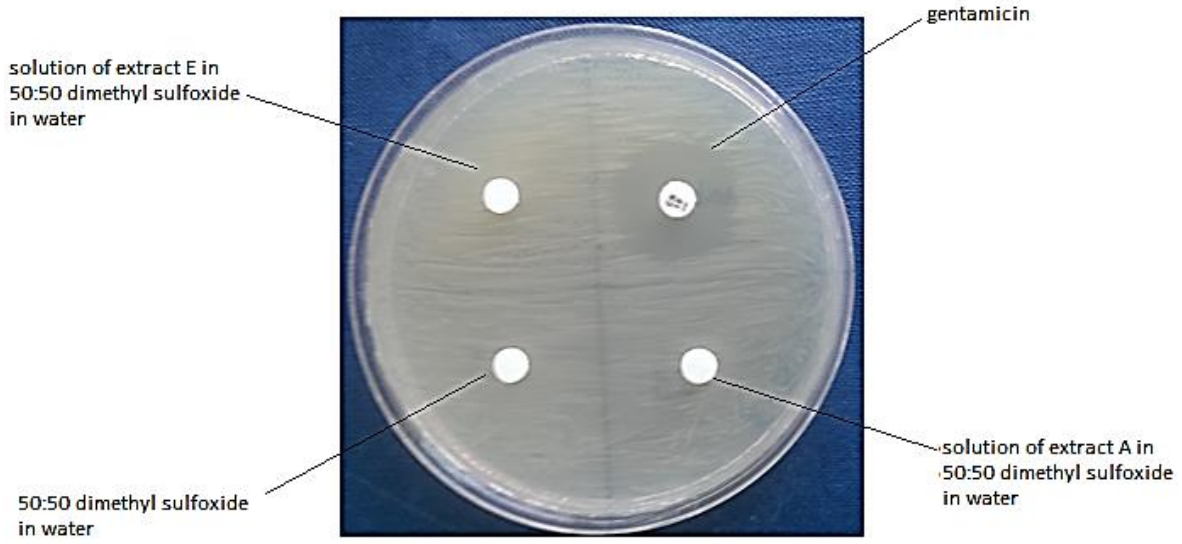


Figure 5.19: Plate A (showing sensitivity testing result for sample A).
Agar plate overlaid with *E. coli* NCTC 4174 with four quadrants each containing an impregnated disc as depicted in the image. No zone of inhibition was obtained for solution of extract A. Culture scale-up and extract concentration would be necessary before future sensitivity testing.

were likely to be present in minute concentrations and therefore unlikely (except for highly potent metabolites) to exhibit *in vitro* activity without culture scale-up and extract concentration.

5.6 Conclusions

All three β-lactone smBGCs in *H. alvei* A23BA and *P. fragi* A13BB encode homologues of 2-IPMS and acyl CoA ligase and synthetase. Other biosynthetic enzymes encoded in at least one of the smBGCs include those involved in fatty acid metabolism or β-oxidation, and those involved in leucine biosynthesis or degradation. Initial reactions of the β-lactone biosynthetic pathway mirror those of leucine biosynthesis where 2-IPMS catalyse the formation of 2-IPM from the condensation of acetyl CoA and 3-methyl-2-oxobutanoate (a degradation product of valine), followed by isomerisation of 2-IPM to 3-IPM. In β-lactone biosynthesis, 3-IPM is activated by an acyl CoA ligase homologue followed by intramolecular cyclisation of the activated intermediate to form the β-lactone warhead in a step that is also catalysed by acyl CoA ligase.

In order to activate the β-lactone smBGCs in the two strains, fermentation conditions that favoured production of substrates of 2-IPMS (i.e. acetyl CoA and 3-methyl-2-oxobutanoate) and acyl CoA ligase (i.e. 3-IPM), as well as induce simultaneous upregulation of many genes found in the clusters were employed. Strain fermentations were carried out with 10.8 mM acetate as the main carbon source in the presence of 0.5% yeast extract while excluding ammonium salts and nitrates, but supplementing fermentation broths with L-valine or L-isoleucine. Both strains were able to utilise these broth components for growth and metabolism during fermentation experiments. *H. alvei* A23BA also appeared to have accumulated plastic-like residues in broths supplemented with L-isoleucine - an observation that is reminiscent of bacterial PHA production and unprecedented in the species. At least 37 compounds or metabolite ions were detected in solutions of crude extracts obtained from inoculated broths. These were not detected in extract solutions obtained from uninoculated or blank broths. The majority of these ions have masses in the range *m/z* 250-730; some were common to all extracts obtained from inoculated broths, and at least two were unique to extracts obtained from fermentation broths supplemented with L-isoleucine.

Reactivity guided screening of crude extracts for β-lactone compounds using cysteine thiol probe was unsuccessful as the probe completely dimerised in the complex matrices of the crude extracts at a rate faster than expected and therefore was not available for labelling reactions. As probe dimer appeared to be the thermodynamically preferred product during labelling reactions, measures to minimise or eliminate dimerization are crucial. Such measures will include degassing all solvents to remove dissolved molecular oxygen, carrying out labelling reactions under inert atmosphere using nitrogen gas, and adding tris (2-chloroethyl) phosphine hydrochloride salt to reaction mixtures. Salinosporamide A can also be added to reaction mixtures as an internal standard.

In vitro screening of crude extracts for antibacterial activity showed no clear zones of inhibition. However, it is very likely that bioactive metabolites in crude extracts were present in very minute quantities. Therefore culture scale-up, fractionation, and extract concentration would be necessary prior to future *in vitro* screening for activity.

CHAPTER SIX

SYNOPSIS AND FUTURE WORK

6. Synopsis and Future Work

The aim of this project was to expand the chemical space of bacterial secondary metabolites that can serve as novel antibiotic lead compounds. The genome-guided approach adopted facilitated a rational bioprospecting effort in which β-lactone smBGCs were selected for activation as their end products bear the hallmarks of an 'ideal antibiotic'. The gene cluster analysis carried out informed fermentation conditions that generated many small molecules that are yet to be characterised.

6.1 Synopsis

AMR continues to be a global public health concern as it threatens the ability to treat or prevent infections. Various policies and strategies have been adopted to stem or even reverse this natural process, with the most crucial strategy being the development of novel antibiotics. The need for novel therapeutics is even more urgent given the exacerbating effects of the COVID-19 pandemic on AMR.[314-317] Bacterial secondary metabolites remain the most promising but underexploited lead compounds, therefore expanding the library of bacterial metabolites that can be screened for activity is key to the fight against AMR. As the ability to biosynthesise therapeutically useful secondary metabolites is not universal to all bacterial species, isolation methods that will preferentially recover secondary metabolite producers are equally important. To that end, 1:100 Ravan medium that can induce starvation stress was used to isolate bacterial strains from a nutrient-dense topsoil sample collected from the rhizosphere, which represents a niche in which antibiosis occurs naturally.[172-174] It was anticipated that recovered strains with genomes larger than 3 Mb would have increased genome capacity to encode secondary metabolites, some of which might have antibiotic properties and potentially serve as novel antibiotic lead compounds.

Another objective of the project was to prioritise species that are not readily associated with secondary metabolite and/or antibiotic production for the bioprospecting work. This is to explore the diverse chemistries encoded by less studied species and to decrease the chances of isolating known metabolites. This objective contrasts with common practice in which species well known for secondary metabolites production, e.g. *Streptomyces* spp, are prioritised for

prospecting for novel compounds with a view to activating silent smBGCs. However, silent smBGCs in a particular species may not necessarily encode novel metabolites as the same smBGCs may be expressed readily in another closely related species. For instance, the smBGC encoding the polyketide antibiotic actinorhodin is silent in *Streptomyces lavendulae* but is readily expressed in *S. coelicolor*.[318,319] Therefore, taxonomic groups that are prolific secondary metabolite producers may be close to being overmined with prospecting efforts frequently yielding known metabolites or their analogues.[319,320] The search for novel compounds will undoubtedly require casting the net wider to include un-/under-exploited taxa that may encode novel chemistries.[319-322] More so as the majority of broad-spectrum antibiotics in clinical use today (with the exemption of β-lactams) were developed from metabolites that are restricted to actinomycetes, which means bioactive metabolites of species belonging to other taxa are largely waiting to be discovered as the latter must have developed their own metabolites to ward off Gram-positive and -negative competitors in order to survive.[319]

This bioprospecting effort led to the recovery of *H. alvei* A23BA and *P. fragi* A13BB, both of which are unexploited species for antibiotic production. The complete genomes of *H. alvei* A23BA and *P. fragi* A13BB were assembled from high-quality sequencing reads, and with genome sizes greater than 3 Mb (4.77 and 4.94 Mb, respectively), both isolates were anticipated to harbour antibiotic-encoding smBGCs given their ecological origin and ability to withstand starvation stress. Genome mining for putative drug-like smBGCs detected thiopeptide, β-lactone, and siderophore smBGCs in the genome of *H. alvei* A23BA. The first two smBGCs show no homology to known smBGCs and as such may encode novel metabolites. Likewise, the genome of *P. fragi* A13BB contains two β-lactone smBGCs that show little homology (20%) to known smBGCs, and a siderophore smBGC. The two β-lactone smBGCs in *P. fragi* A13BB may also encode novel metabolites. Thiopeptide and β-lactone compounds have antibiotic properties, while siderophores are known to enhance or restore antibiotic activity by increasing cellular uptake of antibiotic molecules.[268,273,278] Furthermore, β-lactone compounds have been shown to inhibit several bacteria-specific enzymes, thus potentially exhibiting multiple modes of action when used as antibiotics.[273,276] The ability to inhibit several bacterial processes simultaneously is highly desirable in the development of novel antibiotics, not least because of the effectiveness of such compounds

against persister cells and cells embedded within biofilms, which are arguably more problematic than resistant cells.[16,114,323,324] Although β-lactone compounds hold huge potentials as 'ideal antibiotics', poor cellular penetration may preclude their use as such. This can be overcome by structural optimisation incorporating the rules of permeation already described in section 1.5.3. and other recently proposed design considerations which include a rigid molecular structure with very few rotatable bonds, a flat or elongated molecular shape, and incorporating the positively charged amino group, NH_3^+, into the antibiotic molecule at a position that does not obstruct its site of action.[16,113,325,326] Cellular penetration can also be enhanced by conjugating β-lactone compounds to siderophores, with the siderophore-antibiotic prodrug complex dissociating within bacterial cells with the aid of bacteria-specific enzymes to yield the free antibiotic molecule.[278,280,281] Conjugation of β-lactone compounds to siderophores has added advantages in that unwanted off-target effects of the β-lactone antibiotic in human hosts are prevented and species-specific siderophores can be used to narrow the spectrum of activity of the antibiotic molecule thereby protecting the normal microbiome in the host.[278,282,327]

Given the unexploited potentials of β-lactone compounds as antibiotics, the amide ligase-type β-lactone smBGCs in the two isolates were selected for activation in the first instance. Gene cluster analysis revealed genes encoding two core enzymes: 2-isopropylmalate synthase (2-IPMS) and acyl CoA ligase and synthetase homologues. Other biosynthetic enzymes encoded in the clusters include those involved in fatty acid metabolism, and leucine biosynthesis and degradation. It was speculated that majority of genes in the clusters may be simultaneously upregulated and expressed in fermentations carried out with poor carbon and nitrogen sources - a set of conditions that might induce gene cluster activation and production of β-lactone compounds. Therefore, strain fermentations were carried out with 10.8 mM acetate as the main carbon source and amino acids (5 mM L-valine or L-isoleucine) as the nitrogen source. Crude extracts were generated from fermentation broths and LC-MS analysis detected at least 37 metabolite ions in the extracts, many of which have masses ranging from *m/z* 250-730. Extracts were later screened for β-lactone natural compounds.

A reactivity-guided screening approach was adopted as β-lactone compounds in crude extracts are relatively unstable and are easily missed during rapid screening

of crude extracts for natural compounds.[273] Cysteine thiol probe, which mimics the *in vivo* reaction of cysteine and electrophilic natural compounds, was used for the screening exercise.[287] Prior to crude extract screening, solution of probe in acetonitrile was reacted with solution of salinosporamide A in acetonitrile to assess reactivity of the probe with a known β-lactone compound. Two adducts of probe and salinosporamide A that were MS- and UV-visible were formed within 3 hours. Probe dimer that appeared to be the thermodynamically preferred product was also formed at 3 hours. Solution of probe in acetonitrile was later reacted with solutions of crude extracts generated from fermentation broths. Reaction mixtures were analysed at 3, 4, or 5 hours with the probe shown to have completely dimerised at 3 hours and therefore not available for adduct formation. Even though crude extracts were shown to contain many metabolite ions, it was not possible at this stage to determine if any of these were β-lactone natural compounds. Therefore, optimisation of experimental conditions to minimise or eliminate probe dimerization is needed.

The work carried out here has demonstrated the positive impacts of genomics on bioprospecting efforts. It is now possible to move away from ad hoc discovery platforms to rational experimental designs. Used in combination with other omics approaches, genomics also enables and expedites initiatives that are driven by novel metabolite discovery followed by bioactivity profiling of promising candidates.[328-331] This approach captures novel metabolites which are likely to be missed by traditional prospecting methods that are preceded by bioactivity screening, followed by analysis of bioactive fractions. A transferrable and adaptable genomics and bioinformatics workflow (Figure 6.1) that can expedite similar bioprospecting projects was created here. All tools and pipelines used are freely available either through the web interface or through the Galaxy platform.[332] Furthermore, as amide ligase-type β-lactone smBGCs are likely to be widespread in bacteria given their medium size, the gene cluster analysis carried out here, together with proposed fermentation and extract screening protocols, will serve as a good starting point for similar gene cluster activation projects.

Lastly, apart from expediting this bioprospecting work, the complete genome sequences assembled here are valuable resources to the wider research community. Therefore, the datasets have been deposited in the public repositories

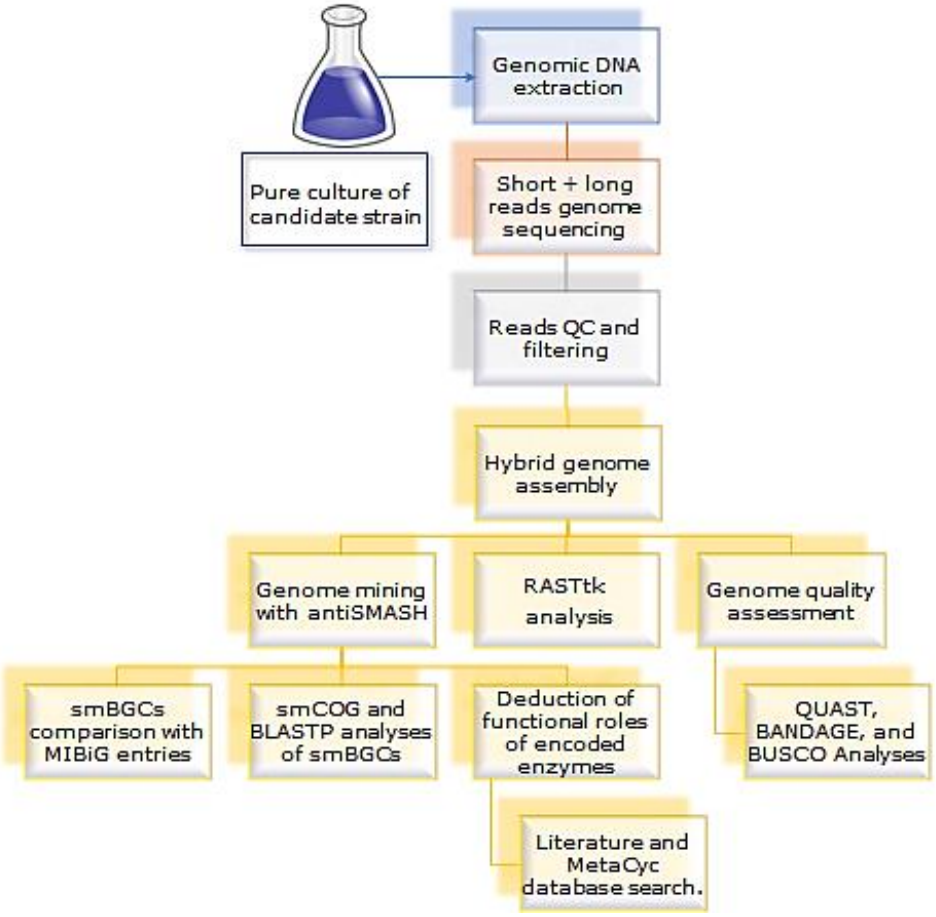


Figure 6.1: Genomics and bioinformatics workflow.

of the International Nucleotide Sequence Database Collaboration.[333,334] The genome assembly of *H. alvei* A23BA has been deposited under accession number GCF_011617105.1, while the genome assembly of *P. fragi* A13BB was deposited under accession number GCA_015767515.1.[335,336] The BioProject accession number for the entire project is PRJNA610978. Archiving the genome sequences of *H. alvei* A23BA and *P. fragi* A13BB in public repositories has significantly increased the number of high-quality genomes of the species that are in the public domain. The complete genome sequence of *H. alvei* A23BA was the first complete genome of a soil isolate to be deposited in the NCBI genome database, and it remains the reference genome of the species in the database at the time of writing. The complete genome sequence of *P. fragi* A13BB was the fourth complete genome of the species to be deposited in the NCBI database. However, the complete genome sequence of strain P121 which was the first to be deposited in 2016 has since had its taxonomy check changed to 'inconclusive', and as such is

no longer the reference genome of the species in the database at the time of writing. Deposition of the complete genome sequence of *P. fragi* A13BB in the database provided the much-needed evidence that the initial taxonomic placement of strain P121 as *P. fragi* was incorrect.

Raw reads data of sequencing projects have been deposited in the Sequence Read Archive (SRA) database under accession number SRP251948.[337] This will enable other researchers who may wish to use the data to repurpose it as required to suit their needs. Furthermore, the hidden metabolic and biosynthetic potentials of both species have been catalogued and placed in the public domain.[237,240,338-340] This has captured the attention of other researchers with requests already made for a culture of *H. alvei* A23BA by researchers at the Max-Planck Institute for Developmental Biology.

6.2 Future work

Six objectives were set out for this work (Section 1.8), with the first five objectives already achieved. The sixth objective is partially achieved with fermentation experiments already conducted and crude extracts generated. Initial LC-MS analysis of extracts detected at least 37 metabolite ions. However, targeted screening of crude extracts for β-lactone natural compound(s) is yet to be achieved. Complete dimerization of the cysteine thiol probe during crude extract labelling reactions has been the main challenge to overcome. Therefore, the next step in the project will involve designing and incorporating measures to minimise or eliminate probe dimerization into the workflow. As discussed in Chapter 5, these measures will include degassing all solvents used for extracts and probe dissolution, carrying out labelling reactions under inert atmosphere using nitrogen gas, adding tris (2-chloroethyl) phosphine hydrochloride salt to reaction mixtures, and spiking crude extract samples with salinosporamide A which will act as an internal standard – an optimised protocol will result in the salinosporamide A added to crude extract matrices being successfully labelled by the probe. Screening of crude extract solutions will be repeated as described, and screening will be extended to include n-butanol and methanol extracts of fermentation broths to ensure relatively polar β-lactone compounds are not missed, if present in the extracts.

Detection of labelled metabolite(s) in a crude extract sample would indicate the presence of β-lactone, β-lactam, or enone natural compounds in the extract. The exact chemical class can later be determined by NMR spectroscopy after large-scale fermentation and compound isolation.[341,342] Alternatively, solutions of crude extracts generated after large-scale fermentation can be reacted with the probe after which any crystalline adduct(s) formed between probe and natural compound is isolated and subjected to X-ray crystallographic analysis in order to deduce the structure of the natural compound.[287,343-345] In fact, the results of these orthogonal structure elucidation strategies can be combined in a validation step. Confirmation of the presence of β-lactone natural compound(s) in crude extracts would mean the β-lactone smBGCs were successfully activated. Isolated compound(s) can then be subjected to bioactivity profiling and structural optimisation.

The reactivity-guided screening adopted here, i.e. crude extract screening for β-lactone compounds using cysteine thiol probe, was designed to simplify the otherwise arduous and time-consuming task of identifying natural compounds in complex crude extract matrices, especially when instrumentation is limited to hyphenated systems utilising single-stage mass spectrometry as was the case here. A disadvantage of reactivity-guided targeted screening however is the lack of global visualisation of metabolites in the sample being analysed. Even though the primary aim of fermentations here was the activation of β-lactone smBGCs, the conditions adopted, i.e. culture broths with poor carbon and nitrogen sources, generally favour secondary metabolites production, including antibiotic production.[346,347] As there are other detected and possibly yet-to-be-detected drug-like smBGCs in the genomes examined, a global visualisation of metabolites in sample matrices is highly desirable. This can be achieved by creating molecular networks for each sample using high resolution tandem mass spectrometry (HR-MS/MS) fragmentation patterns of each parent/precursor ion detected in the samples.[348-350]

The general workflow is illustrated in Figure 6.2, beginning with the generation of HR-MS/MS fragmentation patterns of all MS-detectable ions (i.e. precursor ion) in a sample. The data is then simplified by merging identical MS/MS spectra in the file into one consensus spectrum.[348,349] Consensus spectra are aligned and pairwise similarity scores, depicted by cosine scores 0 to 1, are computed. [348,349]

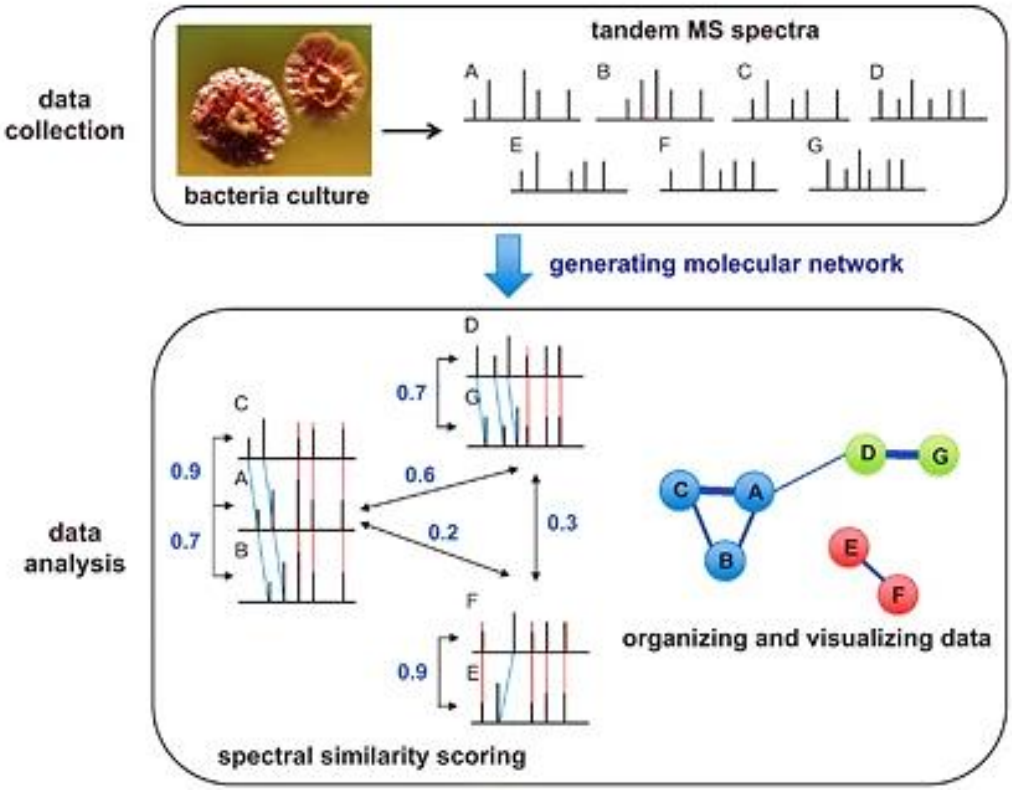


Figure 6.2: A general workflow for creating molecular networks.
The HR-MS/MS fragmentation patterns of all MS-detectable ions in a sample is generated and simplified into consensus spectra. A molecular network is created by clustering spectra with similar patterns into molecular families. Within each family, a consensus spectrum is depicted by a node (i.e. a circle) labelled with the mass of the precursor ion. Similarity of a node to another consensus spectrum is depicted by the edge (i.e. line) joining them, while the thickness of edge indicates the degree of similarity between the two nodes.
Image adapted from Liu *et al*.[349]

A molecular network is created by clustering consensus spectra with similar fragmentation patterns into molecular families based on the assumption that natural products with similar molecular structures share similar MS/MS fragmentation patterns.[351-353] Networks created are then displayed in visualisation programs like Cytoscape.[353-355] Similar molecules (and their analogues) present in a sample are clustered together in a molecular family while potentially unique metabolites usually occur as singletons. Molecular networking can enable rapid dereplication where MS/MS fragmentation patterns obtained for a sample are seeded with those of known compounds obtained from reputable libraries or databases such as the Global Natural Products Social Molecular Networking (GNPS) library, AntiBase® database, and the MassBank database.[356-361] Molecules in samples being analysed that are similar to known compounds used as seeds

will cluster around the latter making it easier to identify potentially novel metabolites. Molecular networks can also be used as the basis for the recently introduced concept of pattern-based genome mining.[349,362]
With access to the right instrumentation, this project can be broadened to include untargeted metabolomics where global visualisation of all metabolites in crude extract samples is enabled by creating molecular networks. A targeted screening can also be performed where MS/MS fragmentation data obtained for each sample is seeded with those of known amide ligase β-lactone compounds.[363]

Other areas of interest identified during this work will also serve as subjects of future projects. These will include activation of the thiopeptide smBGC found in *H. alvei* A23BA, and exploitation of the metabolic potentials of both isolates - especially the potential ability of *H. alvei* A23BA to biosynthesise polyhydroxyalkanoates. Potential use of both strains for bioremediation and/or biocontrol purposes will be other important attributes to explore.

6.3 Conclusions

Bacterial secondary metabolites remain the most versatile but underexploited natural resources in the development of novel antibiotics. Bioprospecting for bioactive bacterial natural compounds has traditionally relied heavily on ad hoc screening programmes that are fast becoming inefficient or unproductive given the scale of the AMR challenge. Genome-guided initiatives, together with other omics approaches, can provide highly efficient and rational alternative strategies to bioprospecting, as demonstrated here.

This work led to the recovery of *H. alvei* A23BA and *P. fragi* A13BB, both of which are unexploited species for antibiotic production. Genome mining of the two isolates detected drug-like smBGCs, i.e. β-lactone and thiopeptide smBGCs, that may encode novel antibiotic lead compounds. Results of gene cluster analysis of detected β-lactone smBGCs informed strain fermentation conditions that generated crude extracts containing many small molecules with masses in the range m/z 250-730. Initial attempts to screen crude extracts for β-lactone compounds were unsuccessful due to unfavourable experimental conditions that accelerated probe dimerization, but measures to improve these have been proposed. Likewise, molecular networking strategies that can afford global

visualisation of metabolites in crude extracts and enable targeted analysis of metabolites of interest have been proposed. It is hoped that these measures and analyses will build on the work already carried out here to result in isolation of novel β-lactone compounds, and indeed other bacterial secondary metabolite(s), that can serve as novel antibiotic lead compounds. In addition, the structural optimisation considerations for β-lactone antibiotic lead compounds put forward in the text may serve to unlock the unrealised potentials of these compounds as 'ideal antibiotics'.

CHAPTER 7

REFERENCES

7. REFERENCES

1. Waksman SA. What is an antibiotic or an antibiotic substance? *Mycologia.* 1947; 39(5):565-569.
2. Waksman SA. Origin and Nature of Antibiotics. *The American Journal of Medicine.* 1949; 7(1):85-99.
3. Bassett E, Keith M, Armelagos G, Martin D, Villanueva A. Tetracycline-Labeled Human-Bone from Ancient Sudanese Nubia (Ad 350). *Science*. 1980; 209(4464):1532-1534.
4. Nelson ML, Dinardo A, Hochberg J, Armelagos GJ. Brief Communication: Mass Spectroscopic Characterization of Tetracycline in the Skeletal Remains of an Ancient Population from Sudanese Nubia 350-550 CE. *American Journal of Physical Anthropology*. 2010; 143(1):151-154.
5. Gould K. Antibiotics: from prehistory to the present day. *The Journal of Antimicrobial Chemotherapy*. 2016; 71(3):572-575.
6. Falkinham, JO 3rd, Wall TE, Tanner JR, Tawaha K, Alali FQ, Li C, *et al*. Proliferation of Antibiotic-Producing Bacteria and Concomitant Antibiotic Production as the Basis for the Antibiotic Activity of Jordan's Red Soils. *Applied and Environmental Microbiology*. 2009; 75(9):2735-2741.
7. Aminov RI. A brief history of the antibiotic era: lessons learned and challenges for the future. *Frontiers in Microbiology*. 2010; 1:134.
8. Schwartz R. Paul Ehrlich's magic bullets. *The New England Journal of Medicine*. 2004; 350(11):1079-1080.
9. Hutchings MI, Truman AW, Wilkinson B. Antibiotics: past, present and future. *Current Opinion in Microbiology*. 2019; 51:72-80.
10. Otten H. Domagk and the development of the sulphonamides. *The Journal of Antimicrobial Chemotherapy*. 1986; 17(6):689-696.
11. Garrod LP. The story of penicillin. *Nature.* 1947; 160:38-39.
12. Waksman SA, Tishler M. The chemical nature of actinomycin, an antimicrobial substance produced by *Actinomyces antibioticus*. *The Journal of Biological Chemistry*. 1942; 142(2):519-528.
13. Waksman SA, Schatz A. Streptomycin - Origin, Nature, and Properties. *Journal of the American Pharmaceutical Association*. 1945; 34(11):273-291.

14. Waksman SA, Lechevalier H. Neomycin, a New Antibiotic Active Against Streptomycin-Resistant Bacteria, Including Tuberculosis Organisms. *Science.* 1949; 109(2830):305-307.
15. Schatz A, Bugie E, Waksman SA. Streptomycin, a substance exhibiting antibiotic activity against Gram-positive and Gram-negative bacteria. *Experimental Biology and Medicine*. 1944; 55(1):66-69.
16. Lewis K. Platforms for antibiotic discovery. *Nature Reviews Drug Discovery*. 2013; 12:371-387.
17. Davies J, Davies D. Origins and evolution of antibiotic resistance. *Microbiology and Molecular Biology Reviews.* 2010; 74(3):417-433.
18. Tenover FC. Mechanisms of antimicrobial resistance in bacteria. *American Journal of Medicine.* 2006; 119(6): S3-S10.
19. Reygaert WC. An overview of the antimicrobial resistance mechanisms of bacteria. *Aims Microbiology.* 2018; 4(3):482-501.
20. Munita JM, Arias CA. Mechanisms of Antibiotic Resistance. *Microbiology Spectrum*. 2016; 4(2):VMBF-0016-2015.
21. Cox G, Wright GD. Intrinsic antibiotic resistance: Mechanisms, origins, challenges and solutions. *International Journal of Medical Microbiology*. 2013; 303(6-7):287-292.
22. Martínez JL, Baquero F, Andersson DI. Predicting antibiotic resistance. *Nature Reviews Microbiology*. 2007; 5(12):958-65.
23. Perry JA, Westman EL, Wright GD. The antibiotic resistome: what's new? *Current Opinion in Microbiology*. 2014; 21:45-50.
24. Martínez JL, Coque TM, Baquero F. What is a resistance gene? Ranking risk in resistomes. *Nature Reviews Microbiology*. 2015; 13(2):116-23.
25. Gordon RJ, Lowy FD. Pathogenesis of methicillin-resistant *Staphylococcus aureus* infection. *Clinical Infectious Diseases*. 2008; 46:S350-S359.
26. Mishra NN, Bayer AS, Weidenmaier C, Grau T, Wanner S, Stefani S, *et al*. Phenotypic and Genotypic Characterization of Daptomycin-Resistant Methicillin-Resistant *Staphylococcus aureus* Strains: Relative Roles of *mprF* and *dlt* Operons. *PLOS ONE*. 2014; 9(9):e107426.
27. Bengtsson-Palme J, Kristiansson E, Larsson DGJ. Environmental factors influencing the development and spread of antibiotic resistance. *FEMS Microbiology Reviews*. 2018; 42(1):68-80.

28. Holmes AH, Moore LSP, Sundsfjord A, Steinbakk M, Regmi S, Karkey A, *et al*. Understanding the mechanisms and drivers of antimicrobial resistance. *The Lancet*. 2016; 387(10014):176-187.
29. Wood TK, Knabel SJ, Kwan BW. Bacterial Persister Cell Formation and Dormancy. *Applied and Environmental Microbiology*. 2013; 79(23):7116-7121.
30. Keren I, Kaldalu N, Spoering A, Wang Y, Lewis K. Persister cells and tolerance to antimicrobials. *FEMS Microbiology Letters*. 2004; 230(1):13-18.
31. Fisher RA, Gollan B, Helaine S. Persistent bacterial infections and persister cells. *Nature Reviews Microbiology*. 2017; 15(8):453-464.
32. Donlan RM. Biofilms: microbial life on surfaces. *Emerging infectious diseases*. 2002; *8*(9):881–890.
33. Vestby LK, Grønseth T, Simm R, Nesse LL. Bacterial Biofilm and its Role in the Pathogenesis of Disease. *Antibiotics-Basel*. 2020; 9(2):59.
34. Stewart PS, Costerton JW. Antibiotic resistance of bacteria in biofilms. *The Lancet.* 2001; 358(9276):135-138.
35. Anderl JN, Franklin MJ, Stewart PS. Role of antibiotic penetration limitation in *Klebsiella pneumoniae* biofilm resistance to ampicillin and ciprofloxacin. *Antimicrobial Agents and Chemotherapy*. 2000; *44*(7):1818–1824.
36. Melnyk AH, Wong A, Kassen R. The fitness costs of antibiotic resistance mutations. *Evolutionary Applications.* 2015; 8(3):273-283.
37. zur Wiesch PS, Engelstaedter J, Bonhoeffer S. Compensation of Fitness Costs and Reversibility of Antibiotic Resistance Mutations. *Antimicrobial Agents and Chemotherapy*. 2010; 54(5):2085-2095.
38. Liu G, Bogaj K, Bortolaia V, Olsen JE, Thomsen LE. Antibiotic-Induced, Increased Conjugative Transfer Is Common to Diverse Naturally Occurring ESBL Plasmids in *Escherichia coli*. *Frontiers in Microbiology.* 2019; 10:2119.
39. Lerminiaux NA, Cameron ADS. Horizontal transfer of antibiotic resistance genes in clinical environments. *Canadian Journal of Microbiology.* 2019; 65(1):34-44.
40. Xiong W, Sun Y, Ding X, Wang M, Zeng Z. Selective pressure of antibiotics on ARGs and bacterial communities in manure-polluted freshwater-sediment microcosms. *Frontiers in Microbiology.* 2015; 6:194.

41. World Health Organization. Antimicrobial Resistance (Fact sheet). World Health Organization, Geneva, Switzerland. 2020. Available at https://www.who.int/news-room/fact-sheets/detail/antimicrobial-resistance. Accessed October 2021.
42. Ventola CL. The antibiotic resistance crisis: part 1: causes and threats. *Pharmacy and Therapeutics*. 2015; 40(4):277-283.
43. Flemming HC, Wuertz S. Bacteria and archaea on Earth and their abundance in biofilms. *Nature Reviews Microbiology*. 2019; 17(4):247-260.
44. Berglund B. Environmental dissemination of antibiotic resistance genes and correlation to anthropogenic contamination with antibiotics. *Infection Ecology & Epidemiology.* 2015; 5:28564.
45. Alvarez-Uria G, Gandra S, Laxminarayan R. Poverty and prevalence of antimicrobial resistance in invasive isolates. *International Journal of Infectious Diseases.* 2016; 52:59-61.
46. Micoli F, Bagnoli F, Rappuoli R, Serruto D. The role of vaccines in combatting antimicrobial resistance. *Nature Reviews Microbiology.* 2021; 19(5):287-302.
47. Lipsitch M, Siber GR. How Can Vaccines Contribute to Solving the Antimicrobial Resistance Problem? *Mbio*. 2016; 7(3):e00428-16.
48. Sihvonen R, Siira L, Toropainen M, Kuusela P, Patari-Sampo A. *Streptococcus pneumoniae* antimicrobial resistance decreased in the Helsinki Metropolitan Area after routine 10-valent pneumococcal conjugate vaccination of infants in Finland. *European Journal of Clinical Microbiology & Infectious Diseases.* 2017; 36(11):2109-2116.
49. John T, Cherian T, Raghupathy P. *Haemophilus influenzae* disease in children in India: a hospital perspective. *The Paediatric Infectious Disease Journal*. 1998; 17(9):S169-S171.
50. Klugman KP, Black S. Impact of existing vaccines in reducing antibiotic resistance: Primary and secondary effects. *Proceedings of the National Academy of Sciences of the United States of America*. 2018; 115(51):12896-12901.
51. Frost I, Van Boeckel TP, Pires J, Craig J, Laxminarayan R. Global geographic trends in antimicrobial resistance: the role of international travel. *Journal of Travel Medicine*. 2019; 26(8):taz036.

52. van der Bij AK, Pitout JDD. The role of international travel in the worldwide spread of multiresistant Enterobacteriaceae. *Journal of Antimicrobial Chemotherapy*. 2012 SEP;67(9):2090-2100.
53. Geissler AL, Carrillo FB, Swanson K, Patrick ME, Fullerton KE, Bennett C, *et al*. Increasing *Campylobacter* Infections, Outbreaks, and Antimicrobial Resistance in the United States, 2004-2012. *Clinical Infectious Diseases.* 2017; 65(10):1624-1631.
54. Post A, Martiny D, van Waterschoot N, Hallin M, Maniewski U, Bottieau E, *et al*. Antibiotic susceptibility profiles among *Campylobacter* isolates obtained from international travelers between 2007 and 2014. *European Journal of Clinical Microbiology & Infectious Diseases.* 2017; 36(11):2101-2107.
55. Dave J, Warburton F, Freedman J, de Pinna E, Grant K, Sefton A, *et al*. What were the risk factors and trends in antimicrobial resistance for enteric fever in London 2005-2012? *Journal of Medical Microbiology.* 2017; 66(6):698-705.
56. Kleine C, Schlabe S, Hischebeth GTR, Molitor E, Pfeifer Y, Wasmuth J, *et al*. Successful Therapy of a Multidrug-Resistant Extended-Spectrum β-Lactamase-Producing and Fluoroquinolone-Resistant *Salmonella enterica* Subspecies *enterica* Serovar Typhi Infection Using Combination Therapy of Meropenem and Fosfomycin. *Clinical Infectious Diseases.* 2017; 65(10):1754-1756.
57. Zhou YP, Wilder-Smith A, Hsu L. The Role of International Travel in the Spread of Methicillin-Resistant *Staphylococcus aureus*. *Journal of Travel Medicine.* 2014; 21(4):272-281.
58. Willems R, Top J, van Santen M, Robinson D, Coque T, Baquero F, *et al*. Global spread of vancomycin-resistant *Enterococcus faecium* from distinct nosocomial genetic complex. *Emerging Infectious Diseases.* 2005; 11(6):821-828.
59. Jain A, Hopkins KL, Turton J, Doumith M, Hill R, Loy R, *et al*. NDM carbapenemases in the United Kingdom: an analysis of the first 250 cases. *The Journal of Antimicrobial Chemotherapy*. 2014; 69(7):1777-1784.
60. Nordmann P, Poirel L, Toleman MA, Walsh TR. Does broad-spectrum beta-lactam resistance due to NDM-1 herald the end of the antibiotic era for treatment of infections caused by Gram-negative bacteria? *The Journal of Antimicrobial Chemotherapy.* 2011; 66(4):689-692.

61. Wang R, van Dorp L, Shaw LP, Bradley P, Wang Q, Wang X, *et al*. The global distribution and spread of the mobilized colistin resistance gene *mcr*-1. *Nature Communications.* 2018; 9:1179.
62. Pendleton JN, Gorman SP, Gilmore BF. Clinical relevance of the ESKAPE pathogens. *Expert Review of Anti-Infective Therapy.* 2013; 11(3):297-308.
63. Nelson RE, Hatfield KM, Wolford H, Samore MH, Scott RD, II, Reddy SC, *et al*. National Estimates of Healthcare Costs Associated with Multidrug-Resistant Bacterial Infections Among Hospitalized Patients in the United States. *Clinical Infectious Diseases*. 2021; 72:S17-S26.
64. Dadgostar P. Antimicrobial Resistance: Implications and Costs. *Infection and drug resistance*. 2019; *12*:3903–3910.
65. O'Neil, J. Tackling drug-resistant infections globally. Final report and recommendations. *The Review on Antimicrobial Resistance*. 2016; 1-77.
66. Bartsch SM, McKinnell JA, Mueller LE, Miller LG, Gohil SK, Huang SS, *et al*. Potential economic burden of carbapenem-resistant *Enterobacteriaceae* (CRE) in the United States. *Clinical Microbiology and Infection.* 2017; 23(1):48.e9-48.e16.
67. Christiansen K, Tibbett P, Beresford W, Pearman J, Lee R, Coombs G, *et al.* Eradication of a Large Outbreak of a Single Strain of *vanB* Vancomycin-Resistant *Enterococcus faecium* at a Major Australian Teaching Hospital. *Infection Control & Hospital Epidemiology.* 2004; *25*(5):384-390.
68. Butler M, Blaskovich M, Cooper M. Antibiotics in the clinical pipeline in 2013. The *Journal of Antibiotics.* 2013; 66:571–591.
69. World Health Organization. WHO priority pathogens list for R&D of new antibiotics. World Health Organization, Geneva, Switzerland. 2017. Available at https://www.who.int/news/item/27-02-2017-who-publishes-list-of-bacteria-for-which-new-antibiotics-are-urgently-needed. Accessed October 2021.
70. McEwen SA, Collignon PJ. Antimicrobial resistance: A One Health perspective. *Microbiology Spectrum.* 2018; 6(2): 1-19.
71. Davies S, Gibbens N. UK five Year Antimicrobial Resistance Strategy 2013–2018. Department of Health. London. 2013; 1-43.
72. Gove M, Hancock M. Tackling Antimicrobial Resistance 2019–2024, The UK's Five-Year National Action Plan. Department of Health and Social Care. London. 2019; 1-98.

73. Renwick M, Mossialos E. What are the economic barriers of antibiotic R&D and how can we overcome them? *Expert Opinion on Drug Discovery*. 2018; 13(10):889-892.
74. Department of Health and Social Care. World-first scheme underway to tackle AMR and protect UK patients (Press Release). Department of Health and Social care. London. 2020. Available at World-first scheme underway to tackle AMR and protect UK patients - GOV.UK (www.gov.uk). Accessed October 2021.
75. Czaplewski L, Bax R, Clokie M, Dawson M, Fairhead H, Fischetti VA, *et al*. Alternatives to antibiotics—a pipeline portfolio review. *The Lancet Infectious Diseases*. 2016; 16(2):239-251.
76. Kumar M, Sarma DJ, Shubham S, Kumawat M, Verma V, Nina PB, *et al*. Futuristic Non-antibiotic Therapies to Combat Antibiotic Resistance: A Review. *Frontiers in Microbiology*. 2021; 12:609459.
77. Motley MP, Banerjee K, Fries BC. Monoclonal antibody-based therapies for bacterial infections. *Current opinion in infectious diseases*. 2019; *32*(3):210–216.
78. World Health Organisation. 2020 antibacterial agents in clinical and preclinical development: an overview and analysis. World Health Organisation, Geneva, Switzerland. 2021; 1-76.
79. Romero-Calle D, Benevides RG, Goes-Neto A, Billington C. Bacteriophages as Alternatives to Antibiotics in Clinical Care. *Antibiotics-Basel.* 2019; 8(3):138.
80. Wittebole X, De Roock S, Opal SM. A historical overview of bacteriophage therapy as an alternative to antibiotics for the treatment of bacterial pathogens. *Virulence*. 2014; 5(1):226-235.
81. Ferriol-Gonzalez C, Domingo-Calap P. Phages for Biofilm Removal. *Antibiotics-Basel.* 2020; 9(5):268.
82. Łusiak-Szelachowska M, Weber-Dąbrowska B, Górski A. Bacteriophages and Lysins in Biofilm Control. *Virologica Sinica.* 2020; *35*(2):125–133.
83. Schooley RT, Biswas B, Gill JJ, Hernandez-Morales A, Lancaster J, Lessor L, *et al*. Development and Use of Personalized Bacteriophage-Based Therapeutic Cocktails to Treat a Patient with a Disseminated Resistant *Acinetobacter baumannii* Infection. *Antimicrobial Agents and Chemotherapy.* 2017; 61(10):e00954-17.

84. Fischetti VA. Bacteriophage lysins as effective antibacterials. *Current Opinion in Microbiology*. 2008; *11*(5):393–400.
85. Love MJ, Bhandari D, Dobson RCJ, Billington C. Potential for Bacteriophage Endolysins to Supplement or Replace Antibiotics in Food Production and Clinical Care. *Antibiotics-Basel*. 2018;7(1):17.
86. Theuretzbacher U, Outterson K, Engel A, Karlen A. The global preclinical antibacterial pipeline. *Nature Reviews Microbiology.* 2020; 18(5):275-285.
87. Leon-Buitimea A, Garza-Cardenas CR, Garza-Cervantes JA, Lerma-Escalera JA, Morones-Ramirez JR. The Demand for New Antibiotics: Antimicrobial Peptides, Nanoparticles, and Combinatorial Therapies as Future Strategies in Antibacterial Agent Design. *Frontiers in Microbiology.* 2020; 11:1669.
88. Mahlapuu M, Hakansson J, Ringstad L, Bjorn C. Antimicrobial Peptides: An Emerging Category of Therapeutic Agents. *Frontiers in Cellular and Infection Microbiology.* 2016; 6:194.
89. Molchanova N, Hansen PR, Franzyk H. Advances in Development of Antimicrobial Peptidomimetics as Potential Drugs. *Molecules*. 2017; 22(9):1430.
90. Dijksteel GS, Ulrich MMW, Middelkoop E, Boekema BKHL. Review: Lessons Learned from Clinical Trials Using Antimicrobial Peptides (AMPs). *Frontiers in Microbiology.* 2021; 12:616979.
91. Sarma P, Mahendiratta S, Prakash A, Medhi B. Specifically targeted antimicrobial peptides: A new and promising avenue in selective antimicrobial therapy. *Indian Journal of Pharmacology*. 2018; 50(1):1-3.
92. Wang G, Li X, Wang Z. APD3: the antimicrobial peptide database as a tool for research and education. *Nucleic Acids Research.* 2016; 44:D1087-D1093.
93. Miró-Canturri A, Ayerbe-Algaba R, Smani Y. Drug Repurposing for the Treatment of Bacterial and Fungal Infections. *Frontiers in Microbiology*. 2019; 10:41.
94. Cheng YS, Williamson PR, Zheng W. Improving therapy of severe infections through drug repurposing of synergistic combinations. *Current Opinion in Pharmacology*. 2019; 48:92-98.
95. Jarada TN, Rokne JG, Alhajj R. A review of computational drug repositioning: strategies, approaches, opportunities, challenges, and directions. *Journal of Cheminformatics.* 2020; 12:46.

96. Konreddy AK, Rani GU, Lee K, Choi Y. Recent Drug-Repurposing-Driven Advances in the Discovery of Novel Antibiotics. *Current Medicinal Chemistry*. 2019; 26(28):5363-5388.
97. Bush K, Bradford PA. Interplay between β-lactamases and new β-lactamase inhibitors. *Nature Reviews Microbiology*. 2019; 17(5):295-306.
98. Shi C, Chen J, Kang X, Shen X, Lao X, Zheng H. Approaches for the discovery of metallo-β-lactamase inhibitors: A review. *Chemical Biology & Drug Design*. 2019; 94(2):1427-1440.
99. Fleitas Martínez O, Cardoso MH, Ribeiro SM, Franco OL. Recent Advances in Anti-virulence Therapeutic Strategies with a Focus on Dismantling Bacterial Membrane Microdomains, Toxin Neutralization, Quorum-Sensing Interference and Biofilm Inhibition. *Frontiers in Cellular and Infection Microbiology*. 2019; 9:74.
100. Defraine V, Fauvart M, Michiels J. Fighting bacterial persistence: Current and emerging anti-persister strategies and therapeutics. *Drug resistance updates: reviews and commentaries in antimicrobial and anticancer chemotherapy*. 2018; *38*:12–26.
101. Hancock REW, Nijnik A, Philpott DJ. Modulating immunity as a therapy for bacterial infections. *Nature Reviews Microbiology.* 2012; 10:243–54.
102. Wong AC, Levy M. New Approaches to Microbiome-Based Therapies. *mSystems.* 2019; 4(3):e00122-19.
103. Kumar M, Curtis A, Hoskins C. Application of Nanoparticle Technologies in the Combat against Anti-Microbial Resistance. *Pharmaceutics*. 2018; 10(1):11.
104. Papp-Wallace KM, Endimiani A, Taracila MA, Bonomo RA. Carbapenems: past, present, and future. *Antimicrobial agents and chemotherapy*. 2011; *55*(11):4943–4960.
105. Jelić D, Antolović R. From Erythromycin to Azithromycin and New Potential Ribosome-Binding Antimicrobials. *Antibiotics-Basel*. 2016; 5(3):29.
106. Olson MW, Ruzin A, Feyfant E, Rush TS 3rd, O'Connell J, Bradford PA. Functional, biophysical, and structural bases for antibacterial activity of tigecycline. *Antimicrobial Agents and Chemotherapy*. 2006; 50(6):2156-2166.
107. Fischbach MA. Combination therapies for combating antimicrobial resistance. *Current Opinion in Microbiology*. 2011; 14(5):519-523.

108. Cassir N, Rolain J, Brouqui P. A new strategy to fight antimicrobial resistance: the revival of old antibiotics. *Frontiers in Microbiology*. 2014; 5:551.
109. Annunziato G. Strategies to Overcome Antimicrobial Resistance (AMR) Making Use of Non-Essential Target Inhibitors: A Review. *International Journal of Molecular Sciences*. 2019; 20(23):5844.
110. Mahmood HY, Jamshidi S, Sutton JM, Rahman KM. Current Advances in Developing Inhibitors of Bacterial Multidrug Efflux Pumps. *Current Medicinal Chemistry*. 2016; 23(10):1062–1081.
111. Payne DJ, Gwynn MN, Holmes DJ, Pompliano DL. Drugs for bad bugs: confronting the challenges of antibacterial discovery. *Nature Reviews. Drug Discovery*. 2007; 6(1):29-40.
112. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced Drug Delivery Reviews*. 2001; 46(1-3):3–26.
113. O'Shea R, Moser HE. Physicochemical properties of antibacterial compounds: implications for drug discovery. *Journal of Medicinal Chemistry*. 2008; 51(10):2871–2878.
114. Gajdács M. The Concept of an Ideal Antibiotic: Implications for Drug Design. *Molecules*. 2019; 24(5):892.
115. Bérdy J. Bioactive microbial metabolites - A personal view. *Journal of Antibiotics.* 2005; 58(1):1-26.
116. Katz L, Baltz RH. Natural product discovery: past, present, and future. *Journal of Industrial Microbiology & Biotechnology.* 2016; 43(2-3):155-176.
117. Demain AL, Fang A. The natural functions of secondary metabolites. In: Fiechter A. (Ed.) History of Modern Biotechnology: Advances in Biochemical Engineering/ Biotechnology. Springer, Berlin, Heidelberg. 2000; 69:1-39.
118. Challis G, Hopwood D. Synergy and contingency as driving forces for the evolution of multiple secondary metabolite production by *Streptomyces* species. *Proceedings of the National Academy of Sciences of the United States of America.* 2003; 100:14555-14561.
119. Bérdy J. Thoughts and facts about antibiotics: Where we are now and where we are heading. The *Journal of Antibiotic*s. 2012; 65:385–395.

120. Sengupta S, Chattopadhyay MK, Grossart HP. The multifaceted roles of antibiotics and antibiotic resistance in nature. *Frontiers in Microbiol*ogy. 2013; 4:47.
121. Awolope OK. Genome-guided Screening of Bacterial Isolates to Identify Potential Antibiotic Producers. Unpublished work. 2018; 1-61.
122. Davies J. Are antibiotics naturally antibiotics? *Journal of Industrial Microbiology and Biotechnology*. 2006; 33(7):496–499.
123. Sun H, Liu Z, Zhao H, Ang EL. Recent advances in combinatorial biosynthesis for drug discovery. *Drug Design, Development and Therapy*. 2015; 9:823–833.
124. Pham VHT, Kim J. Cultivation of unculturable soil bacteria. *Trends in Biotechnology.* 2012; 30(9):475-484.
125. Watve MG, Shejval V, Sonawane CJ, Rahalkar MC, Matapurkar AK, Shouche YS, *et al*. The 'K' selected oligophilic bacteria: a key to uncultured diversity? *Current Science*. 2000; 78(12): 1535-1542.
126. Ling LL, Schneider T, Peoples AJ, Spoering AL, Engels I, Conlon BP, *et al*. A new antibiotic kills pathogens without detectable resistance. *Nature*. 2015; 517(7535):455-472.
127. Walsh CT, Fischbach MA. Natural products version 2.0: Connecting genes to molecules. *Journal of the American Chemical Society.* 2010; 132(8):2469-2493.
128. Cimermancic P, Medema MH, Claesen J, Kurita K, Wieland Brown LC, Mavrommatis K, *et al*. Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters. *Cell.* 2014; 158(2):412-421.
129. Medema MH, Fischbach MA. Computational approaches to natural product discovery. *Nature Chemical Biology.* 2015; 11(9):639-648.
130. Rudd B, Hopwood D. Genetics of actinorhodin biosynthesis by *Streptomyces-coelicolor* A3(2). *Journal of General Microbiology.* 1979; 114:35-43.
131. Hopwood D, Wright H. Cda is a new chromosomally-determined antibiotic from *Streptomyces-coelicolor* A3(2). *Journal of General Microbiology.* 1983; 129:3575-3579.
132. Ikeda H, Kotaki H, Omura S. Genetic-studies of avermectin biosynthesis in *Streptomyces-avermitilis*. *Journal of Bacteriology.* 1987; 169(12):5615-5621.

133. Bentley SD, Chater KF, Cerdeno-Tarraga A, Challis GL, Thomson NR, James KD, *et al*. Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2). *Nature.* 2002; 417(6885):141-147.
134. Weber T, Welzel K, Pelzer S, Vente A, Wohlleben W. Exploiting the genetic potential of polyketide producing S*treptomycetes*. *Journal of Biotechnology.* 2003; 106(2-3):221-232.
135. Ikeda H, Ishikawa J, Hanamoto A, Shinose M, Kikuchi H, Shiba T, *et al*. Complete genome sequence and comparative analysis of the industrial microorganism *Streptomyces avermitilis*. *Nature Biotechnology.* 2003; 21(5):526-531.
136. Ziemert N, Alanjary M, Weber T. The evolution of genome mining in microbes - a review. *Natural Product Reports.* 2016; 33(8):988-1005.
137. Bachmann BO, Van Lanen SG, Baltz RH. Microbial genome mining for accelerated natural products discovery: is a renaissance in the making? *Journal of Industrial Microbiology & Biotechnology.* 2014; 41(2):175-184.
138. Baltz RH. Gifted microbes for genome mining and natural product discovery. *Journal of Industrial Microbiology and Biotechnology*. 2017; 44(4-5):573–588.
139. Weber T. *In silico* tools for the analysis of antibiotic biosynthetic pathways. *International Journal of Medical Microbiology.* 2014; 304(3-4):230-235.
140. Blin K, Shaw S, Steinke K, Villebro R, Ziemert N, Lee SY, *et al.* antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline. *Nucleic Acids Research*. 2019; 47(W1):W81–W87.
141. van Heel AJ, de Jong A, Montalbán-López M, Kok J, Kuipers OP. BAGEL3: Automated identification of genes encoding bacteriocins and (non-) bactericidal posttranslationally modified peptides. *Nucleic Acids Research*. 2013; 41:W448-W453.
142. Reddy BV, Milshteyn A, Charlop-Powers Z, Brady SF. eSNaPD: a versatile, web-based bioinformatics platform for surveying and mining natural product biosynthetic diversity from metagenomes. *Chemistry & Biology*. 2014; *21*(8):1023–1033.
143. Starcevic A, Zucko J, Simunkovic J, Long PF, Cullum J, Hranueli D. ClustScan: an integrated program package for the semi-automatic annotation of modular

biosynthetic gene clusters and *in silico* prediction of novel chemical structures. *Nucleic Acids Research*. 2008; 36(21):6882-6892.

144. Ziemert N, Podell S, Penn K, Badger JH, Allen E, Jensen PR. The natural product domain seeker NaPDoS: a phylogeny based bioinformatic tool to classify secondary metabolite gene diversity. *PLoS One*. 2012; 7(3):e34064.
145. Tietz JI, Schwalen CJ, Patel PS, Maxson T, Blair PM, Tai HC, Zakai UI, Mitchell DA. A new genome-mining tool redefines the lasso peptide biosynthetic landscape. *Nature Chemical Biology*. 2017; 13(5):470–478.
146. Li MH, Ung PM, Zajkowski J, Garneau-Tsodikova S, Sherman DH. Automated genome mining for natural products. *BMC Bioinformatics*. 2009; 10:185.
147. Skinnider MA, Johnston CW, Gunabalasingam M, Merwin NJ, Kieliszek AM, Maclellan RJ, *et al.* Comprehensive prediction of secondary metabolite structure and biological activity from microbial genome sequences. *Nature Communications.* 2020; 11:6058.
148. Boddy CN. Bioinformatics tools for genome mining of polyketide and non-ribosomal peptides. *Journal of Industrial Microbiology & Biotechnology.* 2014; 41(2):443-450.
149. Ren H, Shi C, Zhao H. Computational Tools for Discovering and Engineering Natural Product Biosynthetic Pathways. *iScience*. 2020; 23(1):100795.
150. Sélem-Mojica N, Aguilar C, Gutiérrez-García K, Martínez-Guerrero CE, Barona-Gómez F. EvoMining reveals the origin and fate of natural product biosynthetic enzymes. *Microbial Genom*ics. 2019; 5(12):e000260.
151. Rutledge PJ, Challis GL. Discovery of microbial natural products by activation of silent biosynthetic gene clusters. *Nature Reviews Microbiology.* 2015; 13(8):509-523.
152. Begani J, Lakhani J, Harwani D. Current strategies to induce secondary metabolites from microbial biosynthetic cryptic gene clusters. *Annals of Microbiology.* 2018; 68(7):419-432.
153. Bode H, Bethe B, Hofs R, Zeeck A. Big effects from small changes: possible ways to explore nature's chemical diversity. *Chembiochem.* 2002; 3(7):619-627.
154. Scheffler RJ, Colmer S, Tynan H, Demain AL, Gullo VP. Antimicrobials, drug discovery, and genome mining. *Applied Microbiology and Biotechnology.* 2013; 97(3):969-978.

155. Lautru S, Deeth R, Bailey L, Challis G. Discovery of a new peptide natural product by *Streptomyces coelicolor* genome mining. *Nature Chemical Biology.* 2005; 1(5):265-269.
156. Scherlach K, Hertweck C. Discovery of aspoquinolones A-D, prenylated quinoline-2-one alkaloids from *Aspergillus nidulans*, motivated by genome mining. *Organic & Biomolecular Chemistry.* 2006; 4(18):3517-3520.
157. Liu Z, Zhao Y, Huang C, Luo Y. Recent Advances in Silent Gene Cluster Activation in *Streptomyces*. *Frontiers in Bioengineering and Biotechnology*. 2021; *9*:632230.
158. Martin JF, Santos-Beneit F, Rodriguez-Garcia A, Sola-Landa A, Smith MCM, Ellingsen TE, *et al*. Transcriptomic studies of phosphate control of primary and secondary metabolism in *Streptomyces coelicolor*. *Applied Microbiology and Biotechnology.* 2012; 95(1):61-75.
159. Haferburg G, Groth I, Möllmann U, Kothe E, Sattler I. Arousing sleeping genes: shifts in secondary metabolism of metal tolerant actinobacteria under conditions of heavy metal stress. *Biometals.* 2009; 22(2):225-234.
160. Reuther J, Wohlleben W. Nitrogen metabolism in *Streptomyces coelicolor*: transcriptional and post-translational regulation. *Journal of Molecular Microbiology and Biotechnology.* 2007; 12(1-2):139-146.
161. Ochi K, Hosaka T. New strategies for drug discovery: activation of silent or weakly expressed microbial gene clusters. *Applied Microbiology and Biotechnology.* 2013; 97(1):87-98.
162. Charusanti P, Fong NL, Nagarajan H, Pereira AR, Li HJ, Abate EA, *et al*. Exploiting adaptive laboratory evolution of *Streptomyces clavuligerus* for antibiotic discovery and overproduction. 2012; *Plos One,* 7(3): 1-14.
163. Ochi K, Tanaka Y, Tojo S. Activating the expression of bacterial cryptic genes by *rpoB* mutations in RNA polymerase or by rare earth elements. *Journal of Industrial Microbiology & Biotechnology.* 2014; 41(2):403-414.
164. Ochi K. Insights into microbial cryptic gene activation and strain improvement: principle, application and technical aspects. *Journal of Antibiotics.* 2017; 70(1):25-40.
165. Laureti L, Song L, Huang S, Corre C, Leblond P, Challis GL, *et al*. Identification of a bioactive 51-membered macrolide complex by activation of a silent polyketide synthase in *Streptomyces ambofaciens*. *Proceedings of the*

National Academy of Sciences of the United States of America. 2011; 108: 6258-6263.

166. Katz M, Hover BM, Brady SF. Culture-independent discovery of natural products from soil metagenomes. *Journal of Industrial Microbiology & Biotechnology.* 2016; 43(2-3):129-141.
167. Hover BM, Kim S, Katz M, Charlop-Powers Z, Owen JG, Ternei MA, *et al*. Culture-independent discovery of the malacidins as calcium-dependent antibiotics with activity against multidrug-resistant Gram-positive pathogens. *Nature Microbiology.* 2018; 3(4):415-422.
168. Blin K, Medema MH, Kottmann R, Lee SY, Weber T. The antiSMASH database, a comprehensive database of microbial secondary metabolite biosynthetic gene clusters. *Nucleic Acids Research*. 2017; 45(D1): D555–D559.
169. Kautsar SA, Blin K, Shaw S, Navarro-Muñoz JC, Terlouw BR, van der Hooft JJJ, *et al*. MIBiG 2.0: a repository for biosynthetic gene clusters of known function. *Nucleic Acids Research*. 2020; 48(D1):D454–D458.
170. Adamek M, Alanjary M, Sales-Ortells H, Goodfellow M, Bull AT, Winkler A, *et al*. Comparative genomics reveals phylogenetic distribution patterns of secondary metabolites in *Amycolatopsis* species. 2018; *BMC Genomics,* 19:426-441
171. Schloss PD, Handelsman J. Toward a census of bacteria in soil. *Plos Computational Biology.* 2006; 2(7):786-793.
172. Raaijmakers JM, Paulitz TC, Steinberg C, Alabouvette C, Moënne-Loccoz Y. The rhizosphere: a playground and battlefield for soilborne pathogens and beneficial microorganisms. *Plant Soil*. 2009; 321:341–361.
173. Lugtenberg B, Kamilova F. Plant-growth-promoting rhizobacteria. *Annual Review of Microbiology*. 2009; 63:541-556.
174. Finkel OM, Castrillo G, Paredes SH, Gonzalez IS, Dangl JL. Understanding and exploiting plant beneficial microbes. *Current Opinion in Plant Biology.* 2017; 38:155-163.
175. Martin J, Demain A. Control of antibiotic biosynthesis. *Microbiological Reviews.* 1980; 44(2):230-251.
176. Janda JM, Abbott SL. 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls. *Journal of Clinical Microbiology*. 2007; 45(9):2761-2764.

177. Mignard S, Flandrois JP. 16S rRNA sequencing in routine bacterial identification: a 30-month experiment. *Journal of Microbiological Methods.* 2006; 67(3):574-581.
178. Clarridge J. Impact of 16S rRNA gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases. *Clinical Microbiology Reviews.* 2004; 17(4):840-867.
179. Baker G, Smith J, Cowan D. Review and re-analysis of domain-specific 16S primers. *Journal of Microbiological Methods.* 2003; 55(3):541-555.
180. Treves DS. Review of three DNA analysis applications for use in the microbiology or genetics classroom. *Journal of Microbiology & Biology Education.* 2010; *11*(2):186–187.
181. Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Ostell J, Pruitt KD, *et al*. GenBank. *Nucleic Acids Research.* 2018; 46(D1): D41-D47.
182. Haft DH, DiCuccio M, Badretdin A, Brover V, Chetvernin V, O'Neill K, *et al*. RefSeq: an update on prokaryotic genome annotation and curation. *Nucleic Acids Research.* 2018; 46(D1): D851-D860.
183. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, *et al*. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Research.* 2013; 41(D1): D590-D596.
184. Chun J, Lee J, Jung Y, Kim M, Kim S, Kim BK, *et al*. EzTaxon: a web-based tool for the identification of prokaryotes based on 16S ribosomal RNA gene sequences. *International Journal of Systematic and Evolutionary Microbiology.* 2007; 57:2259-2261.
185. DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, *et al*. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Applied and Environmental Microbiology.* 2006; 72(7):5069-5072.
186. Johnson JS, Spakowicz DJ, Hong BY, Petersen LM, Demkowicz P, Chen L, *et al.* Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. *Nature Communications.* 2019; 10:5029.
187. Kim M, Morrison M, Yu Z. Evaluation of different partial 16S rRNA gene sequence regions for phylogenetic analysis of microbiomes. *Journal of Microbiological Methods.* 2011; 84(1):81-87.
188. Leray M, Knowlton N, Ho S, Nguyen BN, Machida RJ. GenBank is a reliable resource for 21st century biodiversity research. *Proceedings of the National*

Academy of Sciences of the United States of America. 2019; 116 (45) 22651-22656.

189. Priest F, Barker M. Gram-negative bacteria associated with brewery yeasts: reclassification of *Obesumbacterium proteus* biogroup 2 as *Shimwellia pseudoproteus* gen. nov., sp. nov., and transfer of *Escherichia blattae* to *Shimwellia blattae* comb. nov. *International Journal of Systematic and Evolutionary Microbiology.* 2010; 60(4): 828-833.
190. Silby MW, Winstanley C, Godfrey SAC, Levy SB, Jackson RW. *Pseudomonas* genomes: diverse and adaptable. *FEMS Microbiology Reviews*. 2011; 35(4): 652–680.
191. Arrebola E, Tienda S, Vida C, de Vicente A, Cazorla FM. Fitness Features Involved in the Biocontrol Interaction of *Pseudomonas chlororaphis* With Host Plants: The Case Study of PcPCL1606. *Frontiers in Microbiology*. 2019; 10:719.
192. Anzai Y, Kim H, Park JY, Wakabayashi H, Oyaizu H. Phylogenetic affiliation of the pseudomonads based on 16S rRNA sequence. *International Journal of Systematic and Evolutionary Microbiology*. 2000; 50(4): 1563.
193. Holley RW, Apgar J, Everett GA, Madison JT, Marquisee M, Merrill SH, *et al*. Structure of a Ribonucleic Acid. *Science*. 1965; 147(3664):1462–1465.
194. Sanger F, Brownlee GG, Barrell BG. A two-dimentional fractionation procedure for radioactive nucleotides. *Journal of Molecular Biology*. 1965; 13(1):373-398.
195. Heather JM, Chain B. The sequence of sequencers: The history of sequencing DNA. *Genomics*. 2016;107(1):1-8.
196. Giani AM, Gallo GR, Gianfranceschi L, Formenti G. Long walk to genomics: History and current approaches to genome sequencing and assembly. *Computational and Structural Biotechnology Journal*. 2019; 18:9–19.
197. Shendure J, Balasubramanian S, Church GM, Gilbert W, Rogers J, Schloss JA, *et al.* DNA sequencing at 40: past, present and future. *Nature.* 2019; 550**:**345–353.
198. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*. 1977; 74(12):5463–5467.

199. Staden R. A strategy of DNA sequencing employing computer programs. *Nucleic Acids Research*. 1979; *6*(7):2601–2610.
200. Anderson S. Shotgun DNA sequencing using cloned DNase I-generated fragments. *Nucleic Acids Research*. 1981; 9(13):3015-3027.
201. Messing J, Crea R, Seeburg PH. A system for shotgun DNA sequencing. *Nucleic Acids Research*. 1981; 9(2):309–321.
202. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*. 2001; 409**:**860–921.
203. Ronaghi M, Karamohamed S, Pettersson B, Uhlén M, Nyrén P. Real-time DNA sequencing using detection of pyrophosphate release. *Analytical Biochemistry*. 1996; 242(1):84–89.
204. Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A, *et al*. The complete genome of an individual by massively parallel DNA sequencing. *Nature*. 2008; 452(7189):872–876.
205. Voelkerding KV, Dames SA, Durtschi JD. Next-generation sequencing: from basic research to diagnostics. *Clinical chemistry*. 2009; *55*(4):641–658.
206. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, *et al*. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*. 2008; 456(7218):53-9.
207. Alkan C, Sajjadian S, Eichler E. Limitations of next-generation genome sequence assembly. *Nature Methods.* 2011; 8**:**61–65.
208. Goldstein S, Beka L, Graf J, Klassen JL. Evaluation of strategies for the assembly of diverse bacterial genomes using MinION long-read sequencing. *BMC Genomics.* 2019; 20:23.
209. Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, *et al*. Real-time DNA sequencing from single polymerase molecules. *Science.* 2009; *323*(5910):133–138.
210. Deamer D, Akeson M, Branton D. Three decades of nanopore sequencing. *Nature Biotechnol*ogy. 2016; 34:518–524.
211. Bayley H. Nanopore Sequencing: From Imagination to Reality. *Clinical Chemistry*. 2015; 61(1):25-31.
212. Amarasinghe SL, Su S, Dong X, Zappia L, Ritchie M, Gouil Q. Opportunities and challenges in long-read sequencing data analysis. *Genome Biology.* 2020; 21:30.
213. Illumina, Inc. *De Novo* Assembly Using Illumina Reads. Technical Note: Sequencing. Illumina, Inc., San Diego, USA. 2010; 1-8.

214. Lang D, Zhang S, Ren P, Liang F, Sun Z, Meng G, *et al*. Comparison of the two up-to-date sequencing technologies for genome assembly: HiFi reads of Pacific Biosciences Sequel II system and ultralong reads of Oxford Nanopore. *GigaScience*. 2020; 9(12):giaa123.
215. Cock PJ, Fields CJ, Goto N, Heuer ML, Rice PM. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Research*. 2010; 38(6):1767-1771.
216. Andrews S. FastQC: A quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/(2010).
217. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data*. Bioinformatics.* 2014; 30:2114-2120.
218. Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*. 2016; 32(19):3047–3048.
219. De Coster W, D'Hert S, Schultz DT, Cruts M, Van Broeckhoven C. NanoPack: visualizing and processing long-read sequencing data*. Bioinformatics.* 2018; 34:2666-2669.
220. Wick RR, Judd LM, Gorrie CL, Holt KE. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. *PLoS Computational Biology*. 2017; 13:1-22.
221. Wick RR, Judd LM, Cerdeira LT, Hawkey J, Méric G, Vezina B, *et al.* Trycycler: consensus long-read assemblies for bacterial genomes. *Genome Biology.* 2021; 22**:**266.
222. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, *et al*. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology.* 2012; 19(5):455–477.
223. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, *et al*. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One*. 2014; 9(11):e112963.
224. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics.* 2013; 29:1072–1075.
225. Wick RR, Schultz MB, Zobel J, Holt KE. Bandage: interactive visualization of *de novo* genome assemblies. *Bioinformatics*. 2015; 31(20):3350–3352.

226. Larsen MV, Cosentino S, Lukjancenko O, Saputra D, Rasmussen S, Hasman H, *et al*. Benchmarking of methods for genomic taxonomy. *Journal of Clinical Microbiology*. 2014; *52*(5):1529–1539.
227. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, *et al.* The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Research*. 2013; 41:D590–D596.
228. Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nature Communications.* 2018; 9:5114.
229. Ciufo S, Kannan S, Sharma S, Badretdin A, Clark K, Turner S, *et al*. Using average nucleotide identity to improve taxonomic assignments in prokaryotic genomes at the NCBI. *International Journal of Systematic and Evolutionary Microbiology*. 2018; *68*(7):2386–2392.
230. Clark SC, Egan R, Frazier PI, Wang Z. ALE: a generic assembly likelihood evaluation framework for assessing the accuracy of genome and metagenome assemblies. *Bioinformatics*. 2013; 29(4):435–443.
231. Hunt M, Kikuchi T, Sanders M, Newbold C, Berriman M, Otto TD. REAPR: a universal tool for genome assembly evaluation. *Genome Biology.* 2013; 14:R47.
232. Mende D, Sunagawa S, Zeller G, Bork P. Accurate and universal delineation of prokaryotic species. *Nature Methods.* 2013; 10:881–884.
233. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*. 2015; 31(19):3210–3212.
234. Jauhal AA, Newcomb RD. (2021). Assessing genome assembly quality prior to downstream analysis: N50 versus BUSCO. *Molecular Ecology Resources*. 2021; 21(5): 1416–1421.
235. Tatusova T, DiCuccio M, Badretdin A, Chetvernin V, Nawrocki EP, Zaslavsky L, *et al*. NCBI prokaryotic genome annotation pipeline. *Nucleic Acids Research*. 2016; 44:6614–6624.
236. Stothard P, Grant JR, Van Domselaar G. Visualizing and comparing circular genomes using the CGView family of tools. *Briefings in Bioinformatics*. 2019; 20:1576-1582.

237. Awolope OK, O'Driscoll NH, Di Salvo A, Lamb AJ. The complete genome sequence of *Hafnia alvei* A23BA; a potential antibiotic-producing rhizobacterium. *BMC Research Notes.* 2021; 14:8.
238. Carattoli A, Zankari E, García-Fernández A, Voldby Larsen M, Lund O, Villa L, *et al*. In silico detection and typing of plasmids using PlasmidFinder and plasmid multilocus sequence typing. *Antimicrobial agents and chemotherapy*. 2014; *58*(7):3895–3903.
239. Chen YC, Liu T, Yu CH, Chiang TY, Hwang CC. Effects of GC bias in next-generation-sequencing data on de novo genome assembly. *PLoS One*. 2013; 8(4):e62856.
240. Awolope OK, O'Driscoll NH, Di Salvo A, Lamb AJ. De novo genome assembly and analysis unveil biosynthetic and metabolic potentials of *Pseudomonas fragi* A13BB. *BMC Genomic Data*. 2021; 22(1):15.
241. Stanborough T, Fegan N, Powell SM, Singh T, Tamplin M, Chandry PS. Genomic and metabolic characterization of spoilage-associated *Pseudomonas* species. *International Journal of Food Microbiology*. 2018; 268:61–72.
242. Rodríguez LA, Vivas J, Gallardo CS, Acosta F, Barbeyto L, Real F. Identification of *Hafnia alvei* with the MicroScan WalkAway system. *Journal of Clinical Microbiology*. 1999; 37:4186-4188.
243. Legrand R, Lucas N, Dominique M, Azhar S, Deroissart C, Le Solliec MA, *et al*. Commensal *Hafnia alvei* strain reduces food intake and fat mass in obese mice-a new potential probiotic for appetite and body weight management. *International Journal of Obesity*. 2020; 44:1041-1051.
244. Günthard H, Pennekamp A. Clinical significance of extraintestinal *Hafnia alvei* isolates from 61 patients and review of the literature. *Clinical Infectious Diseases*. 1996; 22:1040-1045.
245. Tian B, Moran NA. Genome Sequence of *Hafnia alvei* bta3_1, a Bacterium with Antimicrobial Properties Isolated from Honey Bee Gut. *Genome Announcements*. 2016; 4:e00439-16.
246. Yin Z, Yuan C, Du Y, Yang P, Qian C, Wei Y, *et al*. Comparative genomic analysis of the *Hafnia* genus reveals an explicit evolutionary relationship between the species *alvei* and *paralvei* and provides insights into pathogenicity. *BMC Genomics. 2019; 20:768.*
247. Ercolini D, Casaburi A, Nasi A, Ferrocino I, Monaco RD, Ferranti P, *et al*. Different molecular types of *Pseudomonas fragi* have the same overall

behaviour as meat spoilers. *International Journal of Food Microbiology*. 2010; 142:120–31.

248. Selvakumar G, Joshi P, Nazim S, Mishra P, Bisht J, Gupta H. Phosphate solubilization and growth promotion by *Pseudomonas fragi* CS11RH1 (MTCC 8984), a psychrotolerant bacterium isolated from a high altitude Himalayan rhizosphere. *Biologia*. 2009;64(2):239–45.
249. Farh ME, Kim YJ, Sukweenadhi J, Singh P, Yang DC. Aluminium resistant, plant growth promoting bacteria induce overexpression of Aluminium stress related genes in *Arabidopsis thaliana* and increase the ginseng tolerance against Aluminium stress. *Microbiological Research*. 2017; 200:45-52.
250. Singha LP, Kotoky R, Pandey P. Draft Genome Sequence of *Pseudomonas fragi* Strain DBC, Which Has the Ability To Degrade High-Molecular-Weight Polyaromatic Hydrocarbons. *Genome Announcements*. 2017; 5(49):e01347-17.
251. Brettin T, Davis JJ, Disz T, Edwards RA, Gerdes S, Olsen GJ, *et al*. RASTtk: a modular and extensible implementation of the RAST algorithm for building custom annotation pipelines and annotating batches of genomes. *Scientific Reports*. 2015; *5*:8365.
252. Aziz RK., Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, *et al*. The RAST Server: rapid annotations using subsystems technology. *BMC Genomics*. 2008; *9*:75.
253. Lowe TM, Eddy SR. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Research*. 1997; 25(5):955-964.
254. Delcher AL, Bratke KA, Powers EC, Salzberg SL. Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics*. 2007; 23(6):673-679.
255. Hyatt D, Chen GL, Locascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*. 2010; 11:119.
256. Meyer F, Overbeek R, Rodriguez A. FIGfams: yet another set of protein families. *Nucleic Acids Research*. 2009; 37(20):6643-6654.
257. Overbeek R, Begley T, Butler RM, Choudhuri JV, Chuang HY, Cohoon M, *et al*. The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Research*. 2005; *33*(17):5691–5702.

258. Overbeek R, Olson R, Pusch GD, Olsen GJ, Davis JJ, Disz T, *et al*. (2014). The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). *Nucleic Acids Research*. 2014; 42:D206–D214.
259. Medema MH, Blin K, Cimermancic P, de Jager V, Zakrzewski P, Fischbach MA, *et al*. (2011). antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. *Nucleic Acids Research*. 2011; 39:W339–W346.
260. Baltz RH. Natural product drug discovery in the genomic era: realities, conjectures, misconceptions, and opportunities. *Journal of Industrial Microbiology and Biotechnology*. 2019; 46(3-4):281–299.
261. Weber T, Blin K, Duddela S, Krug D, Kim HU, Bruccoleri R, *et al*. antiSMASH 3.0-a comprehensive resource for the genome mining of biosynthetic gene clusters. *Nucleic Acids Research*. 2015, 43(W1):W237–W243.
262. Kim KR, Kim TJ, Suh JW. The gene cluster for spectinomycin biosynthesis and the aminoglycoside-resistance function of *spcM* in *Streptomyces spectabilis*. *Current Microbiology*. 2008; *57*(4):371–374.
263. Kharel MK, Subba B, Basnet DB, Woo JS, Lee HC, Liou K, *et al*. A gene cluster for biosynthesis of kanamycin from *Streptomyces kanamyceticus*: comparison with gentamicin biosynthetic gene cluster. *Archives of Biochemistry and Biophysics*. 2004; 429(2):204–214.
264. Mishra S, Lin Z, Pang S, Zhang W, Bhatt P, Chen S. Recent Advanced Technologies for the Characterization of Xenobiotic-Degrading Microorganisms and Microbial Communities. *Frontiers in Bioengineering and Biotechnology*. 2021; 9:632059.
265. Donohoue PD, Barrangou R, May AP. Advances in Industrial Biotechnology Using CRISPR-Cas Systems. *Trends in Biotechnology*. 2018; 36(2):134–146.
266. Zheng Y, Li J, Wang B, Han J, Hao Y, Wang S, *et al*. Endogenous Type I CRISPR-Cas: From Foreign DNA Defense to Prokaryotic Engineering. *Frontiers in Bioengineering and Biotechnology*. 2020; 8:62.
267. Champomier-Vergès MC, Stintzi A, Meyer JM. Acquisition of iron by the non-siderophore-producing *Pseudomonas fragi*. *Microbiology*. 1996; 142(5):1191-1199.
268. Vinogradov AA, Suga H. Introduction to Thiopeptides: Biological Activity, Biosynthesis, and Strategies for Functional Reprogramming. *Cell Chemical Biology*. 2020; 27(8):1032–1051.

269. Bennallack PR, Griffitts JS. Elucidating and engineering thiopeptide biosynthesis. *World Journal of Microbiology and Biotechnology.* 2017; 33:119.
270. Ranieri M, Chan D, Yaeger LN, Rudolph M, Karabelas-Pittman S, Abdo H, *et al*. Thiostrepton Hijacks Pyoverdine Receptors To Inhibit Growth of *Pseudomonas aeruginosa*. *Antimicrobial Agents and Chemotherapy*. 2019; 63(9):e00472-19.
271. Imai Y, Meyer KJ, Iinishi A, Favre-Godal Q, Green R, Manuse S, *et al.* A new antibiotic selectively kills Gram-negative pathogens. *Nature*. 2019; 576**:**459–464.
272. Mullane K, Lee C, Bressler A, Buitrago M, Weiss K, Dabovic K, *et al*. Multicenter, randomized clinical trial to compare the safety and efficacy of LFF571 and vancomycin for *Clostridium difficile* infections. *Antimicrobial Agents and Chemotherapy*. 2015; 59(3):1435–1440.
273. Robinson SL, Christenson JK, Wackett LP. Biosynthesis and chemical diversity of β-lactone natural products. *Natural Product Reports*. 2019; 36(3):458–475.
274. Kluge AF, Petter RC. Acylating drugs: redesigning natural covalent inhibitors. *Current Opinion in Chemical Biology*. 2010; 14(3):421-427.
275. De Pascale G, Nazi I, Harrison P, Wright GD. β-Lactone natural products and derivatives inactivate homoserine transacetylase, a target for antimicrobial agents. *The Journal of Antibiotic*s. 2011; 64:483–487.
276. Böttcher T, Sieber SA. Beta-lactones as privileged structures for the active-site labeling of versatile bacterial enzyme classes. *Angewandte Chemie (International ed. in English)*. 2008; 47(24):4600–4603.
277. Hider RC, Kong X. Chemistry and biology of siderophores. *Natural Product Reports*. 2010; 27(5):637–657.
278. Negash KH, Norris JKS, Hodgkinson JT. Siderophore-Antibiotic Conjugate Design: New Drugs for Bad Bugs? *Molecules*. 2019; 24(18):3314.
279. Andrews SC, Robinson AK, Rodríguez-Quiñones F. Bacterial iron homeostasis. *FEMS Microbiology Reviews*. 2003; 27(2-3):215–237.
280. Miller M, Liu R. Design and Syntheses of New Antibiotics Inspired by Nature's Quest for Iron in an Oxidative Climate. *Accounts of Chemical Research*. 2021; 54(7):1646-1661.

281. Braun V, Pramanik A, Gwinner T, Köberle M, Bohn E. Sideromycins: tools and antibiotics. *Biometals*. 2009; 22(1):3-13.
282. Zheng T, Nolan EM. Enterobactin-mediated delivery of β-lactam antibiotics enhances antibacterial activity against pathogenic *Escherichia coli*. *Journal of the American Chemical Society*. 2014; 136(27):9677–9691.
283. Ghosh M, Miller PA, Möllmann U, Claypool WD, Schroeder VA, Wolter WR, *et al*. Targeted Antibiotic Delivery: Selective Siderophore Conjugation with Daptomycin Confers Potent Activity against Multidrug Resistant *Acinetobacter baumannii* Both *in Vitro* and *in Vivo*. *Journal of Medicinal Chemistry*. 2017; 60(11):4577–4583.
284. Zhanel GG, Golden AR, Zelenitsky S, Wiebe K, Lawrence CK, Adam HJ, *et al*. Cefiderocol: A Siderophore Cephalosporin with Activity Against Carbapenem-Resistant and Multidrug-Resistant Gram-Negative Bacilli. *Drugs*. 2019; 79(3):271–289.
285. Naseer S, Weinstein EA, Rubin DB, Suvarna K, Wei X, Higgins K, *et al*. US Food and Drug Administration (FDA): Benefit-Risk Considerations for Cefiderocol (Fetroja®). *Clinical Infectious Diseases*. 2021; 72(12):e1103–e1111.
286. Stanborough T, Fegan N, Powell SM, Tamplin M, Chandry PS. Vibrioferrin production by the food spoilage bacterium *Pseudomonas fragi*. *FEMS Microbiology Letters*. 2018; 365(6): fnx279.
287. Castro-Falcón G, Hahn D, Reimer D, Hughes CC. Thiol Probes To Detect Electrophilic Natural Products Based on Their Mechanism of Action. *ACS Chemical Biology*. 2016; 11(8):2328-36.
288. Wolf F, Bauer JS, Bendel TM, Kulik A, Kalinowski J, Gross H, *et al*. Biosynthesis of the β-Lactone Proteasome Inhibitors Belactosin and Cystargolide. *Angewandte Chemie (International Edition in English)*. 2017; 56(23):6665-6668.
289. Caspi R, Billington R, Keseler IM, Kothari A, Krummenacker M, Midford PE, *et al*. The MetaCyc database of metabolic pathways and enzymes - a 2019 update. *Nucleic Acids Research*. 2020; 48(D1):D445-D453.
290. Amorim Franco TM, Blanchard JS. Bacterial Branched-Chain Amino Acid Biosynthesis: Structures, Mechanisms, and Drugability. *Biochemistry*. 2017; 56(44):5849-5865.

291. Zhang S, Yang W, Chen H, Liu B, Lin B, Tao Y. Metabolic engineering for efficient supply of acetyl-CoA from different carbon sources in *Escherichia coli*. *Microbial Cell Factories*. 2019; 18(1):130.
292. Shi L, Tu BP. Acetyl-CoA and the regulation of metabolism: mechanisms and consequences. *Current Opinion in Cell Biology*. 2015; 33:125–131.
293. Wolfe AJ. The acetate switch. *Microbiology and Molecular Biology Reviews.* 2005; 69(1):12–50.
294. Yang Y, Pollard AM, Höfler C, Poschet G, Wirtz M, Hell R, *et al*. Relation between chemotaxis and consumption of amino acids in bacteria. *Molecular Microbiology*. 2015; 96(6):1272–1282.
295. Bren A, Park JO, Towbin BD, Dekel E, Rabinowitz JD, Alon U*.* Glucose becomes one of the worst carbon sources for *E. coli* on poor nitrogen sources due to suboptimal levels of cAMP. *Scientific Report.* 2016; 6:24834.
296. Fujita Y, Matsuoka H, Hirooka K. Regulation of fatty acid metabolism in bacteria. *Molecular Microbiology*. 2007; 66(4):829–839.
297. Weimar JD, DiRusso CC, Delio R, Black PN. Functional role of fatty acyl-coenzyme A synthetase in the transmembrane movement and activation of exogenous long-chain fatty acids. Amino acid residues within the ATP/AMP signature motif of *Escherichia coli* FadD are required for enzyme activity and fatty acid transport. *The Journal of Biological Chemistry*. 2002; 277(33):29369–29376.
298. Cronan JE, Thomas J. Bacterial fatty acid synthesis and its relationships with polyketide synthetic pathways. *Methods in Enzymology*. 2009; *459*:395–433.
299. Agnihotri G, Liu HW. (2003). Enoyl-CoA hydratase. Reaction, mechanism, and inhibition. *Bioorganic & Medicinal Chemistry*. 2003; 11(1):9–20.
300. Oh MK, Rohlin L, Kao KC, Liao JC. Global expression profiling of acetate-grown *Escherichia coli*. *The Journal of Biological Chemistry*. 2002; 277(15):13175–13183.
301. Ikeda TP, Shauger AE, Kustu S. *Salmonella typhimurium* apparently perceives external nitrogen limitation as internal glutamine limitation. *Journal of Molecular Biology*. 1996; 259(4):589–607.
302. Shin S, Song SG, Lee DS, Pan JG, Park C. Involvement of *iclR* and *rpoS* in the induction of *acs*, the gene for acetyl coenzyme A synthetase of *Escherichia coli* K-12. *FEMS Microbiology Letters*. 1997; 146(1):103–108.

303. Starai VJ, Escalante-Semerena JC. Acetyl-coenzyme A synthetase (AMP forming). *Cellular and Molecular Life Sciences.* 2004; 61(16):2020–2030.
304. Martínez-Blanco H, Reglero Á, Fernández-Valverde M, Ferrero MA, Moreno MA, Peñalva MA, *et al*. (1992). Isolation and characterization of the acetyl-CoA synthetase from *Penicillium chrysogenum*. Involvement of this enzyme in the biosynthesis of penicillins. *The Journal of Biological Chemistry. 1992; 267(8)*:5474-81.
305. Leone S, Sannino F, Tutino ML, Parrilli E, Picone D. Acetate: friend or foe? Efficient production of a sweet protein in *Escherichia coli* BL21 using acetate as a carbon source. *Microbial Cell Factories*. 2015; 14:106.
306. Wang H, Wang F, Wang W, Yao X, Wei D, Cheng H, *et al*. Improving the expression of recombinant proteins in *E. coli* BL21 (DE3) under acetate stress: an alkaline pH shift approach. *PloS One*. 2014; *9*(11):e112777.
307. AAT Bioquest, Inc. *Quest Calculate™ Potassium Phosphate (pH 5.8 to 8.0) Solution Preparation and Recipe*. AAT Bioquest, Inc., Sunnyvale, USA. 2020; available at https://www.aatbio.com Accessed October 2020.
308. Groisman EA, Hollands K, Kriner MA, Lee EJ, Park SY, Pontes MH. Bacterial Mg^{2+} homeostasis, transport, and virulence. *Annual Review of Genetics*. 2013; 47:625–646.
309. Ashino K, Sugano K, Amagasa T, Ying B. Predicting the decision making chemicals used for bacterial growth. *Scientific Reports.* 2019; 9:7251.
310. Kim BJ, Park JH, Park TH, Bronstein PA, Schneider DJ, Cartinhour SW, *et al*. Effect of iron concentration on the growth rate of *Pseudomonas syringae* and the expression of virulence factors in hrp-inducing minimal medium. *Applied and Environmental Microbiology*. 2009; 75(9):2720–2726.
311. Urtuvia V, Villegas P, González M, Seeger M. Bacterial production of the biodegradable plastics polyhydroxyalkanoates. *International Journal of Biological Macromolecules*. 2014; 70:208–213.
312. Verlinden RA, Hill DJ, Kenward MA, Williams CD, Radecka I. Bacterial synthesis of biodegradable polyhydroxyalkanoates. *Journal of Applied Microbiology*. 2007; 102(6): 1437–1449.
313. Surendran A, Lakshmanan M, Chee JY, Sulaiman AM, Thuoc DV, Sudesh K. Can Polyhydroxyalkanoates Be Produced Efficiently From Waste Plant and Animal Oils?. *Frontiers in Bioengineering and Biotechnology*. 2020; *8*:169.

314. Hsu J. How covid-19 is accelerating the threat of antimicrobial resistance. *BMJ (Clinical Research Ed.)*. 2020; 369:m1983.
315. Rawson TM, Ming D, Ahmad R, Moore L, Holmes AH. Antimicrobial use, drug-resistant infections and COVID-19. *Nature Reviews Microbiology*. 2020; 18(8):409–410.
316. Bengoechea JA, Bamford CG. (2020). SARS-CoV-2, bacterial co-infections, and AMR: the deadly trio in COVID-19?. *EMBO Molecular Medicine*. 2020; 12(7):e12560.
317. Nieuwlaat R, Mbuagbaw L, Mertz D, Burrows LL, Bowdish DME, Moja L, *et al*. Coronavirus Disease 2019 and Antimicrobial Resistance: Parallel and Interacting Health Emergencies. *Clinical Infectious Diseases*. 2021; 72(9):1657–1659.
318. Shima J, Hesketh A, Okamoto S, Kawamoto S, Ochi K. Induction of actinorhodin production by rpsL (encoding ribosomal protein S12) mutations that confer streptomycin resistance in *Streptomyces lividans* and *Streptomyces coelicolor* A3(2). *Journal of Bacteriology*. 1996; 178(24):7276–7284.
319. Lewis K. The Science of Antibiotic Discovery. *Cell*. 2020; 181(1):29–45.
320. Miethke M, Pieroni M, Weber T, Brönstrup M, Hammann P, Halby L, *et al.* Towards the sustainable discovery and development of new antibiotics. *Nature Reviews Chemistry.* 2021; 5:726–749.
321. Crits-Christoph A, Diamond S, Butterfield CN, Thomas BC, Banfield JF. Novel soil bacteria possess diverse genes for secondary metabolite biosynthesis. *Nature.* 2018; 558:440–444.
322. Tobias NJ, Wolff H, Djahanschiri B, Grundmann F, Kronenwerth M, Shi Y, *et al.* Natural product diversity associated with the nematode symbionts *Photorhabdus* and *Xenorhabdus*. *Nature Microbiol*ogy. 2017; 2:1676–1685.
323. Lewis K. Persister cells, dormancy and infectious disease. *Nature Reviews Microbiology.* 2007; 5:48–56.
324. Liu J, Gefen O, Ronin I, Bar-Meir M, Balaban NQ. Effect of tolerance on the evolution of antibiotic resistance under drug combinations. *Science.* 2020; 367(6474): 200–204.

325. Parker EN, Drown BS, Geddes EJ, Lee HY, Ismail N, Lau GW, *et al.* Implementation of permeation rules leads to a FabI inhibitor with activity against Gram-negative pathogens. *Nature Microbiology.* 2020; 5:67–75.
326. Richter M, Drown B, Riley A, Garcia A, Shirai T, Svec RL, *et al.* Predictive compound accumulation rules yield a broad-spectrum antibiotic. *Nature.* 2017; 545:299–304.
327. Chairatana P, Zheng T, Nolan EM. Targeting virulence: salmochelin modification tunes the antibacterial activity spectrum of β-lactams for pathogen-selective killing of *Escherichia coli*. *Chemical Science*. 2015; 6(8):4458–4471.
328. Maghembe R, Damian D, Makaranga A, Nyandoro SS, Lyantagaye SL, Kusari S, Hatti-Kaul R. Omics for Bioprospecting and Drug Discovery from Bacteria and Microalgae. *Antibiotics*. 2020; 9(5):229.
329. Palazzotto E, Weber T. Omics and multi-omics approaches to study the biosynthesis of secondary metabolites in microorganisms. *Current Opinion in Microbiology*. 2018; 45:109–116.
330. Maansson M, Vynne NG, Klitgaard A, Nybo JL, Melchiorsen J, Nguyen DD, *et al*. An Integrated Metabolomic and Genomic Mining Workflow To Uncover the Biosynthetic Potential of Bacteria. *mSystems*. 2016; 1(3):e00028-15.
331. Paulus C, Rebets Y, Tokovenko B, Nadmid S, Terekhova LP, Myronovskyi M, *et al*. New natural products identified by combined genomics-metabolomics profiling of marine *Streptomyces* sp. MP131-18. *Scientific Reports*. 2017; 7:42382.
332. Jalili V, Afgan E, Gu Q, Clements D, Blankenberg D, Goecks J, *et al*. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2020 update. *Nucleic Acids Research*. 2020; 48(W1):W395–W402.
333. Awolope OK, Di Salvo A, O'Driscoll NH, Lamb AJ. *Hafnia alvei* strain A23BA chromosome, complete genome. *GenBank.* 2020; https://identifiers.org/insdc:CP050150.
334. Awolope OK, Di Salvo A, O'Driscoll NH, Lamb AJ. *Pseudomonas fragi* strain A13BB chromosome, complete genome. *GenBank.* 2020; https://identifiers.org/insdc:CP065202.
335. National Center for Biotechnology Information. Assembly. https://identifiers.org/ncbi/insdc.gca:GCF_011617105.1 (2020).
336. National Center for Biotechnology Information. Assembly.

https://identifiers.org/insdc.gca:GCA_015767515.1 (2020).

337. National Center for Biotechnology Information. *Sequence Read Archive.* https://identifiers.org/ncbi/insdc.sra:SRP251948 (2020).
338. Awolope OK. Gene clusters potentially associated with environmental adaptation, bioremediation, biocontrol, and plant-growth promotion in *H. alvei* A23BA. Figshare. https://doi.org/10.6084/m9.figshare.12781181.v1 (2020).
339. Awolope OK. Predicted secondary metabolite biosynthetic gene clusters (smBGCs) in *H. alvei* A23BA genome. Figshare. https://doi.org/10.6084/m9.figshare.12781274.v1 (2020).
340. Awolope OK. Gene clusters potentially associated with environmental adaptation, bioremediation, biocontrol, and plant-growth promotion in *P. fragi* A13BB. Figshare. https://doi.org/10.6084/m9.figshare.13507971.v1 (2020).
341. Elyashberg M. Identification and structure elucidation by NMR spectroscopy. *TrAC Trends in Analytical Chemistry*. 2015; 69:88-97.
342. Park J-D, Li Y, Moon K, Han EJ, Lee SR, Seyedsayamdost MR. Structural Elucidation of Cryptic Algaecides in Marine Algal-Bacterial Symbioses by NMR Spectroscopy and MicroED. *Angewandte Chemie (International Ed. in English).* 2022; 61(14)*:*02114022.
343. Deschamps JR. X-ray crystallography of chemical compounds. *Life Sciences*. 2010; 86(15-16):585–589.
344. Krupp F, Frey W, Richert C. Absolute Configuration of Small Molecules by Co-Crystallization. *Angewandte Chemie (International Ed. in English)*. 2020; 59(37):15875–15879.
345. Shigemori H, Komaki H, Yazawa K, Mikami Y, Nemoto A, Tanaka Y, *et al*. Brasilicardin A. A Novel Tricyclic Metabolite with Potent Immunosuppressive Activity from Actinomycete *Nocardia brasiliensis*. *The Journal of Organic Chemistry*. 1998; 63(20):6900–6904.
346. Sánchez S, Chávez A, Forero A, García-Huantè Y, Romero A, Sánchez M, *et al.* Carbon source regulation of antibiotic production. *The Journal of Antibiotics.* 2010; 63:442–459.
347. Sánchez S, Demain AL. Metabolic regulation of fermentation processes. *Enzyme and Microbial Technology.* 2002; 31(7):895-906.

348. Watrous J, Roach P, Alexandrov T, Heath BS, Yang JY, Kersten RD, *et al*. Mass spectral molecular networking of living microbial colonies. *Proceedings of the National Academy of Sciences of the United States of America*. 2012; 109(26):E1743–E1752.
349. Liu WT, Lamsa A, Wong WR, Boudreau PD, Kersten R, Peng Y, *et al*. MS/MS-based networking and peptidogenomics guided genome mining revealed the stenothricin gene cluster in *Streptomyces roseosporus*. *The Journal of Antibiotics*. 2014; 67(1):99–104.
350. Purves K, Macintyre L, Brennan D, Hreggviðsson GÓ, Kuttner E, Ásgeirsdóttir ME, *et al*. Using Molecular Networking for Microbial Secondary Metabolite Bioprospecting. *Metabolites*. 2016; 6(1):2.
351. Nguyen DD, Wu CH, Moree WJ, Lamsa A, Medema MH, Zhao X, *et al*. MS/MS networking guided analysis of molecule and gene cluster families. *Proceedings of the National Academy of Sciences of the United States of America*. 2013; 110(28):E2611–E2620.
352. Frank AM, Bandeira N, Shen Z, Tanner S, Briggs SP, Smith RD, *et al*. Clustering millions of tandem mass spectra. *Journal of Proteome Research*. 2008; 7(1):113–122.
353. Cline MS, Smoot M, Cerami E, Kuchinsky A, Landys N, Workman C, *et al*. Integration of biological networks and gene expression data using Cytoscape. *Nature Protocols*. 2007; 2(10):2366–2382.
354. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, *et al*. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research*. 2003; 13(11):2498–2504.
355. Kohl M, Wiese S, Warscheid B. Cytoscape: software for visualization and analysis of biological networks. *Methods in Molecular Biology.* 2011; 696:291–303.
356. Yang JY, Sanchez LM, Rath CM, Liu X, Boudreau PD, Bruns N, *et al*. Molecular networking as a dereplication strategy. *Journal of Natural Products*. 2013; 76(9):1686–1699.
357. Krug D, Müller R. Secondary metabolomics: the impact of mass spectrometry-based approaches on the discovery and characterization of microbial natural products. *Natural Product Reports*. 2014; 31(6):768–783.
358. Wang M, Carver JJ, Phelan VV, Sanchez LM, Garg N, Peng Y, *et al*. Sharing and community curation of mass spectrometry data with Global Natural

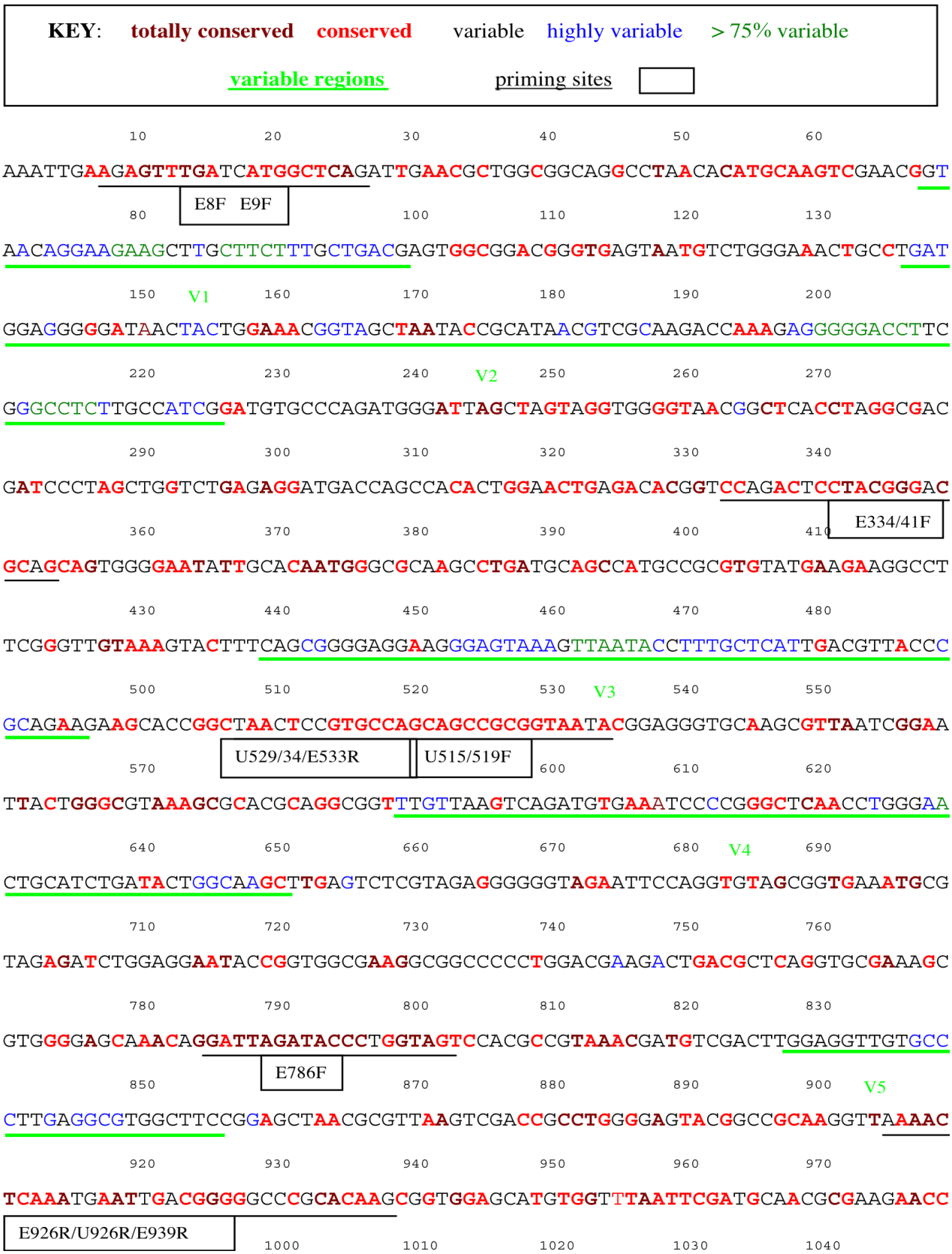
Products Social Molecular Networking. *Nature Biotechnology*. 2016; 34(8):828–837.

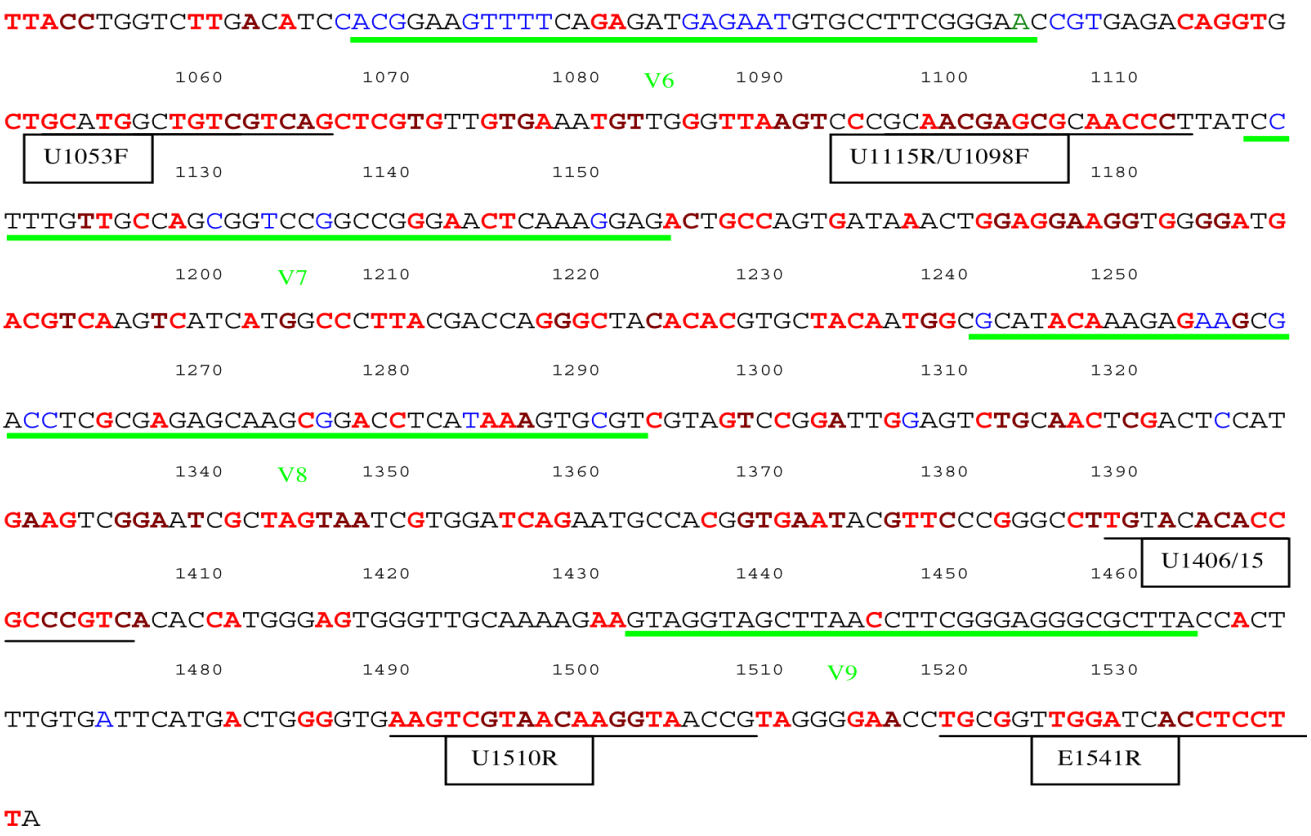
359. Aron AT, Gentry EC, McPhail KL, Nothias LF, Nothias-Esposito M, Bouslimani A, *et al*. Reproducible molecular networking of untargeted mass spectrometry data using GNPS. *Nature Protocols*. 2020; 15(6):1954–1991.
360. Horai H, Arita M, Kanaya S, Nihei Y, Ikeda T, Suwa K, *et al*. MassBank: a public repository for sharing mass spectral data for life sciences. *Journal of Mass Spectrometry.* 2010; 45(7):703–714.
361. Sorokina M, Steinbeck C. Review on natural products databases: where to find data in 2020. *Journal of Cheminformatics*. 2020; 12(1):20.
362. Duncan KR, Crüsemann M, Lechner A, Sarkar A, Li J, Ziemert N, *et al*. Molecular networking and pattern-based genome mining improves discovery of biosynthetic gene clusters and their products from *Salinispora* species. *Chemistry & Biology*. 2015; 22(4):460–471.
363. Vizcaino MI, Engel P, Trautman E, Crawford JM. Comparative metabolomics and structural characterizations illuminate colibactin pathway-dependent small molecules. *Journal of the American Chemical Society*. 2014; 136(26):9244–9247.
364. Padilla D, Acosta F, Ramos-Vivas J, Grasso V, Bravo J, El Aamri F, *et al*. The pathogen *Hafnia alvei* in veterinary medicine: a review. *Journal of Applied Animal Research*. 2015; 43(2):231–235.
365. National Center for Biotechnology Information. Treeviewer JS Version 1.19.3. 2021. Available at NCBI Tree Viewer (nih.gov). Accessed October 2021.
366. Bertelli C, Laird MR, Williams KP, Simon Fraser University Research Computing Group, Lau BY, Hoad G, *et al*. IslandViewer 4: expanded prediction of genomic islands for larger-scale datasets. *Nucleic Acids Research*. 2017; 45(W1):W30–W35.
367. Franzetti L, Scarpellini M. Characterisation of *Pseudomonas* spp. isolated from foods. *Annals of Microbiology.* 2007; 57:39–47.

8. APPENDICES

Appendix 1

Colour-coded Nucleotide Sequence of the 16S rRNA gene in *E. coli* showing primer binding sites.[179]




TTACCTGGTCTTGACATCCACGGAAGTTTTCAGAGATGAGAATGTGCCTTCGGGAACCGTGAGACAGGTG
1060
1070
1080
V6
1090
1100
1110
CTGCATGGCTGTCGTCAGCTCGTGTTGTGAAATGTTGGGTTAAGTCCCGCAACGAGCGCAACCCTTATCC
U1053F
1130
1140
1150
U1115R/U1098F
1180
TTTGTTGCCAGCGGTCCGGCCGGGAACTCAAAGGAGACTGCCAGTGATAAACTGGAGGAAGGTGGGGATG
1200
V7
1210
1220
1230
1240
1250
ACGTCAAGTCATCATGGCCCTTACGACCAGGGCTACACACGTGCTACAATGGCGCATACAAAGAGAAGCG
1270
1280
1290
1300
1310
1320
ACCTCGCGAGAGCAAGCGGACCTCATAAAGTGCGTCGTAGTCCGGATTGGAGTCTGCAACTCGACTCCAT
1340
V8
1350
1360
1370
1380
1390
GAAGTCGGAATCGCTAGTAATCGTGGATCAGAATGCCACGGTGAATACGTTCCCGGGCCTTGTACACACC
U1406/15
1410
1420
1430
1440
1450
1460
GCCCGTCACACCATGGGAGTGGGTTGCAAAAGAAGTAGGTAGCTTAACCTTCGGGAGGGCGCTTACCACT
1480
1490
1500
1510
V9
1520
1530
TTGTGATTCATGACTGGGGTGAAGTCGTAACAAGGTAACCGTAGGGGAACCTGCGGTTGGATCACCTCCT
U1510R
E1541R
TA

Appendix 2

Characteristics of PCR Primers.

	27F (Sigma, UK)	U1510R (Sigma, UK)
Sequence (5'-3')	AGAGTTTGATCCTGGCTCAG	GGTTACCTTGTTACGACTT
Nucleotide Length	20	19
Molecular Weight	6148	5785
T_m (°C)	61.0	53.1
GC%	50	42.1

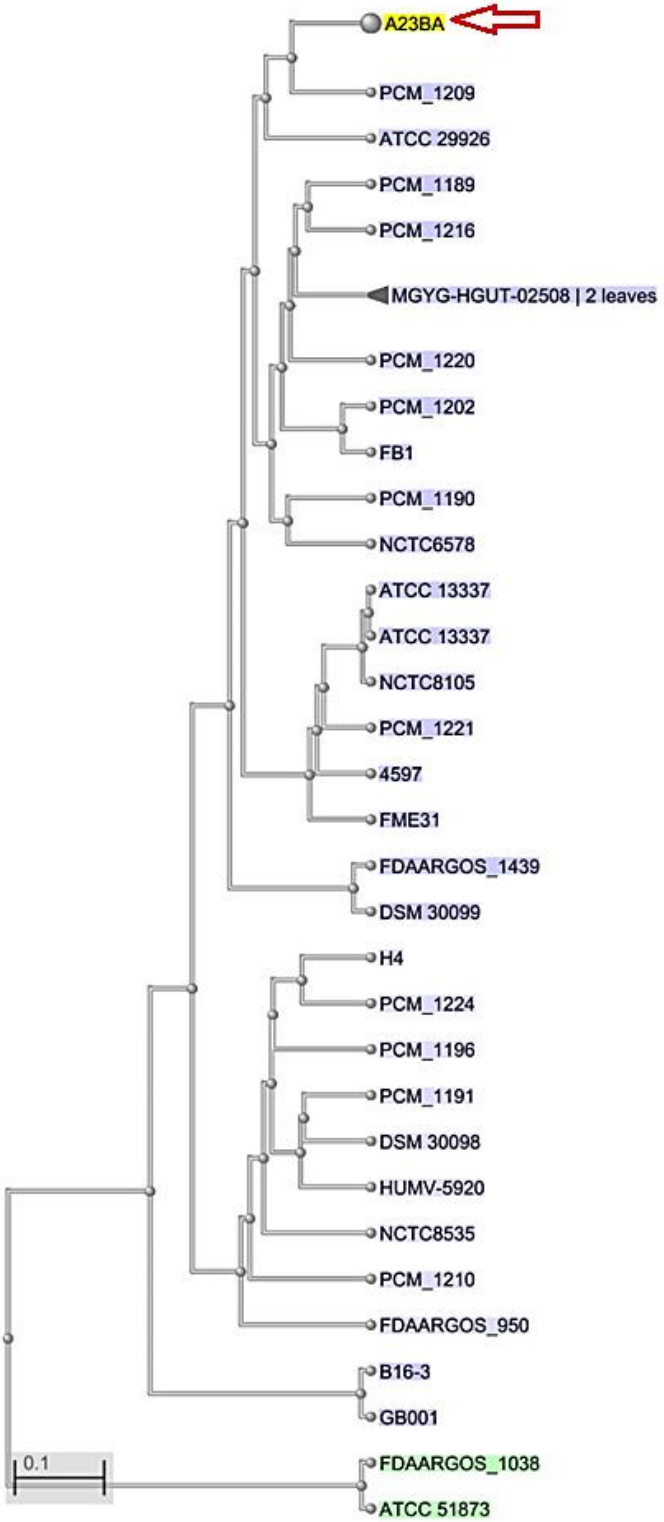
Appendix 3

General and Biochemical Characteristics of *Hafnia alvei*.[242,364]

Characteristic	Description
General Features	
Classification	*Proteobacteria* *Gammaproteobacteria* *Enterobacterales* *Hafniaceae* *Hafnia*
Gram stain	Negative
Cell shape	Rod
Oxygen requirement	Facultatively anaerobic
Growth temperature range	4-44°C
Biochemical Characteristics	
Glucose	Positive
Sucrose	Negative
Oxidase	Negative
Catalase	Positive
Sorbitol	Negative
Indole	Negative
Raffinose	Negative
Rhamnose	Positive
Arabinose	Positive
Inositol	Negative
Adonitol	Negative
Citrate	Negative
Urease	Negative
Lysine	Positive

Appendix 4

Genome Tree of *Hafnia alvei* strains in the NCBI Genome Database

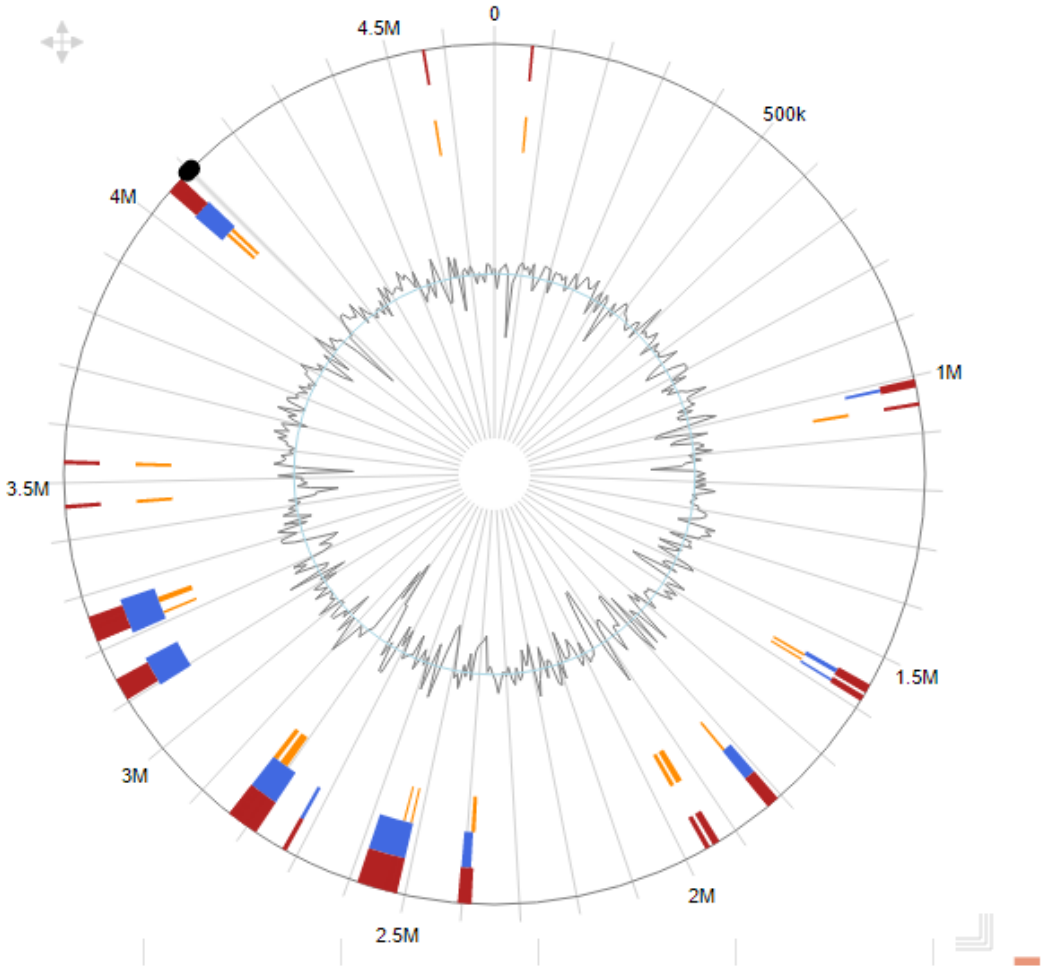


Phylogenetic tree created with NCBI Tree Viewer.[365]

H. alvei A23BA highlighted with the red arrow.

Appendix 5

Genomic Islands in *Hafnia alvei* A23BA



Genomic Islands of *H. alvei* A23BA as predicted by IslandViewer4: Predicted genomic islands are colour-coded according to the detection methods used i.e. **integrated detection, IslandPath-DIMOB, SIGI-HMM.**

Created with IslandViewer 4.[366]

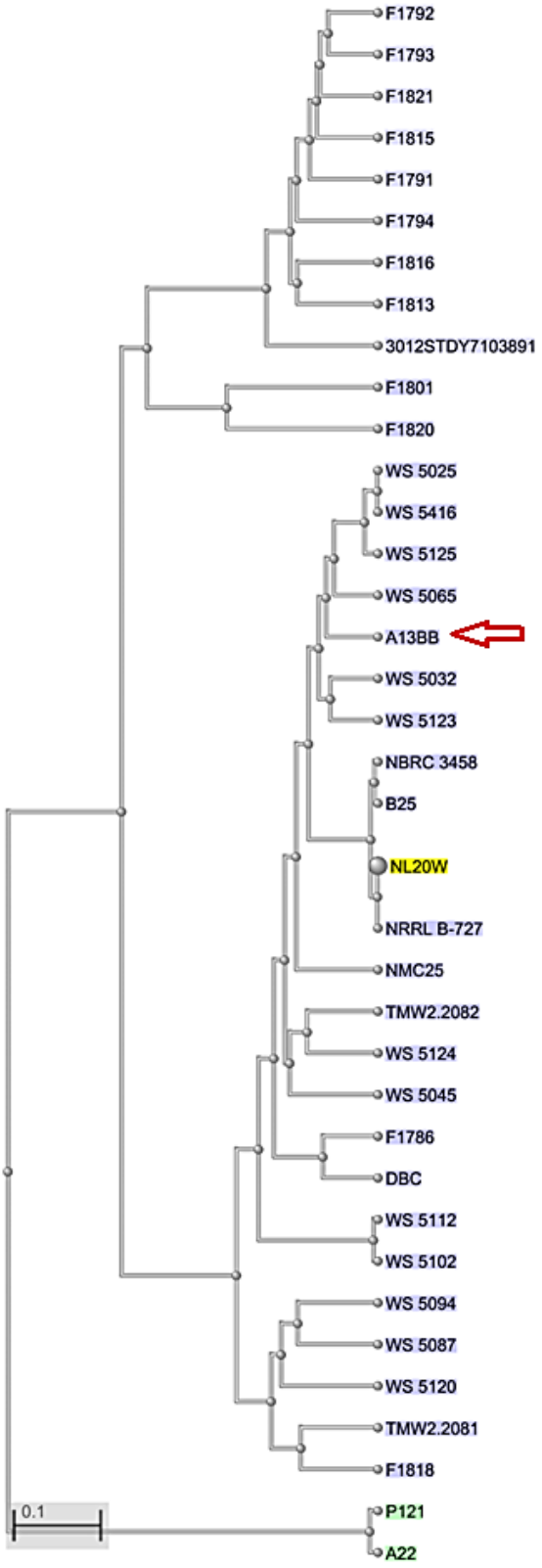
Appendix 6

General and Biochemical Characteristics of *Pseudomonas fragi*.[241,367]

Characteristic	Description
General Features	
Classification	*Proteobacteria* *Gammaproteobacteria* *Pseudomonadales* *Pseudomanadaceae* *Pseudomonas*
Gram stain	Negative
Cell shape	Rod
Oxygen requirement	Aerobic
Growth temperature range	2-35°C
Biochemical Characteristics	
Glucose	Negative
Mannose	Positive
Mannitol	Positive
Indole	Negative
Citrate	Positive
Gluconate	Positive
Adipate	Negative
Maltose	Negative
Arabinose	Positive
Urease	Negative
Catalase	Positive

Appendix 7

Genome Tree of *Pseudomonas fragi* strains in the NCBI Genome Database

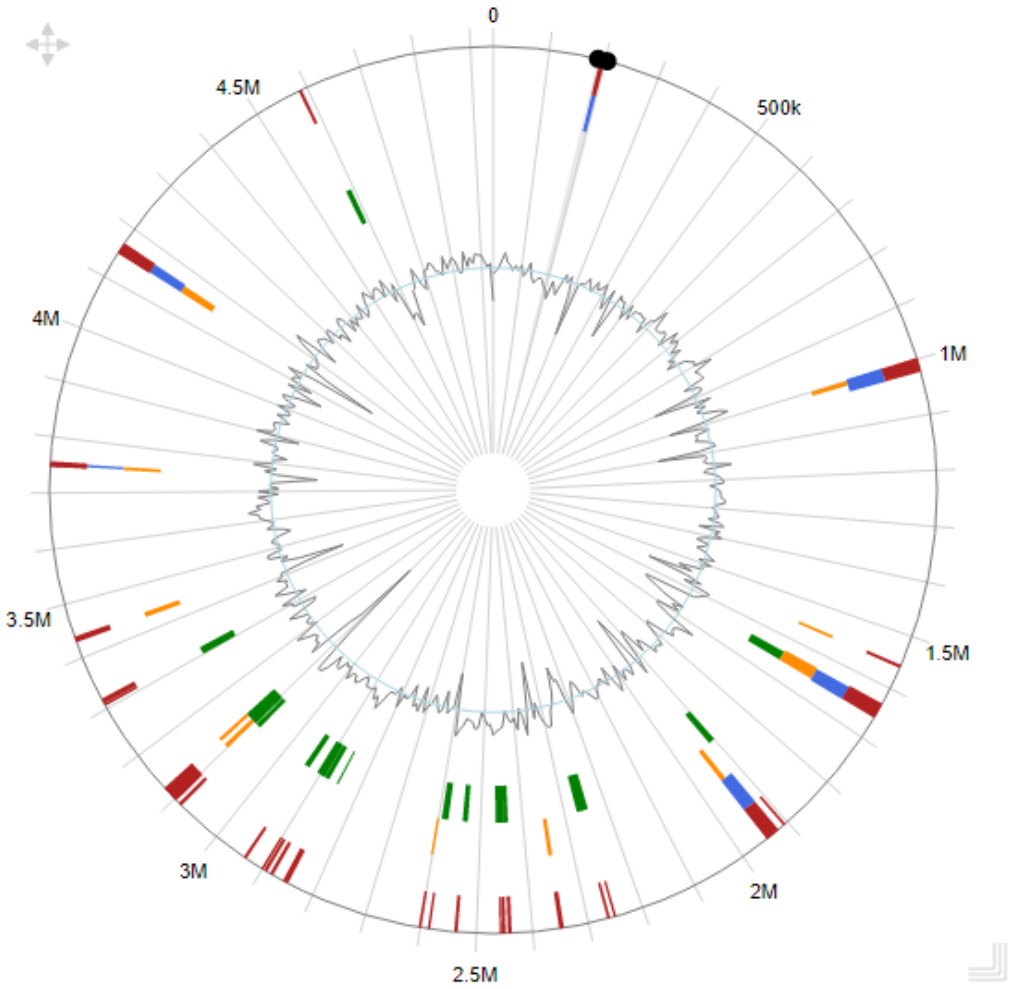


Phylogenetic tree created with NCBI Tree Viewer.[365]

P. fragi A13BB highlighted with the red arrow.

Appendix 8

Genomic Islands in *Pseudomonas fragi* A13BB



Genomic Islands of *P. fragi* A13BB as predicted by IslandViewer4: Predicted genomic islands are colour-coded according to the detection methods used i.e. **integrated detection,** IslandPath-DIMOB, SIGI-HMM, **IslandPick.**

Created with IslandViewer 4.[366]

Appendix 9

A. Preparation of Sterile Fermentation Broth

Fermentation broth was prepared by adding 0.25g of anhydrous sodium acetate (Fisher Scientific, UK) to 250 mL deionised water to give a working concentration of 12 mM acetate in the solution. Thereafter, 1.25g of yeast extract (Sigma Aldrich, USA) was added to the sodium acetate solution, followed by 0.7g of $MgSO_4.7H_20$ (Fisher Scientific, UK) and 0.15g of L-valine (Sigma Aldrich, USA). Finally, 0.1 mL of ferrous sulphate solution (prepared by dissolving 0.25g of $FeSO_4.7H_2O$ (Sigma Aldrich, USA) in 50 mL deionised water) was added to the fermentation broth. The contents of the broth were thoroughly mixed, and 45 mL aliquots were transferred into five 250 mL Erlenmeyer flasks.

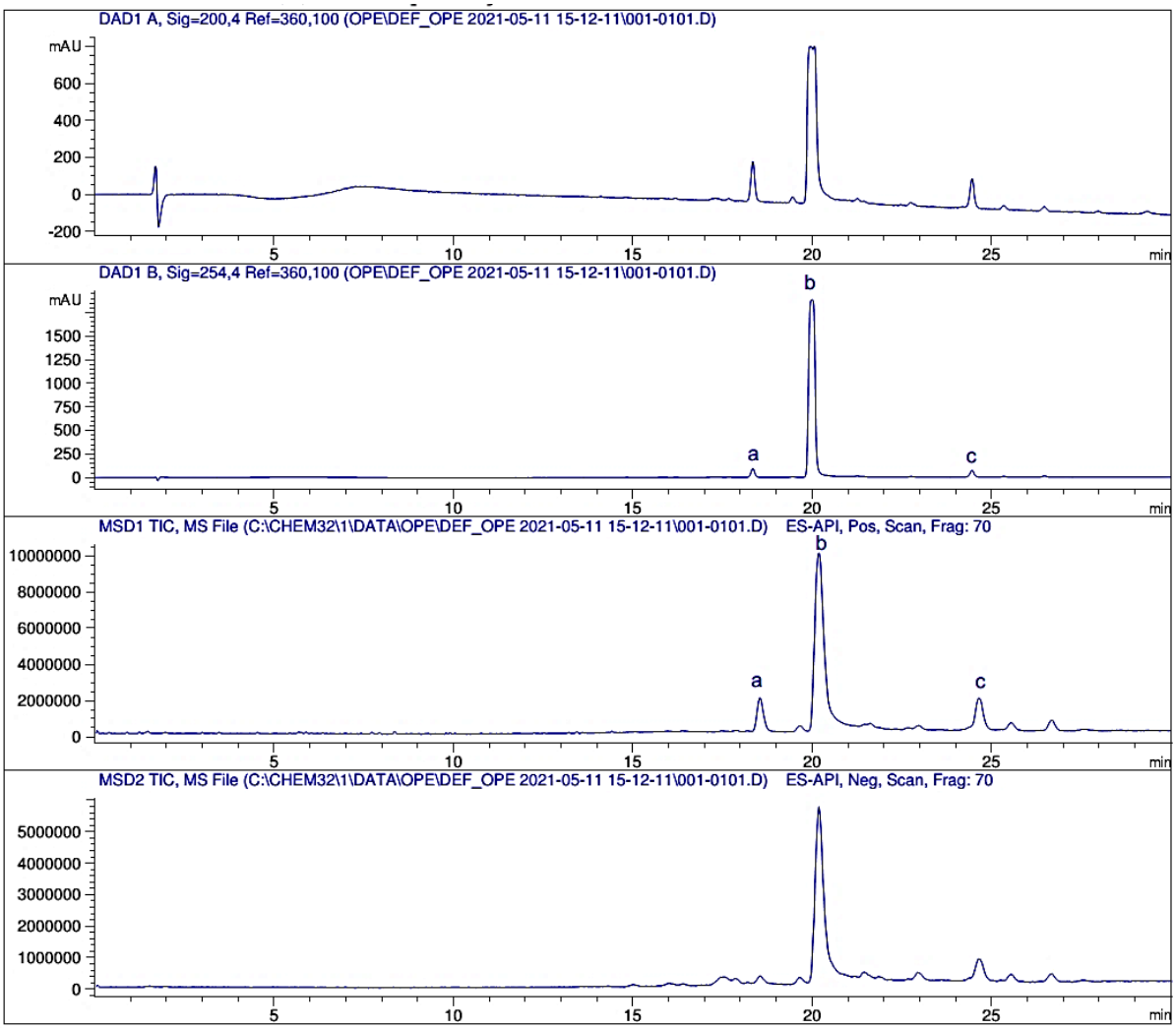
A second batch of fermentation broth was prepared as detailed above but substituting 0.15g of L-valine with 0.16g of L-isoleucine (Sigma Aldrich, USA). The contents of all ten flasks were later sterilised at 121°C for 15 minutes.

B. Preparation of 0.7M Potassium Phosphate Buffer (pH 7.5)

0.7M potassium phosphate buffer solution was prepared by adding 4.49g of dibasic potassium phosphate (K_2HPO_4; Sigma Aldrich, USA) and 1.26g of monobasic potassium phosphate (KH_2PO_4; Sigma Aldrich, USA) to 50 mL deionised water. The solution was thoroughly mixed and sterilised at 121°C for 15 minutes.

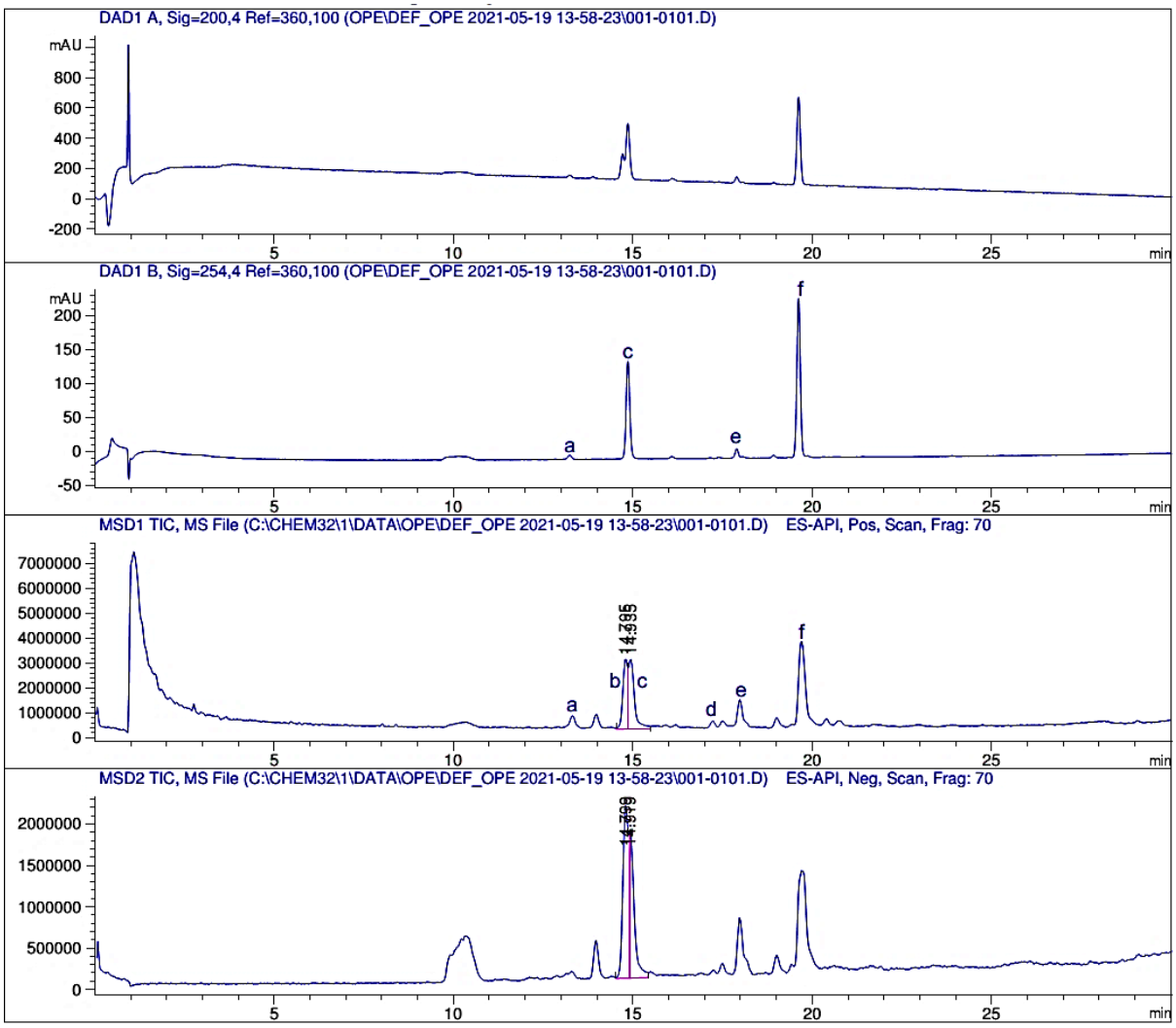
Appendix 10

LC-MS Chromatogram of Probe Solution in Acetonitrile.



Appendix 11

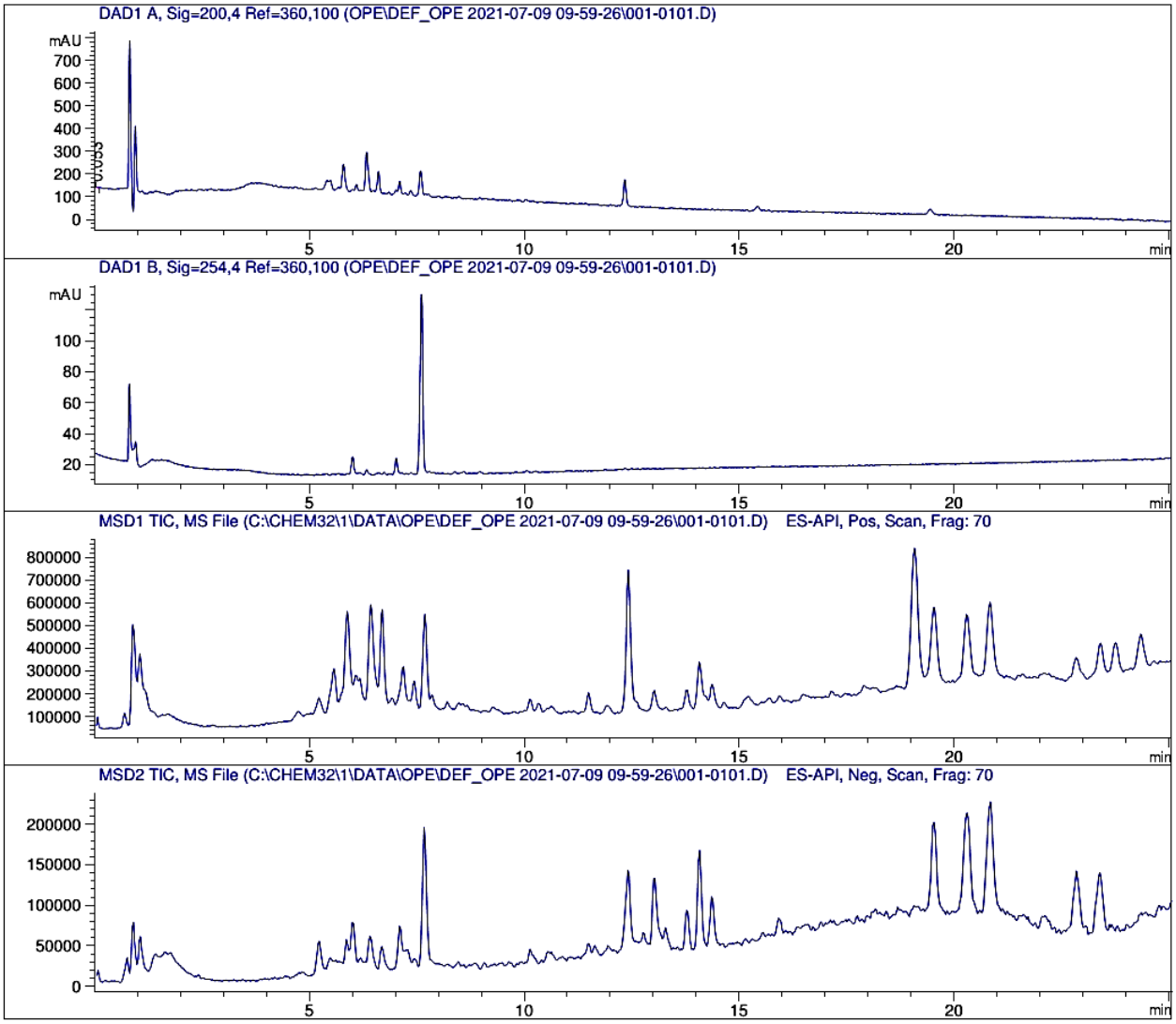
LC-MS Chromatogram of Reaction Mixture of Probe and Salinosporamide A



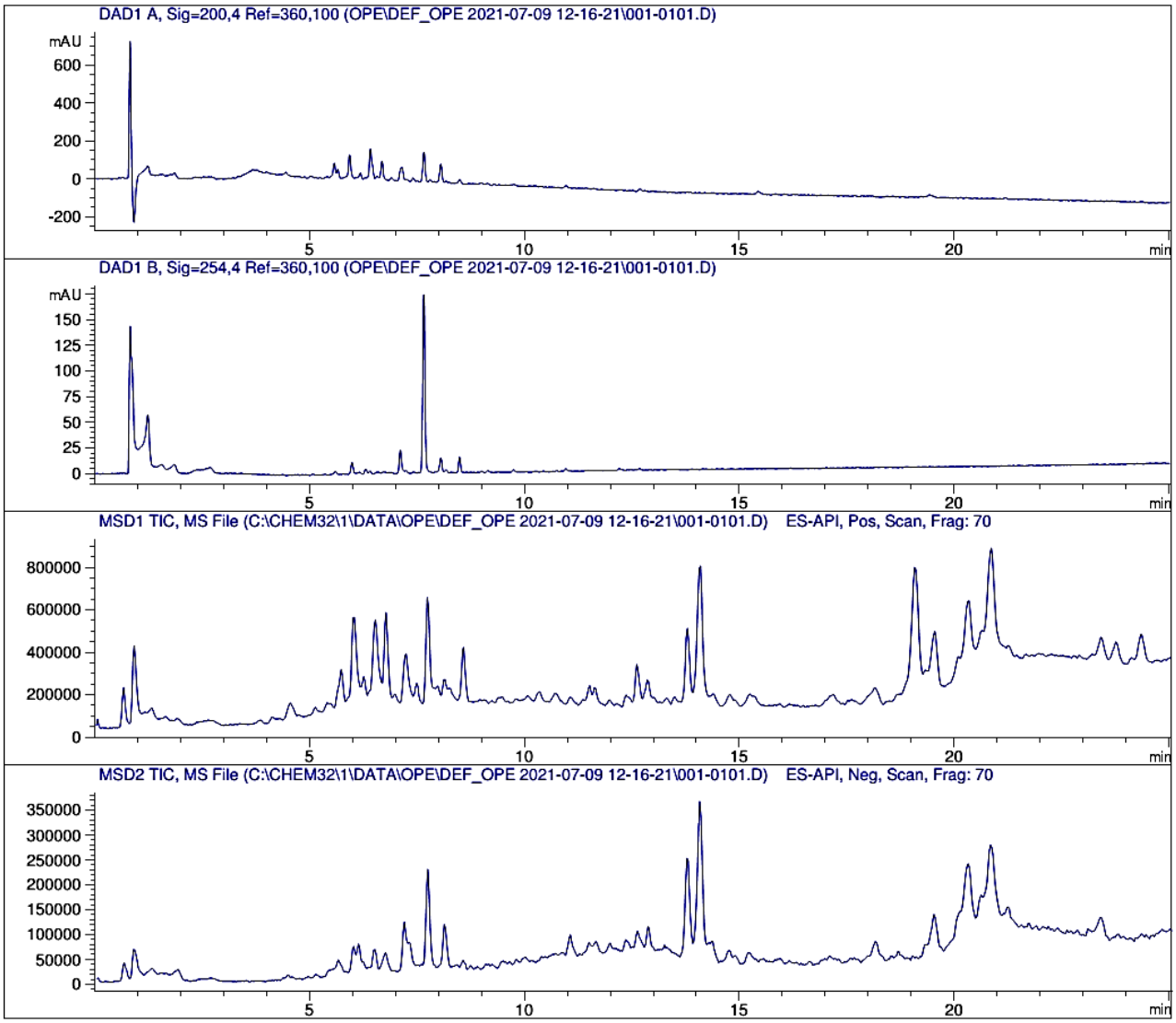
Appendix 12

LC-MS Chromatograms of Solutions of Extracts in Acetonitrile

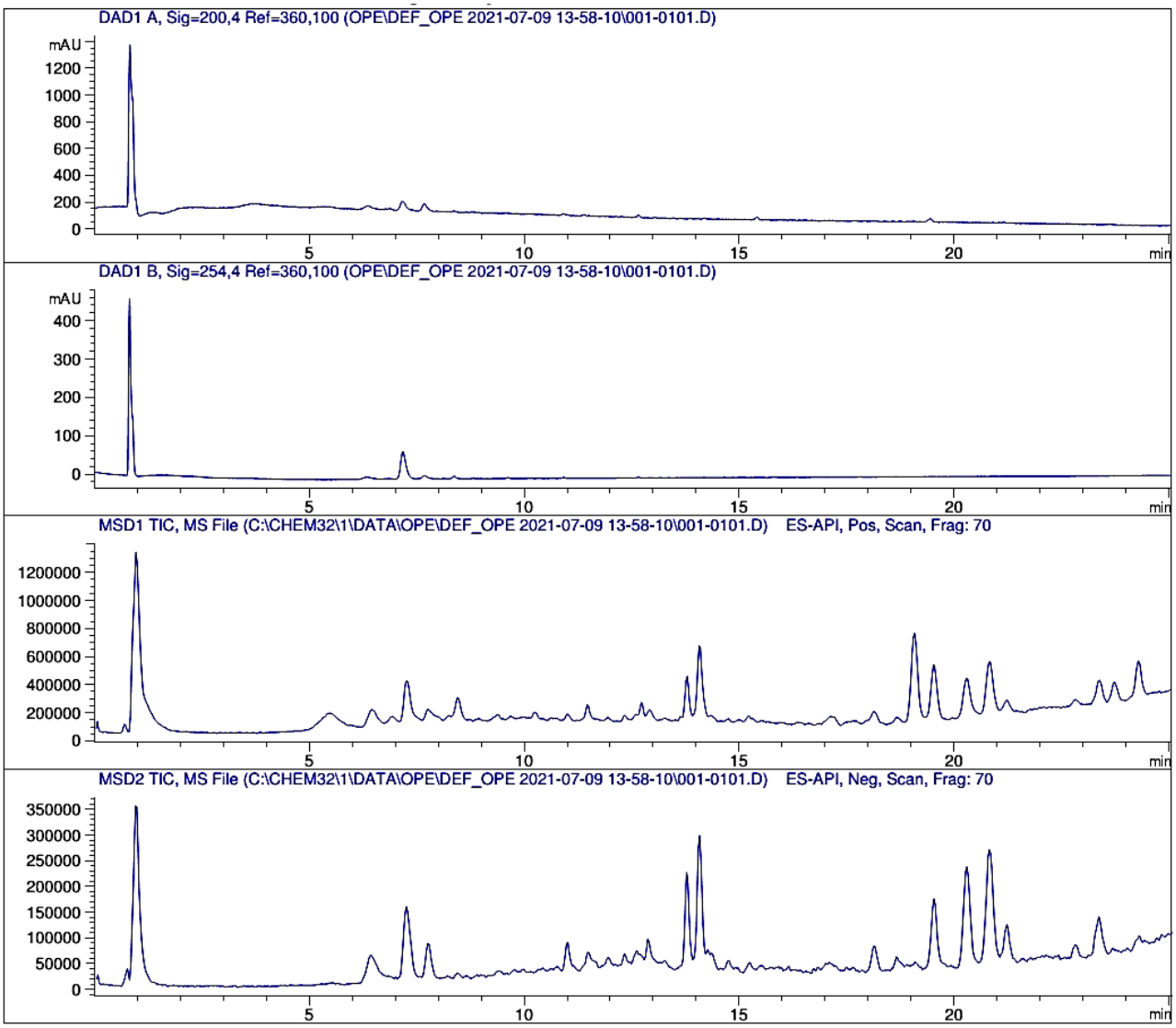
Appendix 12.1: LC-MS chromatogram of sample B (extract of *P. fragi* A13BB inoculated broth with L-valine supplementation)



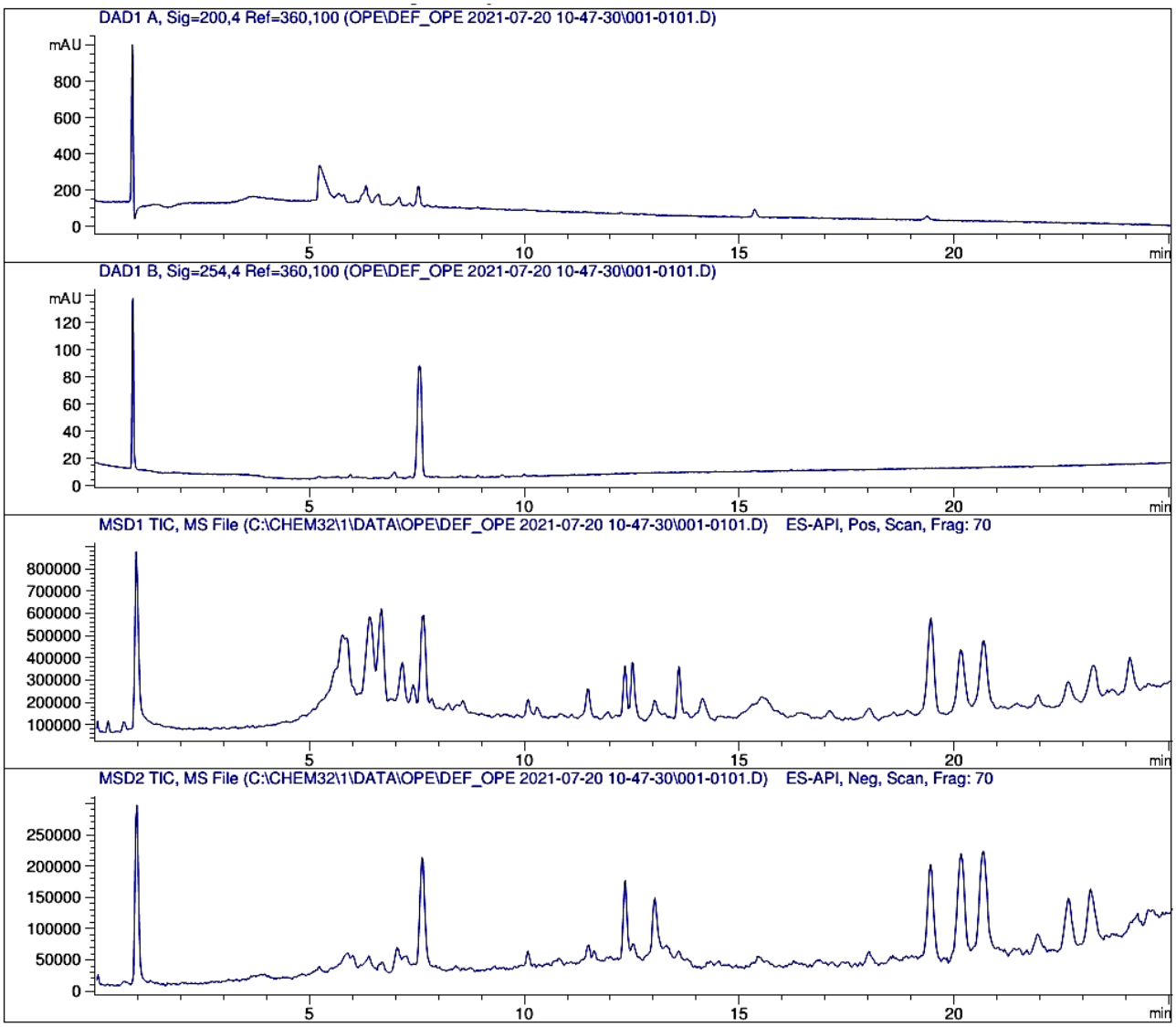
Appendix 12.2: LC-MS chromatogram of sample C (extract of *H. alvei* A23BA inoculated broth with L-valine supplementation)



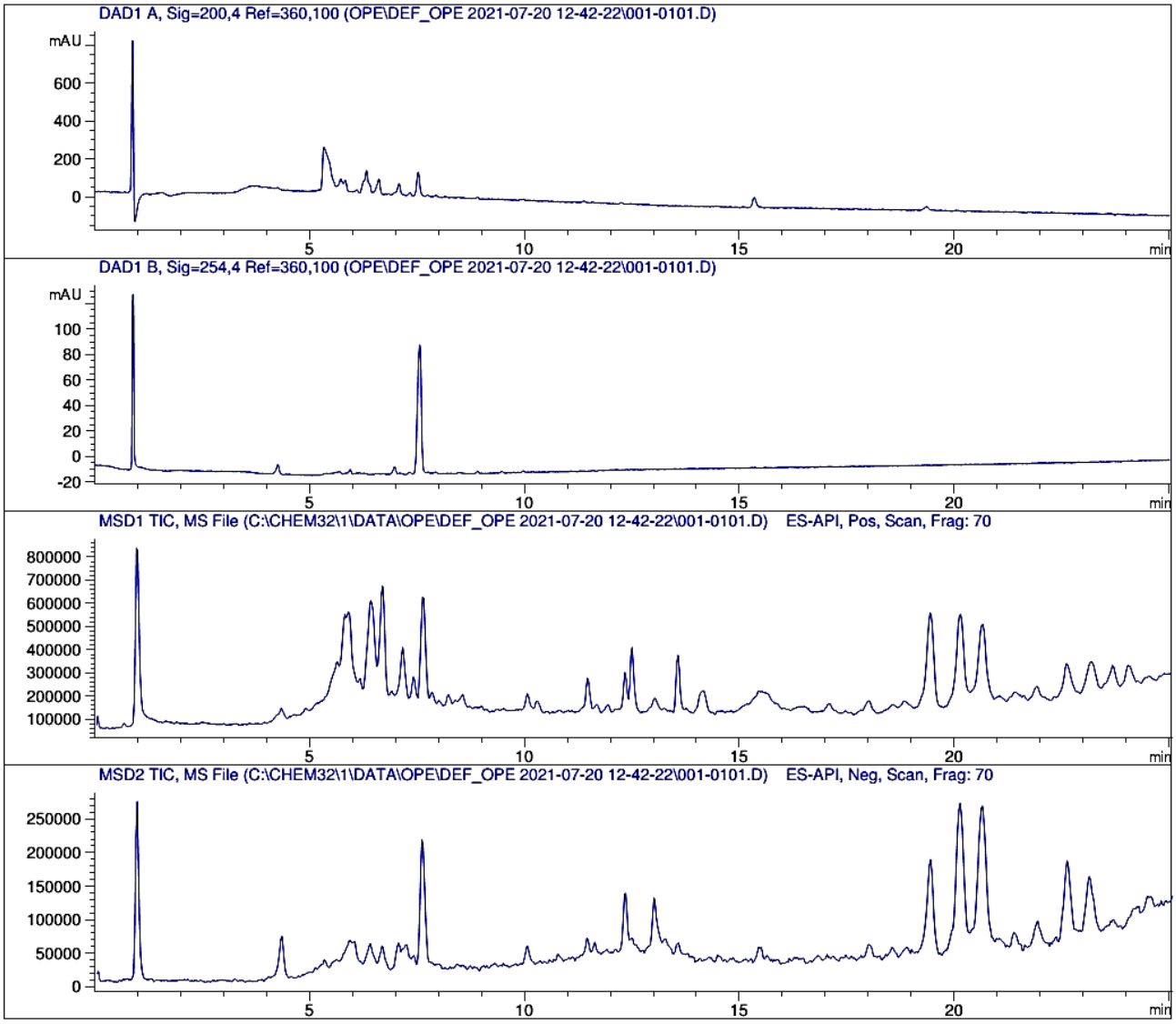
Appendix 12.3: LC-MS chromatogram of sample D (extract of *H. alvei* A23BA inoculated broth with L-valine supplementation)



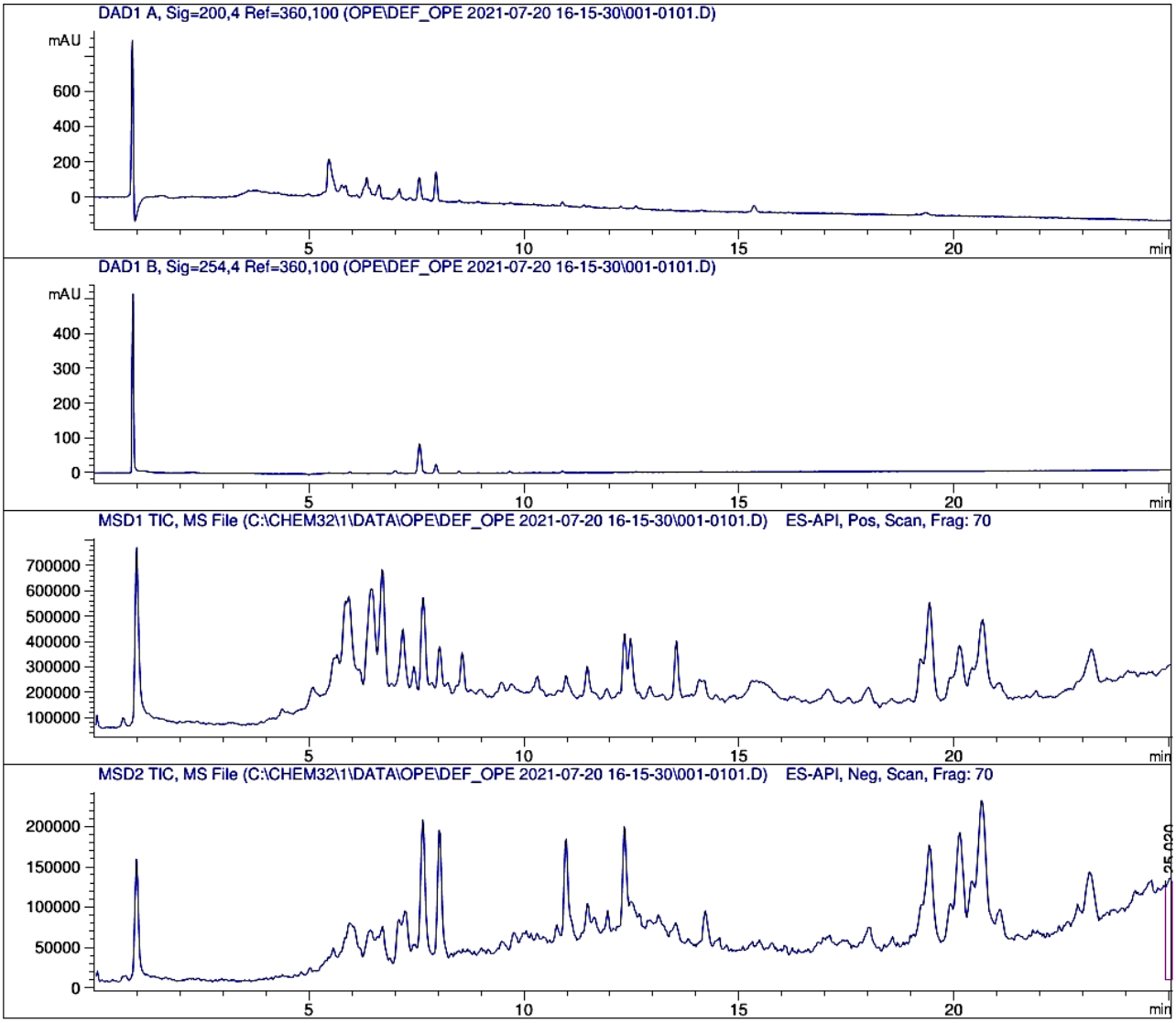
Appendix 12.4: LC-MS chromatogram of sample F (extract of *P. fragi* A13BB inoculated broth with L-isoleucine supplementation)



Appendix 12.5: LC-MS chromatogram of sample G (extract of *P. fragi* A13BB inoculated broth with L-isoleucine supplementation)

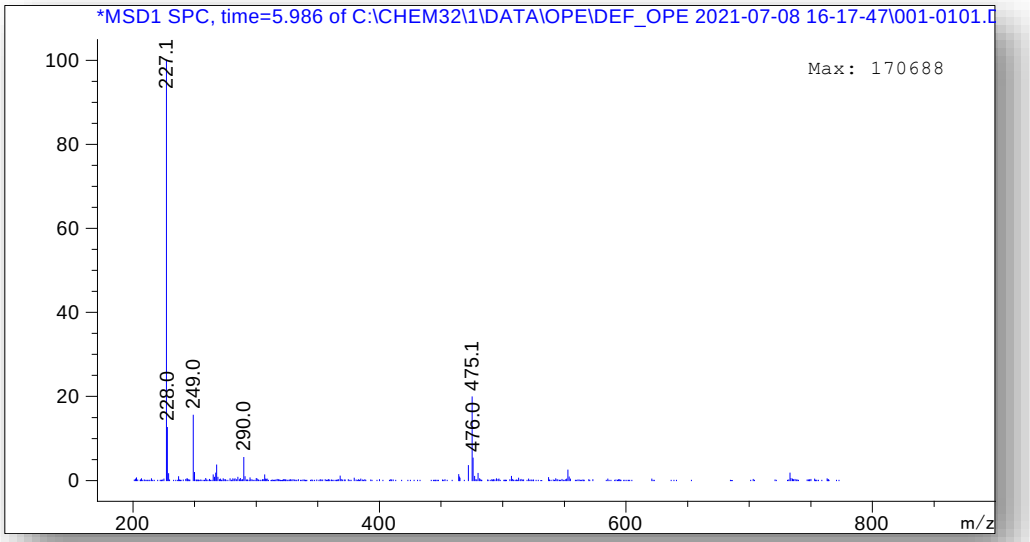


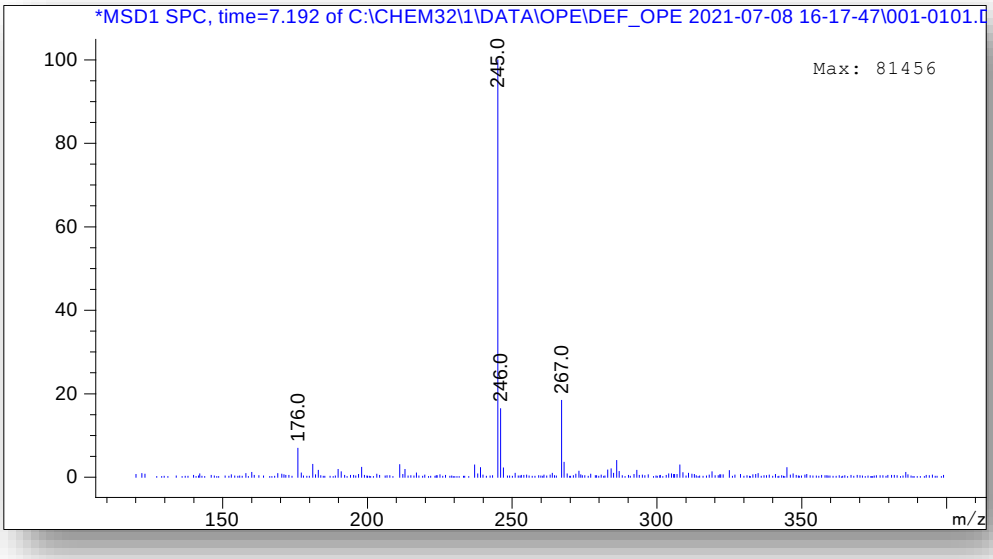
Appendix 12.6: LC-MS chromatogram of sample I (extract of *H. alvei* A23BA inoculated broth with L-isoleucine supplementation)



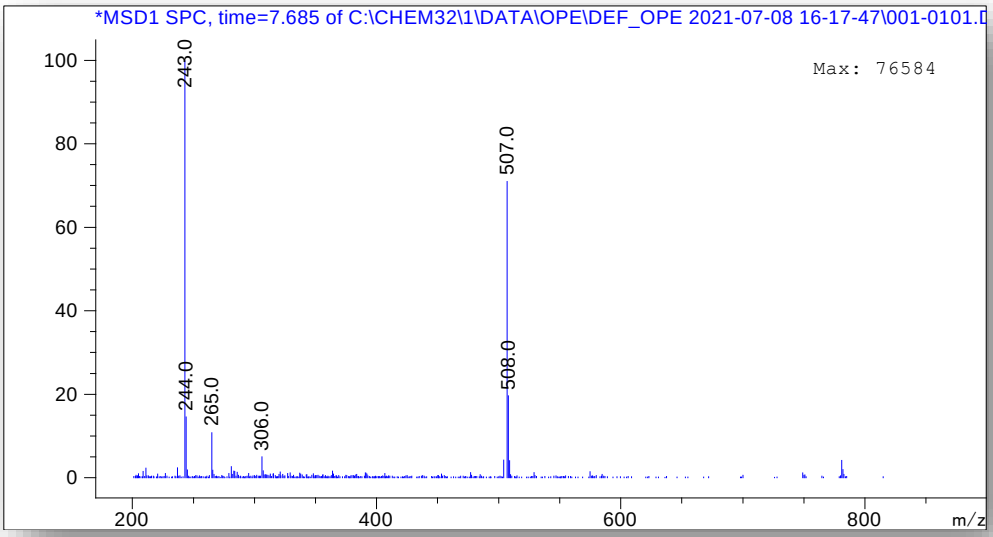
Appendix 13

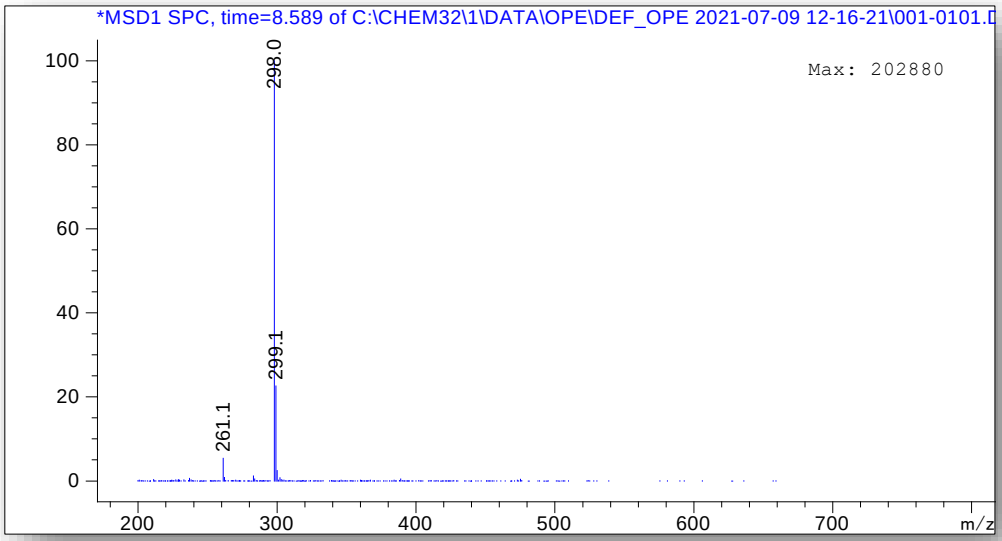
MS Pattern of Metabolite Ions Common to Extracts of Inoculated Broths



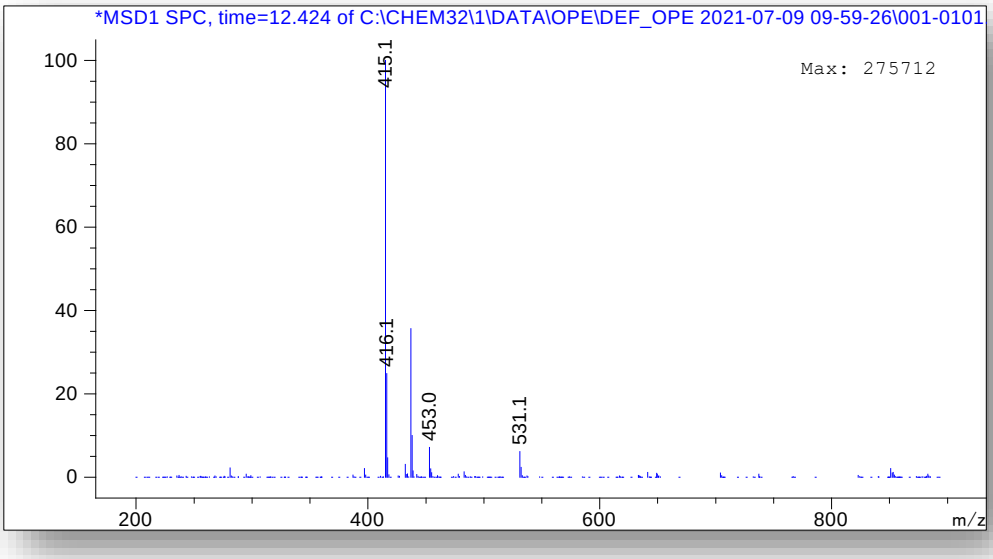


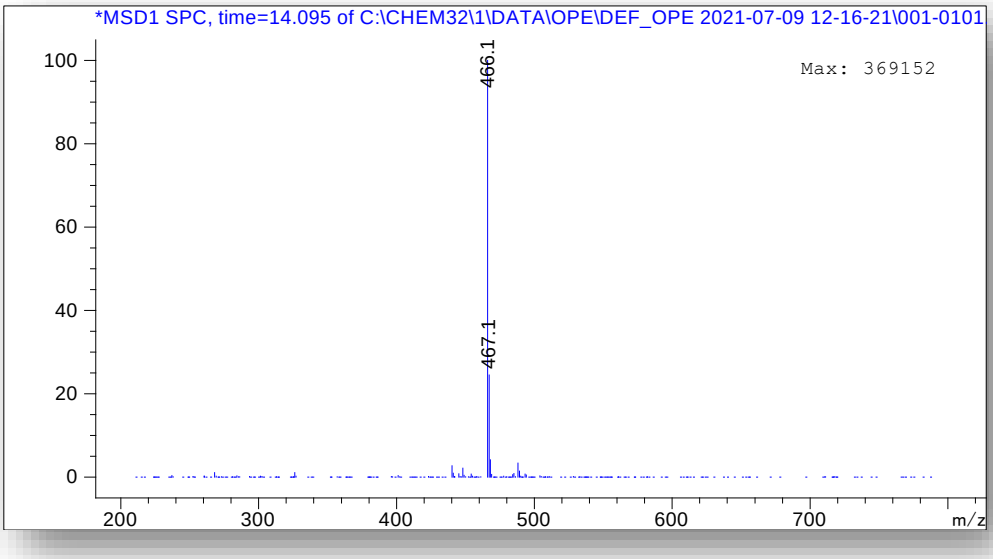
Appendix 13 Cont'd





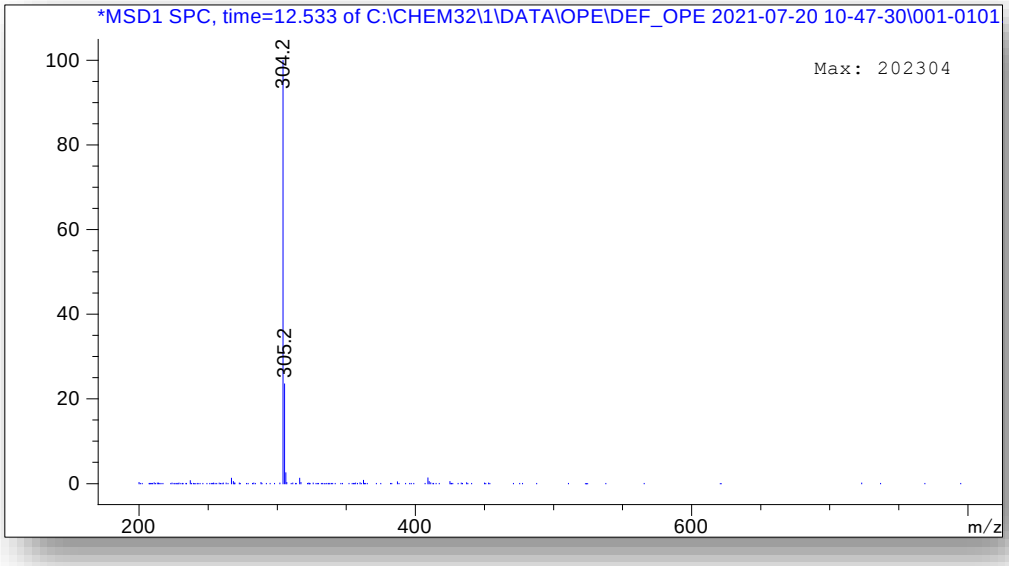
Appendix 13 Cont'd

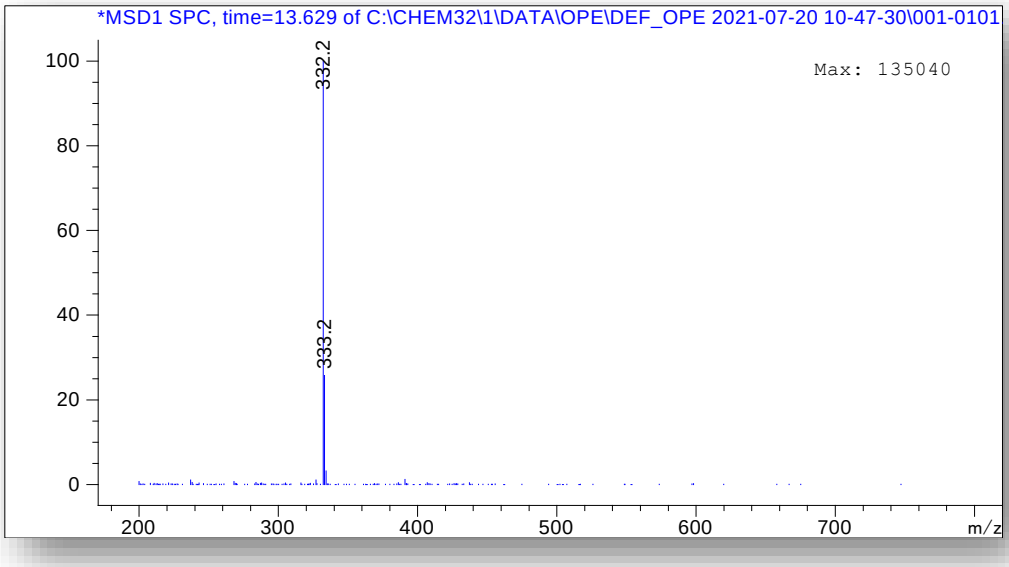




Appendix 14

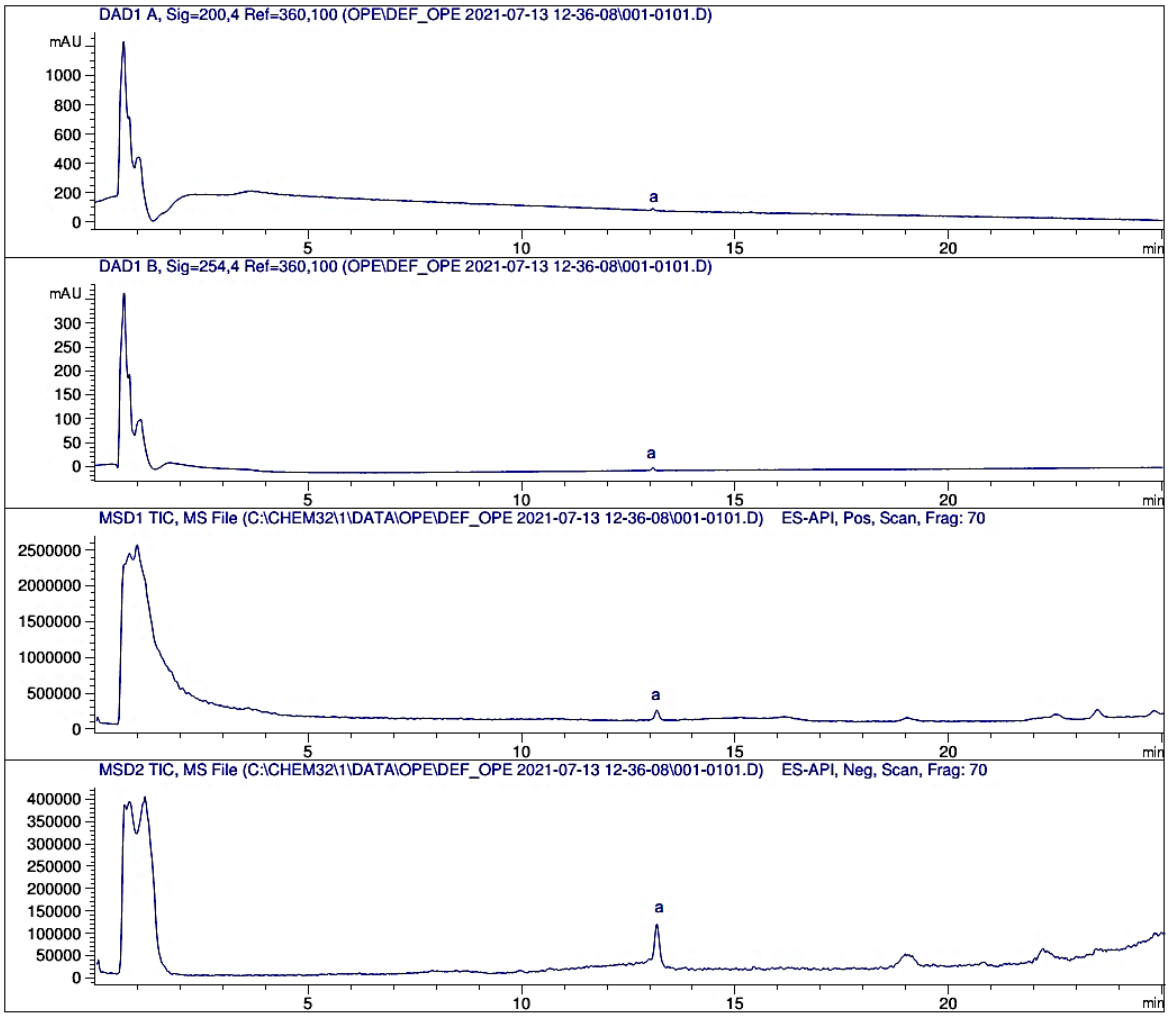
MS Pattern of Metabolite Ions Unique to Extracts of Inoculated Broths with L-isoleucine Supplementation

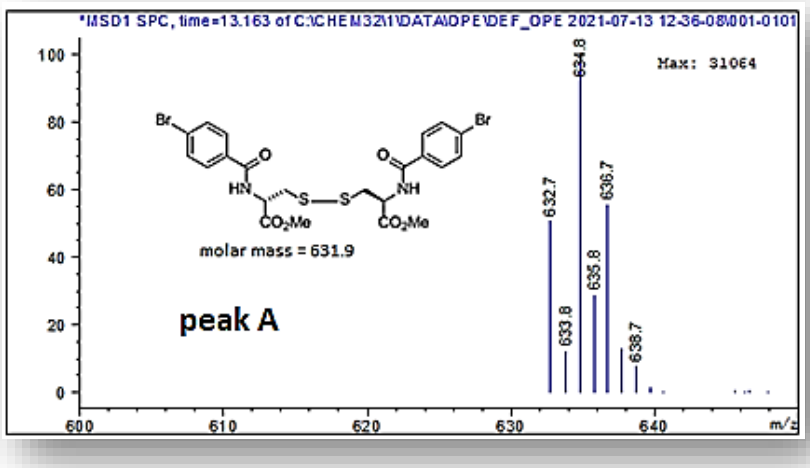




Appendix 15

LC-MS Chromatogram of Reaction Mixture of Sample E (Uninoculated Broth) with Probe.



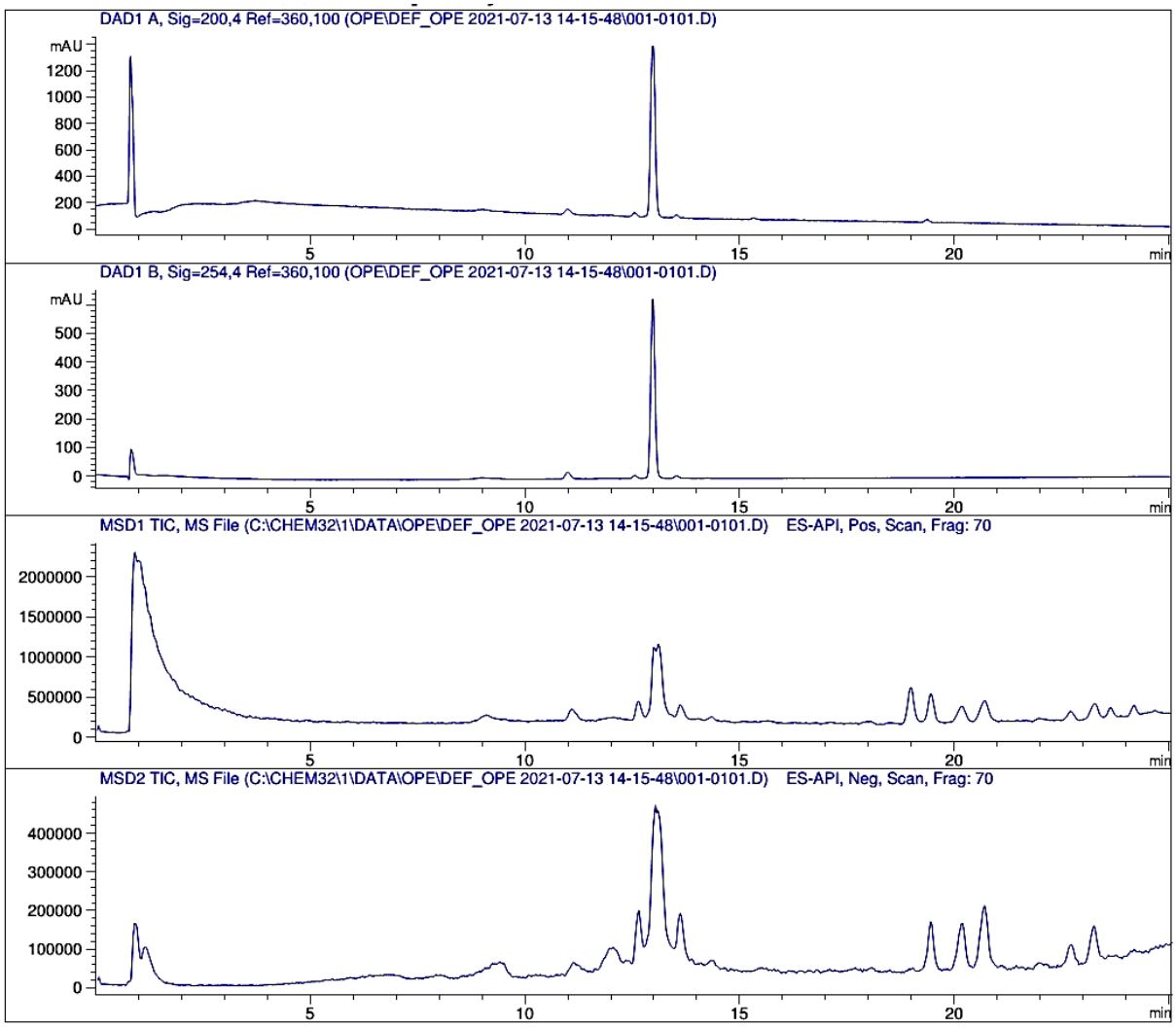


Mass spectrum pattern of peak A (probe dimer).

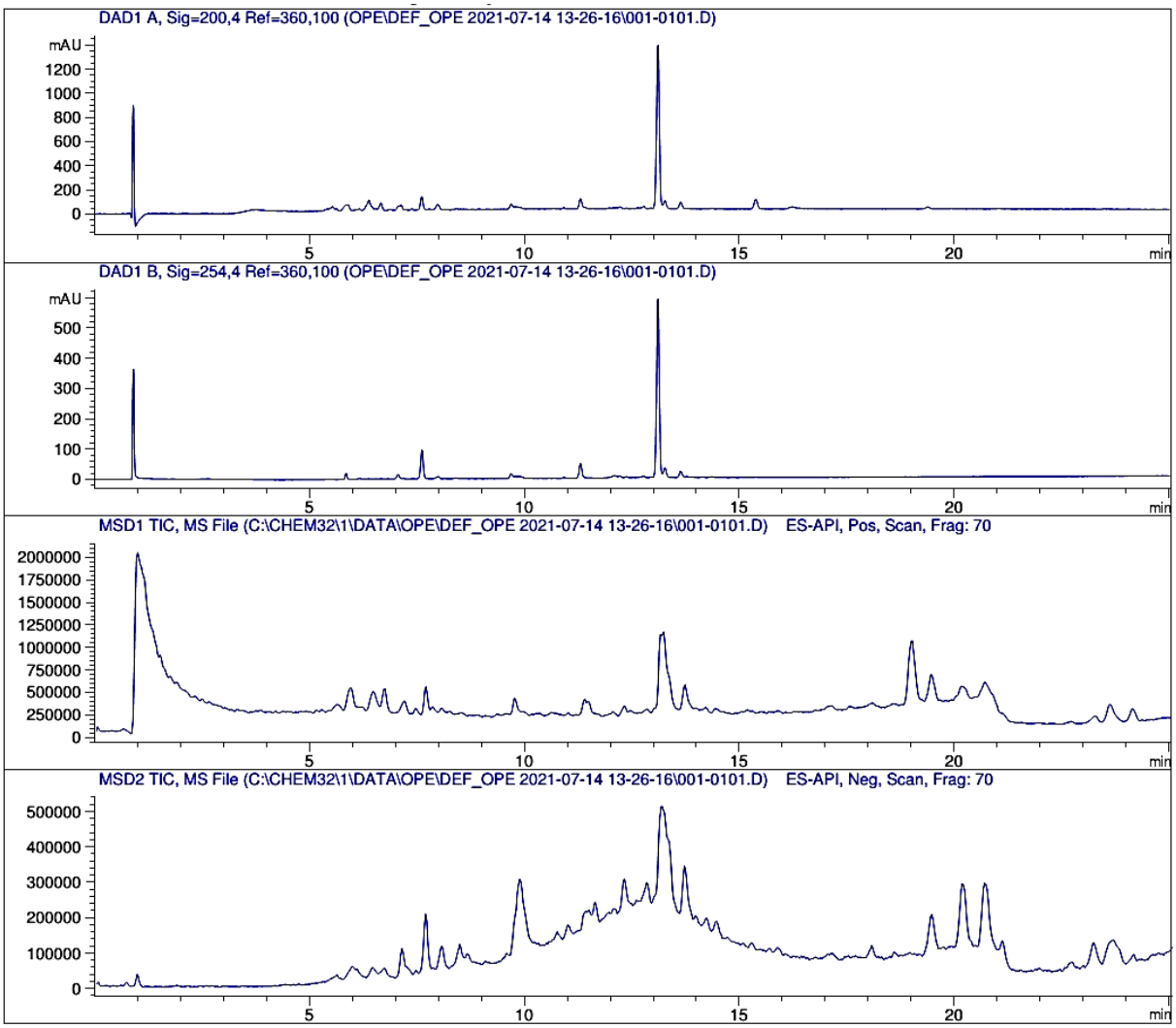
Appendix 16

LC-MS Chromatograms of Reaction Mixtures of Extracts and Probe

Appendix 16.1: LC-MS chromatogram of reaction mixture of sample A (extract of *P. fragi* A13BB inoculated broth with L-valine supplementation) and probe



Appendix 16.2: LC-MS chromatogram of reaction mixture of sample C (extract of *H. alvei* A23BA inoculated broth with L-valine supplementation) and probe



Appendix 16.3: LC-MS chromatogram of reaction mixture of sample D (extract of *H. alvei* A23BA inoculated broth with L-valine supplementation) and probe