

Mining the Genomes of Lichen
Associated Bacteria for Biosynthetic
Gene Clusters Encoding New Secondary
Metabolites

By

Peng Hou

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

Victoria University of Wellington
(2022)

Abstract

Microbial secondary metabolites have made a remarkable contribution to the therapeutics on the market today, particularly in the development of new antimicrobials. The majority of currently employed antimicrobial drugs are either directly biosynthesised by microorganisms or based on the structures of microbial metabolites. The discovery of the antibiotic penicillin in 1928 launched the golden age of antimicrobials. Since that time, bioactive natural products have helped us double our lifespans. They have greatly reduced the mortality and morbidity associated with infectious diseases and revolutionised the pharmaceutical industry. Unfortunately, the introduction of novel antimicrobial drugs into clinical practice has almost always led to resistance to those drugs, which dramatically decreases the useful lifespans of the drugs.

To address this challenge, researchers began to explore new sources for natural product screening, from the traditional phenotypic screening of soil samples to other ecological niches. The microbial assemblages in lichens have become increasingly recognised as an underestimated repertoire of bioactive compounds. This study examined the biosynthetic and chemical diversity of actinobacteria isolated from New Zealand lichen samples using ‘dry lab’ and ‘wet lab’ approaches.

More than 500 actinobacteria were isolated during the course of this study, and an in-house high-throughput NGS library preparation platform was developed to sequence these isolates. Multidimensional analyses were then conducted to explore the phylogenetic and functional divergence in our sequencing datasets. Eight biosynthetic pathways mined from our datasets were cloned and/or refactored. These constructs were then transferred to different heterologous expression hosts and screened for the production of new metabolites using a python-assisted GNPS (Global Natural Product Social Molecular Networking) platform that was developed as part of this work. Four heterologous expression systems were identified as good leads and subjected to further study, leading to the identification and characterisation of chemical entities that have not previously been reported.

This thesis validated that the New Zealand lichens are a novel genetic and chemical reservoir for natural product biosynthesis study. Analysis of over 300 genomes reconstructed from lichen sourced isolates uncovered the phylogenetic divergence and function novelty, shed light on biosynthetic dark matter, and provided a promising direction for future study.

Acknowledgement

I would like to thank my Primary Supervisor, Dr Jeremy Owen, for giving me this PhD opportunity. This is an amazing lab and fantastic team led by you; this thesis would not have been possible without your support.

Special thanks to my Secondary Supervisor, AProf Rob Keyzer, a great mentor. I am impressed with your enthusiasm for research. Thank you for the encouraging words when I needed them and the generous support you provided.

To all the people I have worked with, thank you for being so kind. In particular, to the future Dr Vincent Novak, thanks for teaching me bioinformatics. To Dr Joe Bracegirdle, thanks for all the chemical work you have done and for showing me your talent. To Dr Helen Woolner, thanks for being inspiring, generous and being my eggs on toastie provider. All the best wishes to the MaxMeta team, to future Drs Matt Storey, Manu Blank, Kelly Styles, Frances Kuang and Drs Luke Stevenson, Ethan Wooly, Channel Taylor, Chris Miller, Hungen Lai, Mark Calcott, Liwei Liu. It is a great honour to work with and learn from the MaxMeta team and Ackerley group. And thank all the approachable staff and technicians in the School of Biological Science.

To Dr Lifeng Peng, my dumpling-making supervisor, an elegant lady, thanks for all the care and dinner you and your husband gave me during my stay in Wellington.

To Prof Peiyuan Qian, thanks for your lifelong impact. Your timely reminder of staying out of my comfort zone has prompted me to explore scientific questions without any distractions. Thanks, Wei Ding, Lan Yi, Yang Yi, Xiaoxue Yang, Yanping Yu, Ken Chiang, Lisa Soo and Weipeng Zhang, for your continued sharing and caring.

To Yanan, thanks for helping me make some terribly wrong decisions, staying with me and helping me deal with my chaotic life.

For my mom and dad, thanks for just being so amazing. I can't stop thinking about how proud I am to be your daughter.

Finally, thanks to all my other friends and family for the company and support along my journey.

Table of Contents

Abstract	i
Acknowledgement	iii
Table of Contents	iv
List of Figures	viii
List of Tables	x
Abbreviations	xi
Chapter 1 General Introduction	1
1.1 Using lichens as an alternative source for isolation of actinomycetes	1
1.2 Bacterial NRPs and PKs	2
1.2.1 Bacterial NRPS	3
1.2.2 Bacterial PKS	5
1.3 Genome mining	9
1.3.1 Genome sequencing and assembly	9
1.3.2 Genome mining tools	12
1.3.3 Heterologous expression	12
1.4 Chemical analysis	14
1.4.1 Reverse-phase HPLC	14
1.4.2 Liquid Chromatography Mass Spectrometry	14
1.5 Aims of the study	16
Chapter 2 Materials and Methods	17
2.1 General materials used in this study	17
2.1.1 Strains used in this study	17
2.1.2 Vectors and constructs used in this study	19
2.1.3 Media	20
2.1.4 Antibiotics	22
2.1.5 Enzymes	22
2.2 General experimental procedures	23
2.2.1 Lichen sample collection and processing	23
2.2.2 Isolation of actinomycetes	23
2.2.3 DNA extraction and whole-genome sequencing	24
2.2.4 General molecular biology	25
2.2.5 Polymerase chain reaction	26
2.2.6 Preparation of electrocompetent cells and electroporation	27
2.3 CRISPR/Cas9-mediated TAR (Transformation associated recombination) cloning	28
2.4 Small molecule analysis and characterisation	34
2.4.1 HPLC	34
2.4.2 LCMS/MS	34
2.4.3 NMR	35
2.4.4 Bioinformatic analysis	35
Chapter 3 Genetic Investigation of the Actinomycetes Assemblage in New Zealand lichen ..	36
3.1 Introduction	36
3.2 Results	36
3.2.1 Strain isolation and genome sequencing	36
3.2.2 Taxonomic classification	37
3.2.3 Functional analysis of genome assemblies	39

3.2.4	Identification and categorisation of natural product biosynthetic gene clusters	41
3.2.5	Assignment of BGCs to Gene Cluster Families using BiG-SLiCE	44
3.2.6	"Orphan" membership BGCs from BiG-SLiCE mapping	45
3.2.7	Chemical space-guided gene cluster grouping improves genomics guided congener discovery	49
3.2.8	Chemical similarity networks inspired lichen BGCs mining	57
3.3	Conclusion and discussion	60
3.4	Methods	62
3.4.1	Sampling and isolation of actinomycetes	62
3.4.2	DNA extraction and whole-genome sequencing	62
3.4.3	Genome assembly and annotation	62
3.4.4	Phylogenetic analysis and MASH distance calculation	62
3.4.5	Cheminformatics analysis	63
3.4.6	antiSMASH analysis	63
3.4.7	Gene cluster family (GCF) analysis	64
Chapter 4	Targeted Capture and Heterologous Expression	65
4.1	Introduction	65
4.2	<i>In silico</i> description of cloned pathways	66
4.2.1	BGC004	66
4.2.2	BGC009	68
4.2.3	BGC014	72
4.2.4	BGC016	75
4.2.5	BGC027	78
4.2.6	BGC218-3	80
4.3	Refactoring pathways	83
4.3.1	004 Δ LuxR	83
4.3.2	HygE218-3	85
4.3.3	pIJEF_218-3	85
4.4	Heterologous expression of pathways and identification of the products	86
4.5	Methods	94
4.5.1	Construction of BAC004 Δ LuxR	94
4.5.2	Fermentation	95
Chapter 5	Heterologous Expression of a Puromycin-like Pathway	97
5.1	Introduction	97
5.2	Heterologous expression of BGC031	99
5.3	Characterisation of Compound 1	103
5.4	GNPS molecular networking analysis	103
5.4.1	MS/MS spectra and tentative structural assignments of Compounds 2,3,4,5	104
5.4.2	MS/MS spectrum and tentative structural assignment of Compound 6	107
5.4.3	MS/MS spectra and tentative structural assignments of Compounds 7,8	107
5.5	Conclusion and discussion	109
5.6	Methods	111
5.6.1	General experimental procedures	111
5.6.2	Pool testing (screening) and validation of BGC031	113
5.6.3	Fermentation and metabolites extraction	114

5.6.4	Construction of albus_031 PPTase overexpression strain.....	114
Chapter 6	Novel Type 2 Polyketide Identified by Heterologous Study of BGC005	116
6.1	Introduction	116
6.2	Molecular networking and antiSMASH analysis.....	116
6.3	Characterisation of compounds 9 and 10.....	121
6.4	Strategy for starter unit generation.....	122
6.5	Ring-opening oxygenase.....	124
6.6	Biosynthesis of compounds 9, 10.....	126
6.7	Bioactivity testing.....	127
6.8	Ongoing work toward finding final products of BGC005	127
6.9	Conclusion and discussion.....	128
6.10	Methods.....	129
6.10.1	General experimental procedures.....	129
6.10.2	Fermentation and metabolites extraction	130
6.10.3	Bioassays.....	131
6.10.4	Cloning of the BGC005 cluster.....	131
6.10.5	Construction of the mutant strains.....	131
Chapter 7	Conclusion and Future work	132
7.1	Genomic-driven exploration of lichen associated actinobacteria.....	132
7.2	Gene cluster and metabolites from BGC031	133
7.3	Gene cluster and metabolites of the BGC005.....	134
7.4	MbtH-like protein.....	135
7.5	Culture First vs. Genetic First.....	136
7.5.1	New lanthipeptides detected from the BGC009 producing strain	136
7.5.2	Primary metabolisms of the actinobacteria assemblages in New Zealand lichens	138
	Reference	140
	Appendix 1 Genomic DNA /strains isolated in this study and their corresponding positions on 96-well plates.....	154
	Appendix 2 An Excel workbook containing Sampling_locations, assembly_stats&gt;dbtk.classify, antiSMASH_statistics and BiG-SLiCE_statistics sheets.....	155
	Appendix 3 antiSMASH-json_parser (to generate Fig 3.3).....	156
	Appendix 4 Knowncluster_networking (to generate Fig 3.8b)	159
	Appendix 5 BiG_SLiCE- SQLite3_parser (to generate Fig 3.4)	161
	Appendix 6 BiG_SLiCE-scatter_plot (to generate Fig 3.4a)	163
	Appendix 7 BiG_SLiCE-stacked_plot (to generate Fig 3.5b).....	165
	Appendix 8 Gene fragments and Primers used in this study.....	167
	Appendix 9 GNPS_network analysis python script for positive mode	170
	Appendix 10 GNPS_network analysis python script for negative mode	173
	Appendix 11 BGC218-3 and BGC016 validated ions generated by the python analysis workflow	174
	Appendix 12 ¹H NMR of puromycin B (compound 1).....	175
	Appendix 13 COSY of puromycin B (compound 1).....	176
	Appendix 14 HMQC of puromycin B (compound 1)	177
	Appendix 15 HMBC of puromycin B (compound 1).....	178
	Appendix 16 ¹³C (150 MHz) and ¹H (600 MHz) NMR Data for puromycin B (compound 1, DMSO-d6:MeOD= 1:1)	179
	Appendix 17 Characterisation of compound 9	180
	Appendix 18 ¹H NMR of compound JB1081B (compound 9)	182

Appendix 19 COSY of JB1081B (compound 9).....	183
Appendix 20 HSQC of JB1081B (compound 9).....	184
Appendix 21 HMBC of JB1081B (compound 9).....	185
Appendix 22 ¹³ C (150 MHz) and ¹ H (600 MHz) NMR Data for JB1081B (compound 9, CDCl ₃).....	186
Appendix 23 ¹ H NMR of homorableomycin (compound 10)	187
Appendix 24 COSY of homorableomycin (compound 10).....	188
Appendix 25 HSQC of homorableomycin (compound 10)	189
Appendix 26 HMBC of homorableomycin (compound 10)	190
Appendix 27 ¹³ C (150 MHz) and ¹ H (600 MHz) NMR Data for homoreableomycin (compound 10, CDCl ₃)	191

List of Figures

Fig 1.1 Compounds identified from lichen sourced actinomycetes.....	2
Fig 1.2 Non-Ribosomal Peptide Synthesis.....	3
Fig 1.3 Selected tailoring reactions in non-ribosomal peptide synthesis	4
Fig 1.4 Biosynthesis of DEB (8).....	6
Fig 1.5 T2PKS and T3PKS	8
Fig 1.6 Overview of Illumina and Nanopore methods for whole-genome sequencing	11
Fig 1.7 Genome mining strategy closed the loop from sequencing data to chemistry	13
Fig 1.8 A simplified RP-HPLC system	14
Fig 1.9 Two types of GNPS molecular networking examined in this study.....	15
Fig 2.1 Sampling locations for the lichens used in this study.....	23
Fig 2.2 Large-Scale Low-Cost NGS Library Preparation used in this study.....	25
Fig 2.3 Design, preparation, and evaluation of sgRNAs	30
Fig 2.4 Construction of BGC218-3 pathway-specific capture vector	31
Fig 2.5 Schematic drawing of CRISPR/Cas9-mediated TAR cloning.	32
Fig 2.6 Screening for positive yeast clones	33
Fig 3.1 Whole-genome level taxonomic classification.	39
Fig 3.2 Functional analysis of the lichen sourced actinomycetes genome.	40
Fig 3.3 BGCs characterised by antiSMASH.	42
Fig 3.4 BGC-to-GCF membership assignment and calculation using BiG-SLICE.	45
Fig 3.5 Taxonomic distribution of GCF memberships and comparisons of selected BGCs.	47
Fig 3.6 BiG-SLICE vs. KnownClusterBlast analysis.....	50
Fig 3.7 Overlaying MIBiG GCFs model onto the MIBiG Tanimoto similarity network.	51
Fig 3.8 Chemical space-guided gene clusters grouping.	54
Fig 3.9 BiG-SLICE (a) and BiG-SCAPE (b) analysis of lichen-sourced BGCs (coloured in pink, oval-shaped nodes) along with MIBiG reference BGCs (coloured in blue, rectangle-shaped nodes).	57
Fig 3.10 Two selected Molecular families.	60
Fig 4.1 Bioactive-guided natural products discovery	65
Fig 4.2 BGC004 is a trans-AT type PKS BGC.....	67
Fig 4.3 Detailed PKS/NRPS prediction for BGC009.....	70
Fig 4.4 (a) Detailed NRPS prediction for BGC014 and (b) gene organisation of BGC027	73
Fig 4.5 BGC016 is a trans-AT PKS/NRPS hybrid BGC	78
Fig 4.6 (a) Detailed NRPS domain organisation for BGC218-3 and (b) gene cluster comparison of BGC218-3, daptomycin and A5414	82
Fig 4.7 BGC218-3 pathway refactoring strategies.....	86
Fig 4.8 Comparison of GNPS workflows.....	91
Fig 4.9 Three types of molecular networking analysis results as summarised in Table 4.9.	94
Fig 4.10 Flowchart for the Ctg1_627&Ctg1_628 knockout on the BAC004 in E. coli S17-1.....	95
Fig 4.11 Workflow for GNPS preparation.....	96
Fig 5.1 Structure and mechanism of action of puromycin.....	97
Fig 5.2 (a) Proposed puromycin biosynthetic pathway and (b) gene cluster comparison of BGC031 and pur	98
Fig 5.3 HPLC traces (@UV 268 nm) and UV-vis spectra used in this study.	101
Fig 5.4 Structure and key correlations for compound 1 (600 MHz, DMSO-d ₆ :MeOD= 1:1).....	103
Fig 5.5 Puromycin Cluster from the GNPS Molecular Network.	104
Fig 5.6 MS/MS spectra of compounds 1-5	107
Fig 5.7 MS/MS spectrum of compound 6	107
Fig 5.8 MS/MS spectra of compounds 7,8	108
Fig 5.9 Extracted LC-MS chromatograms of compounds 1-8.....	109
Fig 5.10 Proposed biosynthesis of BGC031.....	111

Fig 5.11 Screening and validation of BGC031	113
Fig 5.12 Construction of pIJ2449.....	115
Fig 6.1 GNPS molecular networking and selected MS profiles.	119
Fig 6.2 Chemical classes and BGC005 identified using antiSMASH.....	120
Fig 6.3 HPLC traces (@UV 430 nm) and UV-vis spectra of compounds 9 and 10	121
Fig 6.4 Structures of compounds 9 and 10.	122
Fig 6.5 Chemical evidence showed that there were two starter unit generation strategies in BGC005.....	123
Fig 6.6 Current knowledge about the enzymes involved in the biosynthesis of different types of atypical angucycline skeletons.....	125
Fig 6.7 BGC comparison and phylogenetic analysis of the ring-opening enzymes	125
Fig 6.8 The detection of compounds 9 and 10 in different strains.	126
Fig 6.9 Proposed biosynthesis of compounds 9 and 10.	127
Fig 6.10 Metabolites from del14_005 and del14_005ΔAbXO.....	128
Fig 7.1 BiG-SLiCE comparison analysis of four environmental datasets.	133
Fig 7.2 Correlation between MtbH-like proteins and NRPS regions of each genome in our datasets	135
Fig 7.3 A GNPS molecular network identified in the BGC009 native producing strain.....	137
Fig 7.4 A candidate lanthipeptide BGC and its proposed structure.....	138
Fig 7.5 Two common metabolites found during fermentation. The structures of these two metabolites were deduced from MS/MS.....	139
Fig 7.6 Phylogenetic distribution of heme related KEGG orthology.	139

List of Tables

Table 2.1 Strains used in this study.....	17
Table 2.2 Vectors and constructs used in this study.....	19
Table 2.3 Media used in this study	20
Table 2.4 Trace elements solution used in this study.....	21
Table 2.5 Antibiotics used in this study.....	22
Table 2.6 General enzymes used in this study.....	22
Table 2.7 Touchdown PCR programme.....	26
Table 2.8 2-step PCR programme	26
Table 2.9 Universal sgRNA primers used in this study.....	29
Table 2.10 HPLC protocol and parameters 1	34
Table 2.11 HPLC protocol and parameters 2	34
Table 2.12 LCMS protocol and parameters.....	35
Table 3.1 Distribution of non-hybrid biosynthetic cluster types for BGCs present in this study....	43
Table 3.2 Predicted functions and sequence alignment of the selected genes in the Li3d-B6_r1c5 region.	48
Table 3.3 Predicted functions of the selected genes in the JGO2c-H7_r10c1 region and A domain amino acid specificity prediction.	48
Table 3.4 Predicted functions, sequence alignment and domain organisation of the selected genes in the Li3d-F5_r28c1 region.....	60
Table 4.1 Predicted functions of the genes on BGC004.....	68
Table 4.2 Predicted functions of the genes on BGC009.....	71
Table 4.3 Predicted functions of the genes on BGC014.....	74
Table 4.4 Predicted functions of the genes on BGC016.....	76
Table 4.5 Predicted functions of the genes on BGC027.....	79
Table 4.6 Stachelhaus analysis of the A domains on BGC218-3, A5145 and Daptomycin.....	81
Table 4.7 Predicted functions of the genes on BGC218-3	83
Table 4.8 Summary of cloned pathways (lengths, chemical classes) and heterologous expression systems examined in this thesis.....	87
Table 4.9 GNPS analysis results of <i>S. albus</i> Del14 related metabolites.	92
Table 5.1 Predicted functions of the genes on BGC031.....	102
Table 5.2 HPLC protocol and parameters	112
Table 6.1 Deduced functions of the genes on BGC005.....	120
Table 6.2 Minimum inhibitory concentration (MIC) values for isolated compound 9 against a range of bacteria.....	127

Abbreviations

A domain	adenylation domain
AAI	average amino acid sequence identity
ACP	acyl carrier protein
ANI	average nucleotide identity
antiSMASH	antibiotics and secondary metabolite analysis shell
<i>apaR</i>	apramycin resistance
APCI	Atmospheric pressure chemical ionization
AT	acyltransferase
BAC	bacterial artificial chromosome
BGC	biosynthetic gene clusters
BiG-SCAPE	Biosynthetic Gene Similarity Clustering and Prospecting Engine
BiG-SLICE	Biosynthetic Genes Super-Linear Clustering Engine
C domain	condensation domain
CDA	Ca ²⁺ -dependent cyclic lipodepsipeptides
CDPS	tRNA-dependent cyclodipeptide synthases
CID	collision-induced dissociation
CLF	chain length factor
COSY	Correlated Spectroscopy
Cy domain	cyclisation domain
CYCs	cyclases/aromatases
DAD	Diode-Array Detection
DEB	6-deoxyerythronolide B
DH	dehydratase
E domain	epimerisation
ER	enoyl-reductase
GCFs	gene cluster families
gDNA	Genomic DNA
GNPS	Global Natural Products Social Molecular Networking
GPAs	glycopeptide antibiotics
HILIC	Hydrophilic interaction chromatography
HMBC	Heteronuclear Multiple Bond Correlation
HMQC	Heteronuclear Single Quantum Coherence
HPLC	High Performance Liquid Chromatography
HRESIMS	high resolution electrospray ionisation mass spectroscopy
<i>hygR</i>	hygromycin resistance
KR	ketoreductase
KS	ketosynthase
LCMS	Liquid Chromatography Mass Spectrometry
MFs	molecular families
MIBiG	Minimum Information about a Biosynthetic Gene cluster
MS	mass spectrometry
MS/MS	tandem mass spectrometry

MT domain	methyltransferase
NAGGN	N-acetylglutaminyglutamine amide
NGS	Next-Generation Sequencing
NMR	Nuclear Magnetic Resonance
NRPs	non-ribosomal peptides
NRPS	non-ribosomal peptide synthetase
Ox domain	oxidation domain
pHMMs	profile Hidden Markov Models
PKs	polyketides
PKS	polyketide synthases
PPtase	phosphopantetheinyl transferase
R domain	reduction domain
RiPPs	ribosomally synthesized and post-translationally modified peptides
sgRNA	single guided RNA
T domain	thiolation domain
T1PKS	type I polyketide synthase
T2PKS	type II polyketide synthase
T3PKS	type III polyketide synthase
TAR cloning	Transformation associated recombination cloning
TE domain	thioesterase domain
YAC	yeast artificial chromosome

Chapter 1 General Introduction

1.1 Using lichens as an alternative source for isolation of actinomycetes

For decades, actinomycetes have served as one of the leading sources for discovering new antibiotics.¹ Through conventional bioassay-guided isolation, a substantial number of drugs have been successfully identified and introduced to the drug market and are still in use today.^{1,2} A good example is the soil screening programme conducted by Eli Lilly. This research programme screened over 200,000 strains and discovered a few renowned marketed antibiotics, including vancomycin, daptomycin, and erythromycin, via isolation and laboratory cultivation of their corresponding actinomycetes producers.³ Over time, bioprospecting the soil samples frequently resulted in the rediscovery of natural products, suggesting scientists' opportunities had dwindled. Recent developments in bacterial natural product discovery converted to some out-of-the-box approaches, including exploring alternative environmental niches, for example, isolation of actinomycetes from mangrove⁴, deep sea⁵, bats⁶ and lichens⁷.

Lichens are self-supporting symbioses, share mutualistic partnerships between heterotrophic mycobiont and autotrophic photobiont, and have internal bacterial communities which have attracted considerable research interests.⁷⁻⁹ Several systematic explorations of actinobacteria assemblages in lichens have uncovered the taxonomic diversity and abundant biosynthetic potential in these microbial communities, which have shed light on using lichens as an alternative source for mining rare actinomycetes and their associated novel metabolites.^{7,10-12} Actinomycetes isolated from lichens have been reported to produce new congeners of biomedically relevant natural products. For example, skyllamycin D (Fig 1.1①), identified from the New Zealand lichen *Pseudocyphellaria dissimilis* associated *Streptomyces anulatus* VUW1, exhibited more potent bioactivity against *Bacillus subtilis* E168 compared to previously reported skyllamycins congener (Fig 1.1②).¹³ A

streptomyces isolated from a British Columbia lichen *Cladonia uncialis* is the producing strain of uncialamycin (Fig 1.1③). This enediyne antibiotic displayed broad-spectrum bioactivity.¹⁴

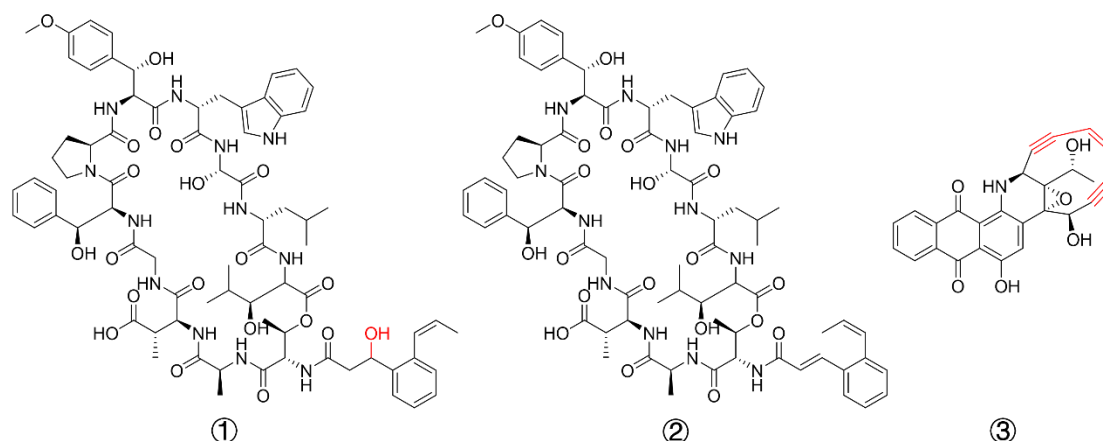


Fig 1.1 Compounds identified from lichen sourced actinomycetes.

Skellamycin D (①) is a new congener of the previously reported skellamycin A (②), where the corresponding β -keto group is reduced. Compounds bear enediyne moiety usually display very potent antitumor activity. Uncialamycin (③) is an enediyne antibiotic isolated from lichen sourced *Streptomyces*.

While New Zealand constitutes only 0.18% of the world's landmass, it is home to approximately 10% of the world's lichen species.⁹ In this study, actinomycetes from New Zealand lichens were obtained and examined. Different data mining tools were used to develop a systematic method to evaluate the abundance of taxonomic, genomic, and/or metabolomic of New Zealand lichens sourced actinomycetes. Molecular biology platforms were then applied to several biosynthetic gene clusters selected from the big data analysis. We completed the initial investigation of the biosynthetic dark matter hidden in the New Zealand lichen sourced actinomycetes by combining the 'dry lab' and 'wet lab' approaches.

1.2 Bacterial NRPs and PKs

The majority of the BGCs (biosynthetic gene clusters) discussed and studied in this work belong to the non-ribosomal peptide (NRP) and polyketide (PK) chemical classes. These two classes are the source of some of the most important drugs currently used in clinical practice, including Cortisporin-TC (Polymyxins E1 and E2, cyclic polypeptide),

Neosporin (Bacitracin, cyclic polypeptide), Adriamycin (Doxorubicin, anthracycline).¹⁵

1.2.1 Bacterial NRPS

NRPs are assembled from mega-synthetases called non-ribosomal peptide synthetases (NRPS) organised in a gene cluster. NRPS can be broken down into modules, where each module is responsible for incorporating a specific amino acid into the growing peptide chain (Fig 1.2a).¹⁶ Modules can be further broken down into discrete domains, each of which catalyses a single reaction. The core domains of NRP biosynthesis are described below.

The first amino acid is selected and activated by the A (adenylation) domain with ATP to form aminoacyl-AMP, which is then transferred to the adjacent T (thiolation) domain (Fig 1.2b). The C (condensation) domain in the next module is involved with peptide chain elongation, catalysing the formation of the amide bond between T1-S~aa1 and downstream T2-S~aa2 (Fig 1.2c). The final module on the NRPS assembly line usually contains a TE (thioesterase) domain, which has hydrolytic and/or macrocyclisation activity and is responsible for the maturation and release of the peptide chain (Fig 1.2a).^{17,18} Prior to NRPs biosynthesis, PPTase (4'-phosphopantetheinyl transferase) is required to convert the T (thiolation) domain from the inactive apo-form to the active holo-form.^{17,19}

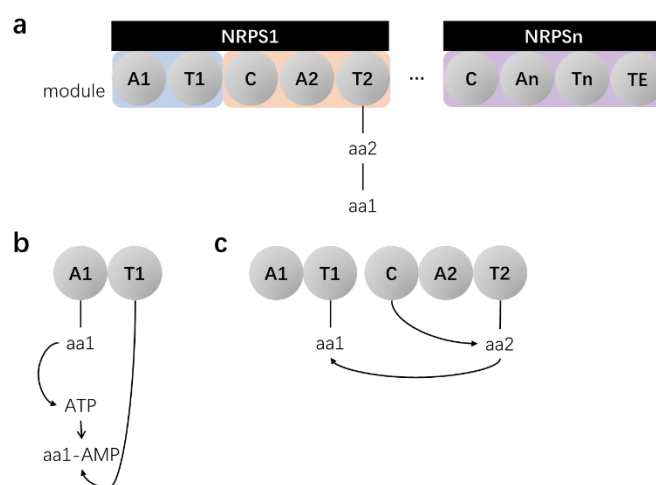


Fig 1.2 Non-Ribosomal Peptide Synthesis

- Modular organisation of NRPS.
- Amino acid activated by the A domain is covalently bound to the T domain.
- Peptide chain elongation catalysed by C domain

Additional tailoring processes can occur during and post NRP biosynthesis. Complestatin (Fig 1.3④) is an extensively modified NRP, which incorporates non-proteinogenic D-amino acids via an E (epimerisation) domain (Fig 1.3, gene and chemical moiety coloured in green). Amino acids are methylated, halogenated via the MT (methyltransferase) domain (Fig 1.3, gene and chemical moiety coloured in cyan) and halogenase (Fig 1.3, gene and chemical moieties coloured in pink). Oxidative crosslinks are formed by Cytochrome P450-like oxygenases (Fig 1.3, genes and chemical moieties coloured in orange).

Other modifications, such as heterocyclisation through Cy (cyclisation) domain, and further oxidation through Ox (oxidation) domain, or reduction through R (reduction) domain on the heterocyclic ring, can be found on compounds such as vibriobactin (Fig 1.3⑤), epothilone (Fig 1.3⑥), pyochelin (Fig 1.3⑦).¹⁸

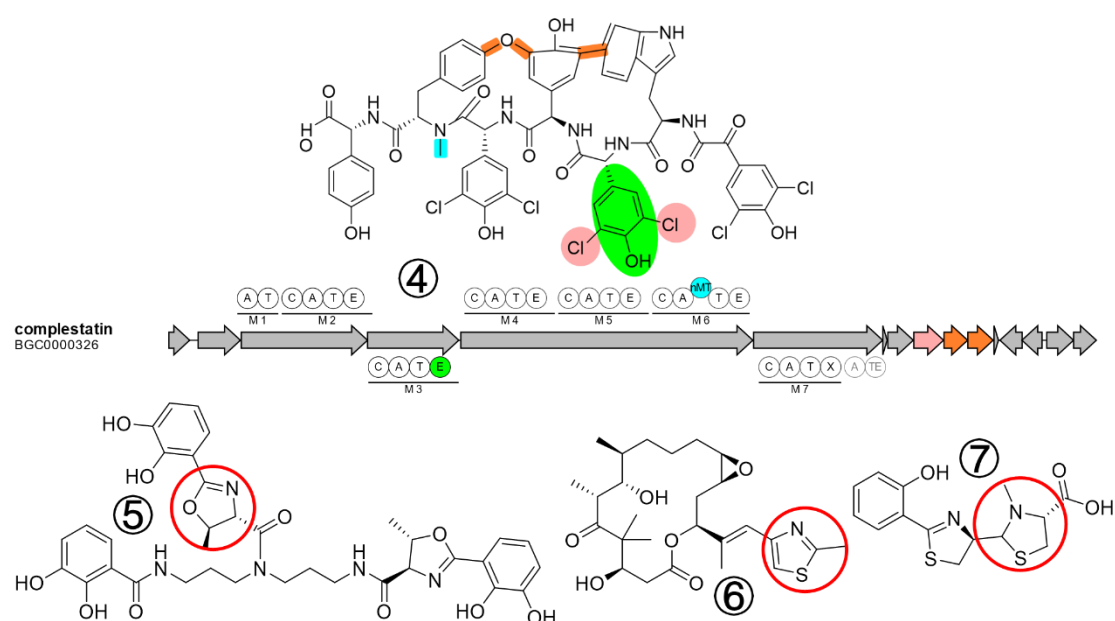


Fig 1.3 Selected tailoring reactions in non-ribosomal peptide synthesis

Structural elements (on complestatin ④) introduced here are colour coded to match the corresponding genetic elements in the assembly line. Module three (M3) contains an epimerisation domain (coloured in green) that catalyses the conversion from L- to D- amino acids. Module six contains an N-methyl transferase domain (highlighted in cyan). Chlorines are introduced by a halogenase (gene coloured in pink), and oxidative crosslinks are introduced by genes highlighted in orange. Cy domain catalyses the cyclisation of threonine to form the oxazoline on vibriobactin (⑤). Ox domain mediates the thiazole formation on the epothilone (⑥). R domain catalyses the formation of the thiazolidine ring on pyochelin (⑦).

1.2.2 Bacterial PKS

Polyketides are derived from acyl CoA precursors and are synthesised by polyketide synthases (PKS). Similar to NRPSs, PKSs are organised into gene clusters that usually contain all of the necessary genes for biosynthesis of a given compound, as well as genes for exporting and conferring resistance. PKS can be categorised into three types based on their mechanism and organisation.

T1PKS (Type 1 polyketide synthase) are modular polyketide synthases, where each module typically governs the incorporation of two or more carbons into the final polyketide skeleton. As with NRPs systems, modules are composed of discrete domains, each with a specific catalytic activity.

Biosynthesis of T1PKS is typically initiated by a starting module that usually has an AT-ACP module structure (Fig 1.4). Various acyl groups can be linked to ACP in the starting module by the AT (acyltransferase). Like the biosynthesis of DEB (6-deoxyerythronolide B, Fig 1.4⑧), propionyl CoA is introduced to the assembly line by the starting module (Fig 1.4). The starter unit is transferred from the ACP domain to the first KS1 domain catalysed by KS1 (ketosynthase), and the extender unit methyl malonyl CoA is incorporated into the ACP1 domain via the AT1 domain. The Claisen-type C-C bond is formed between the ACP1-tethered methyl malonyl CoA and the KS1-bound propionyl CoA, resulting in a free KS1 domain and a growing polyketide chain bound to ACP1. Each elongation module is followed by this cycle. The ketone groups can be further modified via optional domains such as KR (ketoreductase), DH (dehydratase), and ER (enoyl-reductase). The KR domain in module 2 reduces the β -ketone group introduced by module 1 to a β -hydroxy group. In module 4, the KR domain reduces the β -keto group incorporated through module 3 to a β -hydroxy group, the DH domain catalyses the α , β -unsaturated bond formation through a dehydration reaction, and this α - β -double-bond is further reduced to a single bond by the ER domain. Finally, the TE domain adjacent to module 6 releases and cyclises the final product from the pathway via lactonisation (Fig 1.4).^{17,21–23}

AT can also act as one or more standalone proteins (termed *trans*-AT) rather than presenting on every module as a domain.²⁴

Unlike the noniterative modular T1PKS, iterative T1PKS, previously mainly found in fungi, utilises the same elongation monomodule repeatedly during chain length elongation. A recent study reveals the potential wide distribution of iterative T1PKS BGCs in streptomycetes.^{25,26}

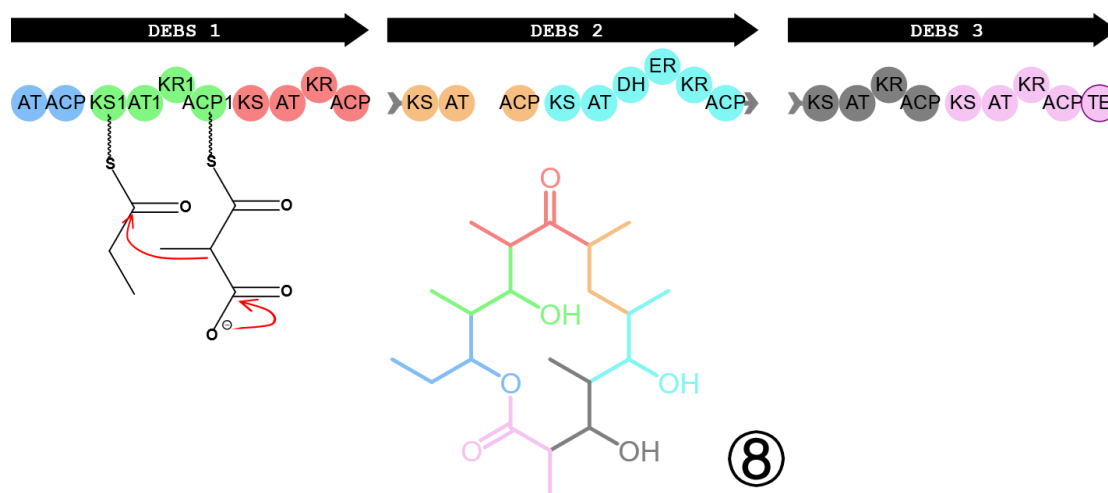


Fig 1.4 Biosynthesis of DEB (8)

DEB is derived from one propionyl-CoA and six methylmalonyl-CoA. Optional tailoring domains can be found in modules 1,2,4,5, and 6. TE domain hydrolyses and cyclises the polyketide chain. DEB structure is colour coded to match the corresponding introducing modules on the assembly line.

T2PKS (Type II polyketide synthase) are iterative biosynthetic systems in which a single set of enzymatic domains catalyses repeated decarboxylative condensation of extender units, usually malonyl-CoA, to yield a linear β -ketone chain. The nascent β -ketone chain is then subjected to reduction, cyclisation, aromatisation, and subsequent derivatisations to yield a vast array of polyaromatic compounds. Biosynthesis of aromatic polyketides begins with the priming of the starter unit onto an ACP, which is usually accomplished by an AT. The starter unit is transferred to and decarboxylated by the ketosynthase complex ($KS\alpha/\beta$). The complex then catalyses the formation of Claisen-type C–C bonds with the incoming extender unit (usually malonyl-ACP) iteratively. The minimal PKS system (ACP and $KS\alpha/\beta$ heterodimer) generates poly- β -ketone backbones, and it is generally believed that the catalysation cycles (backbone length, n) are controlled by chain length factor (CLF, $KS\beta$). The nascent chain is then reduced by KR and cyclised by CYCs (cyclases/aromatases). Subsequent post-modification processes such as oxidation, glycosylation, and methylation give rise to the diverse structures of type II polyketides.^{26–28} The polyaromatic compounds can be subtyped based on the backbone length and cyclisation pattern (Fig 1.5a).

Typical ACP-independent **T3PKS (type III polyketide synthases)** use homodimeric synthases that act directly on the acyl CoA substrates, catalysing the condensation reaction iteratively (Fig 1.5b).^{22,29,30}

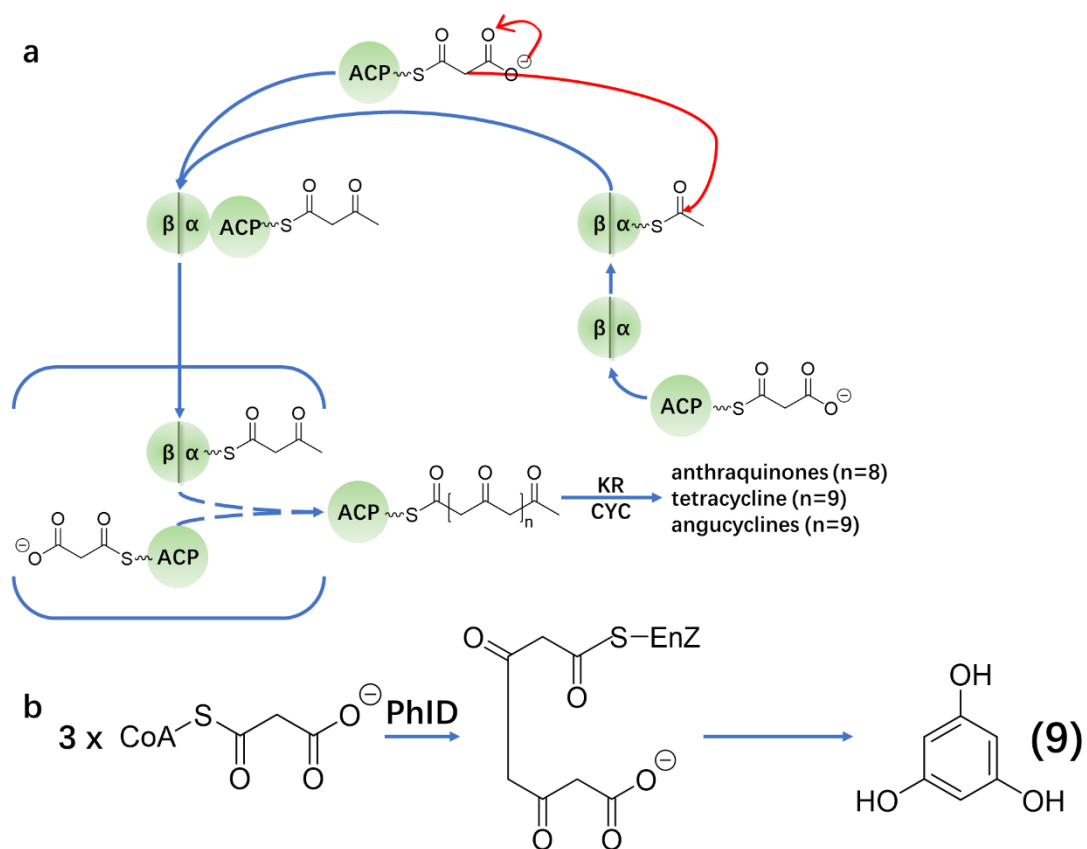


Fig 1.5 T2PKS and T3PKS

- T2PKS biosynthesis is an iterative process that relies on the minimal PKS system (ACP and KS α / β heterodimer)
- The biosynthesis of phloroglucinol (9) is mediated by a T3PKS PhID.

1.3 Genome mining

Many natural secondary metabolites are encoded by a group of genes clustered on a particular genome locus termed BGC (biosynthetic gene cluster).³¹ Genome mining is the process of discovering natural products by first identifying the BGCs that direct their production.³² There are a variety of different genome mining approaches that have found successful applications in modern natural product discovery, and innovations in this space are continuing to grow. The following section outlines the key enabling technologies as well as fundamental approaches in genome mining.

1.3.1 Genome sequencing and assembly

Preparing NGS (Next-Generation Sequencing) libraries is vital to the success of NGS. Different methods/kits are available to meet various needs. A common method for library preparation is tagmentation-based (Fig 1.6①left), in which the template (genomic DNA in this study) is randomly sheared via Tn5 transposons.³³ The adapters that contain complementary sequences to the flow cell and unique index sequences (also known as barcodes) are added to both ends of fragmented DNA. The fragments are enriched using PCR before being applied to the flow cell (Fig 1.6②left). Once bound to the flow cell, each fragment is then amplified to a clonal population through the bridge amplification PCR procedure (cluster generation, Fig 1.6③left). Four fluorescent-tagged, reversible terminator-bound dNTPs are incorporated and bound to the DNA template strand via natural complementarity on each sequencing cycle. Fluorescent signals indicating which nucleotides have been added to the reaction are recorded. A base-calling process is then carried out to identify nucleotides (Fig 1.6③left).³⁴ Followed by sequencing, *de novo* assembly is the process of assembling large numbers of overlapping sequencing reads into contiguous sequences in the absence of a reference genome.³⁵ In the present study, we mainly utilised SPAdes, a de Bruijn graph-based assembler.^{36,37}

In contrast to short-read NGS, Nanopore is a long-read DNA sequencing technique. Nanopore sequencing library construction is also commonly transposase-based (Fig 1.6①right), a simple two-step “cleave and plug” protocol that requires less bench time (Fig 1.6②right).³⁸ Nanopore sequencing flow cells contain an array of tiny protein pores called nanopores. When DNA bases pass through a nanopore, a characteristic current is produced and decoded simultaneously through the base calling process (Fig 1.6③right).³⁹

BGC identification and prediction require fairly long assemblies, however the existing NGS sequencing and SPAdes approaches often result in fragmented assemblies, thus preventing the assembly of larger BGCs into a single contig efficiently. Hybrid assembly using long- and short-reads from Illumina and Oxford Nanopore sequencing datasets is becoming a popular strategy to improve the accuracy and completeness of BGC reconstruction.^{13,40}

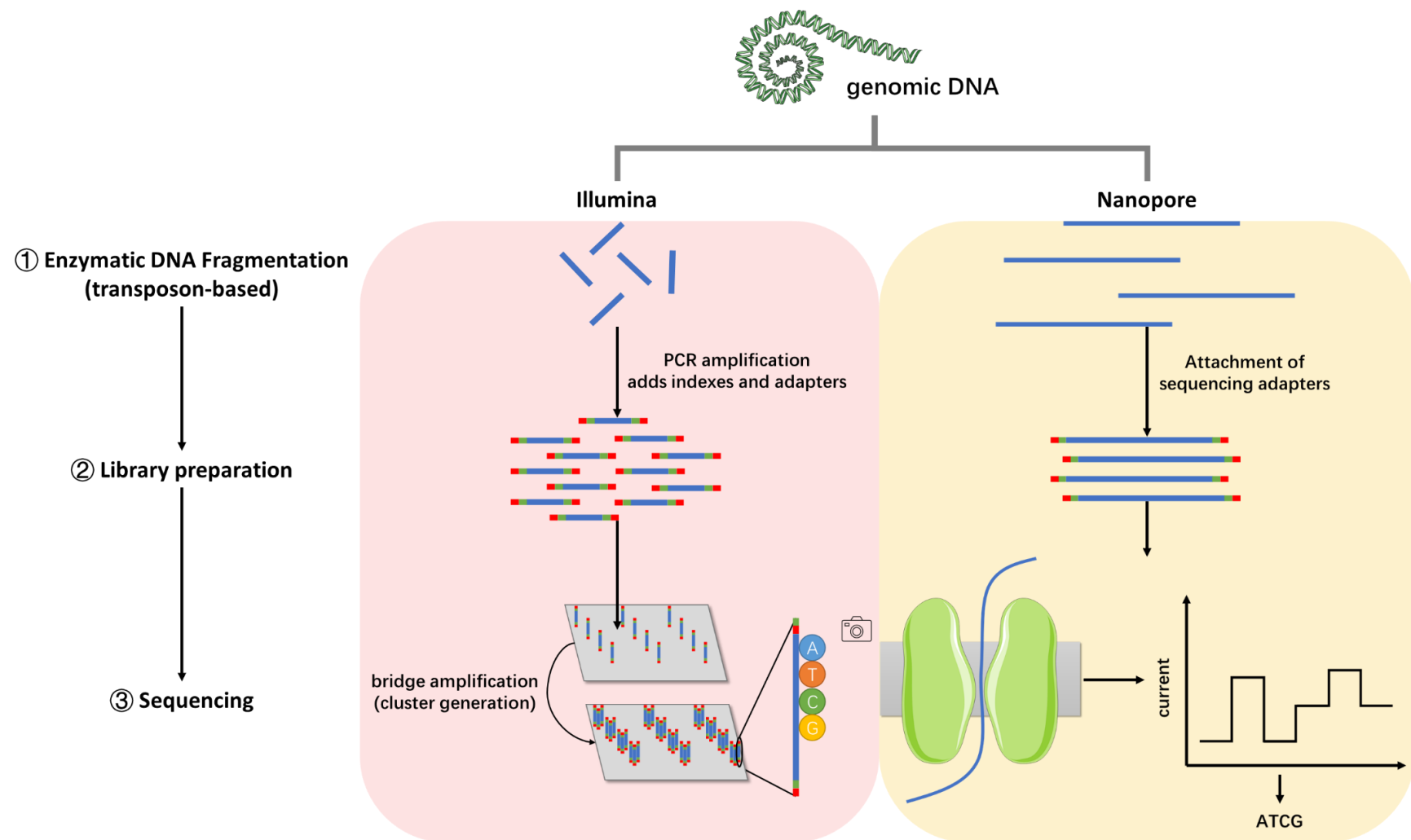


Fig 1.6 Overview of Illumina and Nanopore methods for whole-genome sequencing

1.3.2 Genome mining tools

Although secondary metabolites are well known for their striking structural diversities, the biosynthesis machinery responsible for building and tailoring them is usually derived from relatively conserved enzyme families.⁴¹ This allows for the possibility of detecting particular classes of BGCs based on the co-localisation of conserved enzymatic domains and is the basis of the most commonly employed genome mining tool antiSMASH.⁴² antiSMASH is the gold standard *in silico* tool for mining of BGCs from sequenced genomes, and version 5⁴³ has been extensively used in this study. During antiSMASH analysis, a series of steps are carried out under the control of an automated, intuitive wrapper function. First, genes are annotated, and amino acid sequences are extracted. These are then searched using pHMMs (profile Hidden Markov Models), and gene cluster types are identified using particular combinations of core signature gene pHMMs. Functional annotations are also provided for accessory genes within the gene clusters. Substrate specificity prediction is performed for NRPS and PKS domains based on the active sites and consensus motifs, using algorithms and tools such as Minowa⁴⁴ and NRPSpredictor2⁴⁵ embedded in antiSMASH. There are a number of additional modules, such as KnownClusterBlast, which carries out gene cluster comparative analysis, aligning the identified BGCs against reference BGCs from the MIBiG⁴⁶ database to identify potential known and/or congener clusters. However, we need to note that antiSMASH is a rule-based genome mining tool⁴⁷, which means the identification of BGCs is heavily limited to the current knowledge of biosynthetic machinery. Therefore, antiSMASH is not capable of identifying entirely new BGC classes, the real dark matter.

1.3.3 Heterologous expression

Guided by genome mining, targeted heterologous expression of identified BGCs represents an effective strategy in natural product discovery and provides an efficient way to connect sequencing to chemistry (Fig 1.7). Heterologous expression begins with the transfer of identified BGCs to heterologous hosts. Various DNA cloning platforms are available for BGC capture, including TAR cloning⁴⁸, Gibson assembly⁴⁹ and Golden Gate cloning⁵⁰ (Fig 1.7①).

The constructs are then integrated into heterologous expression hosts such as *Streptomyces albus*^{51,52}, *Streptomyces lividans*⁵³ and *Streptomyces coelicolor*⁵⁴, which have been characterised and engineered for expressing actinomycetes related BGCs (Fig 1.7②).

Following the fermentation process, the resulting metabolites can be profiled using various metabolomics detection and analysis techniques (Fig 1.7⑥, discussed in 1.4). Many molecular platforms have been developed for refactoring biosynthetic gene clusters if they remain silent or poorly expressed under laboratory conditions in the heterologous expression systems. Examples included adding synthetic promoters across a BGC⁵⁵ or upstream of targeted genes⁵⁶ (Fig 1.7③), directly deleting the genes that suppress the expression of the gene cluster⁵⁷ (Fig 1.7④), or introducing multicopy transcription factor decoy to sequester the repressors⁵⁸ (Fig 1.7⑤). There is no one-for-all procedure to activate a BGC, thus activating silent BGCs or optimising the yields of poorly expressed pathways remains a challenging task requiring several build-test-learn cycles. In this thesis, several approaches have been adopted to engineer or refactor some of the selected pathways, which will be discussed in Chapter 4.

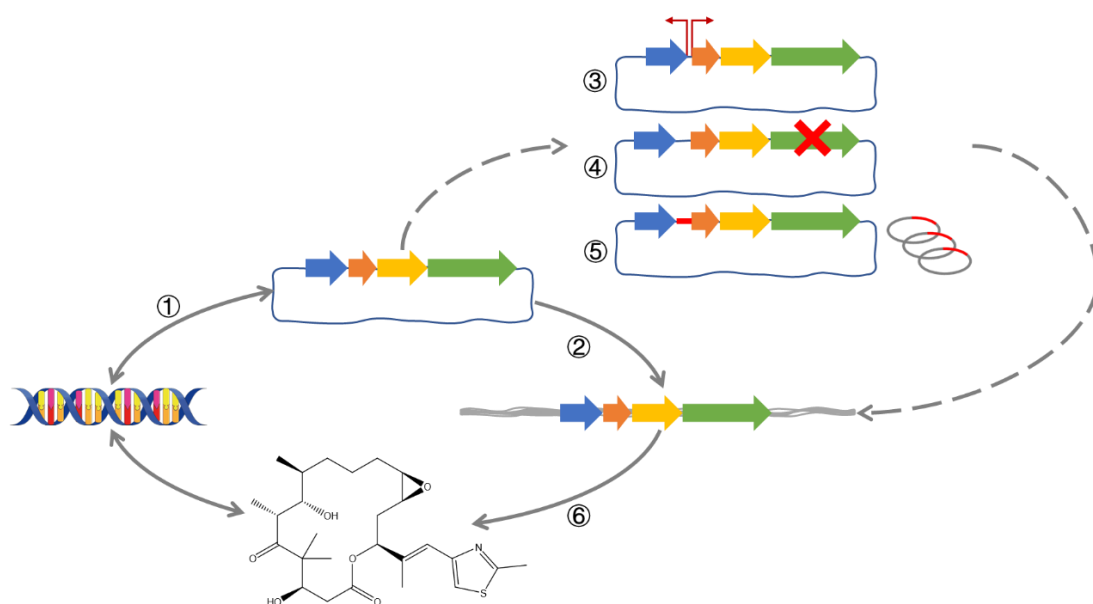


Fig 1.7 Genome mining strategy closed the loop from sequencing data to chemistry. Guided by genome mining, selected BGCs were cloned (①), refactored (③-⑤) and transferred to various heterologous expression systems (②). The heterologous expression system is the bridge between BGCs and metabolites (⑥).

1.4 Chemical analysis

1.4.1 Reverse-phase HPLC

RP-HPLC (Reversed-phase HPLC) is a popular method in natural product chemistry. The RP-HPLC system incorporates a polar mobile phase (usually H₂O/Acetonitrile or H₂O/MeOH) and a nonpolar stationary phase (a C18 column in this thesis, Fig 1.8, coloured in orange). Polar compounds in the sample mixture (Fig 1.8, coloured in blue) have few interactions with the C18 modified silica, and those components will pass through the stationary phase faster and elute earlier. On the other hand, nonpolar compounds will bind to the C18 chain, which results in a slower movement in the stationary phase and a longer retention time. A detector (UV-Vis in this study) is usually used along to detect components in the sample mixture. The components are then recorded and displayed on the recorder if they have chromophores (Fig 1.8).⁵⁹

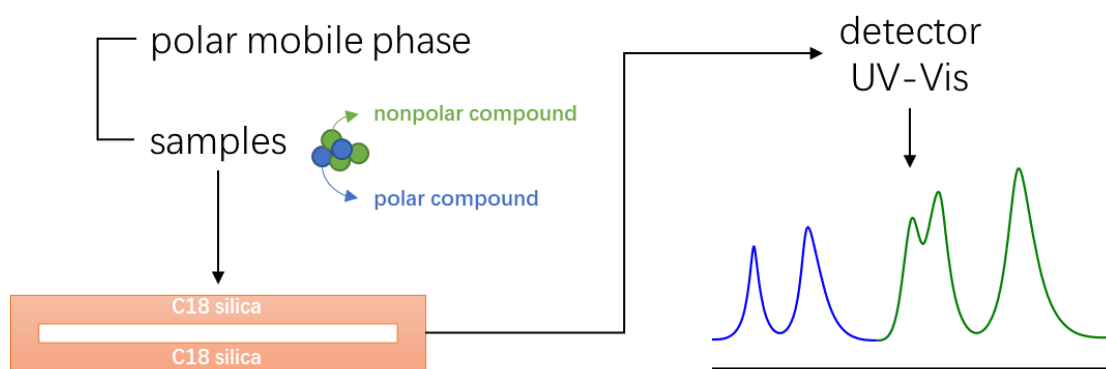


Fig 1.8 A simplified RP-HPLC system

1.4.2 Liquid Chromatography Mass Spectrometry

In Liquid Chromatography Mass Spectrometry (LCMS), an HPLC system is coupled with a detector capable of mass analysis. Compounds eluted from the LC system are channelled into an ionisation source, where they are desolvated and ionised. Following ionisation, compounds are accelerated through an electromagnetic vacuum field and separated based on their mass-to-charge (m/z) ratios. The detected ions are amplified and recorded as m/z ratios together with their relative abundance. Different dissociation methods such as CID

(collision-induced dissociation) can be used to further fragment selected ions. The resultant fragments can be recorded by MS again. By using MS/MS (tandem mass spectrometry), we can obtain additional information about the structural features of a particular ion.^{60–62}

GNPS (Global Natural Products Social Molecular Networking) is a tandem mass-based molecular networking analysis tool. GNPS is based on the concept that similar fragmentation patterns imply similar structures. Thus, compounds (nodes) with similar fragmentation patterns can be grouped into the same molecular family. GNPS molecular networks consist of nodes that represent MS/MS spectra and molecular ions; edges are connected when two nodes exhibit similar (defined by cosine score) MS/MS patterns.⁶³

GNPS is a powerful tool for rapidly annotating complex metabolite profiles. Using a GNPS facilitated heterologous expression strategy, we are particularly interested in two types of clusters in the molecular networks of heterologous expression system:

- New congeners present as extended nodes on a cluster consisting of known product scaffolds (Fig 1.9a, discussed in Chapter 5).¹³
- Novel compounds that form unique clusters, unconnected by the metabolites from other (control) groups (Fig 1.9b, discussed in Chapter 6), as demonstrated by Moore and his co-workers^{64,65}

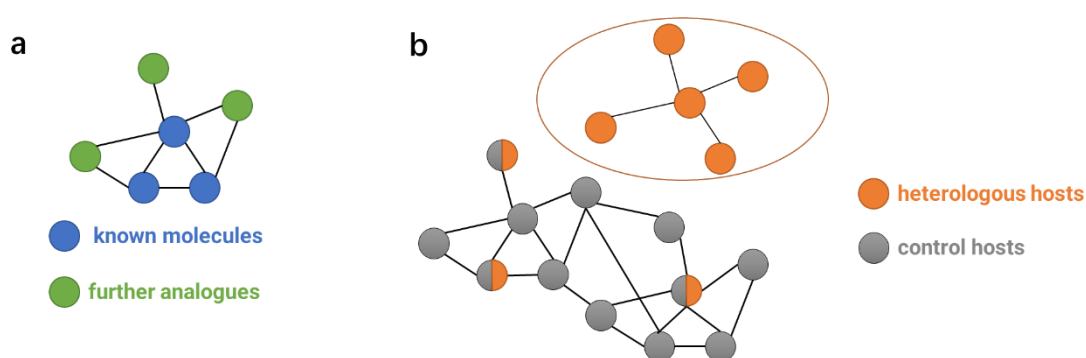


Fig 1.9 Two types of GNPS molecular networking examined in this study

1.5 Aims of the study

The overall aim of this thesis was to investigate the biosynthetic and natural product diversity arising from culturable actinobacteria assemblages present in New Zealand lichens. This overarching aim was further broken down into the following specific sub-aims:

1. To isolate actinomycetes from New Zealand lichens and generate sequencing libraries for the isolates (Chapter 3).
2. To investigate the biosynthetic potential of the sequenced isolates and their potential for producing novel natural products (Chapter 3).
3. To capture selected BGCs, deliver these to heterologous expression hosts, and conduct the metabolite profiling (Chapter 4).
4. To characterise the metabolites from successful heterologous expression systems (Chapters 5&6).

Chapter 2 Materials and Methods

2.1 General materials used in this study

2.1.1 Strains used in this study

Table 2.1 Strains used in this study

Strain	Description	Reference
<i>Escherichia coli</i> EC100	General cloning host	Epicentre
<i>Escherichia coli</i> S17-1	Donor strain for conjugation from <i>E. coli</i> to <i>Streptomyces</i>	66
<i>Saccharomyces cerevisiae</i> BY4727 Δ NHEJ	Host strain for TAR cloning	67
<i>Streptomyces albus</i> J1074	Heterologous host	NCBI: txid457425
<i>Streptomyces albus</i> Del14	Heterologous host	51
<i>Streptomyces lividans</i> TK24	Heterologous host	NCBI: txid457428
<i>Escherichia coli</i> (TolC-deficient)	For antimicrobial assay	DSM 104619
<i>Bacillus subtilis</i> 168	For antimicrobial assay	68
<i>Staphylococcus aureus</i>	For antimicrobial assay	ATCC 25923
del14_004	<i>S. albus</i> Del14 hosting BAC DNA of the 004 gene cluster	This study (Chapter 4.2.1)
del14_004 Δ LuxR	del14_004 with deletion of <i>ctg1_627</i> and <i>ctg1_628</i> gene	This study (Chapter 4.3.1)
del14_005	<i>S. albus</i> Del14 hosting BAC DNA of the 005 gene cluster	This study (Chapter 6)
del14_005 Δ PKS	del14_005 with deletion of <i>ctg4_380</i> , <i>ctg4_381</i> and <i>ctg4_382</i> genes	This study (Chapter 6)
del14_005 Δ AbXO	del14_005 with deletion of <i>ctg4_395</i> gene	This study (Chapter 6)
del14_009	<i>S. albus</i> Del14 hosting BAC DNA of the 009 gene cluster	This study (Chapter 4.2.2)
del14_014	<i>S. albus</i> Del14 hosting BAC DNA of the 014 gene cluster	This study (Chapter 4.2.3)
del14_016	<i>S. albus</i> Del14 hosting BAC DNA of the 016 gene cluster	This study (Chapter 4.2.4)
albus_031	<i>S. albus</i> J1074 hosting BAC DNA of the 031 gene cluster	This study (Chapter 5)
albus_031/pIJ2449	albus_031 PPTase overexpression strain	This study (Chapter 5)
del14_031	<i>S. albus</i> Del14 hosting BAC DNA of the 031 gene cluster	This study (Chapter 5)

del14_218-3	<i>S. albus</i> Del14 hosting BAC DNA of the 218-3 gene cluster	This study (Chapter 4.2.6)
del14_HygE218_3	<i>S. albus</i> Del14 hosting BAC DNA of the 218-3 gene cluster under the control of the <i>ermE</i> * promoter	This study (Chapter 4.3.2)
del14_pIJEF218_3	del14_218-3 lipidation module overexpression strain	This study (Chapter 4.3.3)
liv_004	<i>S. lividans</i> TK24 hosting BAC DNA of the 004 gene cluster	This study (Chapter 4.2.1)
liv_004ΔLuxR	liv_004 with deletion of <i>ctg1_627</i> and <i>ctg1_628</i> genes	This study (Chapter 4.3.1)
liv_009	<i>S. lividans</i> TK24 hosting BAC DNA of the 009 gene cluster	This study (Chapter 4.2.2)
liv_014	<i>S. lividans</i> TK24 hosting BAC DNA of the 014 gene cluster	This study (Chapter 4.2.3)
liv_016	<i>S. lividans</i> TK24 hosting BAC DNA of the 016 gene cluster	This study (Chapter 4.2.4)
liv_218-3	<i>S. lividans</i> TK24 hosting BAC DNA of the 218-3 gene cluster	This study (Chapter 4.2.6)
liv_HygE218_3	<i>S. lividans</i> TK24 hosting BAC DNA of the 218-3 gene cluster under the control of the <i>ermE</i> * promoter	This study (Chapter 4.3.2)
liv_pIJEF218_3	liv_218-3 lipidation module overexpression strain	This study (Chapter 4.3.3)

2.1.2 Vectors and constructs used in this study

Table 2.2 Vectors and constructs used in this study

Vector	Description	Reference
pTARa	TAR cloning <i>Saccharomyces-E. coli-Streptomyces</i> shuttle vector	67
pIJ10257	<i>E. coli-Streptomyces</i> shuttle vector with <i>ermE*</i> promoter and <i>hygR</i> (hygromycin resistance) gene	69
pWHU2449	Plasmid with <i>hygR</i> and the gene cassette <i>ermE*-sfp-svp</i>	19
pRedET (tet)	Red/ET expression plasmid	GeneBridges
YAC004 BAC004	pTARa derivative that carries 004 gene cluster	This study (Chapter 4.2.1)
BAC004 Δ <i>LuxR</i>	BAC004 derivative with <i>ctg1_627-628</i> replaced by <i>hygR</i>	This study (Chapter 4.3.1)
YAC005 BAC005	pTARa derivative that carries 005 gene cluster	This study (Chapter 6)
BAC005 Δ <i>PKS</i>	BAC005 derivative with <i>ctg4_380-382</i> replaced by <i>hygR</i>	This study (Chapter 6)
BAC005 Δ <i>AbXO</i>	BAC005 derivative with <i>ctg4_395</i> replaced by <i>hygR</i>	This study (Chapter 6)
YAC009 BAC009	pTARa derivative that carries 009 gene cluster	This study (Chapter 4.2.2)
YAC014 BAC014	pTARa derivative that carries 014 gene cluster	This study (Chapter 4.2.3)
YAC016 BAC016	pTARa derivative that carries 016 gene cluster	This study (Chapter 4.2.4)
YAC031 BAC031	pTARa derivative that carries 031 gene cluster	This study (Chapter 5)
YAC218-3 BAC218-3	pTARa derivative that carries 218-3 gene cluster	This study (Chapter 4.2.6)
BACHygE218_3	BAC218-3 derivative with <i>ermE*</i> inserted between <i>ctg9_270</i> and <i>ctg9_271</i> positions	This study (Chapter 4.3.2)

2.1.3 Media

Media were prepared following the below recipes. To make solid medium, 20 g/L Micro agar (DUCHEFA, DUC M1002.1000) was added. All media were autoclaved at 121 °C for 20 minutes.

Table 2.3 Media used in this study

Medium	Recipe / Source	Description
LB Broth	DUCHEFA (DUC L1703.2500), 20 g/L	routine growth medium for <i>E. coli</i>
SOC	Dextrose, 3.603 g/L KCl, 0.186 g/L MgSO ₄ , 4.8 g/L Tryptone, 20 g/L Yeast extract, 5 g/L	competent cell recovery medium
TSB	Thermo Fisher (CM0129B), 30 g/L	routine growth medium for <i>Streptomyces sp.</i>
YPD	DUCHEFA (Y1708), 50 g/L	yeast growth medium
ISP4 agar	HIMEDIA (M359), 36.5 g/L	preparing spore stocks, conjugation
MHB	Thermo Fisher (CM0405), 21 g/L	for antimicrobial susceptibility testing
SFM Medium	Mannitol, 20 g/L Soya flour, 20 g/L	preparing spore stocks, conjugation
ISP4	Starch, 10 g/L K ₂ HPO ₄ , 1 g/L MgSO ₄ .7H ₂ O, 1 g/L NaCl, 1 g/L NH ₄ SO ₄ , 2 g/L CaCO ₃ , 2 g/L *Trace elements B, 1 mL/L pH 7.2	fermentation medium used in this study
ISP2	Yeast extract, 4 g/L Malt extract, 10 g/L Dextrose, 4 g/L	fermentation medium used in this study

SMM (for 1 L)	<u>SMM base: 1000 mL</u> Casamino acids, 2 g/L TES Buffer, 5.73 g/L adjust pH to 7.2 with 10N NaOH <u>50 % (w/v) Dextrose: 18 mL</u> <u>1000 x phosphates: 1 mL</u> NaH ₂ PO ₄ .2H ₂ O, 7.8 g/L anhydro K ₂ HPO ₄ , 8.6 g/L <u>^Trace elements A: 2 mL</u> <u>1 M MgSO₄: 20 mL</u>	fermentation medium used in this study
R5a	Sucrose, 100 g/L K ₂ SO ₄ , 0.25 g/L MgCl ₂ , 10.12 g/L glucose, 10 g/L casamino acids, 0.1 g/L MOPS, 21 g/L yeast extract, 5 g/L ^Trace elements A, 2 mL/L pH 6.9	fermentation medium used in this study
AIA (for 1 L)	Sigma-Aldrich (17117-500G), 22 g/L Glycerol, 5 ml	for isolating and cultivating actinomycetes from lichen samples
YENB	Yeast extract, 7 g/L Nutrient Broth, 8 g/L	for preparation of electro-competent <i>E. coli</i>

Table 2.4 Trace elements solution used in this study

^Trace elements A (500x stock)	*Trace elements B (2000x stock)
ZnCl ₂ , 20 g/L FeCl ₃ .2H ₂ O, 100 g/L CuCl ₂ .2H ₂ O, 5 g/L MnCl ₂ .2H ₂ O, 5 g/L Na ₂ B ₄ O ₇ .10H ₂ O, 5 g/L (NH ₄) ₆ Mo ₇ O ₂₄ .4H ₂ O, 5 g/L	CaCO ₃ , 100 g/L FeSO ₄ .7H ₂ O, 100 g/L ZnSO ₄ .7H ₂ O, 100 g/L MNCl ₂ .7H ₂ O, g/L

2.1.4 Antibiotics

Antibiotic stocks were prepared following the below concentration, sterilised by filtration through 0.22 µm syringe filter (Interlab, FPE-204-013) when appropriate and added to cooled autoclaved media when needed.

Table 2.5 Antibiotics used in this study

1000x apramycin	50 mg/mL in ultrapure water (Barnstead EasyPure II water purification system)
1000x chloramphenicol	25 mg/mL in EtOH
1000x hygromycin	100 mg/mL in ultrapure water
1000x Nalidixic acid	100 mg/mL (10 N NaOH was added dropwise to facilitate the dissolving)
1000x Nystatin	100 mg/mL in DMSO
1000x tetracycline	10 mg/ml in MeOH (working concentration for pRedET: 3 µg/ml)

2.1.5 Enzymes

Table 2.6 General enzymes used in this study

restriction endonucleases	purchased from NEB
NEBuilder® HiFi DNA Assembly Master Mix	purchased from NEB E2621L
HiScribe™ T7 High Yield RNA Synthesis Kit	purchased from NEB E2040S
Cas9 Nuclease, <i>S. pyogenes</i>	purchased from NEB M0386S
Q5® High-Fidelity 2X Master Mix	purchased from NEB M0492L
BioMix™ Red	purchased from Bioline BIO-25006

2.2 General experimental procedures

2.2.1 Lichen sample collection and processing

Lichens were sampled from different locations spanning across New Zealand (Fig 2.1, Appendix 2: Sampling_locations). Approximately 200 mg of each lichen sample was ground in a 1.5 mL microfuge tube using a micro pestle and suspended in 0.5 mL sterile 20% glycerol. Samples were collected, processed by Dr Mark Calcott, Dr Liwei Liu, and Dr Jeremy Owen, and stored at -80 °C.



Fig 2.1 Sampling locations for the lichens used in this study.

2.2.2 Isolation of actinomycetes

The lichen stocks were thawed and left to settle for 10 min on the bench. The resulting supernatants with 10-fold serial dilution were then plated on AIA plates incubated at 30 °C for 14 days. The AIA plates were supplemented with nystatin (50 µg/mL) to inhibit fungi and nalidixic acid (50 µg/mL) to inhibit Gram-negative bacteria. Colonies with actinomycetes-like morphologies were restreaked onto fresh AIA plates. Following 14 days of incubation at 30 °C, pure cultures/spores were prepared and stored at -80°C in 30% glycerol as glycerol stocks.

Dr Mark Calcott, Dr Liwei Liu and Dr Jeremy Owen conducted the isolation work.

2.2.3 DNA extraction and whole-genome sequencing

Genomic DNAs of the lichen sourced actinomycetes were extracted from the mycelium after 7-14 days of growth on the ISP4 agar (HIMEDIA) at 30 °C. Mycelium was scraped from plates, and total DNA isolation was performed according to the salting-out protocol.⁷⁰ Genomic DNA were stored in 50 µL of 5 mM Tris-HCl (pH 7.5, sterilised) in 96-well plates at 4 °C. The mappings of isolated strains/gDNA and corresponding positions on the 96-well plates are provided in Appendix 1.

For each genomic DNA in this study, an in-house low-cost NGS Library with an average insert size of 500 bp was prepared based on the protocol described by Hennig.⁷¹ Briefly, the gDNA (genomic DNA) was cleaved through tagmentation (Fig 2.2①), and PCR was conducted to introduce indexes onto the tagmentated gDNA (Fig 2.2②). All amplified and tagmentated gDNA were pooled together and loaded onto a 2% agarose gel, and the gel was then sliced based on size (Fig 2.2③). The size-selected gel slices (500-600 bp) were then recovered and cleaned up using SPRI beads (SpeedBeads™). The cleaned samples were analysed via BioAnalyzer (Fig 2.2④) before being sent to GENEWIZ for NGS in 2x150 bp sequencing using NovaSeq/HiSeq as the sequencing platform (Fig 2.2⑤).

This part of the work was done together with Dr Chanel Taylor.

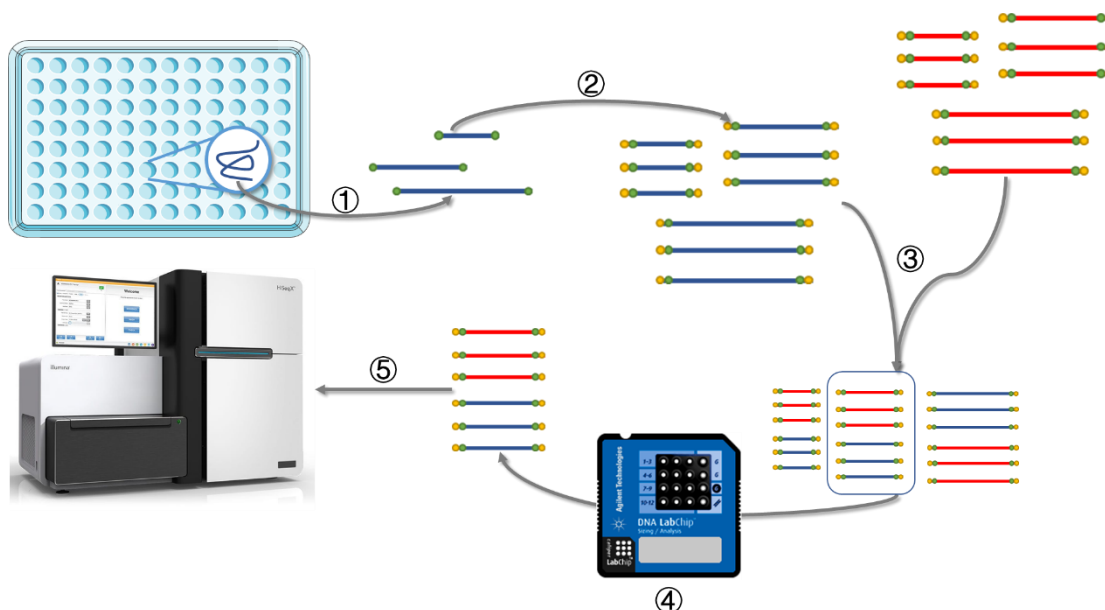


Fig 2.2 Large-Scale Low-Cost NGS Library Preparation used in this study. Genomic DNAs are extracted, sheared (①), barcoded (②) and pooled (③). The resulting fragments are size-selected (③) and analysed via BioAnalyzer (④) before sending out for sequencing (⑤).

2.2.4 General molecular biology

Plasmid DNA purification and gel extraction were prepared, performed using All-in-1 Mini Spin Columns (EconoSpin, 1920-250) and following the manufacturer's instructions (QIAquick® Spin Handbook, April 2015).

Restriction enzyme digestions, NEBuilder HiFi DNA Assembly Reactions were carried out according to the NEB's instructions.

2.2.5 Polymerase chain reaction

All PCRs were conducted on Bio-Rad T100 Thermal Cyclers.

Diagnostic/screening PCRs were performed with BioMix™ Red using a Touchdown PCR protocol (Table 2.7).

Table 2.7 Touchdown PCR programme

STEP	Temperature	Time (min)
Initial Denaturation	95 °C	3:00
Phase 1 (9 cycles)	95 °C	0:30
	68 °C	0:30
	(-1 °C /cycle)	
	72 °C	0:30
Phase 2 (30 cycles)	95 °C	0:30
	58 °C	0:30
	72 °C	0:30
Final Extension	72 °C	3:00
Hold	12 °C	∞

Molecular cloning PCRs were performed with Q5® High-Fidelity 2X Master Mix using a 2-step PCR (Table 2.8).

Table 2.8 2-step PCR programme

STEP	Temperature	Time (min)
Initial Denaturation	98 °C	0:30
30 cycles	98 °C	0:30
	72 °C	30 seconds/kb
Final Extension	72 °C	2:00
Hold	4 °C	∞

2.2.6 Preparation of electrocompetent cells and electroporation

3 mL YENB containing *E. coli* (EC100 or S17-1) was incubated overnight (37 °C). The following morning, this was inoculated into 500 mL of YENB until the OD ~0.5 was reached. The culture was transferred into 50 mL centrifuge tubes and placed on ice to chill. The chilled cells were collected by centrifugation (2500 g, 20 min, 2 °C), the supernatant was discarded, and the pellet was rinsed twice with chilled sterile ultrapure water without disrupting the pellet. The pellet was then resuspended in 50 mL of cold, sterile 10% glycerol using serological pipettes. Cells were again pelleted by centrifugation, and the glycerol washing process was repeated twice. The resulting pellet was resuspended in 500 µL chilled sterile 10 % glycerol, aliquoted into prechilled 1.5 mL microfuge tubes on ice (~40 µL/tube). Tubes were stored at -80 °C until further use.

To perform the electroporation, the desired electrocompetent cell was thawed on ice, mixed with purified DNA constructs, and transferred into a prechilled 2 mm electroporation cuvette (Bulldog Bio, 12358-346). The cells were electroporated (2.5 kV, 25 F, 100 Ω) using the Bio-Rad MicroPulser, and 1 mL prewarmed SOC medium was immediately added to the cuvette and transferred to a 15 mL tube. For large constructs such as BACs, the cells were incubated at 30 °C, 200 rpm for 90 minutes for recovery before being plated onto LB agar plates containing appropriate antibiotics and incubated at 30 °C overnight. In the case of small plasmids, the cells were incubated for 1 hour at 37 °C at 200 rpm for recovery before being plated on LB agar plates containing appropriate antibiotics and incubated at 37 °C overnight.

2.3 CRISPR/Cas9-mediated TAR (Transformation associated recombination) cloning

Traditional BGC recovery and cloning relied on the construction of genomic DNA cosmid libraries. This process is laborious in terms of screening and the need to piece fragmented BGCs into long intact BGCs due to the size limitations of the library.^{72,73} In this thesis, we applied a one-step TAR (Transformation associated recombination) cloning technique and recovered eight pathways ranging from 30 kb to 90 kb.

TAR cloning takes advantage of the *Saccharomyces cerevisiae* *in vivo* homologous recombination environment, and the recombination takes place between the target chromosomal loci (e.g., BGC) and linearised TAR cloning capture vector that contains homologous gene fragments to the targeted region. Coupling CRISPR/Cas9 to generate double-strand breaks on genomes near the target region, TAR cloning enables the direct cloning of chromosomal regions up to 300 kb with ideal successful rates.^{48,74}

A detailed explanation of CRISPR/Cas9-mediated TAR cloning for one pathway- BGC218-3 is depicted below. The CRISPR/Cas9-mediated TAR cloning process following well-established protocols^{57,73,74} has been carried out for a total of 8 biosynthetic pathways in this study.

The programmed sgRNA (single guided RNA, Fig 2.3b) consists of two parts. The crRNA contains 20 nucleotides that complement the target sequence located upstream of the PAM (NGG) motif. The tracrRNA is partially complementary to the crRNA and partially bound to Cas9. In order to obtain sgRNA, a DNA template for transcribing *in vitro* was needed. The sgRNA DNA template sequence contained the T7 promoter sequence, ~20 nucleotides of target-specific sequence, and the crRNA/tracrRNA sequence. sgRNA DNA template was next used to transcribe sgRNA (Fig 2.3b) using an *in vitro* transcription kit (NEB, E2040S).

In the 218_3 cloning project, several sgRNAs targeting the upstream and downstream of the 218_3 biosynthetic pathway (Fig 2.3a) were synthesised (Fig 2.3c). Two DNA fragments, located upstream and downstream of BGC218-3 (Fig 2.3a), were PCR amplified (using primer pairs 218-3_L_PAM_chk_f/r and 218-3_R_PAM_chk_f/r, provided in Appendix 8) and purified

from the genomic DNA of lichen isolate Li4c-G7. The resultant DNA fragments were used as substrates in determining the cutting efficiency of the Cas9-sgRNA complex. The cleavage efficiency of each sgRNA-Cas9 pair was evaluated and visualised on a 1.5% agarose gel (Fig 2.3d). Based on the results, Cas9-sgRNA-L1/L2 and Cas9-sgRNA-R1/R2/R3 were chosen to be co-incubated with genomic DNA of lichen isolate Li4c-G7 at 37 °C for 16 hours.

The universal primers for sgRNA synthesis sgRNA-F/R are listed in Table 2.9, and pathway targeted specific primers used in sgRNA synthesis are provided in Appendix 8.

Table 2.9 Universal sgRNA primers used in this study

sgrna-uni-f	GTTTGTAGAGCTAGAAATAGCAAGTTAAAATAAGGCTAGTC
sgrna-uni-r	AAAAGCACCGACTCGGTGCCACTTTTTCAAGTTGATAACGGACTAGCCTTATTTAACT

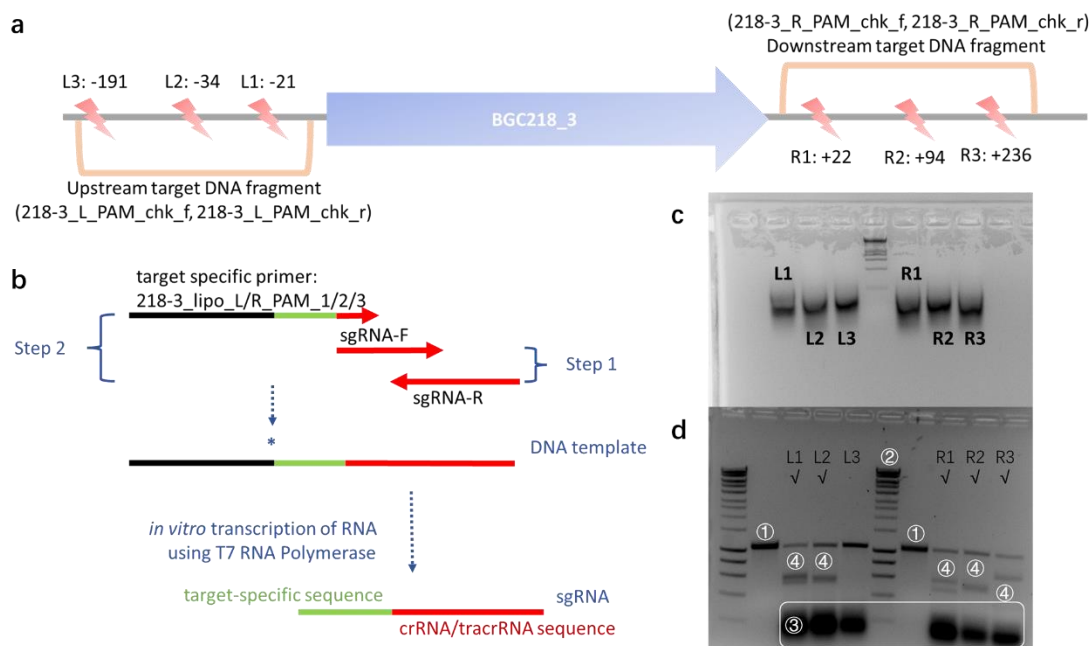


Fig 2.3 Design, preparation, and evaluation of sgRNAs

- Cleavage positions of Cas9-sgRNA complexes near BGC218-3. The cutting sites of Cas9-sgRNA complex L1, L2, and L3 upstream and downstream of R1, R2 and R3 are indicated.
- Preparation of *in vitro* transcription template for sgRNA.
- 3% agarose gel electrophoresis of the resulting *in vitro* transcribed sgRNAs.
- Cutting efficiency of Cas9-sgRNA complex. sg-RNA-L1-3 and sg-RNA-R1-3 were synthesised and used to determine the cutting efficiency of the Cas9-sgRNA complex.
 - sgRNA efficiency test DNA fragments
 - HyperLadder™ 1kb, Bioline
 - sgRNA residues
 - expected fragments

The BGC218-3 pathway-specific capture vector was constructed by introducing two short homology arms (Fig 2.4) corresponding to both ends of BGC218-3. The two short homology arms (~750 bp each) are amplified, then assembled with the NheI linearised pTARa using NEBuilder® HiFi DNA Assembly. The resultant pTARa-based BGC218-3 capture vector has a capture arm separated by the unique PmeI site.

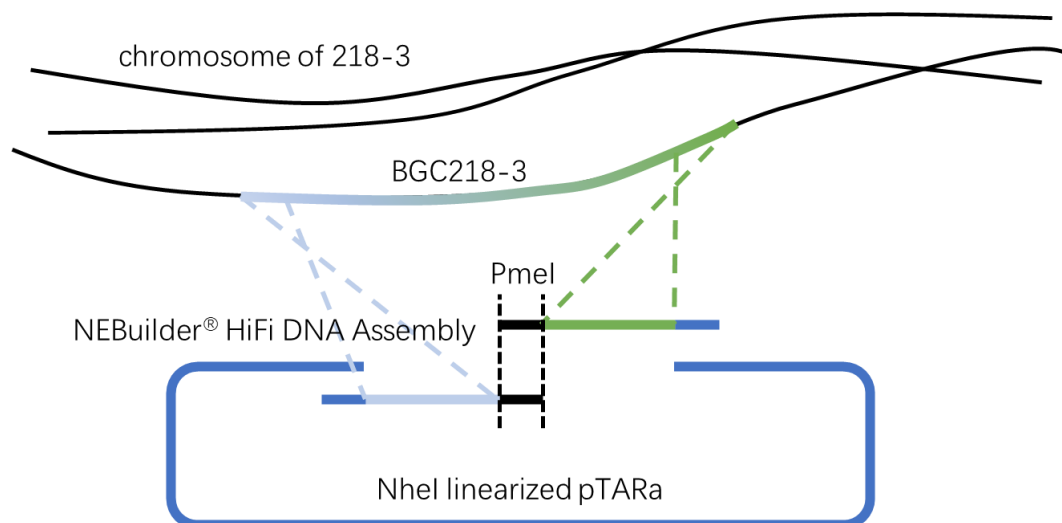


Fig 2.4 Construction of BGC218-3 pathway-specific capture vector
Two short arms (~750 bp each) homologous to the ends of BGC218-3 region are introduced to pTARa using NEBuilder® HiFi DNA Assembly.

Finally, CRISPR/Cas9 treated Li4c-G7 genomic DNA was co-transformed along with the PmeI linearised pathway-specific capture vector into freshly prepared yeast spheroplasts to form YAC218-3 (Yeast Artificial Chromosome 218-3, Fig 2.5).

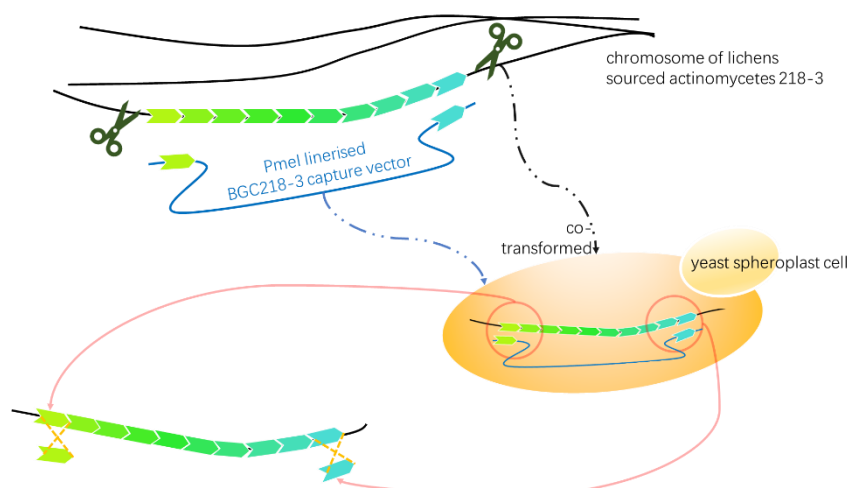


Fig 2.5 Schematic drawing of CRISPR/Cas9-mediated TAR cloning.

TAR takes advantage of the natural *in vivo* homologous recombination environment of *S. cerevisiae* BY4727 Δ NHEJ. CRISPR/Cas9 treated genomic DNA is co-transformed along with the PmeI linearised pathway-specific capture vector into freshly prepared yeast spheroplasts to form YAC.

Fig 2.6a illustrates the pooled testing strategy used for screening positive yeast clones. 100 yeast colonies were picked up using toothpicks and transferred onto 10×10 grid new agar plates (Fig 2.6a, Plate-Replica 1-3). Next, we grouped colonies on Plate-Replica 1 into 10 pools (columns) and tested each pool using one of the diagnostic primers (amplifying certain regions in the cloned cluster, shown in Fig 2.6b). If a pool tested positive (e.g., Fig 2.6a column A in 1st round), we tested each colony from the pool (column) on Plate-Replica 2 using the same diagnostic primers.

Candidate clones (e.g., Fig 2.6a A5 in 2nd round) on Plate-Replica 3 are validated using full sets of diagnostic primers, clones showing all desired PCR signals (four pairs of diagnostic primers amplifying several different regions in the cloned cluster, shown in Fig 2.6b) are assigned as positive yeast clones (Fig 2.6c). The PCR confirmed YAC218-3 is then electroporated into *E. coli* EC100 for BAC218-3 (Bacterial Artificial Chromosome 218-3) construct maintenance and *E. coli* S17-1 for future use. *E. coli* S17-1/ BAC218-3 was then used as a donor strain for transferring into desired streptomycetes (*S. albus* Del14, *S. lividans* TK24) by intergeneric conjugation following standard protocol⁷⁰.

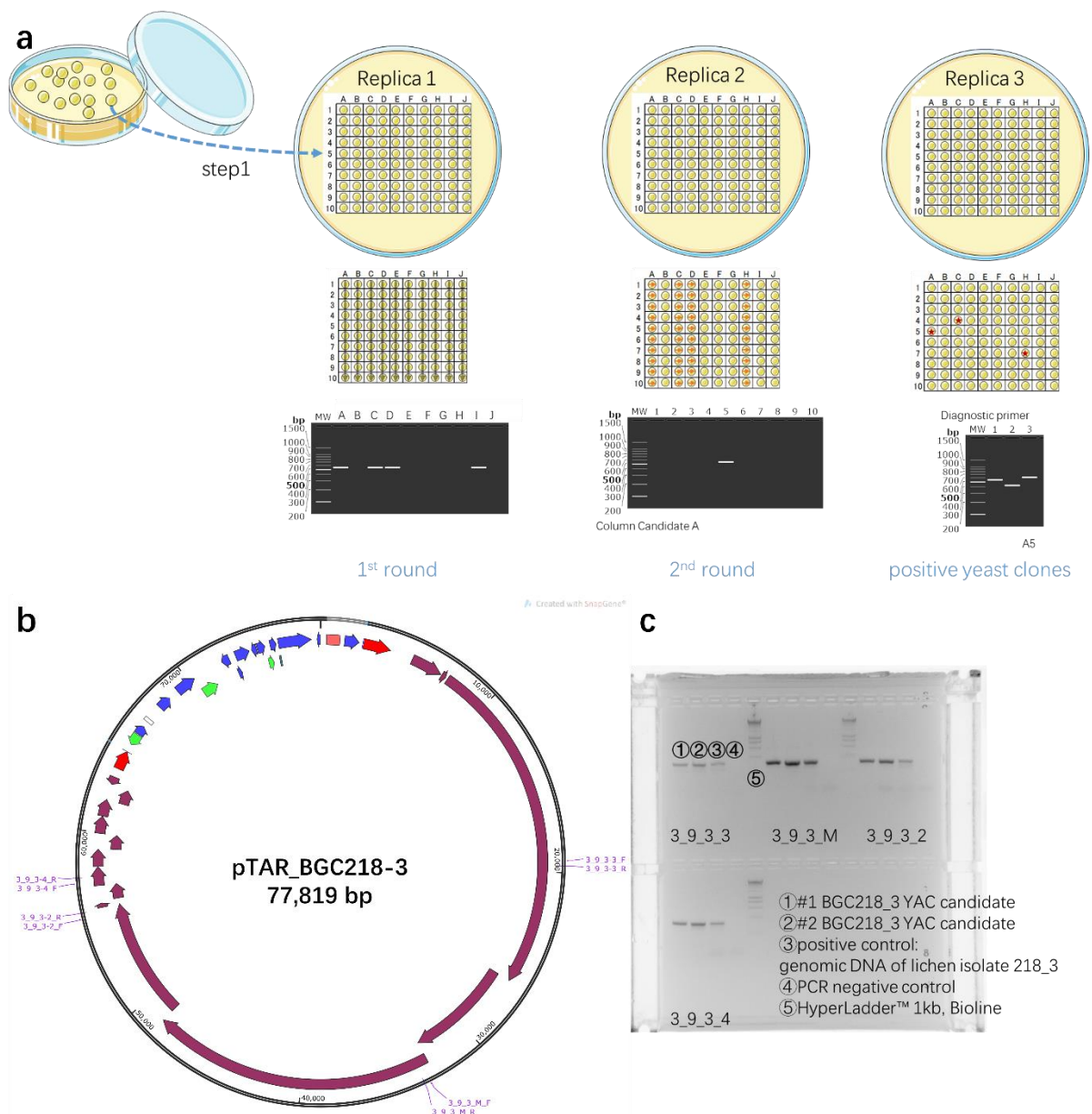


Fig 2.6 Screening for positive yeast clones

- The pooled testing strategy used in yeast clones screening
- Four pairs of diagnostic primers amplifying four different regions in the cloned cluster are indicated on the map. The correct YAC218-3 was screened and verified using these depicted primers.
- PCR amplification of four fragments (3_9_3_3_F/R: 315 bp, 3_9_3_M_F/R: 315 bp, 3_9_3_2_F/R: 315 bp, 3_9_3_4_F/R: 314bp) of the target BGC218_3 performed on individual yeast clones.

2.4 Small molecule analysis and characterisation

2.4.1 HPLC

Agilent Technologies 1200 Series HPLC System was used in this study.

Crude extracts analysis was carried out using PhenoSphere-NEXT™ C18 column (3 µm C18 120 Å, LC Column 150 x 4.6 mm) with protocol and parameters listed in Table 2.10.

Table 2.10 HPLC protocol and parameters 1

Time (Min)	A (%)	B (%)	A: H ₂ O + 0.1% formic acid B: Acetonitrile + 0.1% formic acid DAD (nm): 220, 268, 320, 430 Flow (mL/min): 1.0
0	90	10	
24	0	100	
26	0	100	
26.1	90	10	
28	90	10	

For compound isolation and purification, a semi-preparative column (NUCLEODUR C18 HTec, 5 µm, 125x21 mm) was used with the protocol and parameters listed in Table 2.11. Compounds are usually purified through two rounds.

Table 2.11 HPLC protocol and parameters 2

Time (Min)	A (%)	B (%)	DAD (nm): 220, 268, 320, 430 Flow (mL/min): 1.0 <u>Round 1:</u> A: H ₂ O + 0.1% formic acid B: Acetonitrile + 0.1% formic acid <u>Round 2:</u> A: H ₂ O + 0.1% formic acid B: MeOH + 0.1% formic acid
0	90	10	
24	0	100	
24.1	0	10	
26	90	10	

2.4.2 LCMS/MS

Agilent 6530 Accurate Mass Q-TOF LC-MS coupled with an Agilent 1260 HPLC was used in this study.

The sample was injected to the LC system with PhenoSphere-NEXT™ C18 column (3 µm C18 120 Å, LC Column 150 x 4.6 mm), components within the mixture were separated using the protocol and parameters listed in Table 2.12. The first 1.5 minutes eluents were sent to waste

and eluents of 1.5-25 minutes were sent to the MS system for untargeted CID-MS/MS, where the collision energies were determined based on the equation $[\text{slope}] \times [\text{precursor ions}] / 100 + \text{offset}$ (low energy: slope=2.62, offset=14.75; high energy: slope=3.93, slope=22.13). Each scan's top five intense ions were submitted to CID and automatically excluded after three spectra for 0.3 min.

All samples were running in both positive and negative modes.

Table 2.12 LCMS protocol and parameters

Time (Min)	A (%)	B (%)	A: H ₂ O + 0.1% NH ₄ HCO ₂ B: Acetonitrile + 0.1% formic acid mass range (m/z): 50–2000 capillary voltage: 3500 V nebulizer gas (N ₂) pressure: 35 psi ion source temperature: 300 °C sheath gas temp and flow: 350 °C and 11 L/min
0	95	5	
1	95	5	
20	0	100	
23	0	100	
23.1	95	5	
25	95	5	

2.4.3 NMR

NMR spectra were collected and analysed by Dr Helen Woolner and Dr Joe Bracegirdle.

NMR spectra were recorded on a JEOL JNM-ECZ600R with a 5 mm FG/RO digital autotune probe (600 MHz for ¹H nuclei and 150 MHz for ¹³C nuclei).

MestReNova 14.2.3 was used for NMR data analysis.

2.4.4 Bioinformatic analysis

Briefly, SPAdes Assembler³⁶ was used to assemble genomes from Illumina HiSeq PE reads; GTDB-Tk⁷⁵ was applied to assign taxonomic classifications of the assembled genomes; antiSMASH⁴³ was adopted to search secondary metabolite biosynthetic gene clusters from assembled genomes; BiG-SCAPE⁷⁶ and BiG-SLiCE⁷⁷ were used for grouping BGCs into gene cluster families.

The detailed parameters and scripts will be discussed in Chapter 3 and provided in the appendices.

Chapter 3 Genetic Investigation of the Actinomycetes Assemblage in New Zealand lichen

3.1 Introduction

New Zealand is home to an abundance of lichen species, many of which are not found elsewhere in the world. Here, we use a genetics-first approach to evaluate actinomycetes from New Zealand lichens as a potential source of new bioactive compounds. A collection of 480 putative actinomycete isolates was obtained from lichen samples comprising at least 39 distinct species. Analysis of the resulting 138 Gb genome sequence data revealed over 8600 BGCs, many of which had high genetic divergence compared to a collection of 1.2 million BGCs from previously sequenced bacterial strains and showed low homology to the functionally characterised BGCs present in the MIBiG database. This study demonstrated the feasibility of working with cost-efficient sequencing data, illustrated a pipeline for systematically evaluating large-scale BGC datasets, and highlighted New Zealand lichen sourced actinomycetes, a previously underestimated microbial community, as a powerhouse of genetic resources and secondary metabolites.

3.2 Results

3.2.1 Strain isolation and genome sequencing

The sampling locations for lichens are depicted in Fig 2.1 and are listed in Appendix 2: Sampling_locations. 480 actinomycetes-like bacteria were recovered from the lichen-glycerol mixtures using the dilution plating method.

Genomic DNA was extracted from isolates and sequenced for downstream analysis. Of the 480 isolates for which libraries were constructed, 406 isolates' reads were successfully assembled into contigs. Following assembly, contig number and N50 were used as quality metrics⁷⁸, ranging from 31 to 17,859 contigs and 306 to 1,480,016 bp, respectively (Fig 3.1a,

Appendix 2: assembly_stats>dbtk.classify). Among the 406 assemblies, 74 were of low quality and were excluded from subsequent analysis, resulting in a collection of 332 draft-quality assemblies.

3.2.2 Taxonomic classification

To assign phylogenies to the newly sequenced isolates, GTDB-Tk⁷⁵ was used for taxonomic classification at the whole genome level. All of the 332 isolates for which we obtained genome sequence data could be classified at the genus level. However, only 166 (50%) could be classified at the species level, indicating our search had uncovered a large number of new species (Appendix 2: assembly_stats>gtdbtk.classify). We delineated at least 39 distinct species (Fig 3.1b) from 16 actinobacterial genera in our collection, including *Streptomyces*, *Nocardiopsis*, *Rhodococcus*, *Gordonia*, *Nocardia*, *Microbacterium*, *Oerskovia*, *Amycolatopsis*, *Spirillospora*, *Mycobacterium*, *Streptomyces_B*, *Micromonospora*, *Rothia*, *Kribbella*, *Williamsia_A*, *Embleya*. Ten isolates from five genera were not actinobacteria and were classified as off-target. The current taxonomic grouping and assignment were supported by an independent pairwise genome distance estimation using Mash⁷⁹ (Fig 3.1b).

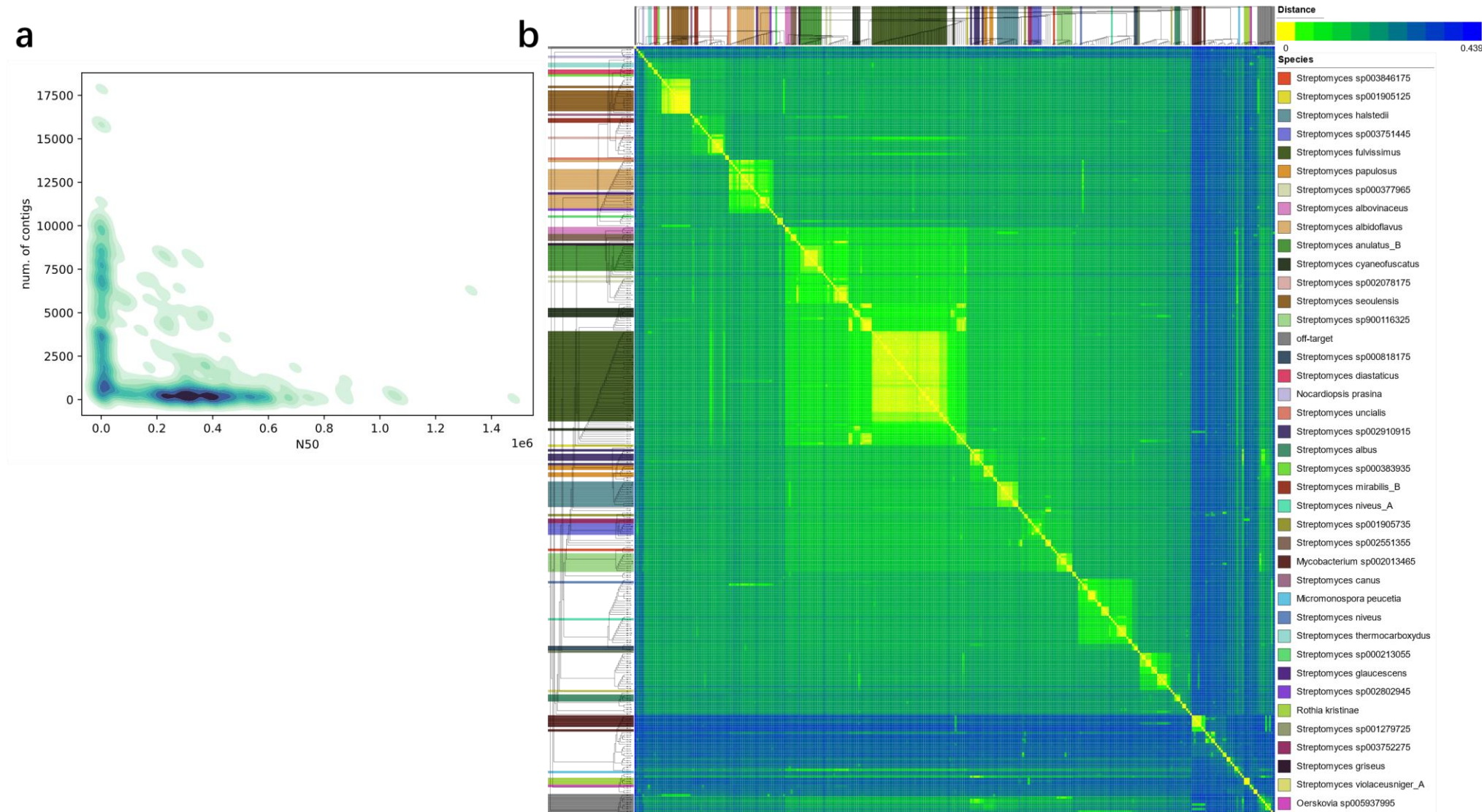


Fig 3.1 Whole-genome level taxonomic classification.

- a. 2D density chart showing the assembly quality assessed by N50 (x-axis) and number_of_contigs (y-axis) parameters.
- b. Taxonomic classification at the whole-genome level using GTDB-Tk. Isolates identified at the species level are indicated in different colour strips. The (colour) trends of all-pairs mash distance (genome similarity) in the matrix, from yellow (identical species) to green and blue (species to genus), supported the current taxonomic grouping and assignment.

3.2.3 Functional analysis of genome assemblies

On the protein-coding gene level, 62.5% of the open reading frames (ORFs) had no hits in the UniProt⁸⁰ and COG⁸¹ databases (Fig 3.2a). 84.6% of annotated genes were from *Streptomyces* (Fig 3.2b), an extensively studied genus. The functional annotation of these genes identified Amino acid transport and metabolism; Carbohydrate transport and metabolism; Translation, ribosomal structure and biogenesis; Lipid transport and metabolism; Transcription as the top 5 functional categories (Fig 3.2b).

We then used RGI⁸² to identify antimicrobial resistance (AMR) genes from our assemblies. In total, we detected 1292 AMR genes from 309 assemblies (Fig 3.2c, Appendix 2: rgi_matrix), belonging to 6 resistance mechanism types (antibiotic efflux; antibiotic inactivation; antibiotic target alteration; antibiotic target protection; antibiotic target replacement and reduced permeability to antibiotic). Isolates Li2c-H1 and Li4c-B5 harbour the most AMR genes (18 each). AMR genes such as β -lactamases which confer resistance to β -lactams, and antibiotic efflux pumps which render resistance to aminoglycosides, were among the most prevalent AMR Gene Families identified in our datasets.

CRISPRCasFinder⁸³ was used for identifying CRISPR arrays and Cas proteins on our assemblies. 165 Cas proteins from 115 isolates were co-detected with the CRISPR-like arrays, 14 Cas proteins from 14 isolates remained untyped. CAS-Type I and CAS-Type III from the Class 1 CRISPR–Cas system were prevalent in our datasets (Fig 3.2d).^{83,84} CAS-Type II is a subclass of the Class 2 CRISPR–Cas system found exclusively in bacteria, with CRISPR-Cas9 serving as the representative⁸⁵. No CAS-Type II can be detected or typed in our datasets (Fig 3.2d).

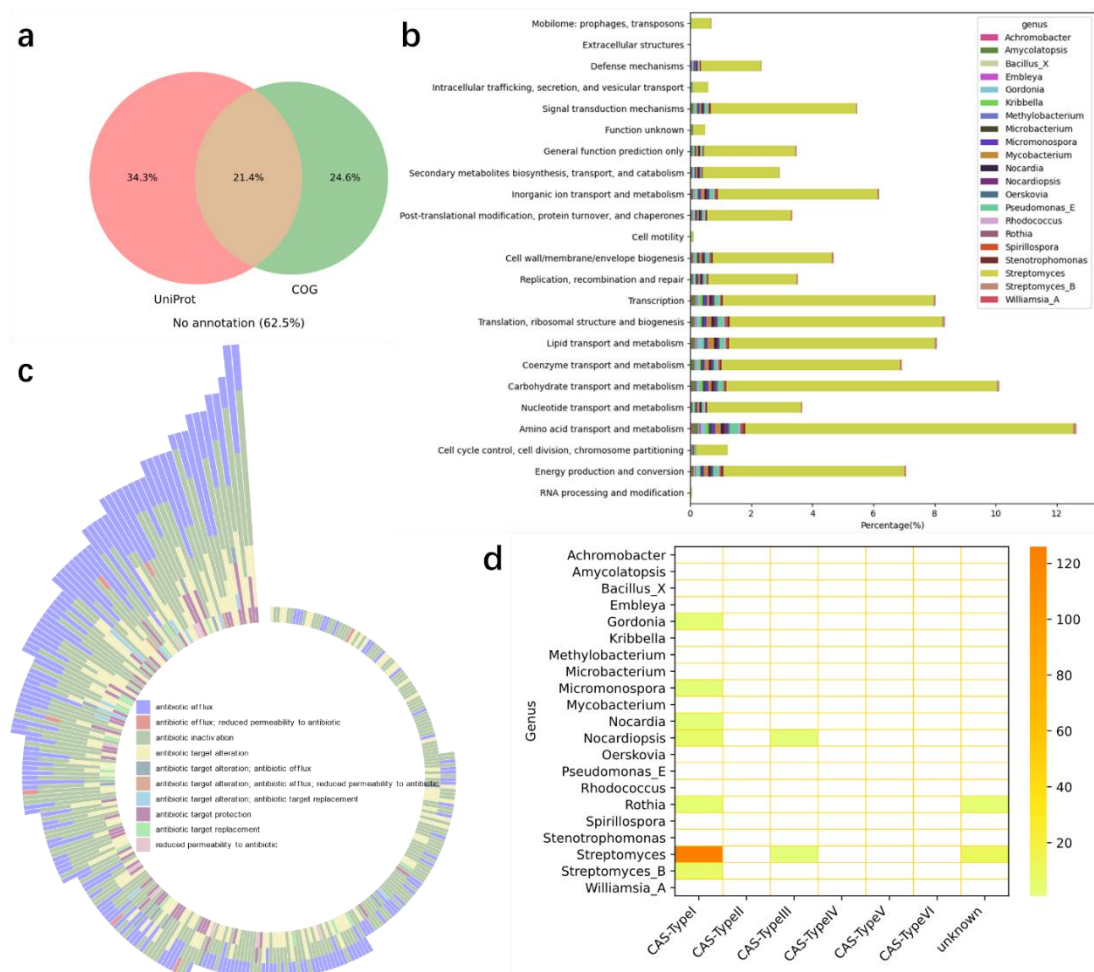


Fig 3.2 Functional analysis of the lichen sourced actinomycetes genome.

- Venn diagram showing the hits in the UniProt and/or COG databases. 62.5% of the genes remained unannotated.
- Classification of the annotated genes into COG functional categories and their taxonomic classification (down to genus level).
- Distribution of resistance genes in 309 isolates (Appendix 2: rgi_matrix).
- Heatmap showing the abundance of CAS-Types at genus level

3.2.4 Identification and categorisation of natural product biosynthetic gene clusters

Actinobacteria are well-known prolific producers of antibiotics, and a majority of these secondary metabolites are encoded by BGCs (biosynthetic gene clusters).¹ As a next step toward exploring the abundance of BGCs, antiSMASH 5.1.0⁴³ was used to identify and characterise BGCs from our 332 draft genomes. In total, we identified 8601 BGCs from 294 isolates (Fig 3.3a①, Appendix 2: antiSMASH_stats). The average length of identified BGCs was 31.8 kb. The longest BGC is 247.9 kb long from the isolate Li4c-A12 (Appendix 2: BiG-SLiCE stats).

Among the identified BGCs, we found 1629 NRPs, 1431 PKs, 1134 terpenes, 1123 RiPPs (ribosomally synthesised and post-translationally modified peptides), 8 saccharides. 1542 BGCs can be classified into the other chemical classes listed in Table 3.1. Another 1727 BGCs have multiple cluster types assigned (Fig 3.3a②, Appendix 2: antiSMASH_stats).

Within the 8601 BGCs, 5390 BGCs (62.7%) regions were not located on the contig edge, indicating the clusters were complete (Fig 3.3b, Appendix 2: antiSMASH_stats). The majority of the BGCs (7113, 82.7%) we discovered had less than 90 % similarity to any BGC in the MIBiG database as assessed by the KnownClusterBlast module embedded in antiSMASH (Fig 3.3c, Appendix 2: antiSMASH_stats).

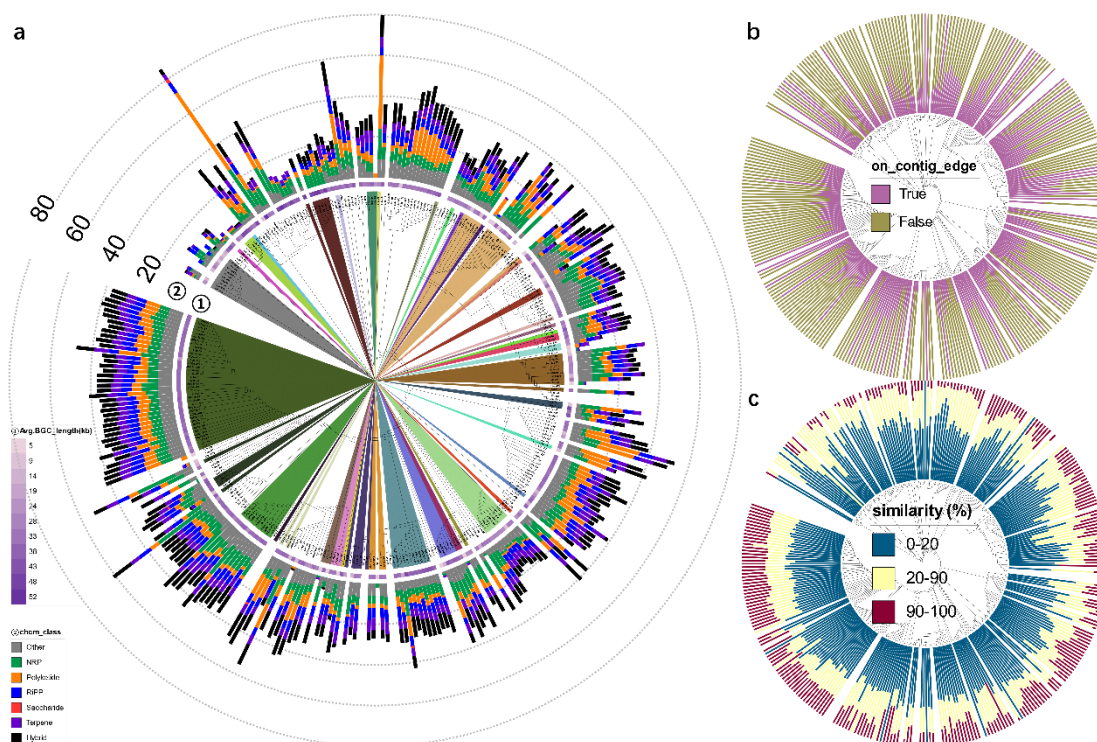


Fig 3.3 BGCs characterised by antiSMASH.

- The GTDB-Tk tree is a circular form and identical to the tree in Fig 3.1b. For 294 isolates, additional information on identified BGCs was provided (① average length and ② chemical classes). Species boundaries are delineated using different colour strips and are identical to those in Fig 3.1b.
- Phylogenetic distribution of the completeness of identified BGCs as defined by 'on_contig_edge'. 5390 BGCs (37.3%) regions were located on the contig edge (on_contig_edge = True), which suggested that they were potentially fragmented.
- Phylogenetic distribution of the similarity levels (queried against the MIBiG database) of identified BGCs. The low similarity level indicated that the BGC did not present clear links with MIBiG reference BGCs and their associated compounds, which likely encode new metabolites.

Table 3.1 Distribution of non-hybrid biosynthetic cluster types for BGCs present in this study.

chem_class	chem_subclass		chem_class	chem_subclass	
NRP	CDPS	14	Terpene	terpene	1134
	NRPS	1315	Other	NAGGN	4
	NRPS-like	300		arylpolyene	18
Polyketide	PKS-like	36		betalactone	86
	T1PKS	853		blactam	18
	T2PKS	158		butyrolactone	274
	T3PKS	351		ectoine	291
	hgIE-KS	24		furan	5
	transAT-PKS	5		fused	4
	transAT-PKS-like	4		hserlactone	2
RiPP	LAP	38		indole	22
	TfuA-related	14		ladderane	8
	bacteriocin	492		melanin	220
	lanthipeptide	355		nucleoside	16
	lassopeptide	151		other	69
	linaridin	62		phenazine	13
	thiopeptide	11		phosphonate	6
Saccharide	amglyccycl	2		resorcinol	1
	oligosaccharide	6		siderophore	485

3.2.5 Assignment of BGCs to Gene Cluster Families using BiG-SLiCE

As a next step toward determining the possible functional diversity and novelty, we used the query model of BiG-SLiCE to place each of the 8601 BGCs into one of the 29,955 pre-computed GCFs (gene cluster families) that are available as part of this software package.⁷⁷ These GCFs were derived from feature vectors extracted from 1.2 million BGCs from 209,000 genomes. Our BGCs mapped to a total of 1098 GCFs. 530 GCFs were found only once in our dataset, and 568 were shared by more than one isolate. GCF membership value (d) for each BGC was also assigned, a metric that indicates the relatedness between BGCs and their associated GCFs (Appendix 2: BiG-SLiCE stats).

The most prevalent GCF in our dataset, GCF_00822, contained 243 ectoine BGCs from 236 isolates (Fig 3.4a, Appendix 2: BiG-SLiCE stats). Ectoine is a compatible solute that protects organisms from osmotic stress and it is synthesised via a biosynthetic pathway that is readily identifiable.⁸⁶ This is interesting given that lichens are subject to cycles of wetting and drying, and the common occurrence of compounds that mitigate osmotic stress might provide a fitness benefit to microbes dwelling in this environment. The S.D. of the d in GCF_00822 was small among our dataset, indicating ectoine related BGCs were relatively conserved in lichen-sourced actinomycetes.

The values of d that are indicative of functional novelty are not precisely defined. The authors of BiG-SLiCE arbitrarily classified BGC-to-GCF relationships as "core" ($d \leq 900$), "putative" ($900 < d \leq 1800$) and "orphan" ($d > 1800$).⁷⁷ 5189 BGCs out of our 8601 BGCs were "core" members of their GCFs. The memberships of 3090 BGCs were "putative", implying that they were only moderately related to the centroid GCFs. 322 (3.74%) BGCs were "orphans", meaning they were distinct from the GCFs to which they were assigned and were good candidates for producing highly novel chemical entities (Fig 3.4b, Appendix 2: BiG-SLiCE stats).

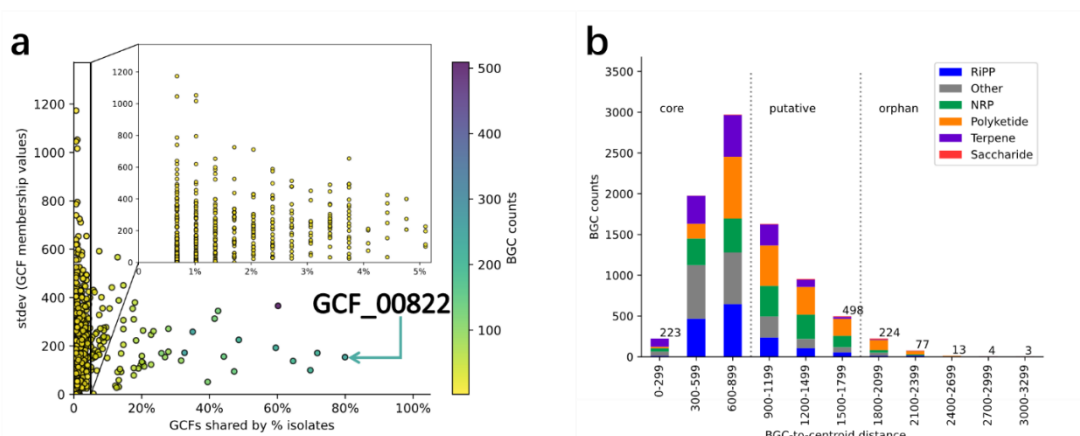


Fig 3.4 BGC-to-GCF membership assignment and calculation using BiG-SLiCE.

- Scatterplot showing the distribution of GCFs among our datasets. The colour of the dots corresponds to the number of BGCs within each GCF.
- The chemical class breakdown and distribution of BGC-to-GCF distance values. BGC-to-GCF memberships can be classified into three types: "core", "putative" and "orphan" based on the cut-off threshold of $T=900$.

3.2.6 "Orphan" membership BGCs from BiG-SLiCE mapping

BiG-SLiCE calculated BGC-to-GCF distance value d , and defined BGC-to-GCF memberships as "orphan" if $d > 2T$.⁷⁷ We have 322 BGCs that fall into "orphan" membership among our datasets when using $T=900$. The "orphan" BGCs span over 12 actinobacterial genera and are relatively abundant in genera *Embleya* (11.9%) and *Williamsia_A* (7.7%) (Fig 3.5a).

In-silico analysis of two cryptic BGCs that were mined using the d value metric is presented below to illustrate how this type of analysis might be helpful to guide the discovery of BGCs for further study.

The Li3d-B6_r1c5 BGC ($d=2051$) is a Type I PKS BGC that had been allocated to GCF_00018. Li3d-B6 was detected in the genus *Amycolatopsis*. The Gene Cluster Family GCF_00018 contains 1,250 core members and 1,427 putative/orphan members, with T1PKS (99.8%) and *Streptomyces* (54.4%) serving as the representative class and taxon, respectively.⁷⁷ BGC0000073.1, a reference BGC from the MIBiG database, is the most relevant member ($d=549$) of GCF_00018.⁷⁷ The BGC0000073 pathway has nine ORFs, seven of which encode Type I polyketide synthases and two of which encode cytochrome P450 monooxygenases. Experimental evidence has previously confirmed that these ORFs are essential for the production of the macrolactone

halstoctacosanolide A.⁸⁷ One starting module and eighteen extender modules spread across the seven polyketide synthases assembled a polyketide chain that was largely consistent with the final structure of halstoctacosanolide A. Two cytochrome P450 monooxygenases contribute additional oxidative transformations to the final compound (Fig 3.5b, Table 3.2).^{87,88}

In silico analysis of the isolate Li-3d genome revealed a 199 kb Li-3d_r1c5 region encoding a T1PKS. Eight Type I polyketide synthases on Li-3d_r1c5 comprising 30 modules were proposed to assemble a long polyketide chain (Fig 3.5b). Synteny⁸⁹ and antiSMASH⁴³ analysis showed that the Li3d-B6_r1c5 ORF organisation is homologous to BGC0000073 (Fig 3.5b, Table 3.2). The presence of KR, ER, and cytochrome P450 monooxygenase on Li-3d_r1c5 BGC was expected to bring further chemical modifications onto the polyketide backbone. Li-3d_r1c5 was assigned an orphan membership and possessed different PKS domain organisations compared to the reference BGC0000073, suggesting a role in novel (macrocyclic) polyketide biosynthesis.

The JGO2c-H7_r10c1 BGC (*d*=2061) was an NRPS BGC identified from the genus *Streptomyces* that was assigned to GCF_11473.⁷⁷ The GCF_11473 is a small gene cluster family with only two core NRP BGC members, and no reference MIBiG BGCs.⁷⁷ Further examination of the 87 kb JGO2c-H7_r10c1 region revealed an NRP region flanked by salicylate synthase and iron ABC transporter, suggesting that this region is involved in siderophore biosynthesis. Stachelhaus et al. proposed that active site residues (also known as Stachelhaus code) of A-domains conferred substrates (amino acids) specificity. One can predict the putative NRPS monomers by querying the newly sequenced A-domains against the active-site signature database using NRSPredictor2.⁴⁵ Stachelhaus prediction/alignment result (Fig 3.5c, Table 3.3) revealed that the proposed JGO2c-H7_r10c1 NRP chain salicylate-(cyclic Ser)-(D-Asn)-(D-Thr)-Ser-(cyclic hydroxyl-ornithine) was different from the chain of the representative siderophore gobichelin A⁹⁰ (MIBiG accession: BGC0000366). The presence of TauD/TfdA family dioxygenase (Ctg10_64, Table 3.3) on the JGO2c-H7_r10c1 BGC was expected to add additional oxidative transformations to the NRP chain. The high *d* value of JGO2c-H7_r10c1-to-GCF_11473 suggested that the JGO2c-

H7_r10c1 region encodes a new siderophore compound family member.

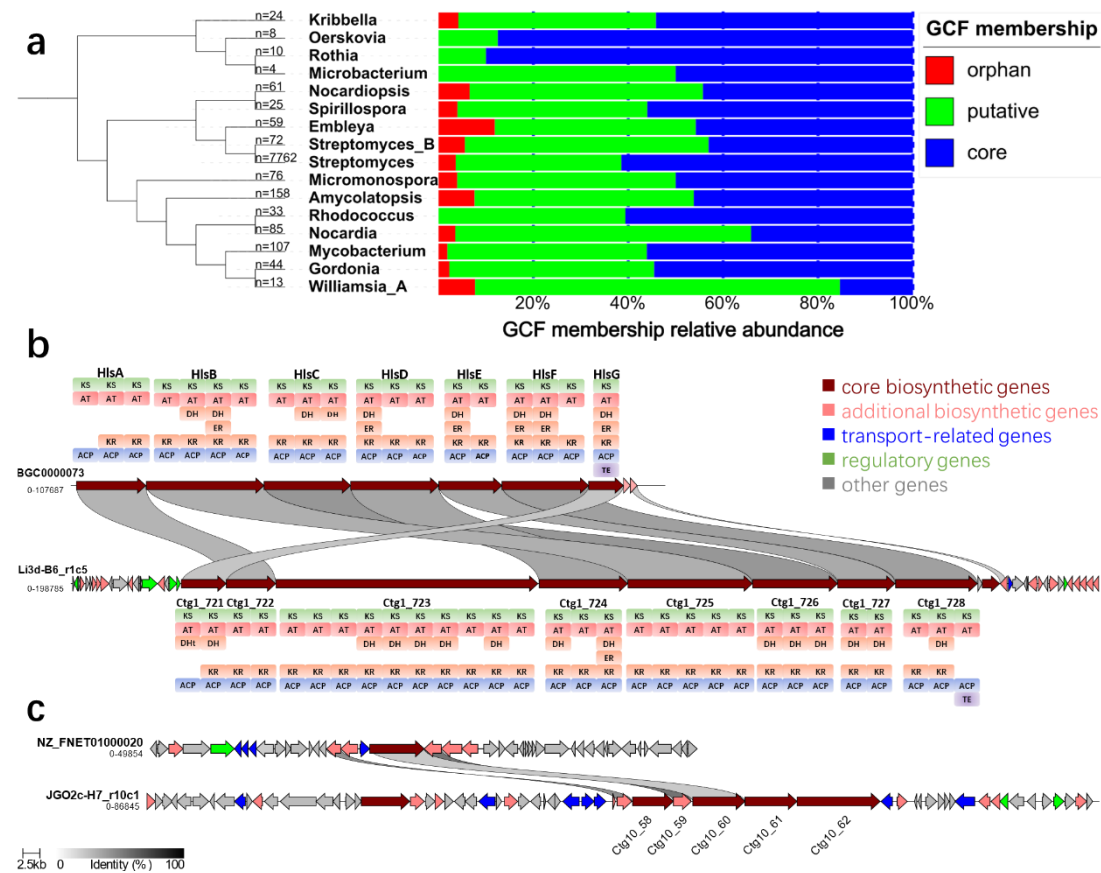


Table 3.2 Predicted functions and sequence alignment of the selected genes in the Li3d-B6_r1c5 region.

Li3d-B6_r1c5 vs BGC0000073					
ORF	aa size	proposed function	homologous BGC0000073 ORF	identity	similarity
<i>ctg1_720</i>	218	LuxR family DNA-binding response regulator			
<i>ctg1_721</i>	2709	PKS	HlsG	0.34	0.42
<i>ctg1_722</i>	2982	PKS	HlsA	0.4	0.51
<i>ctg1_723</i>	15813	PKS			
<i>ctg1_724</i>	5342	PKS	HlsD	0.44	0.58
<i>ctg1_725</i>	7508	PKS	HlsB	0.42	0.55
<i>ctg1_726</i>	5132	PKS	HlsC	0.5	0.63
<i>ctg1_727</i>	3460	PKS	HlsE	0.44	0.56
<i>ctg1_728</i>	4930	PKS	HlsF	0.47	0.61
<i>ctg1_729</i>	269	hypothetical protein			
<i>ctg1_730</i>	1047	NRPS			
<i>ctg1_731</i>	411	cytochrome P450	HlsI	0.3	0.45
<i>ctg1_732</i>	262	ABC transporter ATP-binding protein			

Table 3.3 Predicted functions of the selected genes in the JGO2c-H7_r10c1 region and A domain amino acid specificity prediction.

ORF	aa size	proposed function	ORF	aa size	proposed function
<i>ctg10_53</i>	500	Drug resistance transporter	<i>ctg10_59</i>	543	NRP
<i>ctg10_54</i>	381	transport system permease protein	<i>ctg10_60</i>	1587	NRP
<i>ctg10_55</i>	356	transport system permease protein	<i>ctg10_61</i>	1582	NRP
<i>ctg10_56</i>	78	mbtH-like protein	<i>ctg10_62</i>	2523	NRP
<i>ctg10_57</i>	447	isochorismate synthase	<i>ctg10_63</i>	345	iron compound ABC transporter
<i>ctg10_58</i>	1238	NRP	<i>ctg10_64</i>	309	Dioxygenase TauD/TfdA

	JGO2c-H7_r10c1		BGC0000366		
	Stachelhaus sequence	Substrate Predicted	Stachelhaus sequence	Substrate	
aa1@Ctg10_59	PLPAQGVLSK	salicylate	PLPAQGVLNK	salicylate	aa1@GobK
aa2@Ctg10_58	DLFNLgLIhK	ser	DLFNLgLIhK	ser	aa2@GobJ
aa3@Ctg10_60	DfTKVaEVGK	asn	DAIDgGfVDK	lys	aa3@GobR
aa4@Ctg10_61	DFWNIGMVHK	thr	DtWTIAsvDK	his	aa4@GobR
aa5@Ctg10_62	DVWHLSLIDK	ser	DVWHLSLvDK	ser	aa5@GobS
aa6@Ctg19_62	DVWILGAVNK	hydroxyl-ornithine	DAWeaGLvDK	hydroxyl-ornithine	aa6@GobS

3.2.7 Chemical space-guided gene cluster grouping improves genomics guided congener discovery

Comparison to the pre-computed GCF models using BiG-SLiCE gave us insights into the genetic divergence and diversity of our lichen sourced BGCs. BiG-SLiCE converted the full-length inputting BGCs into BiG-SLiCE features, and the GCF modules were then constructed based on the absence/presence of features. For example, we can immediately observe some common features shared among the BGCs in the ectoine gene cluster family (GCF_00822) from the feature heat map (Fig 3.6a). The numerous additional features present on Li2c-A11_r5c1 compared to the core member BGC0000855 ($d=324$) resulted in the assignment of an orphan membership to Li2c-A11_r5c1 ($d=1938$; Fig 3.6a, boxed in red).

antiSMASH analysis of Li2c-A11_r5c1 showed that Li2c-A11_r5c1 is an interleaved cluster; the protocluster ectoine overlaps with an NRPS protocluster (Fig 3.6b). Hence incorporating genes/features from the NRPS protocluster into the ectoine GCF membership calculation could lead to the overestimation of the BGC novelty (Fig 3.6b). This finding implied that the hybrid state and the boundary of one BGC would greatly affect the distance (d) calculation and potentially lead to inaccurate BGC similarity estimation.

KnownClusterBlast is an analysis module embedded in antiSMASH, which compares identified BGCs with experimentally characterised BGCs in the MIBiG repository and renders similarity scores during comparison. The similarity score is determined based on the equation: $[\text{number of hit genes on the query region}]/[\text{number of hit genes in the reference cluster}]$.⁴² For example, four genes on Li2c-A11_r5c1 BGC showed homology to four genes on BGC0000855, thus the similarity score is 1 (4/4). High similarity score/level indicated that Li2c-A11_r5c1 is likely to encode compound(s) as BGC0000855 produced. We think this region-to-reference comparison can better reflect the BCG similarity (Fig 3.6b).

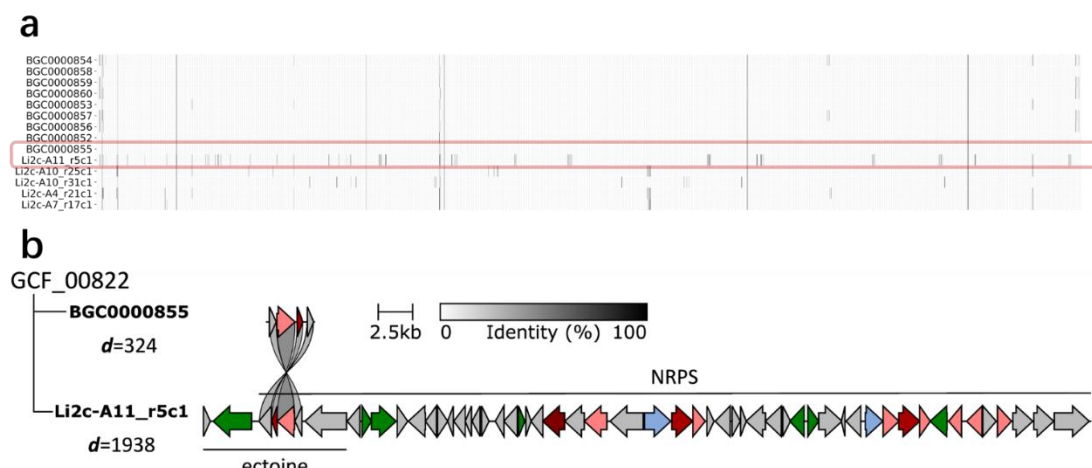


Fig 3.6 BiG-SLiCE vs. KnownClusterBlast analysis

- Feature presence/absence plot of selected BGCs in GCF_00822. This example showed that the hybrid state and the boundary of one BGC would affect the distance (d) calculation. The GCF assignment and distance value calculation are based on biosynthetic feature presence/absence. Non-ectoine related genes/features on Li2c-A11_r5c1 resulted in higher d calculation when compared to BGC0000755.
- Diagram illustrated the similarity score calculation by KnownClusterBlast when comparing Li2c-A11_r5c1 to BGC0000755. Four genes on Li2c-A11_r5c1 BGC have significant hits to genes on the BGC0000755. So the similarity score is 1 (4/4).

We next investigated whether we could infer the chemical diversity directly from the GCF model. The MIBiG database is a repository for the annotated sequences of experimentally characterised BGCs and their corresponding encoded compounds.⁴⁶ We used the genomic part of the MIBiG repository (BGCs) to construct a BiG-SLiCE GCF model (T=900) and organise the chemical part (chemical structures) into MFs (molecular families) based on Tanimoto similarity (tanimoto=0.5). We found that GCFs cannot overlay well with MFs under the current threshold (Fig 3.7). Hence, we cannot directly draw the conclusion of chemical diversity only based on the number of GCFs we have in our datasets.

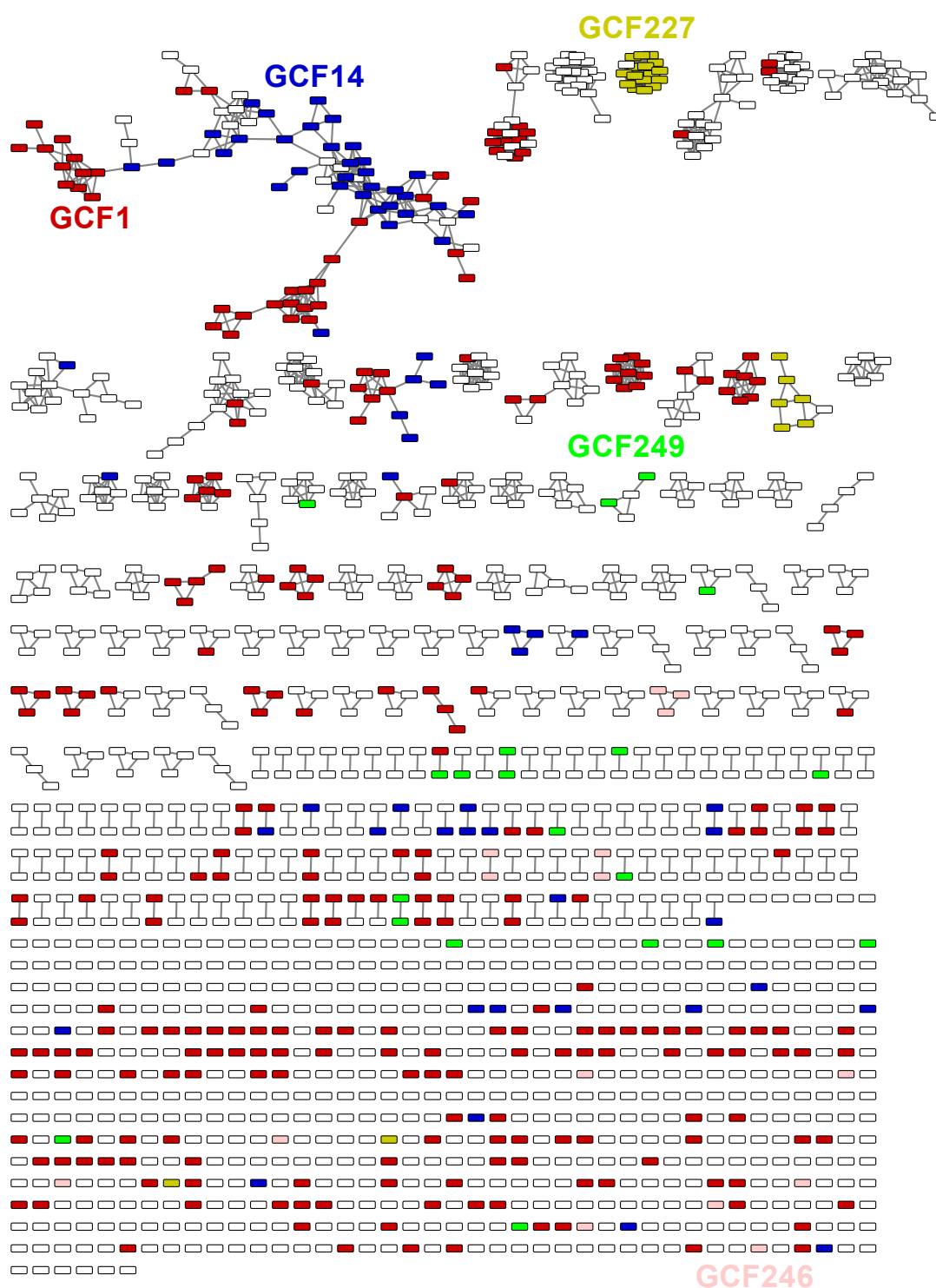


Fig 3.7 Overlaying MIBiG GCFs model onto the MIBiG Tanimoto similarity network. The largest 5 MIBiG GCFs (denoted in different colours) overlapped with the MIBiG Tanimoto similarity network. The overlaying result showed that the GCFs could not be mapped back to the MFs, indicating that GCFs were not equivalent to MFs under certain thresholds (BiG-SLiCE T=900, tanimoto=0.5).

Below, we described the process of inferring chemical diversity in our datasets by combining KnownClusterBlast analysis and the Tanimoto similarity network.

We began by associating our 8601 characterised BGCs (oval-shaped nodes) with the most similar MIBiG BGC using KnownClusterBlast⁴² results with a similarity cut-off of 0.2 (Fig 3.8a①). MIBiG BGCs (rectangle-shaped nodes) were then linked (blue edge, Fig 3.8a②) if their associated metabolites share a chemical similarity⁹¹ above an empirical threshold (Tanimoto similarity score >0.5), thereby increasing the connectivity of the resulting network (Fig 3.8b).

Using our expanded networking approach, 3605 lichen sourced BGCs (oval-shaped nodes) identified in our datasets can be connected (light cyan edge) to 366 MIBiG BGCs (Fig 3.8b) and grouped into 290 compound families. The top 10 largest BGC-to-MIBiG clusters were clusters of ectoine (Fig 3.8b, A), hopene (Fig 3.8b, B), geosmin (Fig 3.8b, C), isorenieratene (Fig 3.8b, D), coelichelin (Fig 3.8b, E), melanin (Fig 3.8b, F), desferrioxamin B (Fig 3.8b, G), SGR PTMs (Fig 3.8b, H), spore pigment related BGCs (Fig 3.8b, I) and streptobactin (Fig 3.8b, J). 4996 lichens sourced BGCs (58.09%) cannot be well associated with MIBiG reference BGCs, implying that a large portion of lichen derived BGCs might be dark matter that are not fully characterised.

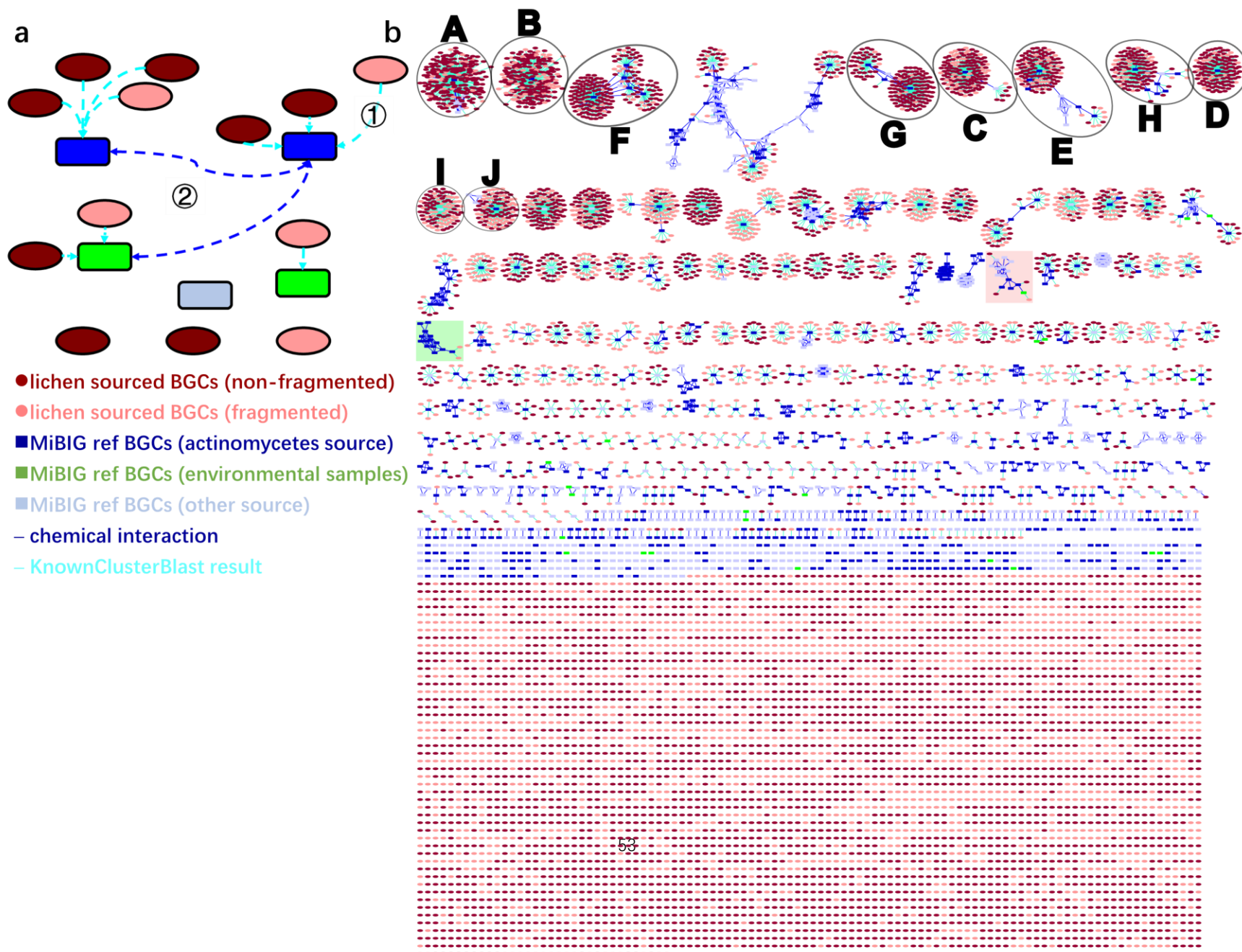
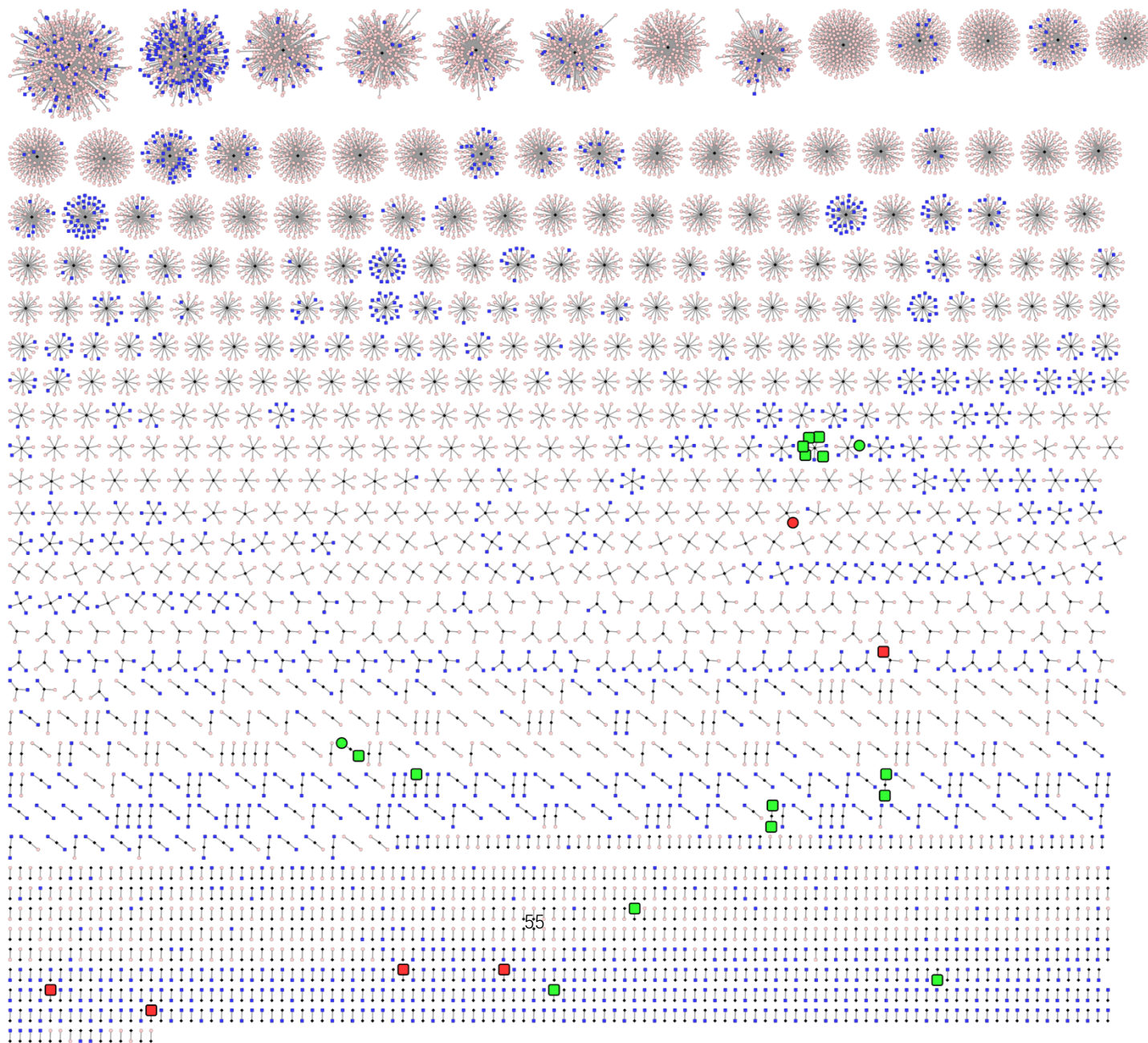


Fig 3.8 Chemical space-guided gene clusters grouping.

- a. An illustrative example of conducting the Chemical space-guided gene clusters grouping.
- b. The KnownClusterBlast based network was extended by linking the centroid MIBiG BGCs based on the chemical similarity score. Nodes and edges were colour coded using the same colour scheme of Fig 3.8a. The top 10 largest BGCs-to-MIBiG clusters are labelled. Glycopeptide MF is green shaded, and lipopeptide MF is highlighted in red.

The genetic-driven approach is playing a very positive role in searching for novel therapeutics to address the antibiotic (resistance) crisis. As such, we sought to explore our newly sequenced data for BGCs potentially encoding new members of these medically valuable compound families. For example, lipopeptides and glycopeptides are two types of tractable targets for genomic and metagenomic driven congener discovery.^{92–94} Our chemical space-guided gene cluster grouping network showed clear and precise connections when linking glycopeptide related BGCs (Fig 3.8b, shaded in green) and clustering lipopeptide related BGCs (Fig 3.8b, shaded in red). However, when using two publicly available GCF models for gene cluster grouping, we found that the linkages among lipopeptide BGCs were weak, such as the link between daptomycin and A54145, two well-characterised lipopeptides, is absent (Fig 3.9). Moreover, it appeared that glycopeptide BGCs were either not well-grouped (Fig 3.9a) or over-grouped with unrelated BGC types (Fig 3.9b).

a



b

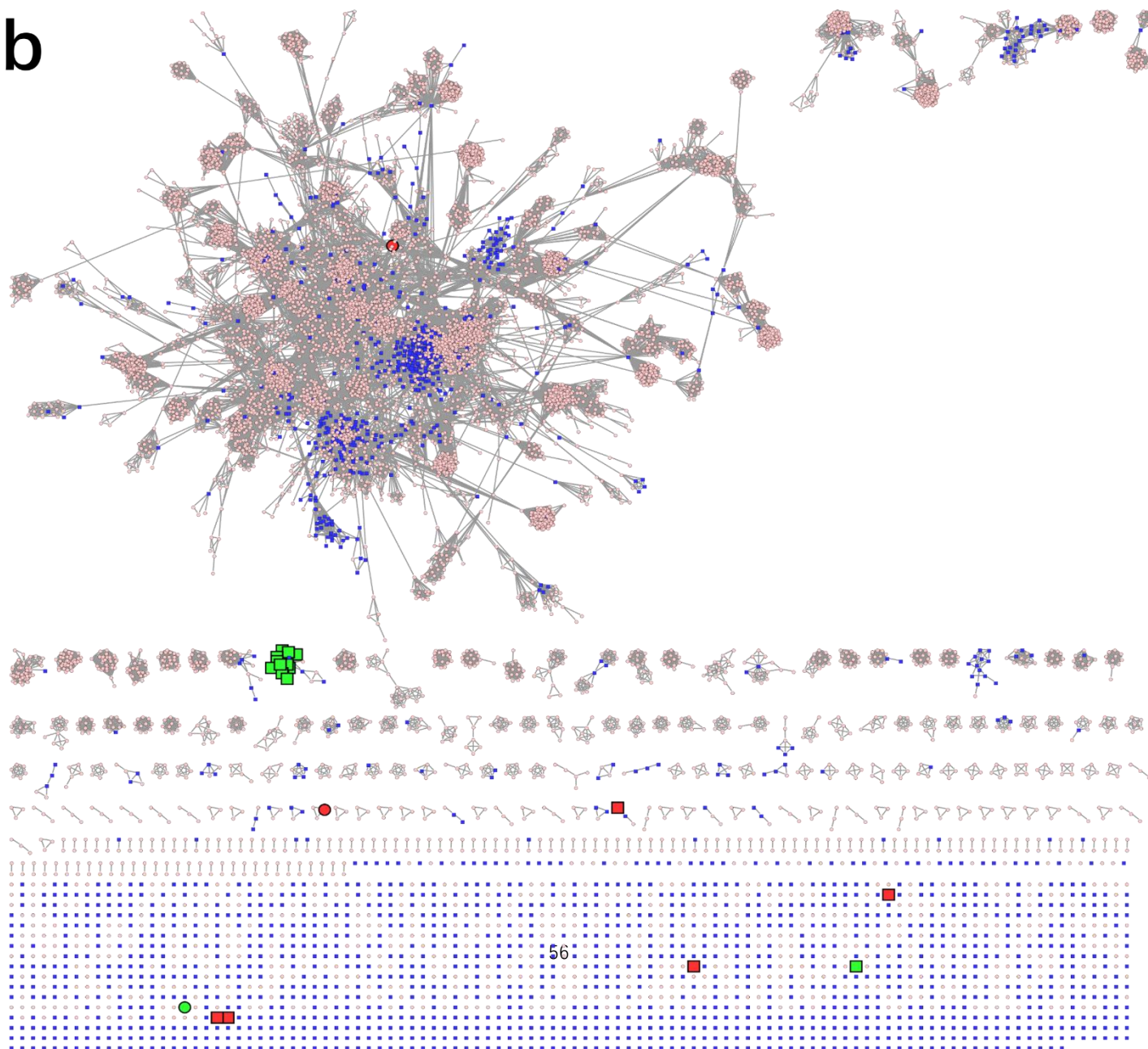


Fig 3.9 BiG-SLiCE (a) and BiG-SCAPE (b) analysis of lichen-sourced BGCs (coloured in pink, oval-shaped nodes) along with MIBiG reference BGCs (coloured in blue, rectangle-shaped nodes). Note that the CDA (Ca^{2+} -dependent cyclic lipodepsipeptide) BGCs marked in red and glycopeptide BGCs marked in green cannot be well linked when using fairly lenient thresholds in both models ($T=900$ for BiG-SLiCE, $c=0.6$ for BiG-SCAPE).

3.2.8 Chemical similarity networks inspired lichen BGCs mining

Here we focus on two cryptic BGCs mined from the chemical space-guided gene cluster grouping network depicted in Fig 3.8.

We began by examining the cluster of lipopeptides (Fig 3.8b, shaded in red; Fig 3.10a). BGC0001103 (mycosubtilin), BGC0001005 (locillomycin), BGC0001614 (hassallidin E), BGC0001127 (jagaricin), BGC0001090 (bacillomycin D), BGC0000402 (paenilarvins), BGC0001098 (iturin), BGC0001095 (fengycin), BGC0000407 (plipastatin) are lipopeptides discovered from non-actinomycetes source, while BGC0000439 (taromycin A), BGC0000315 (CDA1b), BGC0000291 (A54145), BGC0000336 (daptomycin) were lipopeptides discovered from actinomycetes and BGC0001968 (cadaside A) was recovered and expressed from environmental DNA⁹³. Further examination of the actinomycetes/eDNA subcluster revealed that this group of compounds belongs to the CDA (Ca^{2+} -dependent cyclic lipodepsipeptides). A Phylogenetic tree for selected BGCs was created based on AAI (the average amino acid identity).⁷⁹ The alignment of the BGCs within this subcluster was done in parallel using Clinker (Fig 3.10b).⁸⁹ The subcluster contained BGC0000439 (taromycin A), BGC0000291 (A54145), BGC0000336 (daptomycin) and Li4C-G7_218-3 was particularly of interest, thus we decided to add the BGC218-3 pathway to our heterologous expression project and will discuss in detail in Chapter 4.

Another cluster of potential interest was GPAs (glycopeptide antibiotics, Fig 3.8b, shaded in green, Fig 3.10c). GPAs are non-ribosomal synthesised polycyclic hexapeptides that are usually glycosylated. The GPA BGCs were grouped by AAI distance (Fig 3.10c left) and chemical similarity (Fig 3.10c right) in parallel.

GPA can be subclassed into five types based on the amino acid residues at positions 1, 3, and side chains of the peptide backbone⁹⁵. BGC0000326 (isocomplestatin) and BGC0001635 (kistamicin A) containing characteristic tryptophan residue were thus classified into Type V. We then focused on the Type V subclade (Fig 3.10c, right), and found that two lichen sourced BGCs (Li3c-a4_r44c1 and Li3d-F5_r28c1) are linked to the isocomplestatin BGC (MIBiG accession: BGC0000326, Fig 3.10c). The ORF organisations of Li3d-F5_r28c1 (incomplete) and BGC0000326 were strikingly similar (Fig 3.10d): several NRP synthases were flanked by transporter-related genes, halogenase, and cytochrome P450. Two Cytochrome P450 on Li3d-F5_r28c1 (Fig 3.10d, shaded in pink) shared high similarity/identity (Table 3.4) with their isocomplestatin counterparts which are involved in aa6-O-aa4 and aa4-aa2 crosslink generation (Fig 3.10c, shaded in pink).²⁰ Our BGC also contained the X-domain conserved on GPA related BGCs for cytochrome P450 recruitment⁹⁶ (Fig 3.10d, Table 3.4). Further Stachelhaus alignment and domain analysis (Table 3.4) revealed our Li3d-F5_r28c1 BGC incorporates tryptophan, a characteristic moiety of Type V GPAs⁹⁵. However, the methyltransferase domain was absent from our BGC compared to the isocomplestatin BGC (Fig 3.10d, Table 3.4). Although the low-quality sequencing hindered us from obtaining the full-length of Li3d-F5_r28c1 BGC, the current genetic organisation suggested this BGC were involved in the biosynthesis of demethylated isocomplestatin-like compounds, which may contribute to the diversity of the GPA compound family.

Fig 3.10 Two selected Molecular families.

- Several lichen BGCs were found in the lipopeptide molecular family. In this thesis, we further examined the highlighted area (calcium-dependent antibiotic related BGCs).
- BGC alignment analysis (by Clinker⁸⁹) revealed that lichen sourced BGC218-3 shared a similar genetic organisation with other reference CDA BGCs. Detailed analysis is provided in Chapter 4.
- Two lichen BGCs were found to be grouped into the glycopeptide molecular family.
- Clinker⁸⁹ alignment and domain comparison for Li3d-F5_r28c1 and BGC0000326 (isocomplestatin reference BGC) revealed that the biosynthetic organisation of these two pathways is strikingly similar, while one methyltransferase domain was absent from our BGC. The comparison suggested that Li3d-F5_r28c1 is involved in the biosynthesis of demethylated isocomplestatin, which might expand the members of the GPA molecular family.

Table 3.4 Predicted functions, sequence alignment and domain organisation of the selected genes in the Li3d-F5_r28c1 region.

Li3d-F5_r28c1 vs BGC0000326					
ORF	aa size	proposed function	homologous BGC0000326 ORF	identity	similarity
<i>ctg28_1</i> C-A-PCP-E C-A-PCP-E	3337	NRP	AAK81826.1 C-A-PCP-E C-A- Mt -PCP-E	0.4	0.48
<i>ctg28_2</i> C-A-PCP-X-L-TE	2168	NRP	AAK81827.1 C-A-PCP-X-L-TE	0.57	0.7
<i>ctg28_3</i>	75	mbtH-like protein	AAK81828.1	0.77	0.88
<i>ctg28_4</i>	428	sodium/hydrogen exchanger	AAK81829.1	0.67	0.78
<i>ctg28_5</i>	527	Trp_halogenase	AAK81830.1	0.73	0.81
<i>ctg28_6</i>	390	cytochrome P450	AAK81831.1	0.58	0.69
<i>ctg28_7</i>	408	cytochrome P450	AAK81832.1	0.57	0.68
<i>ctg28_8</i>	100	ferredoxin			
<i>ctg28_9</i>	358	alpha-hydroxy-acid oxidizing protein	AAK81834.1	0.73	0.83
<i>ctg28_10</i>	346	4-hydroxyphenylpyruvate dioxygenase	AAK81835.1	0.64	0.77
<i>ctg28_11</i>	444	aminotransferase	AAK81836.1	0.66	0.74
<i>ctg28_12</i>	300	prephenate dehydrogenase	AAK81837.1	0.49	0.59
<i>ctg28_13</i>	190	hypothetical protein			

3.3 Conclusion and discussion

This study demonstrated the feasibility of high-throughput data analysis using cost-effective sequencing. By comparing and aligning our sequencing data to different genetic reference collections, we demonstrate that actinobacteria sourced from New Zealand lichens potentially have the capacity to produce previously uncharacterised secondary metabolites. Our data shed light on New Zealand lichen inhabiting actinomycetes and demonstrate that lichen-sourced actinobacteria assemblages might serve as reservoirs for the discovery of new secondary metabolites.

GCF model is now becoming a widely adopted strategy for grouping BGCs. One can construct a GCF model for thousands of BGCs without prior knowledge of the datasets. Unfortunately, we noticed that our GCF models did not always reflect the underlying chemical reality. For example, given the striking similarities at genetic and chemical

levels, we expected the CDA-related BGCs can cluster together and glycopeptide BGCs to be grouped into one cluster. However, when using BiG-SLiCE⁷⁷ and BiG-SCAPE⁷⁶ for the GCF construction, the above BGCs cannot be well-grouped (Fig 3.9). We should note that sometimes different genetic elements contribute to the biosynthesis of the same chemical moiety. Hence overemphasising the genetic variations at the BGC feature level might lead to inappropriate BGC grouping.

Unlike the existing GCF models, starting from the sequence end, we introduced a new network guided by chemical structures, as we think the biosynthesis study of BGCs should be product oriented. We first pair our query BGCs with the experimentally characterised reference BGCs stored in the MIBiG repository based on the KnownClusterBlast analysis results. The query_BGC-to-MIBiG_BGC pairs will be further grouped if the metabolites encoded by the MIBiG_BGCs are similar enough to trigger a linkage. Through this chemical guided BGCs grouping, we can have a clear and precise image of the chemical potential our newly identified BGCs possess. However, the construction of this network heavily relies on the reference BGCs. Due to the limited reference BGCs published, a large portion of our BGCs cannot be well grouped, thus we are unable to infer their chemical space. Future work can focus on constructing a chemical-guided correction GCFs model.

Downstream studies focused on bringing genetically intriguing BGCs into reality. We selected eight cryptic BGCs mined from our current datasets for the heterologous expression project, which will be discussed in Chapter 4.

3.4 Methods

3.4.1 Sampling and isolation of actinomycetes

See Chapter 2.2.1-2.2.2

3.4.2 DNA extraction and whole-genome sequencing

See Chapter 2.2.3

3.4.3 Genome assembly and annotation

Raw reads were pre-processed using Trimmomatic⁹⁷ for quality trimming and adapter removal. The resulting filtered Illumina PE reads were then assembled using SPAdes 3.12.0³⁶.

Genome assembly quality was examined using assembly_stats 0.1.4⁷⁸ (Fig 3.1a, Appendix 2: assembly_stats>dbtk.classify).

The assembled draft genomes were annotated using Prokka 1.14.5⁹⁸ with default settings.

Resistomes were predicted by searching against RGI 5.2.0⁸² (Perfect and Strict hits only Criteria).

The CRISPRCasFinder⁸³ singularity container was used to detect CRISPR arrays and Cas proteins from our assemblies using Typing Clustering model.

3.4.4 Phylogenetic analysis and MASH distance calculation

Taxonomic classification at the whole genome level was assigned to each isolate using the identify-assign-classify workflow of GTDB-Tk v1.4.0⁷⁵, and the taxonomic classification results were provided in Appendix 2: assembly_stats>dbtk.classify.

MASH (version 2.2.2)⁷⁹ was used for genome-wide ANI calculation for all-pairs comparisons between genomes (Fig 3.1b). MASH (version 2.2.2) was also used for AAI calculation on selected BGCs.

3.4.5 Cheminformatics analysis

All entries in JSON format from MIBiG⁴⁶ (Version 2.0) were downloaded on May 27, 2021. MIBiG accession, Compounds (in SMILES string format), and Species (taxid) information were extracted from JSON files. Taxid from each MIBiG accession was used to fetch taxonomic information when queried back NCBI Taxonomy⁹⁹.

Rdkit (v2020.09.1.0)⁹¹ was used for molecular similarity network construction. Morgan Fingerprints of MIBiG compounds were generated with a radius of 2. Tanimoto similarity score was pairwise calculated between compounds, and an edge between compounds was created when the threshold of 0.5 was reached. Compounds were marked by taxonomy as indicated in the MIBiG molecular similarity network (Fig 3.8b).

3.4.6 antiSMASH analysis

Secondary metabolites were identified from assembled genomes using the standalone antiSMASH 5.1.0⁴³, using prodigal as the genefinding-tool, and only processed sequences larger than 5000 bp. A whole-genome HMMer analysis was also run for each genome. Identified clusters were compared against known reference BGCs from the MIBiG database alongside.

An in-house python (Appendix 3) script was used to extract the following general information: "region"- "contig_edge", "region"- "product" and "region"- "location" from output JSON files for downstream analysis and to produce Fig 3.3. The above information was provided in Appendix 2: antiSMASH_stats.

KnownClusterBlast is a module implemented in antiSMASH 5.1.0 used for comparing each identified cluster against known MIBiG reference BGCs, with an output of MIBiG reference BGCs hits for each lichen query BGC. An in-house python script (Appendix 4) was used to select the highest ranked MIBiG hit(s) from output JSON files ("antismash.modules.clusterblast"- "knowncluster") based on the ranking system described by Medema et al., 2011⁴². Briefly, an edge was created between a Lichen_BGC and top-ranked MIBiG hit(s) when "core_gene_hits" were larger than 1

and "Similarity" larger than 20%.

The whole Lichen_BGCs-MiBiG_BGCs network was then extended with MiBiG molecular similarity network using Cytoscape (Version: 3.8.2) to produce Fig 3.8b.

3.4.7 Gene cluster family (GCF) analysis

BGCs identified from antiSMASH 5.1.0⁴³ were first mapped back to the pre-built GCF model calculated at clustering threshold (T=900) using BiG-SLiCE (version 1.1.0)⁷⁷ query mode. In-house python workflow and scripts (Appendices 5,6,7) were used to extract chemical class, GCF membership and distance values from SQLite3 database raw output file to produce Fig 3.4, Fig 3.5a and Fig 3.9a.

Identified BGCs were also clustered along with MiBiG BGCs (versions 2.1) using BiG-SCAPE 1.1.0⁷⁶ (cut-off = 0.6). Output network raw files were used to generate Fig 3.9b using Cytoscape (Version: 3.8.2).

Chapter 4 Targeted Capture and Heterologous Expression

4.1 Introduction

Microbial natural product discovery has relied primarily on bioactivity as an indicator for decades. Generally, Fig 4.1① crude extracts from microorganisms are fractionated and screened against a panel of testing subjects, such as testing their antibiotic activity against various microorganisms. Fig 4.1② The active fractions are then re-fractionated, re-screened and dereplicated for additional rounds till the Fig 4.1③ fractions/peaks mapping to the activities are pure enough for downstream structure elucidation.¹⁰⁰

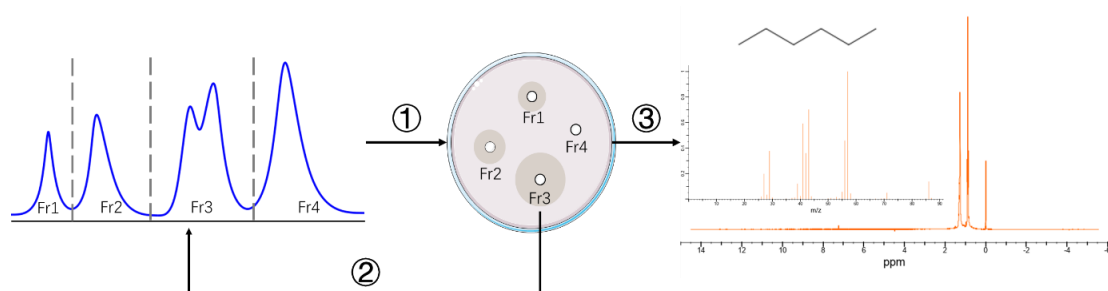


Fig 4.1 Bioactive-guided natural products discovery

Many secondary metabolites are encoded by BGCs. Unfortunately, most BGCs are poorly expressed under conventional laboratory culture conditions or remain silent due to the complex regulatory networks in their native hosts.¹⁰¹ Using heterologous expression strategies, we can transfer cryptic BGCs to a genetically amendable host to improve production by pathway refactoring and generating derivatives.

In this chapter, eight pathways (BGC004, BGC005, BGC009, BGC014, BGC016, BGC027, BGC031, BGC218-3) have been cloned and/or refactored. The BGCs of interest were identified from the genome of New Zealand lichen sourced actinomycetes (depicted in Chapter 3). Selected BGCs were then cloned using ① CRISPR/Cas9-mediated TAR cloning (depicted in Chapter 2.3), ② conjugated into *S. albus* Del14 and *S. lividans* TK24 (depicted in Chapter 2.3), ③ the resulting solid fermentation metabolites were

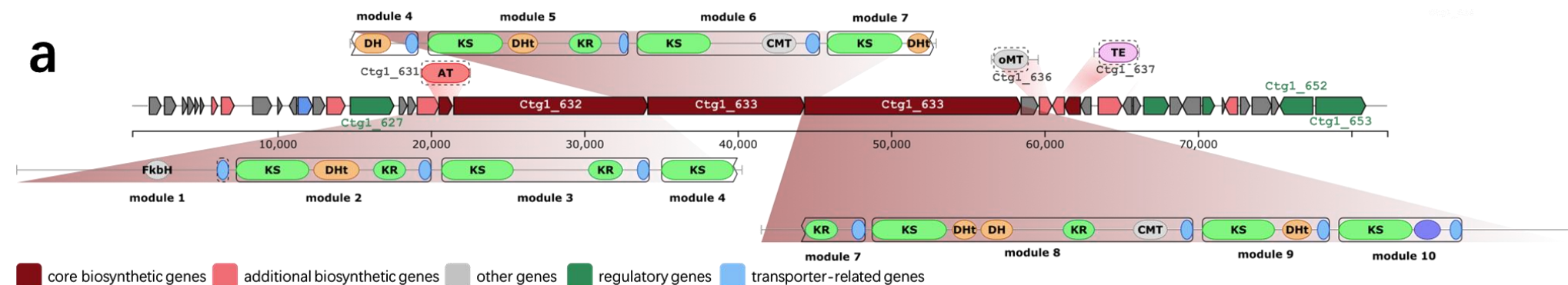
profiled and analysed through molecular networking.

4.2 *In silico* description of cloned pathways

4.2.1 BGC004

An antiSMASH analysis of BGC004 revealed a 9-module *trans*-AT PKS organisation (Fig 4.2a). It was predicted that a free-standing AT present on a single protein, Ctg1_631, initiates the proposed biosynthesis by selecting acyl-CoA as building blocks, which will be translocated to ACP domains embedded in the downstream modules to participate in the chain extension cycles. However, deducing the *trans*-AT biosynthesis can be challenging, as the presence of non-canonical modules such as module splitting (modules 4 and 7, Fig 4.2a) and non-elongating modules (KS5⁰ in module 5 and KS9⁰ in module 9, Fig 4.2b)¹⁰². Furthermore, non-standard biosynthesis events like domain duplication and unusual domain order pose difficulties and add uncertainties for *in silico* retrobiosynthetic predictions.^{102,103} The presence of FkbH in the loading module indicated that it could also incorporate unusual building blocks and introduce them to the elongating chain, a biosynthesis process similar to tetronomycin.¹⁰⁴

The pathway was captured with shuttle vector pTARa using CRISPR/Cas9-mediated TAR cloning method into *S. cerevisiae* BY4727 Δ NHEJ (YAC004), transferred to *E. coli* S17-1 (BAC004) and conjugated into *S. albus* Del14 and *S. lividans* TK24 to yield del14_004 and liv_004. The predicted functions of the genes on BGC004, gene fragments and primers used in cloning BGC004 are listed in Table 4.1 and Appendix 8, respectively.



b

```

KS2  GTDPSTIGYVEAHGTGTPLGDPIEIRGLGSAFAT-----VP-SSSEIPIGSVKGNIGHT  340
KS3  GVDARTVSYVETHGTGTRLGDPIEIAGLTAAFGSPAEG-----EPPWCRLGAVKPNIGHL  332
KS4  GVSASSVSYIEAHGTGTMIGDPMELRALTAFAE-----STDERGFCGVSVKSNVGH  339
KS50 GVSPTIGYLEAHASGTPLGDQIEVEALTAVHRR-----TSDDVGYCAIGSAKPVIGHL  346
KS6  GVDPAHIGYLEAHGTGTALGDPVEIEGAVAAFR-----HTTASAFCAVGSAKAHLGHA  344
KS7  GVEPRTVTYVEAHGTGTEVGDPVIEVRGLRKAFALDGRDDEVEPWCLLGTVKSNVGHT  353
KS8  GVDPGTISCVEAHGTGTELGDPIEIEGLARVLGR-----STERPAPVALGSLKPNIGHA  351
KS90 GADARSVQYVEAQATGSPVGDPELAALRRAYDV-----RA-DGHHLRLGSKVPNTGHL  343
KS10 GIRPETVGYVEAHGTGTALGDPIEVTGLRTAFDPPAGSQDHATPRQYCGLGSVKTNVGH  338
      *      :      :*: :*: :*: :*: :*      :*      **

```

Fig 4.2 BGC004 is a *trans*-AT type PKS BGC.

- Detailed PKS domain annotation. Modules 4 and 7 are proposed splitting modules.
- The KS domain alignment reveals the HGTGT motif changes in the KS5 and KS9 domains. HGTGT motif is essential for decarboxylative condensation, and changes in this consensus motif indicated that KS5 and KS9 are KS⁰ type condensation-free domains that do not play a role in chain elongation.

Table 4.1 Predicted functions of the genes on BGC004

orf	aa size	proposed function	orf	aa size	proposed function
<i>ctg1_613</i>	254	Isochorismatase family protein	<i>ctg1_634</i>	4699	SDR family oxidoreductase
<i>ctg1_614</i>	244	branched-chain amino acid ABC transporter substrate-binding protein	<i>ctg1_635</i>	366	LLM class flavin-dependent oxidoreductase
<i>ctg1_615</i>	102	N-acetyltransferase	<i>ctg1_636</i>	271	methyltransferase domain-containing protein
<i>ctg1_616</i>	136	(4Fe-4S)-binding protein	<i>ctg1_637</i>	262	alpha/beta fold hydrolase
<i>ctg1_617</i>	76	NADP-dependent oxidoreductase	<i>ctg1_638</i>	345	ketoacyl-ACP synthase III family protein
<i>ctg1_618</i>	76	NADPH:quinone reductase	<i>ctg1_639</i>	224	hypothetical protein
<i>ctg1_619</i>	129	nuclear transport factor 2 family protein	<i>ctg1_640</i>	515	alanine:cation symporter family protein
<i>ctg1_620</i>	279	alpha/beta hydrolase	<i>ctg1_641</i>	193	TetR/AcrR family transcriptional regulator
<i>ctg1_621</i>	410	glutaminase	<i>ctg1_642</i>	158	VOC family protein
<i>ctg1_622</i>	95	hypothetical protein M877_08590	<i>ctg1_643</i>	541	serine/threonine protein kinase
<i>ctg1_623</i>	157	MarR family transcriptional regulator	<i>ctg1_644</i>	266	SAM-dependent methyltransferase
<i>ctg1_624</i>	309	ABC transporter ATP-binding protein	<i>ctg1_645</i>	398	FAD-dependent monooxygenase
<i>ctg1_625</i>	256	ABC transporter permease subunit	<i>ctg1_646</i>	257	TetR/AcrR family transcriptional regulator
<i>ctg1_626</i>	396	cytochrome P450	<i>ctg1_647</i>	63	NAD(P)-binding domain-containing protein
<i>ctg1_627</i>	947	LuxR family transcriptional regulator	<i>ctg1_648</i>	276	aminoglycoside phosphotransferase family protein
<i>ctg1_628</i>	167	hypothetical protein	<i>ctg1_649</i>	203	hypothetical protein
<i>ctg1_629</i>	170	hypothetical protein	<i>ctg1_650</i>	425	hypothetical protein GA0115243_104331
<i>ctg1_630</i>	473	PfaD family polyunsaturated fatty acid/polyketide biosynthesis protein	<i>ctg1_651</i>	162	hypothetical protein
<i>ctg1_631</i>	287	ACP S-malonyltransferase	<i>ctg1_652</i>	719	tetratricopeptide repeat protein
<i>ctg1_632</i>	4215	SDR family NAD(P)-dependent oxidoreductase	<i>ctg1_653</i>	1085	tetratricopeptide repeat protein
<i>ctg1_633</i>	3406	SDR family NAD(P)-dependent oxidoreductase			

4.2.2 BGC009

The *in silico* analysis of the BGC009 identified one PKS (Ctg27_306, Fig 4.3) and three NRPS (Ctg27_307, Ctg27_308 and Ctg27_314) that spanned over seven modules which had low similarity to any known characterised BGC in the MIBiG database and sequenced genomes in NCBI.⁴⁶ The proposed biosynthesis is initiated by a monomodular PKS (Ctg27_306, Fig 4.3). The monomer prediction for the AT domain indicates that this PKS module will incorporate Malonyl-CoA as building block.⁴⁴ The following monomer is from an NRPS module (on Ctg27_307). However, substrate prediction for the A1 domain based on the Stachelhaus code remains inconclusive.⁴⁵ A C-N bond is then formed between the PKS and NRPS module. The downstream modules (present on Ctg27_307, Ctg27_308 and Ctg27_314) are proposed to assemble the Gly-Thr-Gly-Orn-X. The standalone TEII (type II thioesterase) is anticipated to release the projected compound from the assembly line either in cyclic

or linear form.¹⁰⁵

The gene *ctg27_309* encodes an MbtH homolog that is frequently found in NRPS gene clusters. The *ctg27_309* gene is located just downstream of three PKS/NRPS genes (Fig 4.3). *Ctg27_309* exhibited 45% amino acid identity with DptG (GenAccession: AAX31560.1) and 52 % with LptG (GenAccession: AAZ23079.1), two well-studied MbtH-like proteins. Recent studies indicate that MbtH proteins play an essential role in secondary metabolite biosynthesis, but their specific functions vary from pathway to pathway. *Ctg27_309* may represent a promising target. It would be interesting to see whether expressing this MbtH homolog under a promoter will derepress and/or enhance the yield of the BGC009 targeted metabolisms.

The pathway was captured with shuttle vector pTARa using CRISPR/Cas9-mediated TAR cloning method into *S. cerevisiae* BY4727 Δ NHEJ (YAC009), transferred to *E. coli* S17-1 (BAC009) and conjugated into *S. albus* Del14 and *S. lividans* TK24 to yield *del14_009* and *liv_009*. The predicted functions of the genes on BGC009, gene fragments and primers used in cloning BGC009 are listed in Table 4.2 and Appendix 8.

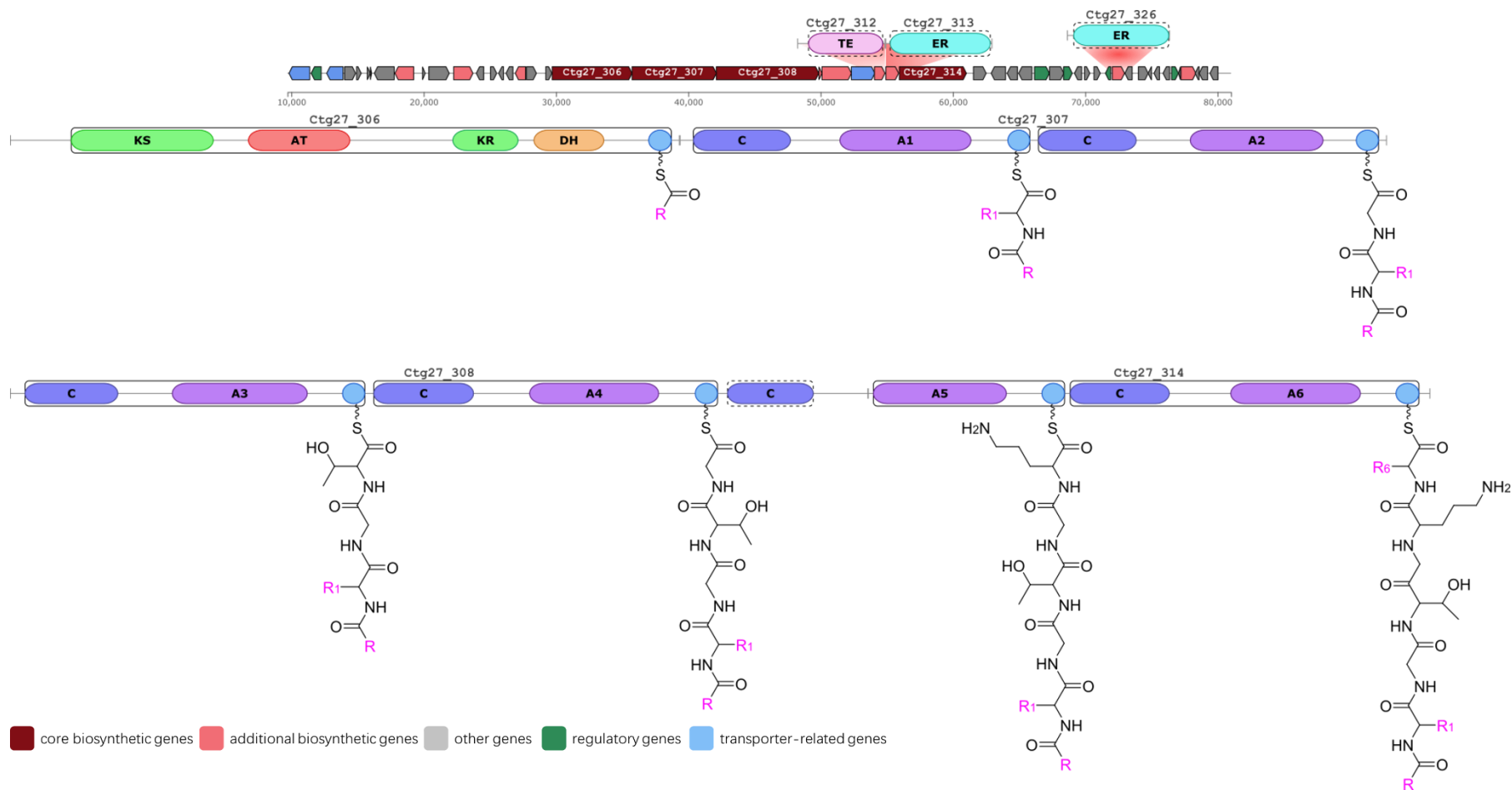


Fig 4.3 Detailed PKS/NRPS prediction for BGC009.

BGC009 is a PKS/NRPS hybrid pathway. The PKS and NRPS monomers are predicted using Minowa⁴⁴ and NRPSpredictor2⁴⁵ embedded in antiSMASH⁴³.

Table 4.2 Predicted functions of the genes on BGC009

orf	aa size	proposed function	orf	aa size	proposed function
<i>ctg27_275</i>	119	DUF4190 domain-containing protein	<i>ctg27_308</i>	2580	amino acid adenylation domain-containing protein
<i>ctg27_276</i>	145	No significant similarity found	<i>ctg27_309</i>	77	MbtH family NRPS accessory protein
<i>ctg27_277</i>	83	hypothetical protein	<i>ctg27_310</i>	742	lantibiotic dehydratase
<i>ctg27_278</i>	174	DDE-type integrase/transposase/recombinase	<i>ctg27_311</i>	591	ABC transporter G family ATP-binding protein/permease
<i>ctg27_279</i>	132	DDE-type integrase/transposase/recombinase	<i>ctg27_312</i>	263	alpha/beta fold hydrolase
<i>ctg27_280</i>	190	alanine racemase	<i>ctg27_313</i>	319	NADPH:quinone oxidoreductase family protein
<i>ctg27_281</i>	219	alanine racemase	<i>ctg27_314</i>	1690	amino acid adenylation domain-containing protein
<i>ctg27_282</i>	422	serine hydrolase	<i>ctg27_315</i>	320	LuxR C-terminal-related transcriptional regulator
<i>ctg27_283</i>	383	Sensor histidine kinase LiaS	<i>ctg27_316</i>	363	peptidoglycan DD-metalloendopeptidase family protein
<i>ctg27_284</i>	226	response regulator transcription factor	<i>ctg27_317</i>	259	D-Ala-D-Ala carboxypeptidase family metallohydrolase
<i>ctg27_285</i>	182	DUF2087 domain-containing protein	<i>ctg27_318</i>	334	Zinc D-Ala-D-Ala carboxypeptidase precursor
<i>ctg27_286</i>	110	DUF6204 family protein	<i>ctg27_319</i>	375	sensory histidine kinase UhpB
<i>ctg27_287</i>	523	MFS transporter	<i>ctg27_320</i>	362	hypothetical protein
<i>ctg27_288</i>	251	transcriptional regulator, TetR family	<i>ctg27_321</i>	218	response regulator transcription factor
<i>ctg27_289</i>	411	MFS transporter	<i>ctg27_322</i>	181	amylo-alpha-1,6-glucosidase
<i>ctg27_290</i>	278	helix-turn-helix domain-containing protein	<i>ctg27_323</i>	133	hypothetical protein SCWH03_58550
<i>ctg27_291</i>	117	cytochrome P450	<i>ctg27_324</i>	158	DUF4265 domain-containing protein
<i>ctg27_292</i>	78	hypothetical protein GCM10010207_75760	<i>ctg27_325</i>	133	MarR family winged helix-turn-helix transcriptional regulator
<i>ctg27_293</i>	30	transposase	<i>ctg27_326</i>	307	zinc-binding dehydrogenase
<i>ctg27_294</i>	499	DUF6056 family protein	<i>ctg27_327</i>	182	IS5 family transposase
<i>ctg27_295</i>	471	glycosyltransferase	<i>ctg27_328</i>	246	hypothetical protein
<i>ctg27_296</i>	66	site-specific integrase	<i>ctg27_329</i>	85	hypothetical protein GA0115234_105628
<i>ctg27_297</i>	514	hypothetical protein	<i>ctg27_330</i>	139	RidA family protein
<i>ctg27_298</i>	486	GlcNAc-PI de-N-acetylase	<i>ctg27_331</i>	166	cupin domain-containing protein
<i>ctg27_299</i>	185	Transposase	<i>ctg27_332</i>	174	MarR family transcriptional regulator
<i>ctg27_300</i>	155	MarR family winged helix-turn-helix transcriptional regulator	<i>ctg27_333</i>	47	hypothetical protein GCM10010207_86980
<i>ctg27_301</i>	122	IS630 family transposase	<i>ctg27_334</i>	374	GNAT family N-acetyltransferase
<i>ctg27_302</i>	173	IS630 family transposase	<i>ctg27_335</i>	100	transposase
<i>ctg27_303</i>	260	type I methionyl aminopeptidase	<i>ctg27_336</i>	213	methyltransferase domain-containing protein
<i>ctg27_304</i>	250	helix-turn-helix domain-containing protein	<i>ctg27_337</i>	200	IS1380 family transposase
<i>ctg27_305</i>	154	glycosyltransferase family 2 protein	<i>ctg27_338</i>	330	CU044_5270 family protein
<i>ctg27_306</i>	2012	SDR family NAD(P)-dependent oxidoreductase	<i>ctg27_339</i>	216	RNA polymerase sigma factor
<i>ctg27_307</i>	2125	amino acid adenylation domain-containing protein			

4.2.3 BGC014

antiSMASH classified BGC014 as an NRPS containing BGC, which showed low similarity to any BGCs in the MIBiG repository and sequenced genomes deposited in NCBI at the time of analysis. The BGC014 possesses three genes (*ctg1_87*, *ctg1_85*, *ctg1_81*) that encode standard NRPS. A total of five A domains were identified: two on Ctg1_87, two on Ctg1_85, and one on Ctg1_81. However, the Stachelhaus codes of some A domains failed to render confident NRPS monomer predictions. The two NRPS Ctg1_87 and Ctg1_85 are separated by the *ctg1_86* encoding MbtH protein. Moreover, the NRPS Ctg1_81 is separated from Ctg1_85-87 by genes (Fig 4.4a region highlighted in yellow) involved in Dab (diaminobutyric acid) synthesis. Similar Dab biosynthesis gene sets can be found on friulimicin¹⁰⁶ and laspartomycin¹⁰⁷ BGCs.

Further analysis of the upstream genes revealed that Ctg1_88 lacks an A domain while the C and PCP domains remain (Fig 4.4a). The unconventional module structure (loss of the critical A domain) suggests that the substrates incorporated in this module should be supplemented in *trans*. There are multiple routes for supplying substrates to this monomodule. For example, via a *trans*-acting A domain located on (Ctg1_71, Ctg1_92, Ctg1_93, Table 4.3) or outside the cloned region; through intramodular substrate movement as reported in thalassospiramide A biosynthesis¹⁰⁸; or by uptaking the Dab synthesised from Ctg1_84-Ctg1_82 in a way similar to that observed in friulimicin¹⁰⁶ and laspartomycin¹⁰⁷. The interesting monomodular *ctg1_88* is proposed to increase the structural diversity.

The pathway was captured with shuttle vector pTARa using CRISPR/Cas9-mediated TAR cloning method into *S. cerevisiae* BY4727 Δ NHEJ (YAC014), transferred to *E. coli* S17-1 (BAC014) and conjugated into *S. albus* Del14 and *S. lividans* TK24 to yield del14_014 and liv_014. The predicted functions of the genes on BGC014, gene fragments and primers used in cloning BGC014 are listed in Table 4.3 and Appendix 8.

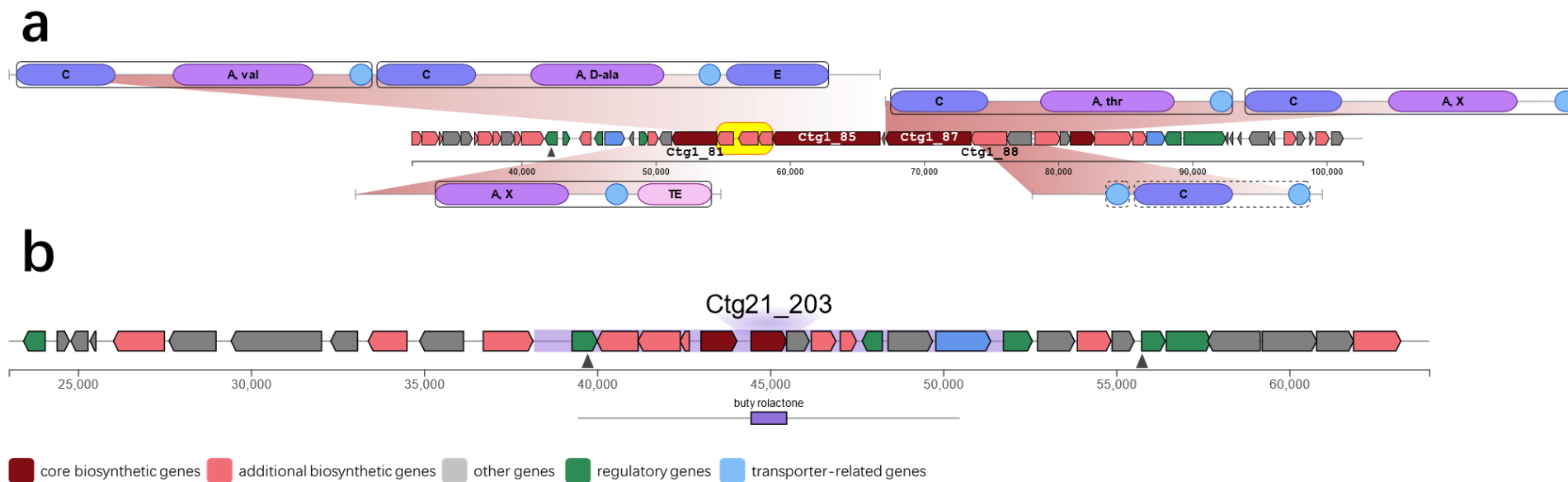


Fig 4.4 (a) Detailed NRPS prediction for BGC014 and (b) gene organisation of BGC027

Table 4.3 Predicted functions of the genes on BGC014

orf	aa size	proposed function	orf	aa size	proposed function
<i>ctg1_49</i>	243	transcription antitermination regulator	<i>ctg1_82</i>	411	ATP-grasp domain-containing protein
<i>ctg1_50</i>	142	VOC family protein	<i>ctg1_83</i>	490	argininosuccinate lyase
<i>ctg1_51</i>	410	ATP-grasp domain-containing protein	<i>ctg1_84</i>	356	pyridoxal-phosphate dependent enzyme
<i>ctg1_52</i>	428	MFS transporter	<i>ctg1_85</i>	2671	non-ribosomal peptide synthetase
<i>ctg1_53</i>	417	ATP-grasp domain-containing protein	<i>ctg1_86</i>	67	MbtH family protein
<i>ctg1_54</i>	421	ATP-grasp domain-containing protein	<i>ctg1_87</i>	2125	hypothetical protein GCM10018790_13500
<i>ctg1_55</i>	426	aminotransferase class III-fold pyridoxal phosphate-dependent enzyme	<i>ctg1_88</i>	889	condensation domain-containing protein
<i>ctg1_56</i>	526	amino acid adenylation domain-containing protein	<i>ctg1_89</i>	612	hypothetical protein
<i>ctg1_57</i>	392	acyl-CoA/acyl-ACP dehydrogenase	<i>ctg1_90</i>	626	asparagine synthase (glutamine-hydrolyzing)
<i>ctg1_58</i>	88	acyl carrier protein	<i>ctg1_91</i>	255	aspartate/glutamate racemase family protein
<i>ctg1_59</i>	417	pyridoxal phosphate-dependent aminotransferase	<i>ctg1_92</i>	599	non-ribosomal peptide synthetase
<i>ctg1_60</i>	438	lysine 2,3-aminomutase	<i>ctg1_93</i>	934	non-ribosomal peptide synthetase 3
<i>ctg1_61</i>	249	thioesterase domain-containing protein	<i>ctg1_94</i>	345	TauD/TfdA family dioxygenase
<i>ctg1_62</i>	416	glutamate-5-semialdehyde dehydrogenase	<i>ctg1_95</i>	457	MFS transporter
<i>ctg1_63</i>	93	acyl carrier protein	<i>ctg1_96</i>	415	helix-turn-helix domain-containing protein
<i>ctg1_64</i>	449	TrpB-like pyridoxal phosphate-dependent enzyme	<i>ctg1_97</i>	1037	tetratricopeptide repeat protein
<i>ctg1_65</i>	280	hypothetical protein	<i>ctg1_98</i>	63	hypothetical protein
<i>ctg1_66</i>	86	acyl carrier protein	<i>ctg1_99</i>	82	hypothetical protein
<i>ctg1_67</i>	371	acyl-CoA dehydrogenase	<i>ctg1_100</i>	81	chaplin
<i>ctg1_68</i>	200	4'-phosphopantetheinyl transferase superfamily protein	<i>ctg1_101</i>	500	multicopper oxidase domain-containing protein
<i>ctg1_69</i>	323	hypothetical protein	<i>ctg1_102</i>	125	VOC family protein
<i>ctg1_70</i>	174	flavin reductase family protein	<i>ctg1_103</i>	296	SDR family oxidoreductase
<i>ctg1_71</i>	553	amino acid adenylation domain-containing protein	<i>ctg1_104</i>	190	DNA starvation/stationary phase protection protein
<i>ctg1_72</i>	306	LysR family transcriptional regulator	<i>ctg1_105</i>	86	DUF5133 domain-containing protein
<i>ctg1_73</i>	170	MarR family winged helix-turn-helix transcriptional regulator	<i>ctg1_106</i>	332	SDR family NAD(P)-dependent oxidoreductase
<i>ctg1_74</i>	282	SDR family NAD(P)-dependent oxidoreductase	<i>ctg1_107</i>	290	SigB/SigF/SigG family RNA polymerase sigma factor
<i>ctg1_75</i>	201	TetR/AcrR family transcriptional regulator	<i>ctg1_108</i>	500	SulP family inorganic anion transporter
<i>ctg1_76</i>	495	MFS transporter	<i>ctg1_109</i>	128	MerR family transcriptional regulator
<i>ctg1_77</i>	107	glutaminase	<i>ctg1_110</i>	179	transcriptional regulator
<i>ctg1_78</i>	214	response regulator transcription factor	<i>ctg1_111</i>	438	MFS transporter
<i>ctg1_79</i>	258	alpha/beta fold hydrolase	<i>ctg1_112</i>	177	TerD family protein
<i>ctg1_80</i>	308	kinase	<i>ctg1_113</i>	437	Na ⁺ /H ⁺ antiporter NhaA
<i>ctg1_81</i>	1120	amino acid adenylation domain-containing protein	<i>ctg1_114</i>	320	DUF389 domain-containing protein

4.2.4 BGC016

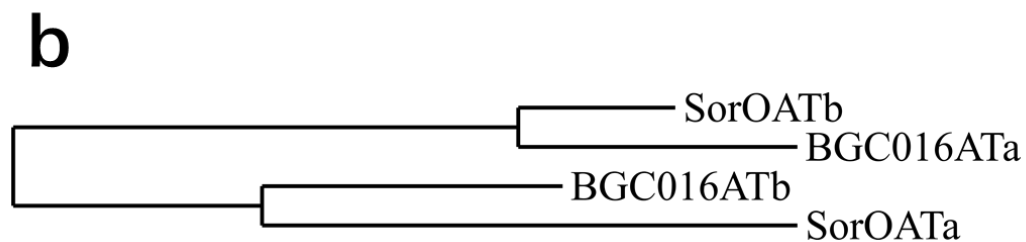
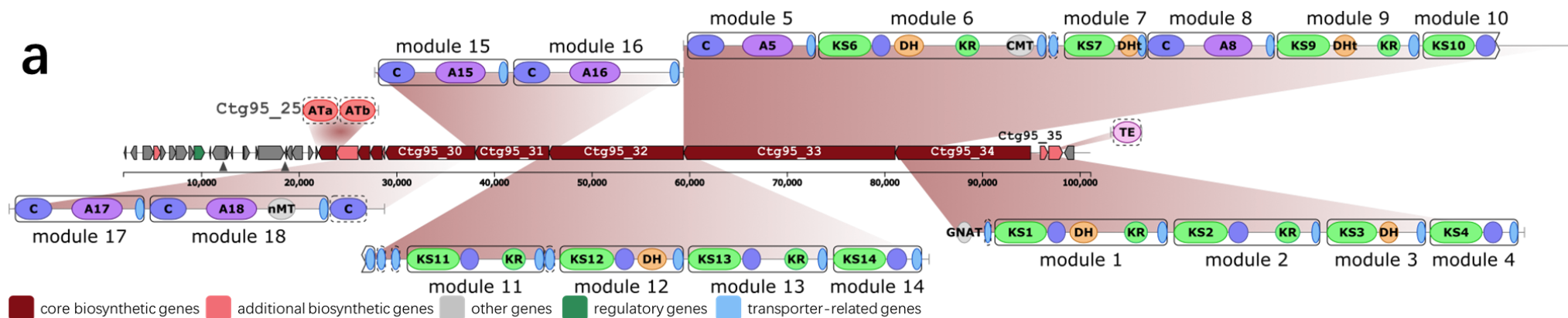
An antiSMASH analysis identified five core biosynthetic genes on BGC016, including two AT-less PKS (Ctg95_34, Ctg95_32); a hybrid AT-less PKS/NRPS (Ctg95_33); two NRPS (Ctg95_31, Ctg95_30) (Fig 4.5a). The presence of two standalone ATs (ATa and ATb) within a single protein (Ctg95_25) suggests that this BGC is a *trans*-AT NRPS/PKS hybrid (Fig 4.5a). Similar tandem AT architecture can be found in SorO on sorangicin BGC. The alignment of the two BGC016 ATs with Sor ATs (Fig 4.5b) revealed that BGC016 ATa clades with Sor ATb, an AT domain exhibiting a substrate preference for malonate. And BGC016 ATb clustered with Sor ATa, an AH (acyl hydrolase) domain that is responsible for proofreading.¹⁰³

The presence of the GNAT domain on protein Ctg95_34 indicates that this protein initiates the PKS part biosynthesis. The GNAT-like domain was first identified and characterised from curacin BGC, which displayed decarboxylation activity involved in stater group generation.¹⁰⁹ Other *trans*-AT characteristics, such as module splitting (module 10, Fig 4.5a), and non-extending modules (Fig 4.5c), can also be found in the PKS part. However, *in silico* analysis cannot deduce whether the NRPS or PKS part initiated the biosynthesis. It may be possible that nonlinear assembly processes take place during biosynthesis.

The captured pathway was conjugated into *S. albus* Del14 and *S. lividans* TK24 to yield del14_016 and liv_016. The predicted functions of the genes on BGC016 are listed in Table 4.4.

Table 4.4 Predicted functions of the genes on BGC016

orf	aa size	proposed function	orf	aa size	proposed function
<i>ctg95_12</i>	36	IS701 family transposase	<i>ctg95_28</i>	407	polyketide beta-ketoacyl:ACP synthase
<i>ctg95_13</i>	441	IS701 family transposase	<i>ctg95_29</i>	83	acyl carrier protein
<i>ctg95_14</i>	99	hypothetical protein	<i>ctg95_30</i>	3059	non-ribosomal peptide synthetase
<i>ctg95_15</i>	57	helix-turn-helix domain-containing protein	<i>ctg95_31</i>	2516	non-ribosomal peptide synthetase
<i>ctg95_16</i>	48	No significant similarity found	<i>ctg95_32</i>	4596	SDR family NAD(P)-dependent oxidoreductase
<i>ctg95_17</i>	181	tyrosine-type recombinase/integrase	<i>ctg95_33</i>	7223	non-ribosomal peptide synthetase
<i>ctg95_18</i>	52	hypothetical protein	<i>ctg95_34</i>	4624	Polyketide synthase PksL
<i>ctg95_19</i>	874	DUF262 domain-containing protein	<i>ctg95_35</i>	252	alpha/beta fold hydrolase
<i>ctg95_20</i>	57	hypothetical protein	<i>ctg95_36</i>	471	amidase
<i>ctg95_21</i>	182	hypothetical protein	<i>ctg95_37</i>	319	pentapeptide repeat-containing protein
<i>ctg95_22</i>	356	ParA family protein	<i>ctg95_38</i>	38	No significant similarity found
<i>ctg95_23</i>	147	tyrosine-type recombinase/integrase	<i>ctg95_39</i>	108	IS1380 family transposase
<i>ctg95_24</i>	64	hypothetical protein EES40_09560	<i>ctg95_40</i>	123	IS4 family transposase
<i>ctg95_25</i>	618	ACP S-malonyltransferase	<i>ctg95_41</i>	201	IS4 family transposase
<i>ctg95_26</i>	739	enoyl-CoA hydratase/isomerase	<i>ctg95_42</i>	221	transposase
<i>ctg95_27</i>	402	hydroxymethylglutaryl-CoA synthase family protein	<i>ctg95_43</i>	240	NUDIX hydrolase



c

KS1	LADAGVEPSGVSYVEVHGTGTALGDPIEVQGLTRAFAGS-----ATPCGLGSVKT	334
KS2	LDEAGVDPRAISYVEAHGTGTLGDPIEITGLSQAFGAATGS-----ADNQYCHLGS	344
KS3 ⁰	LRQAGRAPRDVRYLEVNGSGSQLTDLLELKAVQAVYRDGEPT-----AAPLHLGSMKP	324
KS4	LDRAGVDPASLGYVEAHGTGTEIGDLMEVEALERAFTARD-----RTGFCALGSVKS	337
KS6	LERTGVDARSIGYVEAHGTGTLGDPVELAALTEAYREATDE-----TGYCGIGSVKS	341
KS7 ⁰	LEKAGVEPAGVSYVEAHGTATALGDSIEISALSRAFGTAR-----AGQYCAVGS	318
KS9	LDRAGVHPETIGYLEAHGTGTRLGDPPIEVMAVSEAYRRYTDR-----TGFCGIGSVKS	340
KS10	YERGGIDPRDLGHLVTHGTGTLGDPVEVNALRAAFRTDAGV-----TERGFCALTSTKT	331
KS11	YDKFGIDPADIRLVEAHGTGTALGDPVEIDGLVESFGAHTEE-----RGYCAIGSVKS	338
KS12	YRQAGIDPRTVTYVEAHGTGTELGDPVELGGLKSAFAALAGETPDGLGEPARCGIGSVKS	347
KS13	LAEAEVPPAALSYLEAHGTGTALGDPIEIAGLVKAFFRAGGG-----ALPQSLAIGSVKS	332
KS14 ⁰	LERAGAVPESIGYVEAAANGSALGDAVEFSALREVFGAV-----PEPVALGSVKS	336
	: : : * : * . . : : : * *	

Fig 4.5 BGC016 is a *trans*-AT PKS/NRPS hybrid BGC

- a. Detailed PKS domain annotation.
- b. Alignment of the two BGC016 ATs against Sor ATs
- c. The alignment of the KS domain reveals the changes of the HGTGT motif in the KS3, KS7 and KS14 domains

4.2.5 BGC027

antiSMASH revealed a butyrolactone-like region on BGC027 as well as a homolog of AfsA, Ctg21_203. AfsA-like proteins are the key enzymes involved in the biosynthesis of γ -butyrolactones and γ -butenolides that are widely distributed among *Streptomyces*.^{110,111} γ -butyrolactones and γ -butenolides are diffusible signal molecules that exhibit regulatory functions that mediate biological processes such as morphological differentiation and antibiotic production. The regulatory cascades triggered by these compounds are diverse and complex.¹¹⁰

BGC027 harbour a candidate system that possesses multiple additional biosynthesis genes (Fig 4.4b, genes coloured in pink), suggesting the capability of generating highly modified signalling molecules. Understanding how these signalling molecules function may provide insights into eliciting/derepressing silent biosynthetic pathways.

The pathway was captured with shuttle vector pTARa using CRISPR/Cas9-mediated TAR cloning method into *S. cerevisiae* BY4727 Δ NHEJ (YAC027), transferred to *E. coli* S17-1 (BAC027). Unfortunately, this pathway was failed to be conjugated into *S. albus* J1074, *S. albus* Del14 or *S. lividans* TK24, even after multiple attempts, so it was excluded from the metabolic profiling. It is worth noting that this pathway is not particularly large, so perhaps the failure to transfer into a streptomyces host indicates that it might produce a toxic product that is killing any successful exconjugants. Future work might involve placing the pathway under the control of inducible promoters and then attempting conjugation into a heterologous host.

The predicted functions of the genes on BGC027, gene fragments and primers used in cloning BGC027 are listed in Table 4.5 and Appendix 8, respectively.

Table 4.5 Predicted functions of the genes on BGC027

orf	aa size	proposed function	orf	aa size	proposed function
<i>ctg21_181</i>	673	endo-beta-N-acetylglucosaminidase	<i>ctg21_209</i>	527	MFS transporter
<i>ctg21_182</i>	418	cellulase family glycosylhydrolase	<i>ctg21_210</i>	278	hypothetical protein EF917_16495
<i>ctg21_183</i>	300	carbohydrate ABC transporter permease	<i>ctg21_211</i>	357	SMP-30/gluconolactonase/LRE family protein
<i>ctg21_184</i>	314	sugar ABC transporter permease	<i>ctg21_212</i>	326	aldo/keto reductase
<i>ctg21_185</i>	430	extracellular solute-binding protein	<i>ctg21_213</i>	214	NAD(P)H-binding protein
<i>ctg21_186</i>	348	LacI family transcriptional regulator	<i>ctg21_214</i>	226	response regulator transcription factor
<i>ctg21_187</i>	209	hypothetical protein GCM10010512_04220	<i>ctg21_215</i>	430	HAMP domain-containing histidine kinase
<i>ctg21_188</i>	118	DoxX family protein	<i>ctg21_216</i>	506	histidine ammonia-lyase
<i>ctg21_189</i>	164	GNAT family N-acetyltransferase	<i>ctg21_217</i>	524	pyridine nucleotide-disulfide oxidoreductase
<i>ctg21_190</i>	57	hypothetical protein	<i>ctg21_218</i>	357	alpha-hydroxy-acid oxidizing protein
<i>ctg21_191</i>	491	SDR family NAD(P)-dependent oxidoreductase	<i>ctg21_219</i>	455	cytochrome P450
<i>ctg21_192</i>	453	wax ester/triacylglycerol synthase family O-acyltransferase	<i>ctg21_220</i>	422	cation:proton antiporter
<i>ctg21_193</i>	871	PA14 domain-containing protein	<i>ctg21_221</i>	483	long-chain fatty acid--CoA ligase
<i>ctg21_194</i>	258	(2Fe-2S)-binding protein	<i>ctg21_222</i>	514	NAD(P)-binding domain-containing protein
<i>ctg21_195</i>	373	epoxide hydrolase	<i>ctg21_223</i>	152	MarR family transcriptional regulator
<i>ctg21_196</i>	422	transglycosylase family protein	<i>ctg21_224</i>	416	MFS transporter
<i>ctg21_197</i>	479	NADP-dependent phosphogluconate dehydrogenase	<i>ctg21_225</i>	340	sigma-70 family RNA polymerase sigma factor
<i>ctg21_198</i>	243	response regulator transcription factor	<i>ctg21_226</i>	375	ABC transporter permease
<i>ctg21_199</i>	395	hypothetical protein EF917_16550	<i>ctg21_227</i>	238	ABC transporter ATP-binding protein
<i>ctg21_200</i>	407	beta-ketoacyl-[acyl-carrier-protein] synthase family protein	<i>ctg21_228</i>	203	TetR/AcrR family transcriptional regulator
<i>ctg21_201</i>	88	acyl carrier protein	<i>ctg21_229</i>	409	hypothetical protein
<i>ctg21_202</i>	345	ketoacyl-ACP synthase III family protein	<i>ctg21_230</i>	41	No significant similarity found
<i>ctg21_203</i>	342	hypothetical protein	<i>ctg21_231</i>	334	Gfo/Idh/MocA family oxidoreductase
<i>ctg21_204</i>	218	HAD family phosphatase	<i>ctg21_232</i>	300	sugar phosphate isomerase/epimerase
<i>ctg21_205</i>	234	3-oxoacyl-ACP reductase FabG	<i>ctg21_233</i>	302	ATP-binding cassette domain-containing protein
<i>ctg21_206</i>	154	ester cyclase	<i>ctg21_234</i>	356	ABC transporter permease
<i>ctg21_207</i>	194	TetR family transcriptional regulator	<i>ctg21_235</i>	337	sugar ABC transporter substrate-binding protein
<i>ctg21_208</i>	427	alpha/beta hydrolase			

4.2.6 BGC218-3

antiSMASH analysis of BGC218-3 revealed four NRPS-encoding core biosynthetic genes (*ctg9_273*, *ctg9_274*, *ctg9_275* and *ctg9_276*), which composed a total of 13 modules (Fig 4.6a). The BGC218-3 shares many similarities with daptomycin¹¹² and A54145¹¹³, two extensively studied CDAs (Ca²⁺-dependent cyclic lipopeptides). The fatty acyl-AMP ligase (Ctg9_271) and the acyl carrier protein (Ctg9_272) are located upstream of the four NRPS genes. The *ctg9_217* and *ctg9_272* counterparts on daptomycin BGC (*dptE* and *dptF*) and A54145 BGC (*lptEF*) are involved in the fatty acylation of the first amino acid incorporated during NRPS biosynthesis. Auxiliary genes such as ABC transporters and MbtH protein can also be found among all three BGCS (Fig 4.6b).^{112–114}

Detailed Stachelhaus analysis of the A domains revealed several structural features conserved in BGC218-3, daptomycin, and A54145: (1) amino acids at positions 2,9,10 are conserved in D-configuration, Gly and Asp analogues, respectively; (2) the presence of a canonical tetrapeptide motif Asp-X-Asp-Gly for Ca²⁺ binding (Table 4.6).^{112–114}

Discrepancies in A domain substrate specificity and the common features shared among BGC218-3, daptomycin and A54145 suggested BGC218-3 might encode a series of compounds that are closely related to the reported CDAs which have different amino acid compositions in the peptide backbone.

The pathway was captured with shuttle vector pTARa using CRISPR/Cas9-mediated TAR cloning method into *S. cerevisiae* BY4727 Δ NHEJ (YAC218-3), transferred to *E. coli* S17-1 (BAC218-3) and conjugated into *S. albus* Del14 and *S. lividans* TK24 to yield del14_218-3 and liv_218-3. The predicted functions of the genes on BGC218-3, and primers used in cloning BGC218-3 are listed in Table 4.7 and Appendix 8, respectively.

Table 4.6 Stachelhaus analysis of the A domains on BGC218-3, A5145 and Daptomycin
Stachelhaus sequences were extracted using NRPSPredictor2. The Stachelhaus sequences of A5145 and Daptomycin were used to predict the A domain substrate of BGC218-3.
Amino acids written in lowercase indicated that the amino acids in these positions are not conserved.

position	(Stachelhaus sequence) Substrate		
	BGC218-3	A5145	Daptomycin
1	(DAycVAAVeK) Ala	(DVALVGVVQK) Trp	(DVSSIGAVEK) Trp
2	(DmaKVASVNK) D-X	(DLVKVASVNK) D-Glu	(DLTKLGDVNK) D-Asn
3	(DLTKVGDVNK) hAsn	(DLTKVGDVNK) hAsn	(DLTKLGAVNK) Asp
4	(DFWSVGMVHK) Thr	(DFWSVGMVHK) Thr	(DFWSVGMVHK) Thr
5	(DILQLGVIWK) Sar	(DILQLGVIWK) Sar	(DILQLGVIWK) Gly
6	(DvFcVAAVyK) Ala	(DvFnLaIVFK) Ala	(DtwDmGyVDK) Orn
7	(DLTpVGAVNK) Asp	(DLTKvGAVNK) Asp	(DLTKLGAVNK) Asp
8	(DAwDaGtVDK) D-Lys	(DAwDaGtVDK) D-Lys	(DvwSaAfVYK) D-Ala
9	(DLTKIGAVNK) Asp	(DLTKIGAVNK) MeAsp	(DLTKLGAVNK) Asp
10	(DILQLGLVWK) Gly	(DILQLGLVWK) Gly	(DILQVGmiWK) Gly
11	(DLTKVGDVsK) D-Asn	(DLTKVGDVsK) D-Asn	(DVWHISLVDK) D-ser
12	(DIGKTGVvnK) Glu	(DIGKTGVvnK) 3-MeGlu	(DIGKTGVinK) 3-MeGlu
13	(DVQYvgHVVK) Pro	(DglFvGiAVK) Ile	(DAWTttGVgK) Kyn
	monomer predicted	monomer observed	

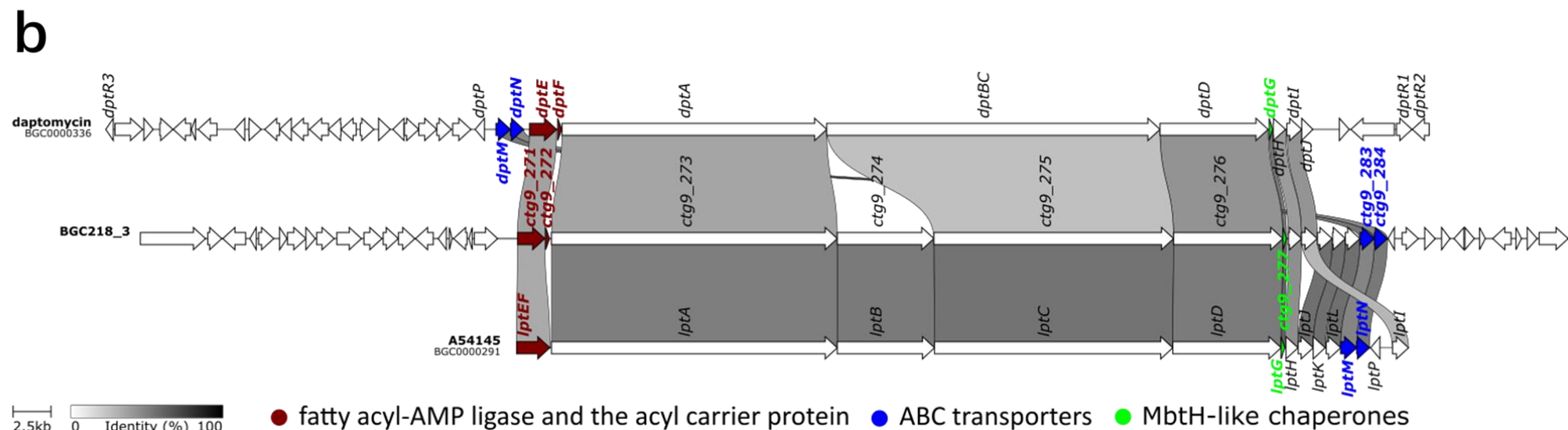
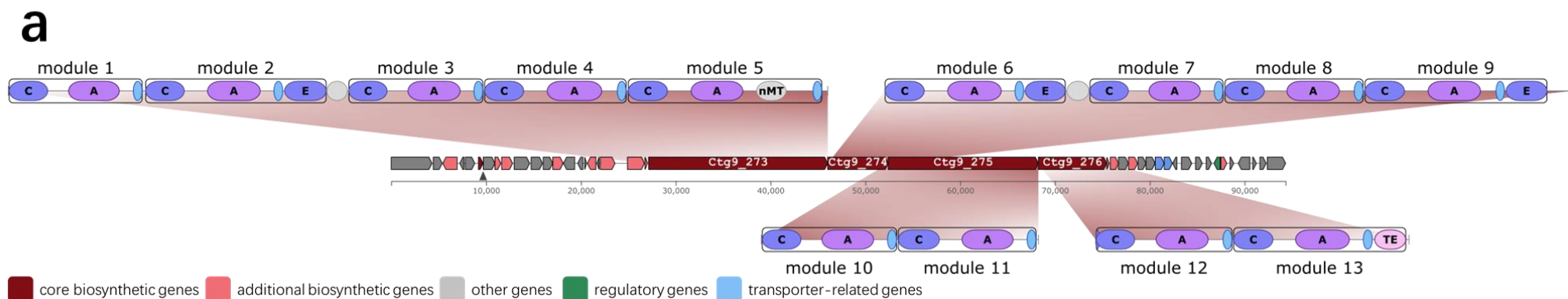


Fig 4.6 (a) Detailed NRPS domain organisation for BGC218-3 and (b) gene cluster comparison of BGC218-3, daptomycin and A5414

Table 4.7 Predicted functions of the genes on BGC218-3

orf	aa size	proposed function
<i>ctg9_271</i>	599	fatty acyl-AMP ligase
<i>ctg9_272</i>	85	acyl carrier protein
<i>ctg9_273</i>	6292	non-ribosomal peptide synthetase
<i>ctg9_274</i>	2128	non-ribosomal peptide synthetase
<i>ctg9_275</i>	5277	amino acid adenylation domain-containing protein
<i>ctg9_276</i>	2406	non-ribosomal peptide synthetase
<i>ctg9_277</i>	78	protein mbtH
<i>ctg9_278</i>	289	alpha/beta hydrolase
<i>ctg9_279</i>	347	methyltransferase domain-containing protein
<i>ctg9_280</i>	320	TauD/TfdA family dioxygenase
<i>ctg9_281</i>	280	FkbM family methyltransferase
<i>ctg9_282</i>	319	TauD/TfdA family dioxygenase
<i>ctg9_283</i>	315	ATP-binding cassette domain-containing protein
<i>ctg9_284</i>	286	transport permease protein
<i>ctg9_285</i>	140	hypothetical protein

4.3 Refactoring pathways

During our first capture-fermentation-test cycle (discussed in chapter 4.4), not many pathways produce plausible ions. In addition to attempting the expression of cloned pathways, two cloned pathways were further refactored to induce the production of the metabolites they encode.

4.3.1 004ΔLuxR

Further analysis of the cluster BGC004 revealed four putative regulators. Two of these, Ctg1_652 and Ctg1_653, were annotated as SARPs (Streptomyces antibiotic regulatory proteins). Ctg1_627 was annotated as a LuxR transcriptional regulator, and *ctg1_628* encodes a member of the well-studied MarR repressor family of proteins (Fig 4.2, Table 4.1).¹¹⁵

Several pathways are known to be positively regulated by their pathway specific SARPs. However, we could not detect any SARP binding site pattern (heptameric repeat located upstream of the -10 region) along this pathway⁵⁶, suggesting these SARPs are not part of the pathway's regulation system.

LuxR can function as an activator or a repressor.¹¹⁶ We hence generated a Ctg1_627-628 (LuxR-MarR) knockout BAC model (the process is depicted in 4.5.1, Fig 4.10) to

examine the role LuxR plays in this pathway and try to increase/derepress the heterologous expression. The refactored BAC was then conjugated to *S. albus* Del14 and *S. lividans* TK24 to yield del14_004ΔLuxR and liv_004ΔLuxR.

4.3.2 HygE218-3

Our first attempt at awakening BGC218-3 biosynthesis was to add the strong constitutive *ermE** promoter directly in front of *ctg9_271* (the start of the BGC) using the Red/ET system (a similar process is depicted in 4.5.1, Fig 4.7a). The refactored BAC was then conjugated to *S. albus* Del14 and *S. lividans* TK24 to yield *del14_HygE218-3* and *liv_HygE218-3*, respectively.

4.3.3 pIJEF_218-3

An earlier study showed that overexpression of the DptE and DptF increased daptomycin production.¹¹⁷ Therefore, in the present study, we also tried overexpression of DptEF homologs *Ctg9_271&272* in *S. albus* and *S. lividans* heterologous system.

We cloned *ctg9_271* and *ctg9_272* into the ϕ BT1-based integration conjugation vector pIJ10257 and integrated the resultant plasmid pIJEF (Fig 4.7b) onto the chromosomes of *del14_BGC218-3* and *liv_BGC218-3*. The co-integration of BGC218-3 with *Ctg9_271/272* allowed for the expression of BGC218-3 and overexpressing the lipidation module under a strong consecutive promoter simultaneously (Fig 4.7c).

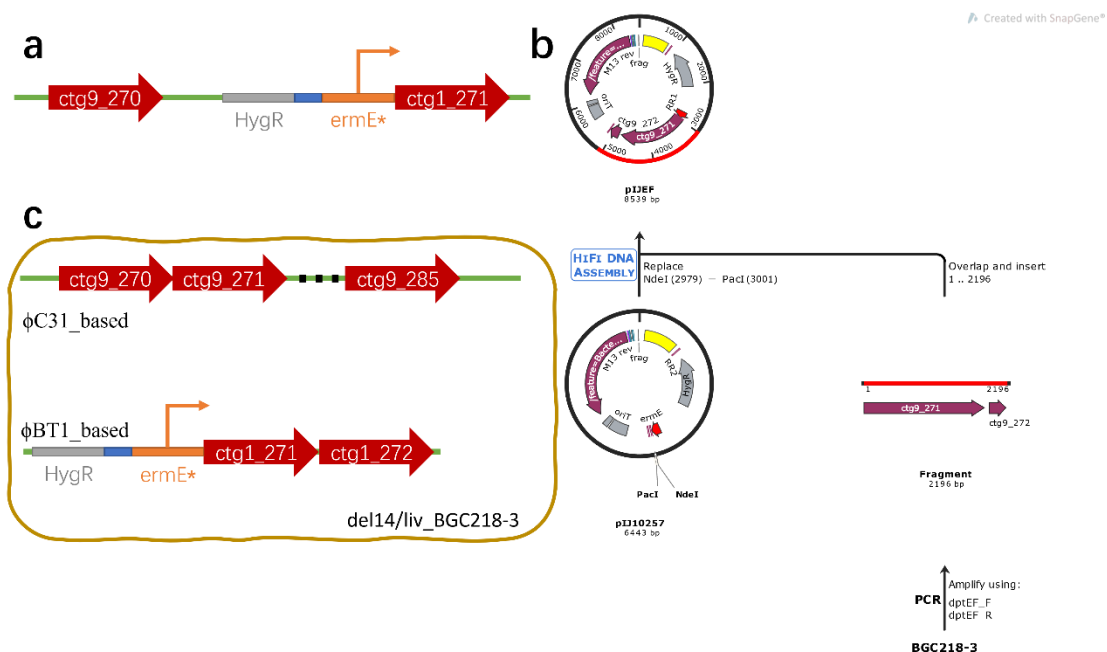


Fig 4.7 BGC218-3 pathway refactoring strategies

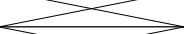
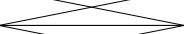
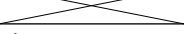


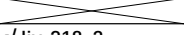
- Constitutive *ermE** promoter was added directly in front of *ctg9_271* (the start of the BGC) to yield HygE218-3
- DptEF homologs *Ctg9_271*&*272* were cloned to yield pIJEF
- The resultant pIJEF was integrated onto the chromosomes of *del14_BGC218-3* and *liv_BGC218-3* at ϕ BT1 site.

4.4 Heterologous expression of pathways and identification of the products

In order to determine whether any of the cloned pathways or their refactored derivatives were producing new compounds, a comparative LC-MS/MS approach was adopted.

Each of the pathways was introduced into one of the two hosts (*S. albus* Del14, *S. lividans* TK24) and cultivated in triplicate on four different solid media. An empty vector control was also included. The crude extracts from desired heterologous expression hosts (listed in Table 4.8) were subjected to HRESIMS/MS for GNPS molecular networking. An automated, python-assisted analytical procedure was then run to identify molecular features (unique ions in unique clusters) that were found in all three triplicates for a particular pathway but not in any other pathways. Such features were deemed to be potential products of the pathway and were subject to further investigation.

Table 4.8 Summary of cloned pathways (lengths, chemical classes) and heterologous expression systems examined in this thesis

	cloned region, length (kb)	chem_class	in <i>S. albus</i> Del14?	in <i>S. lividans</i> TK24?
004	ctg1_631-ctg1_653, 80.216	transAT-PKS,PKS-like	√ del14_004	√ liv_004
004ΔLuxR			√ del14_004ΔLuxR	√ liv_004ΔLuxR
005	ctg4_359-ctg4_427, 77.603	T2PKS,PKS-like	√ del14_005	
005ΔPKS			√ del14_005ΔPKS	
005ΔAbXO			√ del14_005ΔAbXO	
009	ctg27_275-ctg27_339, 81.413	T1PKS,NRPS	√ del14_009	√ liv_009
014	ctg1_48-ctg1_114, 92.151	NRPS	√ del14_014	√ liv_014
016	ctg95_13-ctg95_43, 94.250	transAT-PKS,T3PKS,PKS-like,NRPS	√ del14_016	√ liv_016
027	ctg21_181-ctg21_235, 67.339	PKS-like,butyrolactone		
031	ctg9_28-ctg9_62, 35.397	nucleoside	√ del14_031	
218-3	ctg9_271-ctg9_285, 62.115	NRPS	√ del14_218-3	√ liv_218-3
HygE218_3			√ del14_HygE218_3	√ liv_HygE218_3
pIJEF218_3			√ del14_pIJEF218_3	√ liv_pIJEF218_3

The left part of figure 4.8 illustrates our original workflow for GNPS analysis. The LCMS/MS files (Fig 4.8①) were grouped (based on the constructs and fermentation medium) and submitted to GNPS for molecular networking, and the resulting network file (graphml, Fig 4.8②) was visualised via Cytoscape. The molecular network was used to group metabolites (detected ions) from pathways and the negative control (pTARa), and we aimed to find clusters formed only by pathways metabolites (circled clusters in Fig 4.8③). The molecular ions from those "stood out" clusters were then manually extracted and inspected from the raw LC-MS/MS data using MassHunter (Fig 4.8④).

The molecular networks generated in this study were large and complicated, making visual analysis time-consuming and prone to errors. Therefore, we developed a workflow in Python to automatically inspect the network and extract the picked-out ions from the LC-MS data for downstream analysis.

With the Python assisted analysis workflow, instead of using a visual graphml file (Fig 4.8③) of the GNPS generated network, we used a clustersummary file generated by GNPS analysis, a "text version" of the molecule network (Fig 4.8⑤). The clustersummary can be parsed as a CSV file. The original clustersummary file is quite informative but a bit complicated (Fig 4.8⑤), so we used Python to extract the information we needed for downstream analysis. A row in the simplified summary file corresponds to a node in the graphical network, and columns of each row refer to the node's attributes (Fig 4.8⑥). The componentindex cell records the composition of the cluster. When the componentindex is equal to -1, this indicates that the node is isolated and is not connected to any other nodes on the network. On the other hand, nodes within the same cluster will have the same componentindex. The default groups indicate the source of a node. For example, Cluster246 has four nodes, where the metabolites from G2 (Pathway 1) constitute the nodes (Fig 4.8⑥).

Our algorithm, written in Python, seeks to find clusters (componentindex $\neq -1$) composed by a single group (clusters constituted from the same colour, like the circled clusters in Fig 4.8③). The ions of the picked-out clusters were then automatically extracted from the mzXML files, and the spectra were plotted with the script adapted from Pyteomics¹¹⁸ (Fig 4.8⑦). Analysis of those plotted spectra (Fig 4.8⑧) can help to confirm:

- The ion is reproducible over replicas
- The ion inspected is "meaningful": not a false-positive signal resulting from any algorithmic artifacts of GNPS nor abundance differences among LC-MS data.
- The production level of this ion is plausible for downstream studies.

The Python workflow enabled us to quickly locate and verify feature metabolites from the complex molecular network, avoid the neglect of ions in the visual inspection process, and save efforts such as manually examining and transferring parameters among different software (e.g., Cytoscape and MassHunter Qualitative Analysis) during the analysis.

① .mzXML

G1: pTARa_medium_R1 pTARa_medium_R2 pTARa_medium_R3	G3: pathway2_medium_R1 pathway2_medium_R2 pathway2_medium_R3
G2: pathway1_medium_R1 pathway1_medium_R2 pathway1_medium_R3	G4: pathway3_medium_R1 pathway3_medium_R2 pathway3_medium_R3

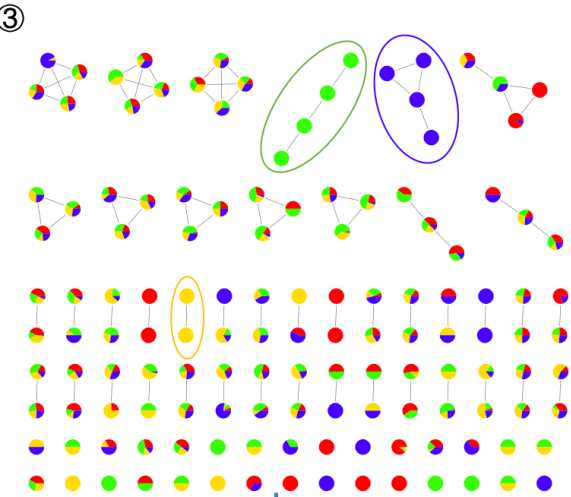
DefaultGrp	EvenOdd	G1	G2	G3	G4	G5	G6	GNPSLink	GNPSLinkLibraryID	MQScore	MZErrorPF	MassDiff
G1,G2,G3,G4	0	8	30	18	10	0	0	https://gnps.org/https://gnps.org/	N/A	N/A	N/A	N/A
G2,G3,G4	0	0	2	17	4	0	0	https://gnps.org/https://gnps.org/	N/A	N/A	N/A	N/A
G2,G4	1	0	1	0	1	0	0	https://gnps.org/https://gnps.org/	N/A	N/A	N/A	N/A
G2,G3,G4	1	0	4	4	1	0	0	https://gnps.org/https://gnps.org/	N/A	N/A	N/A	N/A
G3	1	0	0	3	0	0	0	https://gnps.org/https://gnps.org/	N/A	N/A	N/A	N/A
G2,G3,G4	1	0	7	15	6	0	0	https://gnps.org/https://gnps.org/	N/A	N/A	N/A	N/A
G1,G2,G3,G4	0	53	53	68	73	0	0	https://gnps.org/https://gnps.org/	N/A	N/A	N/A	N/A
G1,G2,G3,G4	0	3	6	9	3	0	0	https://gnps.org/https://gnps.org/	N/A	N/A	N/A	N/A

RTMean	RTMean_n	RTStdErr	Smiles	SpectrumID	UniqueFile	UniqueFile	cluster	ind	componer	number	parent	ma	precursor	precursor	sum	precu
904.5397	15.07566	1.796697	N/A	N/A	pTAR_ISP2	10	1	-1	66	105.071	1	105.071	2.98E+06			
536.8524	8.94754	1.58512	N/A	N/A	016_ISP2_1	6	2	-1	23	108.045	1	108.045	765425			
483.6845	8.061408	5.8915	N/A	N/A	016_ISP2_2	2	75	-1	2	110.061	0	110.061	16979			
487.6389	8.127315	4.066947	N/A	N/A	009_ISP2_3	5	77	-1	9	110.061	1	110.061	129569			
487.9473	8.132456	8.800026	N/A	N/A	009_ISP2_1	3	97	-1	3	110.061	1	110.061	35154			
619.5758	10.32626	1.219449	N/A	N/A	016_ISP2_3	7	183	-1	28	111.046	1	111.046	284330			
299.9027	4.998378	26.67465	N/A	N/A	pTAR_ISP2	12	212	-1	247	112.051	1	112.051	1.08E+07			
333.7907	5.563178	16.96153	N/A	N/A	pTAR_ISP2	11	216	-1	21	112.051	1	112.051	532669			

② .graphml



⑤ .clustersummary



Agilent .d

④ MassHunter Qualitative Analysis

⑥

DefaultGroups	RTMean_min	precursor mass	componentindex
G3	xx.xxxxxx	xxx.xxx	componentindex=-1
G2,G3,G4	xx.xxxxxx	xxx.xxx	componentindex=123
G2	xx.xxxxxx	xxx.xxx	componentindex=246
G2,G3,G4	xx.xxxxxx	xxx.xxx	componentindex=123
G2	xx.xxxxxx	xxx.xxx	componentindex=246
G2,G3,G4	xx.xxxxxx	xxx.xxx	componentindex=123
G2	xx.xxxxxx	xxx.xxx	componentindex=246
G1,G2,G3,G4	xx.xxxxxx	xxx.xxx	componentindex=-1
G2	xx.xxxxxx	xxx.xxx	componentindex=246
G1,G2,G3,G4	xx.xxxxxx	xxx.xxx	componentindex=-1
⋮	⋮	⋮	⋮

Cluster 246



⑦ Pyteomics

⑧ mz=xx.xxx

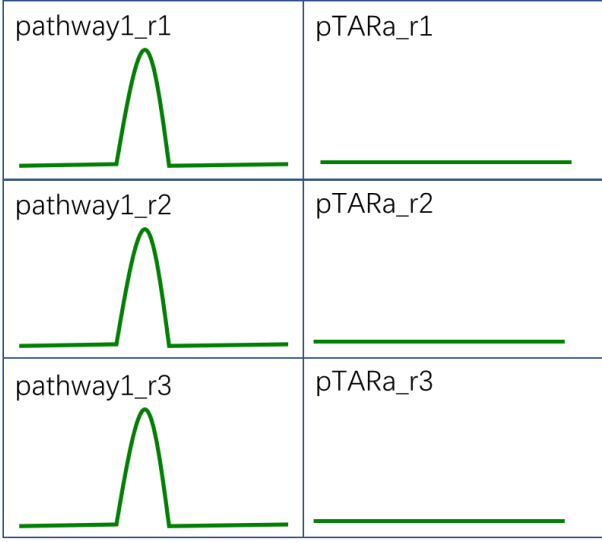


Fig 4.8 Comparison of GNPS workflows.

The left part illustrated our original GNPS manual analysis workflow (①-④): The GNPS molecular networking was visualised through Cytoscape (②), where we manually went through and picked out pathway “unique” ions/clusters (③). We then validated those ions by extracting them in the Agilent analysis software (④).

The right part (①, ⑤-⑧) demonstrated our python-based GNPS automatic analysis workflow. This python script provided a one-step analysis pipeline that integrated the molecular networking analysis (⑤-⑥) and ion validation (⑦). The extracted chromatograms (⑧) were generated automatically.

The python script was tested and run on Victoria University's Rāpoi HPC Cluster. The scripts for this analysis are attached in Appendix 9 (for positive mode, also available via GitHub) and Appendix 10 (for negative mode, available via GitHub). This analysis led to the identification of potentially new metabolites being produced by four pathways. Results are exemplified in Fig 4.9 and summarised in Table 4.9.

Two of the pathways that were putatively producing new compounds under heterologous expression conditions were selected for follow-up work discussed in Chapters 5 and 6.

There are some nodes that stood up from the other two pathways. One is from the refactored 218-3 pathway, HygE218-3. *In silico* analysis revealed that the targeted compounds produced from this pathway should be a lipopeptide, so we targeted the high molecular weight ions. All the pathway-specific ions were just above 1000, suggesting these compounds cannot be the final products, but some intermediates this pathway encoded during the biosynthesis process (such as cyclic peptides without the fatty acid chains). Another set of unique signals was detected from BGC016; however, the ions fell out of the mass range of the proposed compounds produced from BGC016. These results suggest that future work aiming at refactoring and optimising the BGC016 and BGC218-3 heterologous expression systems might be fruitful. The detected and verified ions from BGC016 and BGC218-3 were provided in Appendix 11.

Table 4.9 GNPS analysis results of *S. albus* Del14 related metabolites.

Examples of the three types of molecular networking analysis results (a-c) are provided in Fig 4.9.

	ISP2	ISP4	R5a	SMM
del14004_pos	b	b	b	b
del14004ΔLuxR_pos	a	b	a	b
del14009_pos	a	b	a	b
del14014_pos	a	b		b
del14016_pos	b	b	b	b
del14031_pos	c: Followed up and discussed in Chapter 5			
del14218-3_pos	a	a	a	a
del14HygE218_3_pos	a	c: Discussed in Chapter 4		
del14pIJEF218_3_pos	a	a	a	a
del14004_neg	b	b	b	b
del14004ΔLuxR_neg	b	a	b	b
del14005_neg	c: Followed up and discussed in Chapter 6			
del14009_neg	b	a	b	b
del14014_neg	b	b	b	b
del14016_neg	b	b	c: Discussed in Chapter 4	
del14218-3_neg	a	a	a	a
del14HygE218_3_neg	b	b	c: Discussed in Chapter 4	
del14pIJEF218_3_neg	a	b	a	b
a: no unique nodes and/or clusters				
b: only false positive nodes and/or clusters				
c: followed up and/or discussed				

a

```
list(result_dict2.values())
```

```
Out[8]: []
```

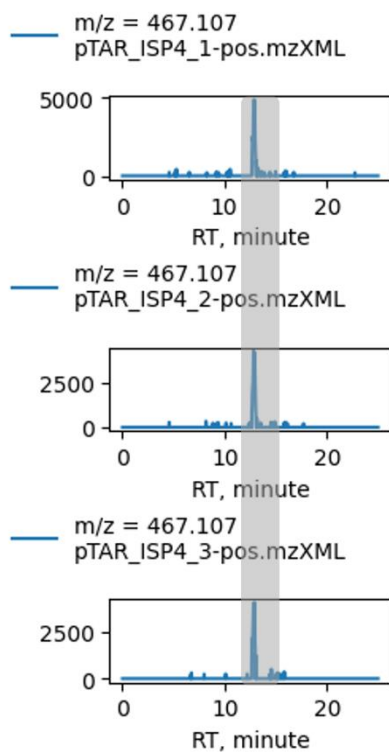
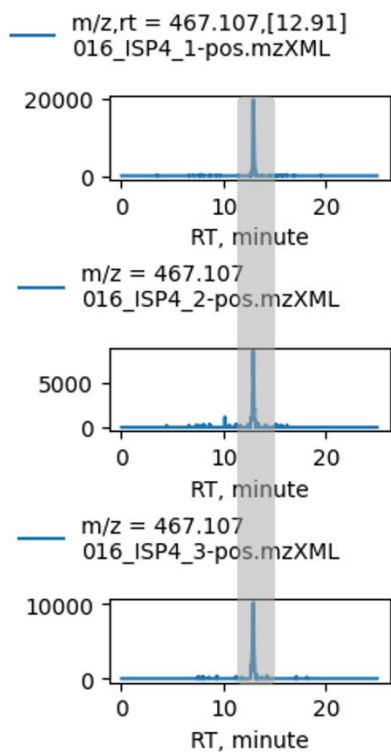
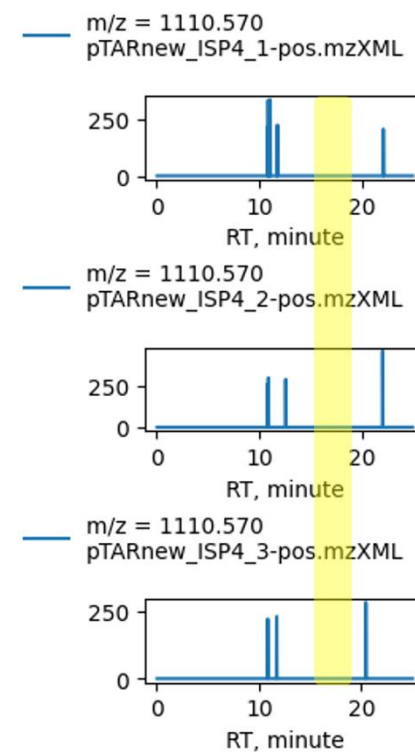
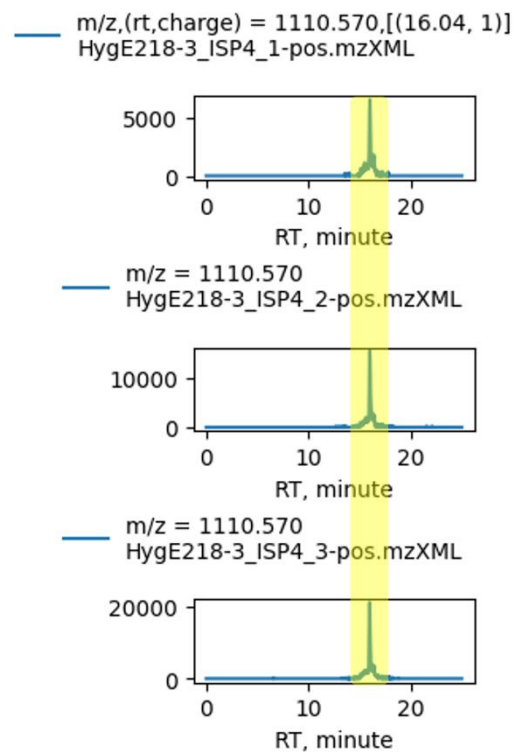
b**c**

Fig 4.9 Three types of molecular networking analysis results as summarised in Table 4.9.

- a. No unique nodes and/or clusters: the analysis (Fig 4.8 ⑤-⑥) did not identify any pathway unique nodes and/or clusters from molecular networking. No extracted chromatograms will be generated in this case (Fig 4.8 ⑧).

Once the python pipeline identifies pathway unique nodes and/or clusters from molecular networking, extracted chromatograms (Fig 4.8 ⑧) will be generated:

- b. False positive results: ions (grey shaded) can be found both in the pathway (b, left) and control (b, right) groups. These ions will be excluded from the downstream studies.
- c. Valid positive results: ions (highlighted in yellow) can only be found in the pathway

4.5 Methods

4.5.1 Construction of BAC004ΔLuxR

A PCR product containing *hygE* (selection marker) and *ermE** (strong constitutive promoter) gene cassette flanked by two homology arms targeting the *ctg1_627* upstream region and *ctg1_629* downstream region was generated using the primers 004_HygE_F/R (Appendix 8) and the template pIJ10257. We transferred the PCR product to the induced *E. coli* s17-1 containing pRedET and BAC004 to generate BAC004ΔLuxR by *in vivo* Red/ET reaction (Fig 4.10). The knockout process was conducted according to the manufacturer's instructions (GENE BRIDGES, Quick & Easy *E. coli* Gene Deletion Kit).

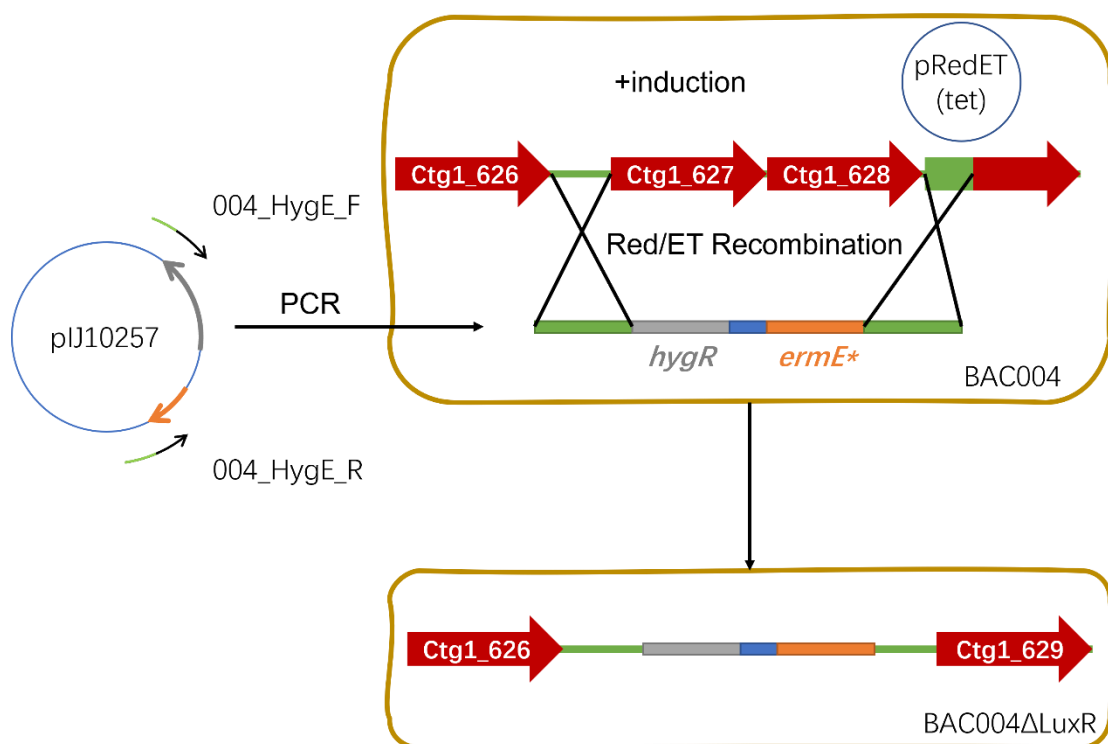


Fig 4.10 Flowchart for the *Ctg1_627*&*Ctg1_628* knockout on the BAC004 in *E. coli* S17-1

4.5.2 Fermentation

Spores (in *S. albus* Del14 or *S. lividans* TK24) of pathways (listed in Table 4.8) and pTARa (empty vector, control group) were streaked onto 12-well plates containing four types of media (ISP2, ISP4, R5a, SMM, three replicas per medium; Fig 4.11①). Following 14 days of incubation (30 °C, 200 rpm), each well was extracted with 25 mL MeOH, dried, and redissolved in MeOH (Fig 4.11②). The prepared samples were injected through a C18 column (PhenoSphere-NEXT™ 3 μm C18 120 Å, LC Column 150 x 4.6 mm), and the production of metabolites was analysed by HPLC-HR-ESI-MSMS as described in Chapter 2.4.2 and Table 2.12 (Fig 4.11③). Recorded data were submitted to the GNPS for molecular networking generation (Fig 4.11④), and the resulting networks were analysed using in-house python workflows (Fig 4.11⑤).

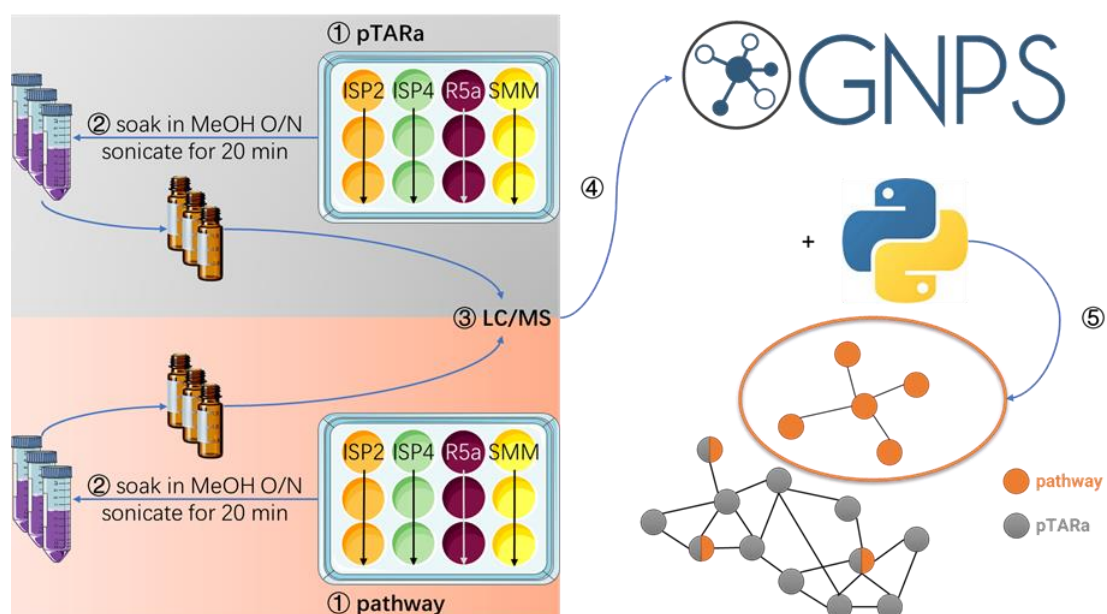


Fig 4.11 Workflow for GNPS preparation

Chapter 5 Heterologous Expression of a Puromycin-like Pathway

5.1 Introduction

Puromycin is an aminonucleoside antibiotic produced by *Streptomyces alboniger*.¹¹⁹ The overall structure of puromycin (Fig 5.1a) consisted of two parts, a modified nucleoside covalently bound to a modified tyrosine, which resembles that of tyrosyl-tRNA (Fig 5.1b). During protein biosynthesis, mRNA-tRNAs are moved along the ribosome by translocation, leaving an empty A site. The elongation cycle of translation is repeated if another aminoacyl-tRNA enters the empty A site. When the A-site is occupied by puromycin, the elongating peptidyl-tRNA in the P site undergoes puromycylation, resulting in the release of premature puromycylated peptide (Fig 5.1c).^{120,121} Based on the unique characteristics of puromycin, various puromycin derivatives have been generated for the labelling and imaging of proteins.¹²¹

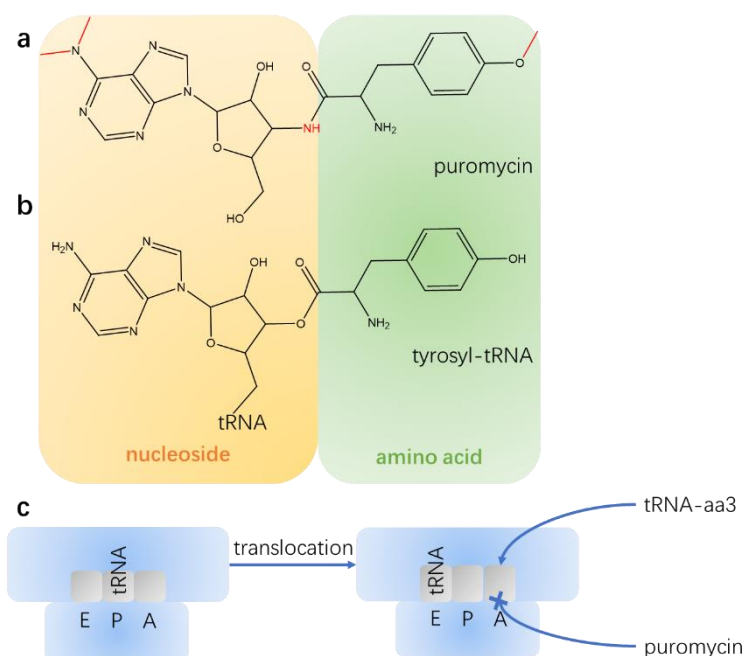


Fig 5.1 Structure and mechanism of action of puromycin
Puromycin is structurally similar to tyrosyl-tRNA. When occupied in the A site, the puromycin would cause the termination of protein biosynthesis.

In the puromycin BGC (*pur*), the initial precursor ATP is dehydrated via NAD(P)-dependent oxidoreductase (*Pur10*) to produce 3'-keto-3'-dATP, which is then converted into 3'-amino-3'-dA by pyrophosphatase (*Pur7*), aminotransferase (*Pur4*), and monophosphatase (*Pur3*). A tyrosine moiety is transferred to 3'-amino-3'-dA by *Pur6*, which is then acylated (*Pac*) and methylated (*Pur5*, *Dmp*, two SAM-dependent methyltransferases) to produce N-acetylpuromycin. The puromycin prodrug is secreted by *Pur8*¹²² and hydrolysed by *NapH* to complete the maturation of bioactive puromycin (Fig 5.2a).^{123–125}

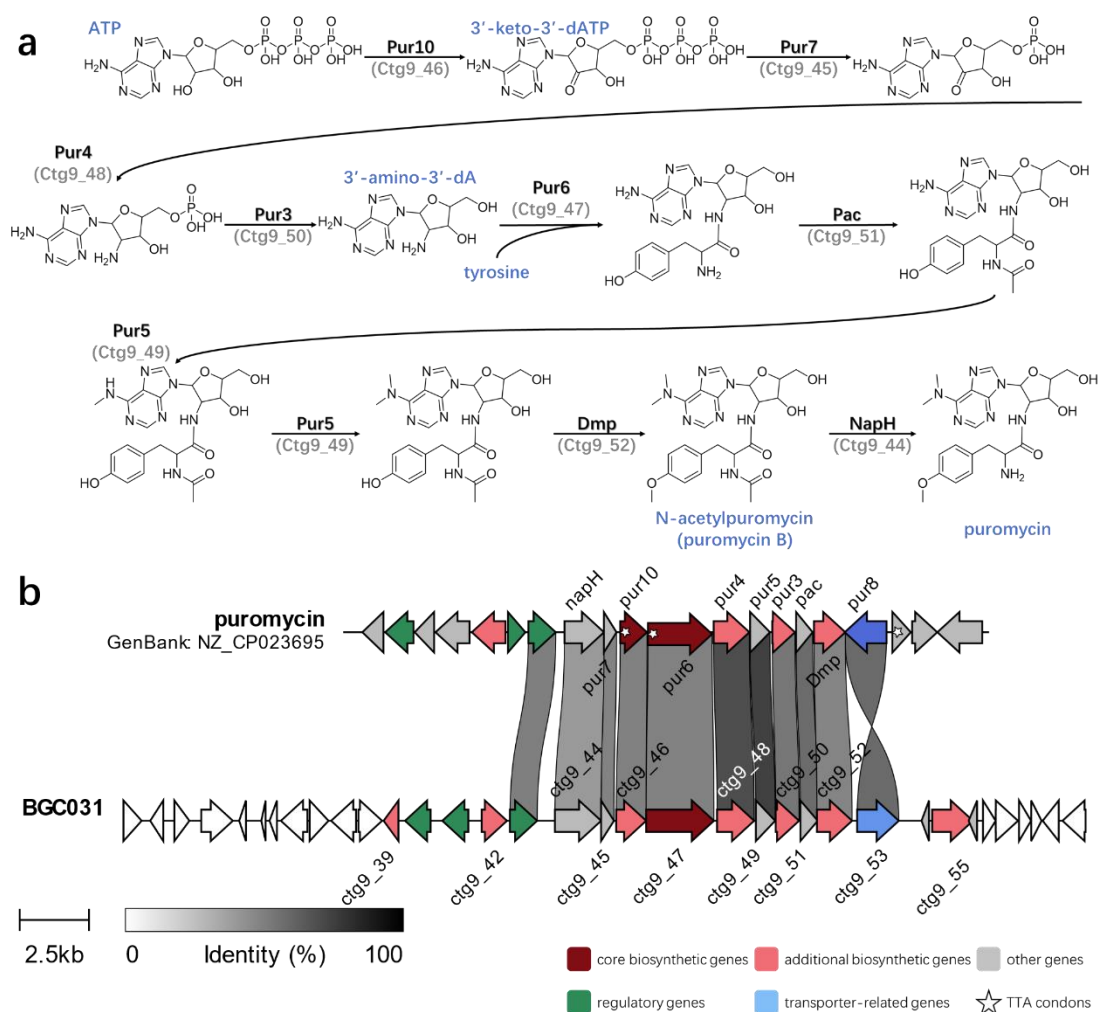


Fig 5.2 (a) Proposed puromycin biosynthetic pathway and (b) gene cluster comparison of BGC031 and *pur*

5.2 Heterologous expression of BGC031

antiSMASH identified BGC031 as a nucleoside BGC that resembled the well-studied puromycin BGC. Previous studies have confirmed that the pur BGC boundaries are marked by the *napH* and *pur8* genes (Fig 5.2b).¹²⁵ In comparison with the pur pathway, BGC031 contained homologs of all the structural genes required for puromycin production (Fig 5.2b, Table 5.1). Due to the similarity in gene organisations of the BGC031 and the pur pathways, we proposed that BGC031 should direct the biosynthesis of puromycin-like compounds (Fig 5.2a). In addition to homologues of all puromycin biosynthetic genes, BGC031 contains several additional biosynthetic genes that do not have counterparts on pur BGC (Fig 5.2b). For example, oxidoreductase (Ctg9_37), methyltransferase (Ctg9_39), and decarboxylase (Ctg9_55) may contribute to further tailoring during BGC031 biosynthesis (Fig 5.2b, Table 5.1). Thus, we decided to clone this BGC031 starting from *ctg9_28* to *ctg9_62*, well beyond the reported pur BGC boundaries.

The 36 kb region was captured with shuttle vector pTARa using CRISPR/Cas9-mediated TAR cloning method into *S. cerevisiae* BY4727 Δ NHEJ (YAC031), transferred to *E. coli* S17-1 (BAC031). The predicted functions of the genes on BGC031, gene fragments and primers used in cloning BGC031 are listed in Table 5.1 and Appendix 8, respectively. The BAC was first conjugated to *S. albus* J1074 to yield *albus_031*.

Ethyl acetate extracts from agar cultivation revealed that the heterologous expression strain *albus_031* (Fig 5.3③) was not producing any new metabolites relative to the *empty vector control* (Fig 5.3④).

A literature search revealed that Prof. Xudong Qu and his colleagues were able to activate a silent puromycin biosynthetic pathway located on the genome of *Streptomyces alboniger* NRRL B-1832 by co-expression of PPTase genes. The reasons for this activation are not obvious, as the biosynthesis of puromycin is not dependent on PPTase; however, given that empirical evidence suggested this strategy was effective, it was decided that this might be a fruitful approach for activation the BGC031. Upon request, Qu Lab provided pWHU2449.¹⁹ We integrated the two broad-

selective PPTase genes *sfp-svp* derived from pWHU2449 along with the *ermE** promoter to the ϕ BT1 site of the *albus_031* heterologous expression strain, yielding *albus_031*-pIJ2449. Following the fermentation-extraction procedure, we analysed the metabolites from *albus_031*-pIJ2449 (Fig 5.3②), *albus_031* (Fig 5.3③), *albus_pTARa* (Fig 5.3④) and a *puromycin standard* (Fig 5.3①). The HPLC profiles suggested that a new compound/peak was produced only when the BGC031 and the PPTase-based activation system were both presented. This compound had a retention time that differed from puromycin's, and production of puromycin was not detected in the same strain. The UV-Vis spectrum of the new compound produced by *albus_031*-pIJ2449 was similar to that of a *puromycin standard*, indicating that it may be a new puromycin analogue (Fig 5.3⑧). However, the production of peak11.5 was very low, preventing its downstream isolation and characterisation.

In an attempt to increase the production of the new compound observed in *albus_031*-pIJ2449, the BAC containing the BGC031 was also conjugated to *S. albus* Del14, an optimised strain of *S. albus* J1074 in which endogenous secondary metabolite clusters have been knocked out.⁵¹ The resulting strain (*del14_031*) was cultivated on agar, and the extracted metabolites were compared to those from an empty vector control (Fig 5.3⑥⑦). This analysis led to the identification of the same small puromycin analogue peak11.5 we observed in *albus_031*-pIJ2449; this time, the peak was produced by *del14_031*, a heterologous expression system without any refactoring. Finally, in a last effort to boost production, the medium was supplemented with 0.1 g/L Arginine and 10 mM MgCl₂ (two components shown to propel spore germination). This trial resulted in an increased production of the peak11.5 (Fig 5.3⑤⑨).

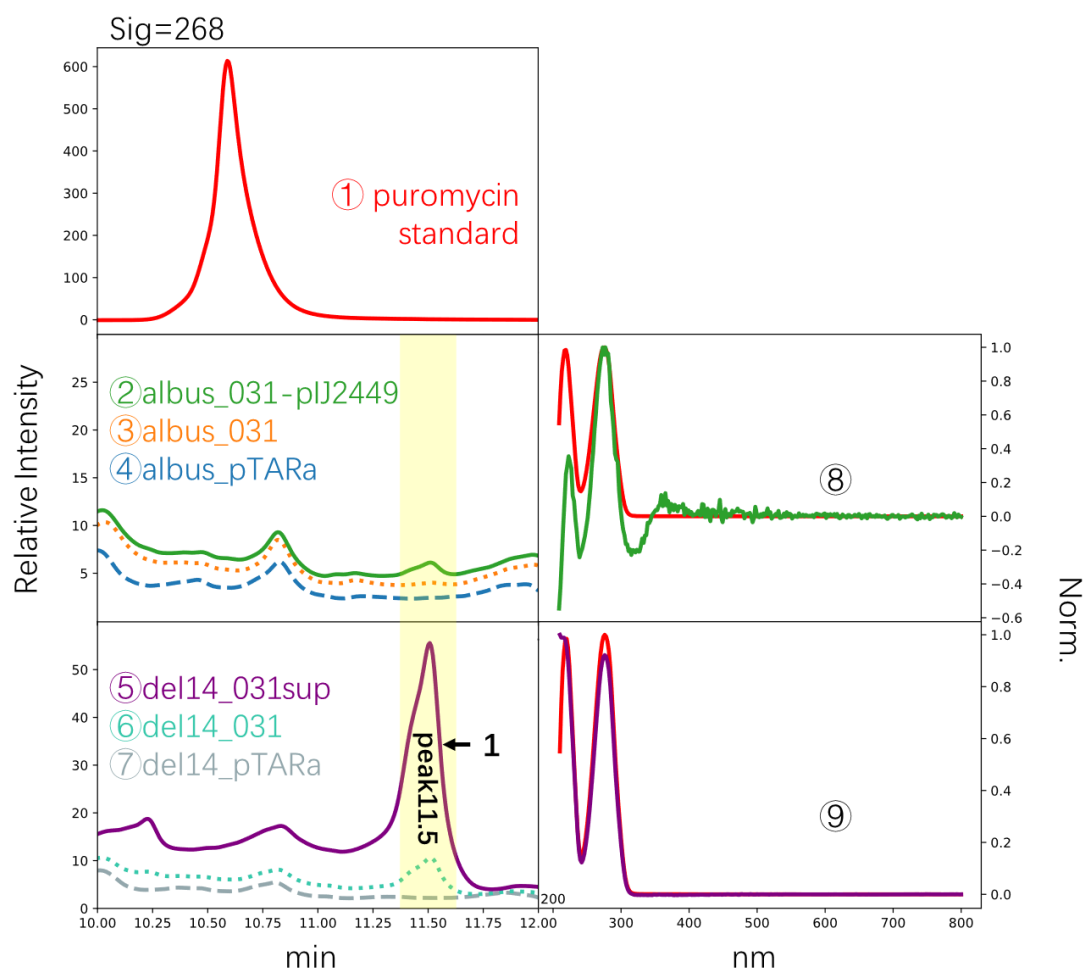


Fig 5.3 HPLC traces (@UV 268 nm) and UV-vis spectra used in this study.

The BGC031 was first introduced to *S.albus* J1074 (③) and compared to the UV trace of vector control (④) at 268 nm, indicating that this pathway remained silent. To activate this pathway, different activation strategies have been carried out, for example, pathway refactoring strategy (②), changing host (⑥), and optimising fermentation conditions (⑤). UV-Vis spectra of new peaks from successful heterologous expression systems (②_⑧, ⑤_⑨) were similar to that of a puromycin standard (①).

Table 5.1 Predicted functions of the genes on BGC031

orf	aa size	proposed function	homologous Pur enzymes	identity/similarity(%)
<i>ctg9_28</i>	218	dihydrofolate reductase family protein	-	-
<i>ctg9_29</i>	156	hypothetical protein	-	-
<i>ctg9_30</i>	180	hypothetical protein	-	-
<i>ctg9_31</i>	380	hypothetical protein	-	-
<i>ctg9_32</i>	80	hypothetical protein	-	-
<i>ctg9_33</i>	51	DUF6131 family protein	-	-
<i>ctg9_34</i>	88	aldo/keto reductase	-	-
<i>ctg9_35</i>	317	aldo/keto reductase	-	-
<i>ctg9_36</i>	226	TetR/AcrR family transcriptional regulator	-	-
<i>ctg9_37</i>	300	SDR family oxidoreductase	-	-
<i>ctg9_38</i>	284	helix-turn-helix domain-containing protein	-	-
<i>ctg9_39</i>	178	class I SAM-dependent methyltransferase	-	-
<i>ctg9_40</i>	307	LysR family transcriptional regulator	-	-
<i>ctg9_41</i>	313	LysR family transcriptional regulator	-	-
<i>ctg9_42</i>	305	SDR family oxidoreductase	-	-
<i>ctg9_43</i>	327	LysR family transcriptional regulator	WP_167532747.1	58.66/70.21
<i>ctg9_44</i>	556	M28 family peptidase	NapH	54.11/63.58
<i>ctg9_45</i>	154	NUDIX domain-containing protein	Pur7	57.24/66.45
<i>ctg9_46</i>	360	Gfo/Idh/MocA family oxidoreductase	Pur10	63.66/72.98
<i>ctg9_47</i>	816	hypothetical protein	Pur6	61.14/71.11
<i>ctg9_48</i>	446	DegT/DnrJ/EryC1/StrS family aminotransferase	Pur4	78.55/85.31
<i>ctg9_49</i>	231	methyltransferase	Pur5	80.67/85.29
<i>ctg9_50</i>	281	histidinol-phosphatase	Pur3	63/69.6
<i>ctg9_51</i>	208	GNAT family N-acetyltransferase	Pac	66.83/76.38
<i>ctg9_52</i>	422	methyltransferase	Dmp	59.57/69.41
<i>ctg9_53</i>	502	MFS transporter	Pur8	65.01/72.96
<i>ctg9_54</i>	89	GlsB/YeaQ/YmgE family stress response membrane protein	-	-
<i>ctg9_55</i>	491	diaminopimelate decarboxylase	-	-
<i>ctg9_56</i>	101	hypothetical protein	-	-
<i>ctg9_57</i>	152	hypothetical protein	-	-
<i>ctg9_58</i>	267	hypothetical protein	-	-
<i>ctg9_59</i>	143	roadblock/LC7 domain-containing protein	-	-
<i>ctg9_60</i>	124	hypothetical protein	-	-
<i>ctg9_61</i>	192	hypothetical protein	-	-
<i>ctg9_62</i>	278	MurR/RpiR family transcriptional regulator	-	-

5.3 Characterisation of Compound 1

To determine the chemical structure of the predominant heterologously expressed **peak11.5**, we inoculated del14_031 onto 40 agar plates, each containing 50 mL of ISP4 agar supplemented with 0.1 g/L Arginine and 10mM MgCl₂. From the crude extracts, we purified compound 1. Compound 1 was obtained as a white powder. Analysis of the ¹H NMR, COSY, HSQC, HMBC (Appendices 12-16), and ESI-HRMS data revealed that compound 1 was puromycin B, a puromycin congener reported by Qu's group (Fig 5.4).¹⁹

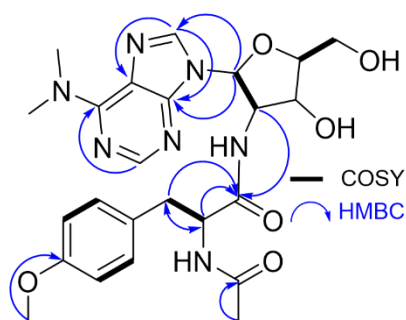


Fig 5.4 Structure and key correlations for compound 1 (600 MHz, DMSO-d₆:MeOD= 1:1)

5.4 GNPS molecular networking analysis

Preliminary HPLC analysis of heterologous expression strains suggested that a number of congeners of puromycin might be produced. Unfortunately, the amount produced in each case was too low to permit isolation and structure elucidation using NMR. In order to obtain additional information about the potential structures of the new metabolites being produced, we subjected an aliquot of crude extract from section 5.3 to LC-MS/MS using the protocol and parameters described in Table 2.11. The collected data was submitted to GNPS for molecular networking.⁶³ A cluster containing the node corresponding to compound 1 (peak11.5, puromycin B) stood out immediately (Fig 5.5). The corresponding MS/MS spectra were manually inspected on each node in the puromycin putative analogue containing cluster (Fig 5.6-5.8).

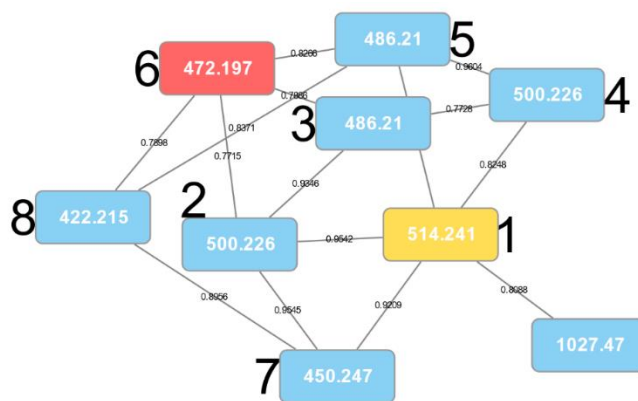


Fig 5.5 Puromycin Cluster from the GNPS Molecular Network.
All nodes are labelled with the corresponding precursor ions. All edges are labelled with Cosine score.
Analysed nodes are marked with a number (compounds and MS/MS spectra share the same number).

5.4.1 MS/MS spectra and tentative structural assignments of Compounds 2,3,4,5

We found two sets of 500.226 (spectra/nodes 2 and 4) and 486.21 (spectra/nodes 3 and 5) nodes directly connected to the characterised puromycin B node (spectrum/node 1, yellow) (Fig 5.5). The mass difference of 14 among those nodes/compounds suggested they are demethylated puromycin B analogues.

The absence of N-dimethylated adenine moiety fragment (164.1016, MS/MS spectra 1,4), the presence of N-acylated-O-methylated tyrosine moiety fragment (192.1019) and mono-methylated adenine moiety fragment (150.8) on MS/MS spectrum 4 supported the structure assignment of compound **4** (Fig 5.6).

The presence of N-acylated-O-methylated tyrosine moiety fragment (192.1019) and the fact that MS/MS spectrum 5 has a higher cosine score with MS/MS spectrum 4 than MS/MS spectrum 3 (Fig 5.5) supported the structure assignment of compound **5**, a demethylated analogue of compound **4** (Fig 5.6).

The presence of N-dimethylated adenine moiety fragment (164.1016, MS/MS spectra 1,2), the tyrosine moiety fragment (136.07, MS/MS spectrum 2), N-acylated tyrosine moiety fragment (178.08) and the absence of N-acylated-O-methylated tyrosine moiety (192.1019, MS/MS spectrum 1) supported the structure assignment of compound **2** (Fig 5.6).

The presence of the tyrosine moiety fragment (136.07, MS/MS spectrum 3), N-acylated tyrosine moiety fragment (178.08, MS/MS spectrum 3), the absence of N-dimethylated adenine moiety fragment (164.1016, MS/MS spectra 1,2) and the fact that MS/MS spectrum 3 is directly connected to spectrum 2 rather than spectrum 5 (Fig 5.5) supported the assignment of the structure of compound **3** as a demethylated analogue of compound **2** (Fig 5.6).

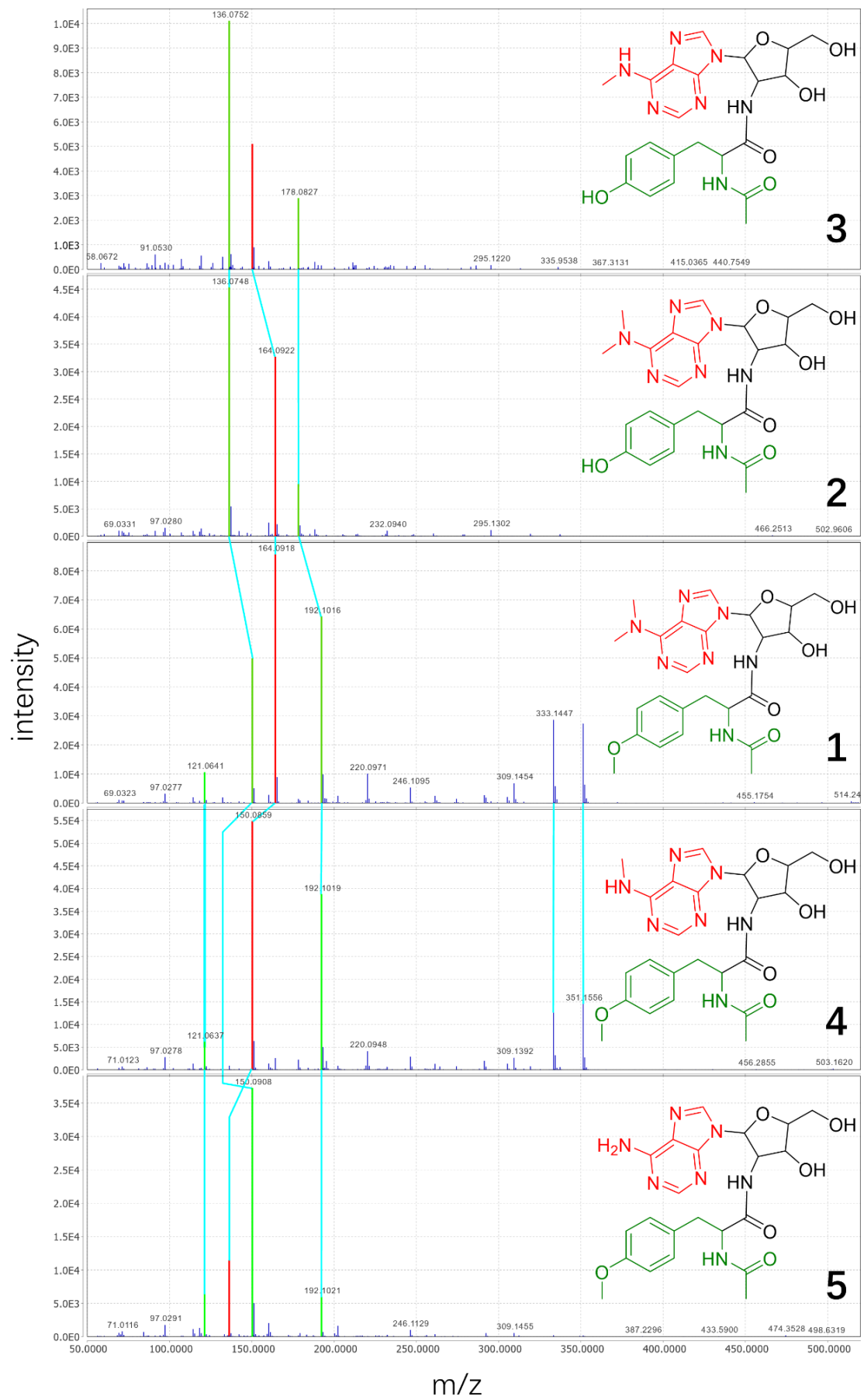


Fig 5.6 MS/MS spectra of compounds 1-5

MS/MS diagnostic fragments (coloured in red and green) were aligned (indicated by cyan lines) to deduce the metabolite/congener transformations. Metabolites/congeners were colour coded to match the MS/MS diagnostic fragments.

5.4.2 MS/MS spectrum and tentative structural assignment of Compound 6

The presence of nucleoside moiety fragments (319.12), tyrosine moiety fragments (136.07, 178.08) and the precursor mass ions suggested that compound **6** is corresponding to puromycin, the deacylated analogue of compound **1** (Fig 5.7).

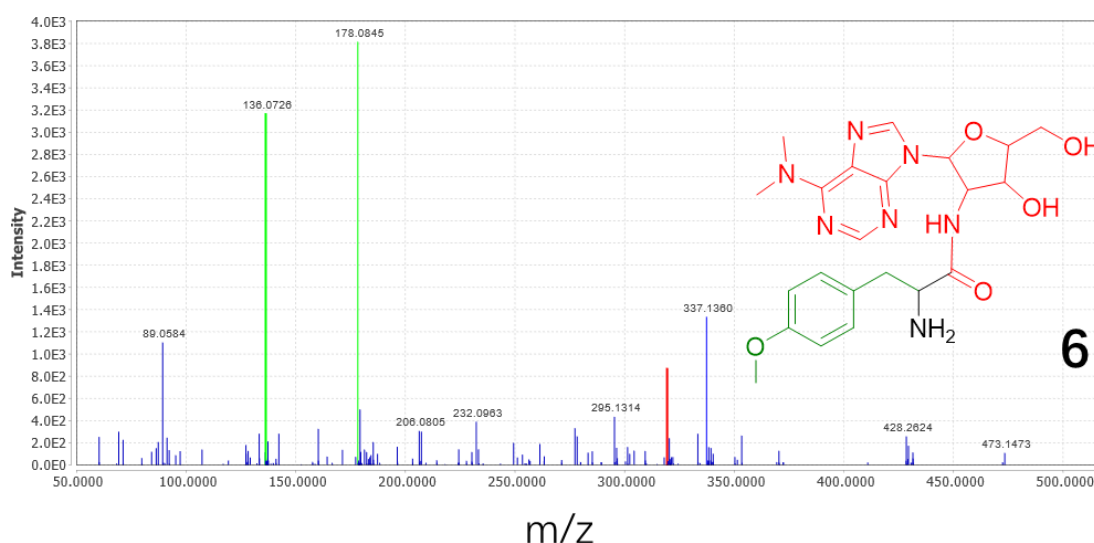


Fig 5.7 MS/MS spectrum of compound 6

5.4.3 MS/MS spectra and tentative structural assignments of Compounds 7,8

The mass difference between nodes 7 and 1 is 63.994, and GNPS annotated this edge as a fragment corresponding to C₄O, suggesting the decomposition of the phenol moiety on tyrosine (Fig 5.5). The absence of tyrosine moiety related fragments on spectrum 7 further supported this hypothesis. The spectra for compounds **7** and **8** contain two new diagnostic fragments with masses of 128.1095 and 86.0970, which should be assigned to the amino acid constituent of the respective structures. The Thorson group have previously reported a new puromycin analogue where the tyrosine moiety of compound **1** is replaced by leucine.¹²⁶ HRESIMS and fragmentation

patterns suggested that compound **7** could be the puromycin B reported by Thorson's group (Fig 5.8).

The mass difference between nodes 7 and 8 is 28.032, suggesting a loss of two methyl groups. The presence of 128.1095 and 86.0970 on spectrum 8, the absence of the N-dimethylated adenine moiety fragment (164.0922), the presence of adenine fragment (136.06) and the fact that node 8 is directly connected to node 5 (a puromycin analogue in non-methylated adenine form), rather than node 3 (a puromycin analogue of monomethylated adenine form) suggested the compound **8** is an analogue of compound **7** in non-methylated adenine form (Fig 5.8).

However, we cannot exclude other possible structural assignments (for example, Fig 5.7: 7',8') of compounds **7** and **8** based on our current MS/MS spectra results (Fig 5.8).

The extracted LC-MS chromatograms of compounds **1-8** are provided in Fig 5.9.

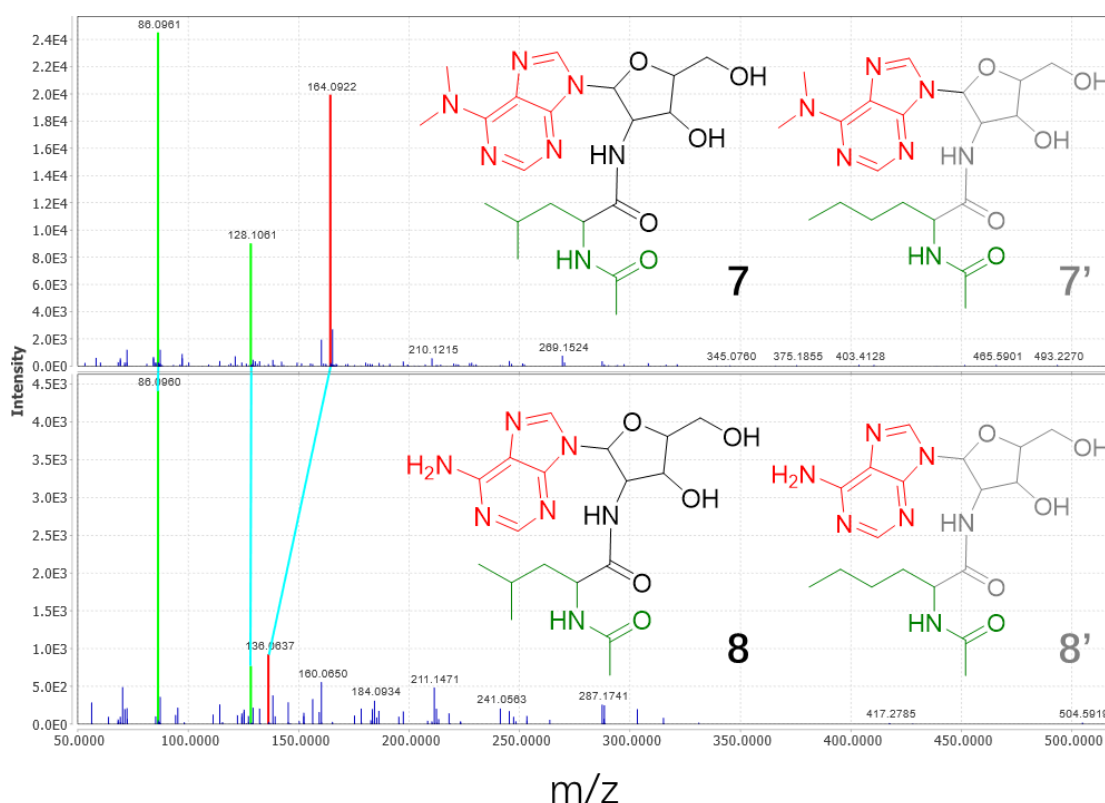


Fig 5.8 MS/MS spectra of compounds **7,8**

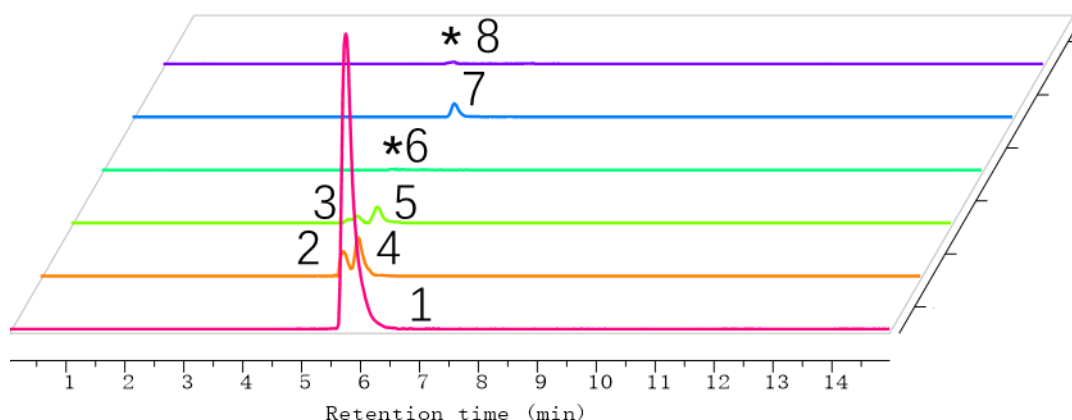


Fig 5.9 Extracted LC-MS chromatograms of compounds 1-8

5.5 Conclusion and discussion

In this study, we combined the heterologous expression platform with molecular networking, identified several potential new analogues of existing natural products scaffolds, and assigned tentative structures for those analogues. However, the production of these analogues was low, thus preventing compound isolation and structure elucidation by NMR.

We attempted several strategies to activate the pathway BGC031, which was remained silent in the *S. albus* J1074 host. We first introduced the two broad-selective PPTases to *S.albus* J1074 harboured BGC031, and successfully activated the silent heterologous expression system *albus_031*. PPTases catalyse the conversion of carrier proteins (T or ACP) from inactive to active forms, which is an essential step for initiating the biosynthesis of NRPS and PKS.^{17,19} BGC031 is a nucleoside type BGC, which means that no carrier protein is needed during the biosynthesis. However, we noted that the expression of BGC031 in *S. albus* J1074 is PPTase-dependent, suggesting that PPTases may have participated in broader biological processes than initially thought (Fig 5.3).

Next, we transferred this pathway to another heterologous expression platform, *S. albus* Del14, a “cleaner” version of *S. albus* J1074. The results showed that this pathway can be readily expressed without any further pathway engineering. This finding is consistent with the claim that a cleaner secondary metabolite background

can improve detection limits and yields.⁵¹ The fermentation system was also supplemented with two trace elements believed to facilitate sporulation. We observed improved production yields of the target compound (Fig 5.3). Certain types of antibiotic production are closely correlated with sporulation, either during germination or after¹²⁷, so factors that facilitate successful germination may also contribute to successful fermentation.

pur10 and *pur6* genes on the original puromycin BGC (Genbank: X92429.1) contain TTA codons. TTA codon is a rare codon usage in GC rich genomes, and the availability of *bldA* gene product tRNA^{UUA} is a limiting factor for BGCs expression. We did not detect any TTA codons on our BGC0031, suggesting that the regulation system of BGC031 is different from the original puromycin BGC, independent of the *bldA* gene product (Fig 5.2).^{125,128}

There is a good understanding of the biosynthesis events of puromycin, but its intermediates are only proposed based on enzymatic logic rather than confirmed through chemical evidence.^{122,124,125,129,130} The GNPS molecular networking enabled us to identify several puromycin congeners (Fig 5.5). We are able to map compounds **3,2,1** and **6** back to the well-characterised puromycin BGC (Fig 5.10a, upper), thus compounds **5,4** could either be the shunt products due to the promiscuity of methyltransferases or indicate the formation of a bypass during the biosynthesis of puromycin (Fig 5.10a, lower).

In addition, we identified two puromycin analogues that cannot be placed in the pur pathway. Thorson and co-workers determined a new cryptic puromycin analogue, on which the appended amino acid is leucine. Our HRESIMS/MS analysis showed that compound **7** could be the puromycin B Thorson group reported. If this is the case, we can place compounds **7,8** to a new divergent pathway, incorporating leucine rather than tyrosine (Fig 5.10b).

Table 5.2 HPLC protocol and parameters

Time (Min)	A (%)	B (%)	DAD (nm): 268 Flow (mL/min): 4.0 Round 1: A: H ₂ O+5 mM NH ₄ Ac B: ACN Round 2: A: H ₂ O+5 mM NH ₄ Ac B: MeOH
0	95	5	
30	0	100	
33	0	100	
34	95	5	
40	95	5	

5.6.2 Pool testing (screening) and validation of BGC031

80 yeast transformants were picked up using toothpicks and transferred onto 8×10 grid new agar plates (as depicted in Fig 2.6a). Colonies from Plate-Replica 1 were then grouped into 10 pools (columns) and each pool was tested using primer 031_0 FORWARD/REVERSE. Positive pool(column) F was selected and entered the second round of screening (Fig 5.11a). Primer 031_0 FORWARD/REVERSE was used to screen the 10 individual yeast colonies in column F (Fig 5.11b). Positive yeast colonies were verified using three sets of primers (031_0-2 FORWARD/REVERSE) in the third round of screening (Fig 5.11b). The verified YAC was then transferred to *E. coli* EC100. The resultant BAC was verified by digestion pattern analysis (Fig 5.11c).

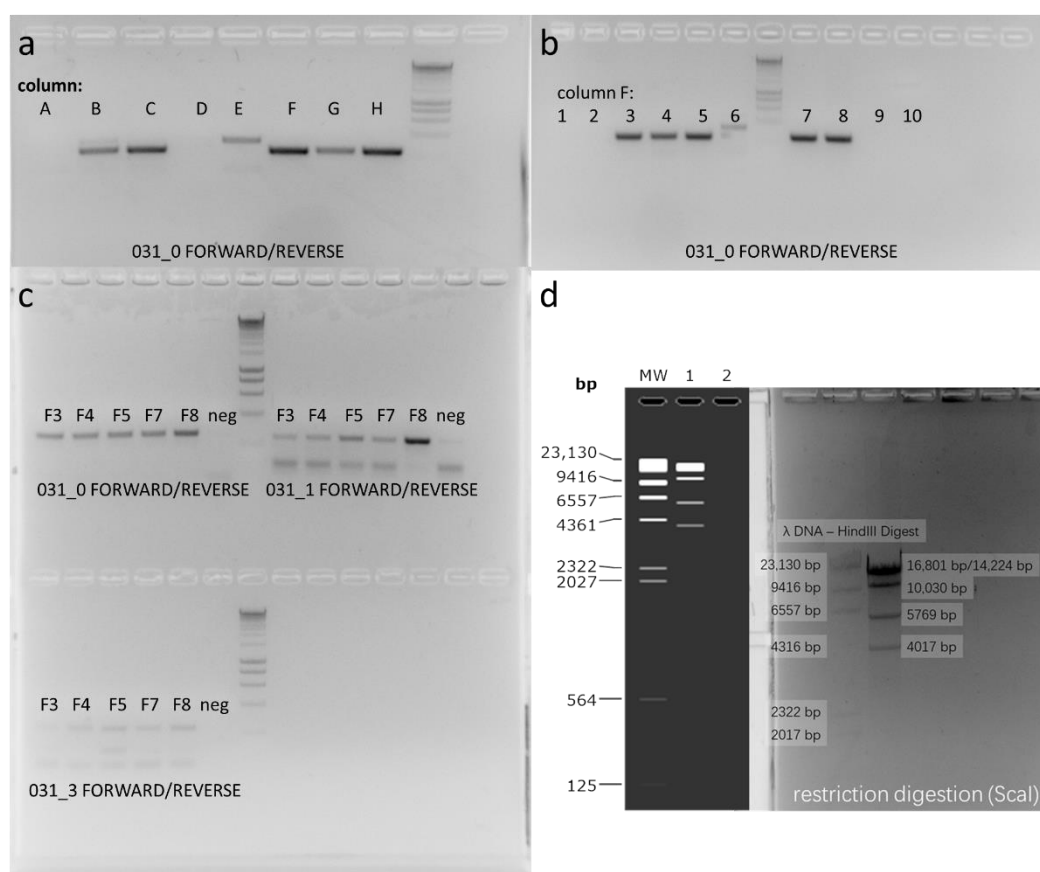


Fig 5.11 Screening and validation of BGC031

- The first round of pool testing (screening) of yeast colonies using primer 031_0 FORWARD/REVERSE. 80 colonies were grouped into 8 pools at this round.
- 10 Yeast colonies from the positive pool (column F) were individually screened using primers 031_0 FORWARD/REVERSE.
- The positive individual colonies were verified using full sets of primers.
- BAC031 was digested using Scal, and the digestion pattern matched the simulated agarose gel prediction generated by SnapGene.

5.6.3 Fermentation and metabolites extraction

The desired hosts (albus_pTARa, albus_031, albus_031-pIJ2449, del14_pTARa, del14_031) was plated on 12-well plate (Interlab, TCP-000-012) or 150mm Petri Dish (Interlab, PD-141) containing ISP4 agar. Following 10 days (30 °C) of incubation, the agar was harvested, soaked overnight in equal volumes of ethyl acetate, sonicated for 20 minutes, and then vacuum dried (30 °C). The dried extract was then redissolved in MeOH, and filtered through PTFE Syringe Filters (Sterlitech, 1470424) to remove particles. Crude extracts were subjected to HPLC for metabolite profile analysis, compound isolation, and LC-MS/MS for GNPS molecular networking.

5.6.4 Construction of albus_031 PPTase overexpression strain

pWHU2449 is a vector equipped with an *aprR* (apramycin resistance marker), the phage ϕ C31 integrase and two broad-selective PPTase genes, *sfp* from *B. subtilis* and *svp* from *S. verticillus*, under the control of the *ermE** promoter (Fig 5.12).¹⁹ As pTARa and pWHU2449 possess the same resistance marker (*aprR*) and integration system (ϕ C31), it was not possible to use pWHU2449 directly. Therefore, the gene fragment *sfp-svp* amplified from pWHU2449 was cloned into NdeI-PacI linearised pIJ10257 to generate pIJ2449 (Fig 5.12). The resultant gene cassette *ermE*-sfp-svp* was conjugated to the ϕ BT1 of the albus_031 heterologous expression strain and selected with hygromycin to yield albus_031-pIJ2449.

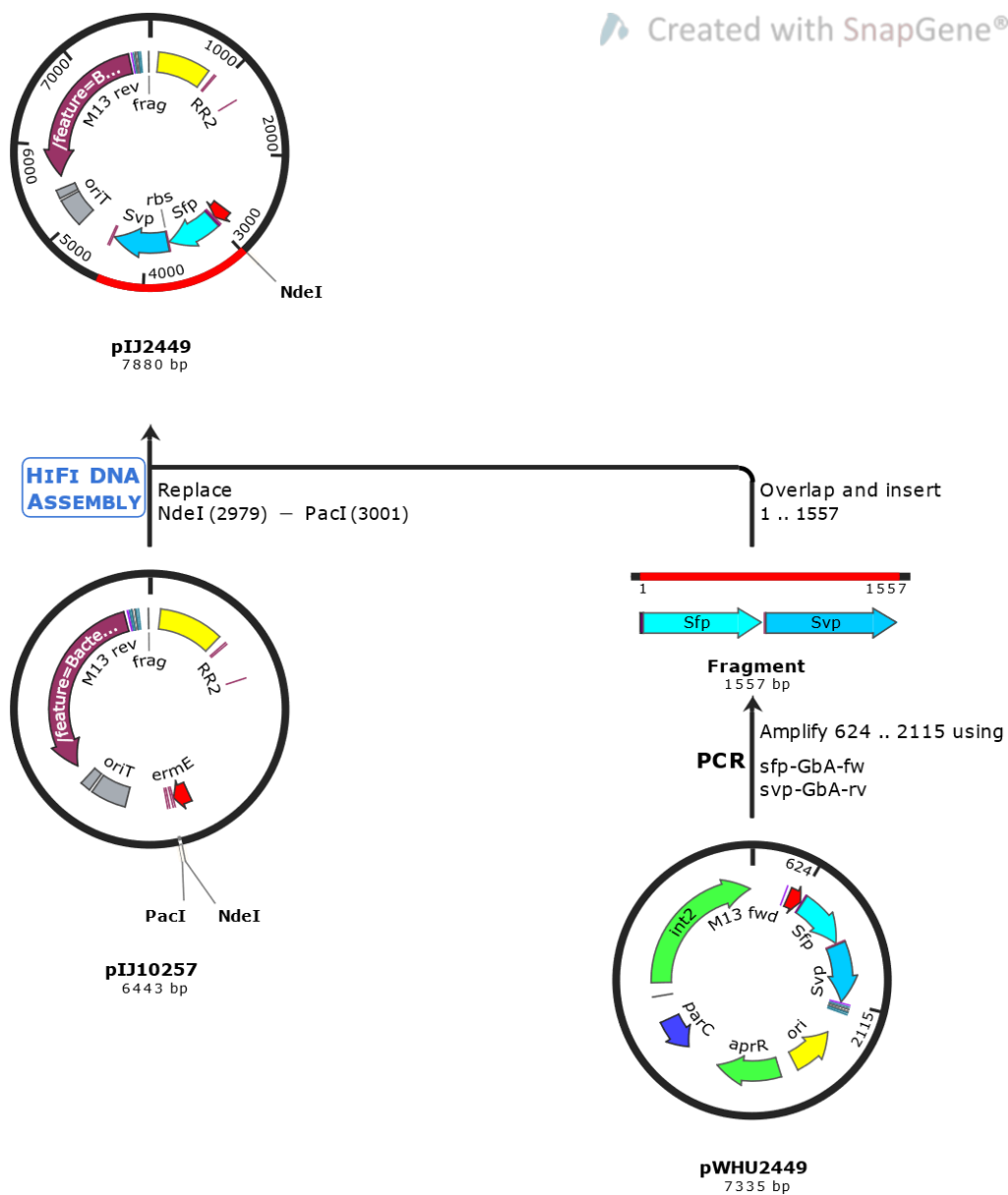


Fig 5.12 Construction of pIJ2449

Chapter 6 Novel Type 2 Polyketide Identified by Heterologous Study of BGC005

6.1 Introduction

Angucyclines are the most prevalent natural product class biosynthesised by T2PKS, which exhibit a variety of chemical scaffolds and bioactivities.¹³¹ For instance, marmycin A, isolated from a marine sediment-derived actinomycete, exhibited significant cytotoxicity.¹³² Lomaiviticin A, discovered from *Salinispora pacifica*, remains one of the most cytotoxic substances to date.^{133,134}

Angucyclines typically possess a four-ring (rings A-D) aromatic polyketide skeleton.¹³⁵ A unique family of oxygenases can initiate the B ring C–C bond cleavage and subsequent rearrangement during the post-tailoring process, leading to the formation of atypical angucyclines with intriguing chemical structures.¹³⁶ The current biosynthesis knowledge of atypical angucyclines has been primarily gained from studies of jadomycin¹³⁷, gilvocarcin¹³⁸, and lomaiviticin¹³⁹. In the present study, we identified two compounds produced by an unreported atypical type 2 polyketide biosynthetic pathway found on the genome of New Zealand lichen sourced *Streptomyces sp.* 438-3. We herein describe the isolation, structure elucidation, biological activity evaluation, and biosynthesis of the new bacterial aromatic polyketide JB1081B (**9**).

6.2 Molecular networking and antiSMASH analysis

During our large-scale heterologous expression project (discussed in Chapter 4), our python assisted GNPS workflow helped us identify and locate several clusters of nodes mainly composed of precursor ions that were pathway BGC005 specific (Fig 6.1).

antiSMASH⁴³ analysis of the sequenced genome of the producing strain *Streptomyces sp.* 438-3 revealed 34 BGCs; among which, BGC005 was the only Type II polyketide

pathway identified (Fig 6.2a). BGC005 spans a DNA region of 77 kb and contains 69 ORFs (Fig 6.2b). The genes on the BGC005 can be subclassified into several functions (Fig 6.2b), including Type II PKS (minimal PKS system and starter unit generation) and modification enzymes, amino sugar biosynthesis, transporters, and regulators (Fig 6.2b, Table 6.1).

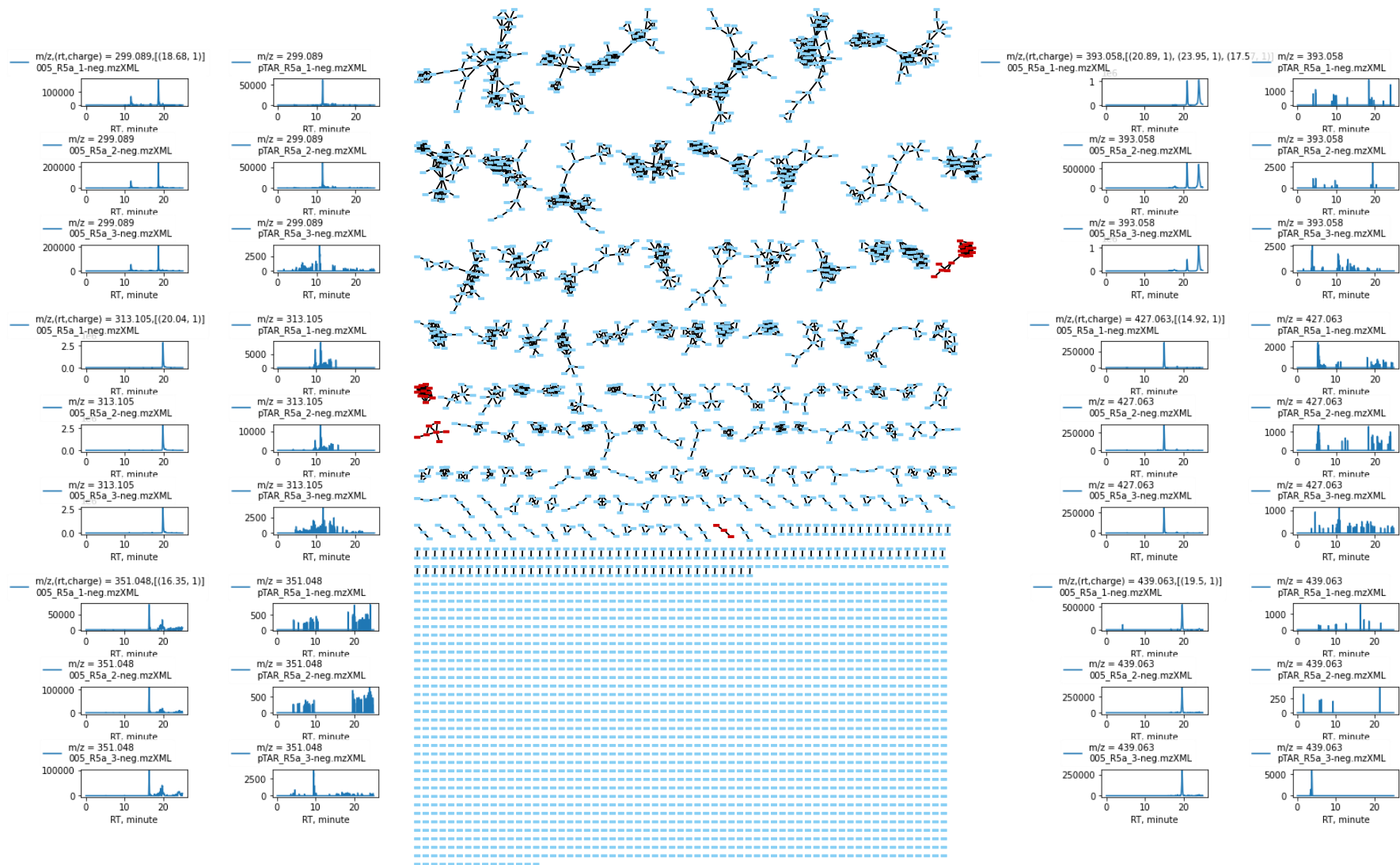


Fig 6.1 GNPS molecular networking and selected MS profiles.

Molecular networking (central) was generated using crude extracts of del14_pTARa (negative control), del14_004, del14_005, del14_009, del14_014 and del14_016. Through the python assisted GNPS analysis workflow, several BGC005 specific clusters (highlighted in red) were identified. Extracted chromatographs (left and right) were generated along the python pipeline and used to confirm that those nodes/ions are true positive signals that were only present in the del14_005 metabolic profile but not in del14_pTARa.

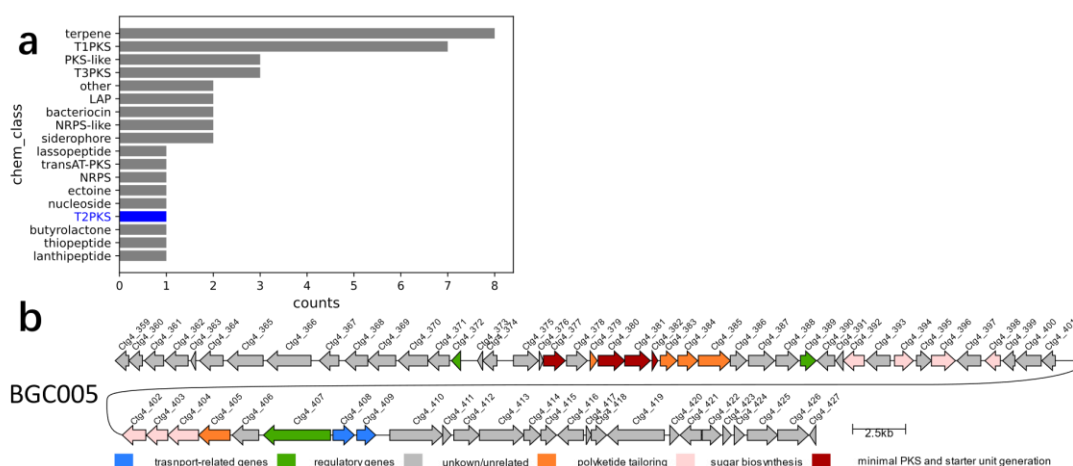


Fig 6.2 Chemical classes and BGC005 identified using antiSMASH

- BGC and chem_classes presented on the chromosome of strain *Streptomyces sp.* 438-3.
- Gene cluster organisation of the only T2PKS BGC005

Table 6.1 Deduced functions of the genes on BGC005

Gene	aa size	proposed function	Gene	aa size	proposed function
ctg4_359	213	low molecular weight phosphatase family protein	ctg4_394	298	glucose-1-phosphate thymidyltransferase RfbA
ctg4_360	215	threonylcarbamoyl-AMP synthase	ctg4_395	234	antibiotic biosynthesis monooxygenase
ctg4_361	301	Release factor glutamine methyltransferase	ctg4_396	377	glycosyltransferase
ctg4_362	363	peptide chain release factor 1	ctg4_397	373	acyltransferase
ctg4_363	75	50S ribosomal protein L31	ctg4_398	241	class I SAM-dependent methyltransferase
ctg4_364	366	LCP family protein	ctg4_399	193	NAD(P)H-dependent oxidoreductase
ctg4_365	563	trypsin-like serine protease	ctg4_400	400	glycosyltransferase
ctg4_366	694	transcription termination factor Rho	ctg4_401	209	dTDP-4-dehydrorhamnose 3,5-epimerase
ctg4_367	306	homoserine kinase	ctg4_402	369	DegT/DnrJ/EryC1/StrS family aminotransferase
ctg4_368	356	threonine synthase	ctg4_403	336	NAD-dependent epimerase/dehydratase
ctg4_369	435	homoserine dehydrogenase	ctg4_404	484	NDP-hexose 2,3-dehydratase family protein
ctg4_370	463	diaminopimelate decarboxylase	ctg4_405	497	FAD-dependent monooxygenase
ctg4_371	333	hypothetical protein	ctg4_406	416	class I SAM-dependent methyltransferase
ctg4_372	141	response regulator	ctg4_407	1054	AAA family ATPase
ctg4_373	81	hypothetical protein	ctg4_408	337	ATP-binding cassette domain-containing protein
ctg4_374	224	hypothetical protein	ctg4_409	297	ABC transporter permease
ctg4_375	392	cytochrome P450	ctg4_410	832	nitrate- and nitrite sensing domain-containing protein
ctg4_376	64	ferredoxin	ctg4_411	134	roadblock/LC7 domain-containing protein
ctg4_377	343	3-oxoacyl-ACP synthase	ctg4_412	401	ABC transporter substrate-binding protein
ctg4_378	326	acyltransferase domain-containing protein	ctg4_413	697	ABC transporter permease
ctg4_379	112	TcmI family type II polyketide cyclase	ctg4_414	264	ABC transporter ATP-binding protein
ctg4_380	418	beta-ketoacyl-[acyl-carrier-protein] synthase family protein	ctg4_415	245	ABC transporter ATP-binding protein
ctg4_381	403	ketosynthase chain-length factor	ctg4_416	405	FAD-dependent monooxygenase
ctg4_382	89	acyl carrier protein	ctg4_417	79	hypothetical protein
ctg4_383	261	3-oxoacyl-ACP reductase FabG	ctg4_418	241	hypothetical protein
ctg4_384	310	aromatase/cyclase	ctg4_419	894	glycoside hydrolase family 97 catalytic domain-containing protein
ctg4_385	504	FAD-dependent monooxygenase	ctg4_420	144	VOC family protein
ctg4_386	256	SDR family oxidoreductase	ctg4_421	334	Ku protein
ctg4_387	409	FAD-dependent oxidoreductase	ctg4_422	294	non-homologous end-joining DNA ligase
ctg4_388	356	hypothetical protein	ctg4_423	133	zinc-ribbon domain-containing protein
ctg4_389	248	response regulator transcription factor	ctg4_424	160	SH3 domain-containing protein
ctg4_390	277	aldo/keto reductase	ctg4_425	476	FtsW/RodA/SpoVE family cell cycle protein
ctg4_391	117	hypothetical protein	ctg4_426	485	penicillin-binding protein
ctg4_392	330	dTDP-glucose 4,6-dehydratase	ctg4_427	98	hypothetical protein
ctg4_393	376	acyl-CoA dehydrogenase family protein			

6.3 Characterisation of compounds **9** and **10**

Previous studies reported that crosstalk can take place between the heterologous expression host and the introduced T2PKS BGC, resulting in the occurrence of unusual congeners.^{140,141} So, as a next step toward characterising the “genuine” BGC005 related metabolites, crude extracts from the fermentation of the native producing strain were obtained. We purified compound **9** from solid fermentation (Fig 6.3①③) and compound **10** (Fig 6.3②④) from liquid fermentation. Two compounds displayed similar UV absorption patterns to previously reported T2PKs.¹⁴²

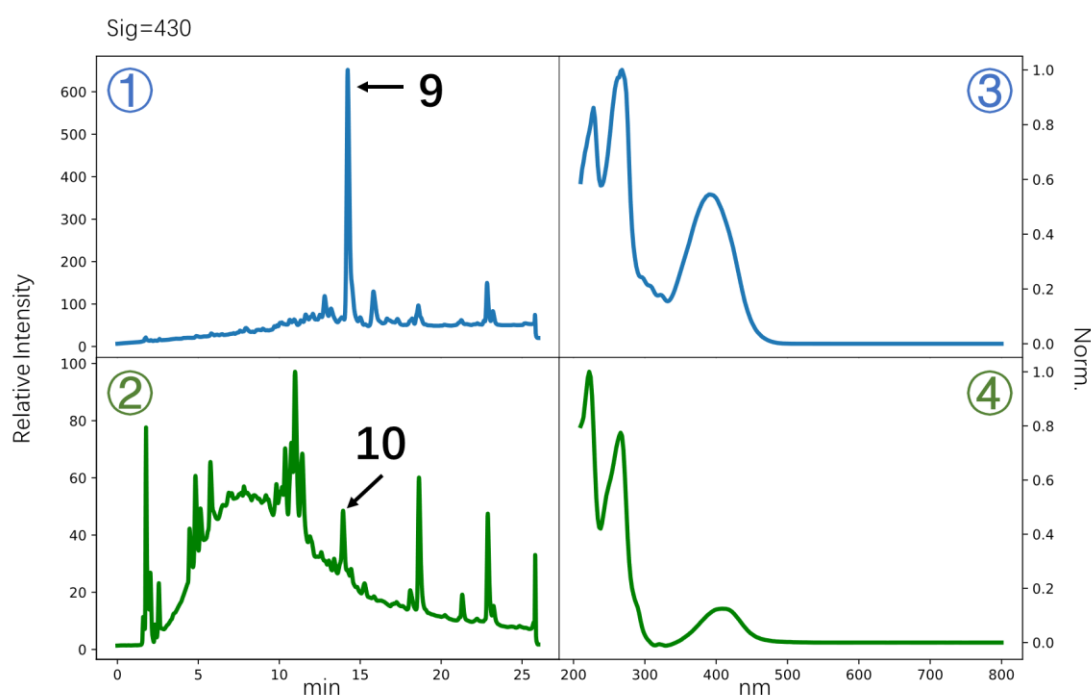


Fig 6.3 HPLC traces (@UV 430 nm) and UV-vis spectra of compounds **9** and **10**

Structural elucidation work of compound **9** (JB1081B, Fig 6.4) was done by Dr Helen Woolner and Dr Joe Bracegirdle and provided in Appendices 17-22.

Compound **10** was obtained as a yellow film. Analysis of the ¹H NMR, COSY, HSQC, HMBC (provided in Appendices 23-27) and HRESIMS data revealed that compound **10** was homorabelomycin (Fig 6.4), a shunt product reported from studies on gilvocarcin biosynthesis by Rohr's group.¹⁴³

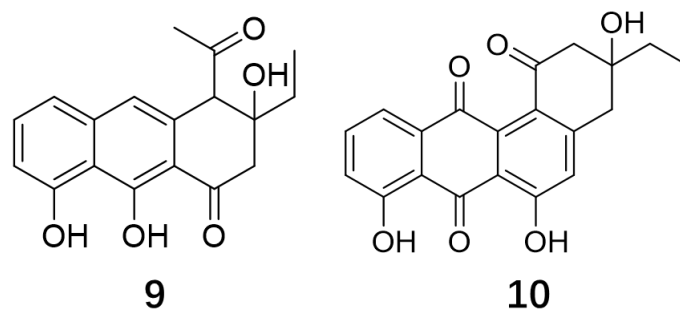


Fig 6.4 Structures of compounds **9** and **10**.

6.4 Strategy for starter unit generation

Malonyl-CoA is a commonly used starter unit primed onto the ACP in the minimal PKS system. Consequently, most polyketides constructed in this way possess a terminal methyl group.²⁷ The ethyl tails of compounds **9** and **10** (Fig 6.4) suggest the usage of alternate starter units. Apart from the conventional ACP (Ctg4_382, Fig 6.2b, coloured in maroon) in the minimal PKS system, *ctg4_377* (Fig 6.2b, Table 6.1) encodes a second ACP, showing 66.67% identity to CosE (ABC00733.1). CosE (Fig 6.5a, highlighted in yellow) is a ketoacylsynthase III-type condensation enzyme responsible for loading the propionyl-CoA starter unit in cosmomycin biosynthesis.¹⁴⁴ Homologous analysis revealed that the BGC005 pathway does not contain homologs of Lom62 (Fig 6.5a, coloured in green), a bifunctional acyltransferase/decarboxylase that is involved in A-ring ethyl group installation during lomaiviticin biosynthesis.¹⁴⁵ These implied that the biosynthesis of compounds **9** and **10** employs a starter unit generation strategy similar to that of cosmomycin, which catalyses an initial Claisen condensation between propionyl-CoA and malonyl-ACP. The resultant diketide is then introduced onto the KS α / β complex, where it participates in the downstream chain elongation process (Fig 6.5b).^{144,145}

Inspection of the GNPS molecular networking identified a node (m/z 299.09) directly connected to compound **9** (Fig 6.5c, left). The (-)-HRESIMS fragmentation of compound **9** mainly produced two types of fragments: naphthalene-type fragments (daughter ions m/z 173.061 and 215.072) and anthracene-type fragments (m/z 253.086 and 295.099) (Fig 6.5c, right). Four major ions were observed in the negative mode MS/MS of node_299.09: 173.060, 215.071, 239.070 and 281.080. Alignment of

the MS/MS spectra of node_299.09 with compound **9** suggested that the first two rings on these two compounds are identical. The mass difference of 14 (methyl group) took place in the starter unit region of the third ring. Taken together, the MS/MS spectra indicated that node_299.09 corresponds to a demethylated homologue of compound **9** (Fig 6.5c, right).

The above results suggested the existence of two starter unit generation strategies during biosynthesis: one adopts the extra ctg4_377 ACP (Fig 6.5b), and one uses the conventional ACP (Fig 6.5d).

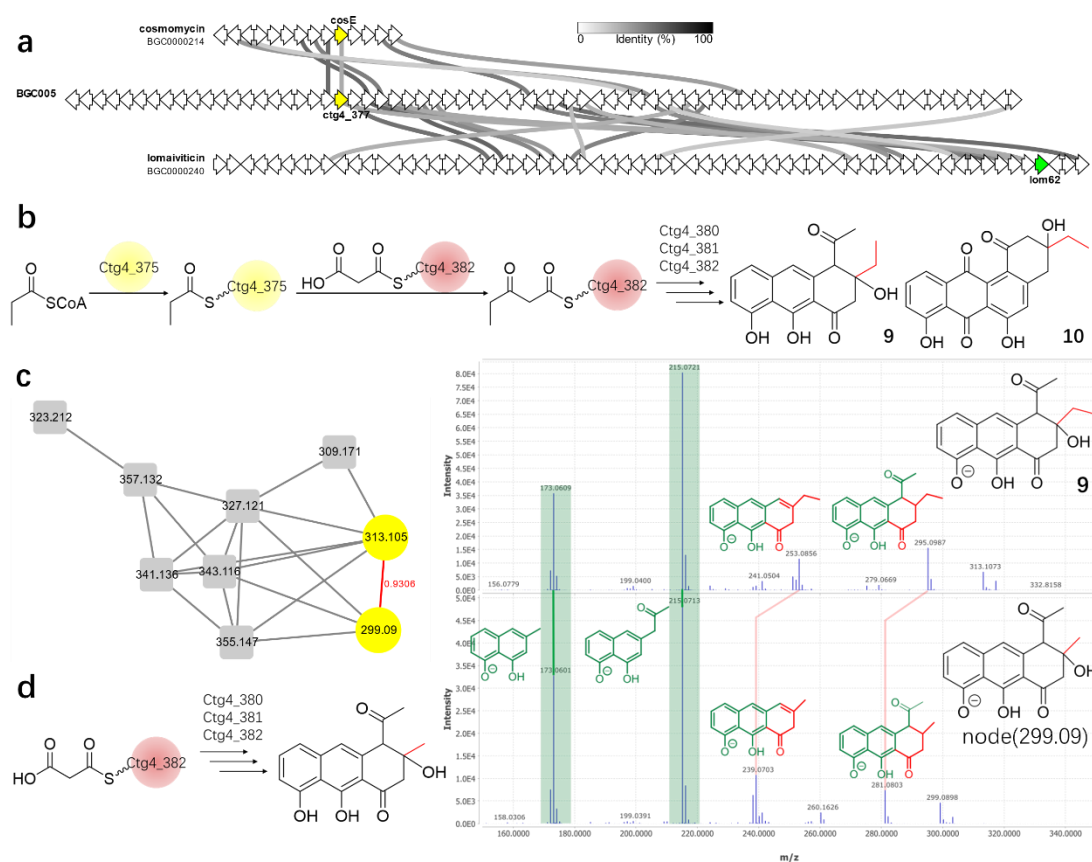


Fig 6.5 Chemical evidence showed that there were two starter unit generation strategies in BGC005

- Homologous alignment of BGC005 and two other T2PKS BGCs encoding polyaromatic compounds possessed ethyl tails
- Starter unit generation strategy for compounds **9** and **10** using a second ACP (Ctg4_375)
- Negative ion mode HRESIMS GNPS molecular network contained compound **9** and spectra alignment of compound **9** and node_299.09. The edge is labelled in red by the cosine score.
- Starter unit generation strategy for node_299.09 using the conventional ACP (Ctg4_382) in the minimal PKS system

6.5 Ring-opening oxygenase

BLAST analysis revealed that Ctg4_395 shared the highest homology with an uncharacterised antibiotic biosynthesis monooxygenase (89.52%, WP_132820917.1). This gene family is found exclusively on atypical T2PKS BGCs, mediates B-ring opening, and in conjunction with other enzymes, initiates the subsequent B-ring modification during biosynthesis of atypical angucyclines (Fig 6.6).^{136–139} There are three types of B ring modifications known so far: ring rearrangement (Fig 6.6, blue route), non-enzymatic amino acid incorporation (Fig 6.6, green route) and ring contraction (Fig 6.6, pink route).

The homology analysis of BGC005 revealed no genes involved in diazo assembly, precluding a mechanism of B ring contraction similar to that deduced for the biosynthesis of lomaiviticin¹³⁹ and other benzofluorene-containing angucyclines^{146–148} (Fig 6.6 pink route & Fig 6.7a, region highlighted in pink). In addition, BGC005 has no homologs like GilM (Fig 6.6 blue route & Fig 6.7a, gene in blue), a key multifunctional enzyme catalysing B-ring rearrangement during biosynthesis of gilvocarcin-type compounds.^{138,149} A phylogenetic analysis (Fig 6.7b) of these ring-opening oxygenases confirmed they formed a unique clade, and each subclade within this clade corresponded to distinct core structures of atypical angucyclines. The homology and phylogenetic analyses indicated that the final product of the BGC005 pathway is likely to be an atypical angucycline that either underwent a jadomycin-like modification (Fig 6.6 green route & Fig 6.7a, gene in purple) or a ring-opening process not previously reported.

6.6 Biosynthesis of compounds 9, 10

To further confirm that **9** and **10** are products of BGC005, we checked the metabolic profiles of the heterologous expression strains and two knockout models. Both **9** and **10** were readily detectable in the heterologous expression strain (Fig 6.8), indicating that compounds **9** and **10** are “genuine” BGC005 related metabolites and can be biosynthesised both natively in the producing strain and heterologously in *S. albus* Del14.

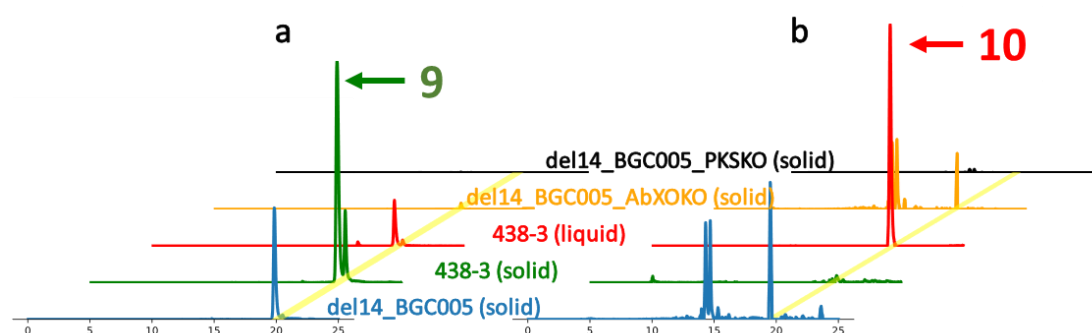


Fig 6.8 The detection of compounds **9** and **10** in different strains.

- Extracted LC-MS chromatograms (from formula $C_{18}H_{17}O_5$, corresponding to compound **9**) of the crude extracts from heterologous expression strain (del14_BGC005), native producing strain (438-3, solid and liquid fermentation), ring-opening enzyme mutant strain (del14_BGC005_AbXOKO), minimal PKS system mutant strain (del14_BGC005_PKSKO).
- Extracted LC-MS chromatograms (from formula $C_{20}H_{15}O_6$, corresponding to compound **10**) of the crude extracts from heterologous expression strain (del14_BGC005), native producing strain (438-3, solid and liquid fermentation), ring-opening enzyme mutant strain (del14_BGC005_AbXOKO), minimal PKS system mutant strain (del14_BGC005_PKSKO).

The heterologous expression and PKS knockout model demonstrated that the nonaketide (Compounds **9**) and the decaketide (Compounds **10**) were constructed by the collective action of Ctg4_380 - Ctg4_382 from eight malonyl-CoAs (Fig 6.9a) and nine malonyl-CoAs (Fig 6.9b), respectively. Following this, both nascent chains were subjected to C9-ketoreduction through the ketoreductase Ctg4_383. Cyclases Ctg4_379 and Ctg4_384 subsequently catalysed the cyclisation to convert the intermediates into **9** (cyclisation pattern: C7/C12, C5/C14, C4/C17) and **10** (cyclisation pattern: C7/C12, C5/C14, C4/C17, C2/C19).

The co-occurrence of compounds **9** and **10** in the native producing strain (Fig 6.8) indicated that the KS β (the chain length control enzyme Ctg4_381) exhibited a loose

control over chain length, allowing the formation of nonaketide and decaketide simultaneously.

The detection of compounds **9** and **10** in the ring-opening oxygenase (Ctg4_395) knockout model (del14_BGC005_AbXOKO) suggested that the biosynthesis of these two compounds (**9** and **10**) did not undergo any ring-opening/rearrangement (Fig 6.8).

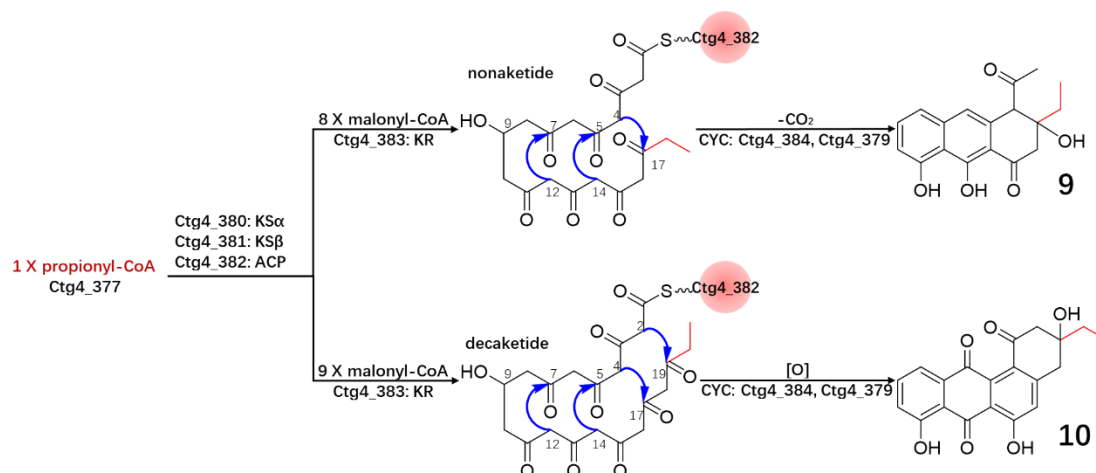


Fig 6.9 Proposed biosynthesis of compounds **9** and **10**.

The divergent biosynthetic routes for nonaketide and decaketide indicated the loose control over chain length during T2PKs biosynthesis.

6.7 Bioactivity testing

The antibacterial activity of newly isolated compound **9** was evaluated against strains of *Bacillus subtilis* E168, *Escherichia coli* (TolC deficient), and *Staphylococcus aureus*. A strong inhibitory effect of compound **9** was observed against *S. aureus* (2 µg/mL, Table 6.2).

Table 6.2 Minimum inhibitory concentration (MIC) values for isolated compound **9** against a range of bacteria

Test Culture	MIC (µg/ml)	
	compound 9	kanamycin
<i>B. subtilis</i> E168	32	2
<i>E. coli</i> (TolC-deficient)	32	2
<i>S. aureus</i>	2	1

6.8 Ongoing work toward finding final products of BGC005

In order to search for the final products of BGC005, metabolites of BGC005 and BGC005_ΔAbXO were eluted from HP20 co-incubated in the R5A liquid (14 days, 30 °C)

and fractionated on a benchtop LH20 column.

Although the colour differences (Fig 6.10) between metabolites of del14_005 and the corresponding AbXO knockout model (del14_005 Δ AbXO) suggested that the biosynthesis of BGC005 went much further than we currently obtained, since the compounds we isolated did not require the action of this gene (compounds **9** and **10** were still present when *abXO* gene was knocked out, Fig 6.8). After extensive investigation, we still failed to observe any differences (peaks) when running HPLC or MS, thus unable to identify the final products. These failures indicated that future work should focus on alternative chromatography or separation mode to isolate and characterise the final products of this BGC.

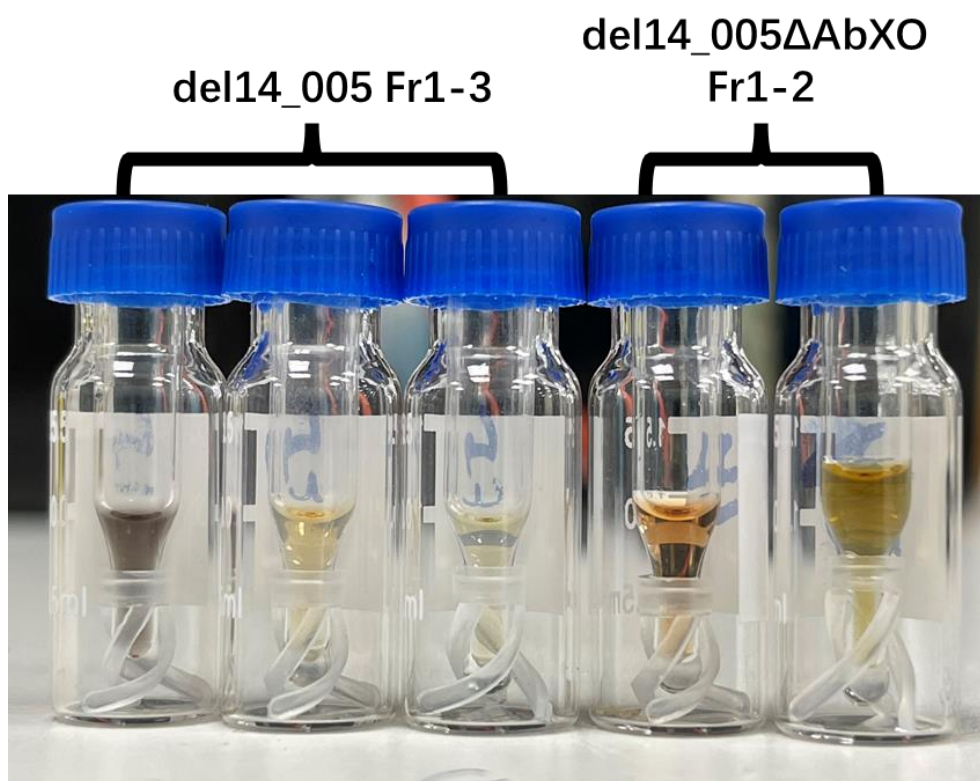


Fig 6.10 Metabolites from del14_005 and del14_005 Δ AbXO.

Both metabolites were eluted from HP20 co-incubated in R5A liquid (14 days, 30 °C) and fractionated on a benchtop LH20 column. The colour differences indicated del14_005 produced additional metabolites when compared to knockout control models (del14_005 Δ AbXO).

6.9 Conclusion and discussion

In conclusion, JB1081B (compound **9**) is a new polyaromatic natural product synthesised from a previously unreported, atypical T2PKS pathway and exhibited

significant antibacterial activity against *S. aureus*. The KS β (also known as chain length factor) has long been recognised as the key factor in polyketide chain length modulation.²⁸ However, sloppy chain length control has been reported previously. Using a heterologous expression strategy, Moore's, and Zhang's groups, respectively, have shown that expressing T2PKS systems in heterologous hosts can form a set of aromatic polyketides with diverse sizes and cyclisation patterns. These findings suggested that unidentified/unrelated enzymes in the host may crosstalk with the introduced T2PKS system and affect the backbone length and folding pattern.^{140,141} Several knockout/inactivation studies have also identified that certain oxygenases can play a role in determining chain length.^{150,151} These facts indicate that KS β alone cannot fully dictate the length of nascent poly- β -ketone chains. Instead, a more complicated and sophisticated/complex system governs the chain length control. The identification of JB1081B (**9**) and homorableomycin (**10**) provides additional support for the hypothesis that promiscuous chain length control also exists in native/non-manipulated host systems.

Unfortunately, the detection of JB1081B (**9**) and homorableomycin (**10**) in the ring-opening knockout model indicated that these two compounds are early metabolites rather than post-PKS tailoring products (Fig 6.8). Work is ongoing to characterise the final products of this pathway.

Finally, New Zealand has a wide variety of lichen species, the microbiomes of which remain understudied for their potential to produce novel natural products.^{9,13} Our data suggested that this natural resource is a fruitful avenue for discovering novel antimicrobial compounds and intriguing biosynthesis events.

6.10 Methods

6.10.1 General experimental procedures

Reversed-phase column chromatography was obtained using PhenoSphere-NEXT™ C18 column (3 μ m C18 120 Å, LC Column 150 x 4.6 mm) according to the parameters described in Table 2.10. UV/vis spectra were extracted from HPLC chromatograms.

A semi-preparative column (NUCLEODUR C18 HTec, 5 μ m, 125x21 mm) was used for compound isolation and purification following the protocol and parameters listed in Table 2.11. Compounds are purified in two rounds.

The molecular network was generated using the workflow described in Chapter 2.4.2.

The NMR data were recorded and processed following Chapter 2.4.3's description. The residual solvent peak was used as an internal chemical shift reference (CDCl₃: δ C 77.2; δ H 7.26).

6.10.2 Fermentation and metabolites extraction

For liquid fermentation:

To establish the seed culture, glycerol stock of *Streptomyces sp.* 438-3 was inoculated into 10 mL R5A liquid in a centrifuge tube. Two 2.5 L Ultrayield flasks (Thompson) containing R5a medium (1 L) with Diaion HP20 (30 g) were inoculated with 5 mL of seed culture after 72 hours of growth (30 °C, 200 rpm). HP20 beads were then harvested after 14-day cultivation (30 °C, 200 rpm), washed extensively with water, and extracted with MeOH (300 mL), which was then dried under vacuum at 30 °C. The dried extract was then reconstituted in MeOH. From the crude extracts, we purified compound **10**. Compound **10** was obtained as a yellow film.

For solid-phase fermentation:

The desired hosts (del14_005, del14_pTARa, del14_005 Δ PKS, del14_005 Δ AbXO, *Streptomyces sp.* 438-3) was plated on 12-well plate (Interlab, TCP-000-012) or 150mm Petri Dish (Interlab, PD-141) containing R5a agar. Following 14 days (30 °C) of incubation, the agar was harvested, soaked overnight in equal volumes of ethyl acetate, sonicated for 20 minutes, and then vacuum dried (30 °C). The dried extract was then redissolved in MeOH, and filtered through PTFE Syringe Filters (Sterlitech, 1470424) to remove particles. Compound **9** was purified and obtained as a yellow film from the crude extract of *Streptomyces sp.* 438-3 solid fermentation.

6.10.3 Bioassays

An established minimum inhibitory concentration (MIC) protocol¹⁵² was used to test the antimicrobial activity of compound **9** against *B. subtilis* E168, *E. coli* (TolC-deficient), and *S. aureus*. Three single colonies of each test strain were inoculated in LB and incubated overnight (30 °C, 200 rpm), which were then diluted (1:500) using Mueller–Hinton broth. The stock solution of Compound **9** was prepared at 50 mg/mL in DMSO and was further diluted in 2-fold serial dilutions in 96-well plates. Each 96-well plate well contained 190 µL diluted testing strain culture and 10 µL diluted Compound **9**. The plates were then shaken at 30 °C, 200 rpm for 16–20 h. The microbial growth was assessed by measuring the absorbance of each well at 600 nm (Enspire 2300 multilabel reader, PerkinElmer; Waltham, MA, USA). The tests were carried out in biological triplicate. The positive control was kanamycin, while the negative controls were DMSO and media-only.

6.10.4 Cloning of the BGC005 cluster

See Chapter 2.3.

Gene fragments and primers are provided in Appendix 8.

6.10.5 Construction of the mutant strains

Knockout models were obtained by *in vivo* Red/ET reaction. The knockout process was conducted according to the manufacturer's instructions (GENE BRIDGES, Quick & Easy *E. coli* Gene Deletion Kit).

Chapter 7 Conclusion and Future work

7.1 Genomic-driven exploration of lichen associated actinobacteria

A major focus of this thesis was exploring the possibility of using New Zealand lichens as an alternative ecological niche in search of new bioactive natural products. We began our journey by isolating actinobacteria from lichens spanning New Zealand and then prepared NGS libraries using a high-throughput workflow that we developed in-house. 332 of the 480 sequencing libraries yielded qualified data that were suitable for downstream analysis.

First, we evaluated the taxonomic diversity of the 332 sequenced isolates. Our assigned isolates (166) have relatively low ANI values ($> 95\%$, $\leq 98\%$) compared to their closest GTDB-Tk reference genomes.⁷⁵ The majority (166, 50%) of our sequenced isolates remain unclassified and can only be categorised as unassigned species within their respective genera. The results demonstrate the novelty of actinobacteria assemblages in New Zealand lichens.

The resistome predictions revealed the abundance and diversity of resistance genes in the actinomycetes from New Zealand lichens. The high abundance of β -lactamases was consistent with the fact that certain soil microorganisms were β -lactam producers.¹⁵³ Further analysis of the evolutionary distance of lichen sourced resistance genes to known AMR genes can help us infer the potential impact of the lichen sourced AMR gene on human health.

To evaluate the BGC diversity and novelty, we mapped our identified BGCs to a pre-built GCF model.⁷⁷ Among our 8601 BGCs, 3.74% BGCs are "orphans" in the GCF model, indicating these BGCs are genetically divergent from the 1.2 million BGCs. However, genetic divergence does not always reflect chemical novelty, so we next grouped the BGCs based on the potential chemical space they occupied by associating our query BGCs to the MIBiG reference BGCs using a new analysis method. Finally, we selected

several cryptic BGCs mined from our current datasets to carry out the functional study.

This study demonstrated the feasibility of working with cost-efficient sequencing data, and illustrated a pipeline for systematically evaluating large-scale BGC datasets. The preliminary comparison result of our dataset (Fig 7.1: lichen) with *Streptomyces* from insects (Fig 7.1: insect), oceanic biofilm-forming microbiome (Fig 7.1: marine) and New Zealand soil environmental DNA cosmid libraries (Fig 7.1: soil) highlighted the previously understudied cohort of actinobacteria associated with New Zealand lichens as a potentially rich and unique (Fig 7.1a) biosynthetic resource for the discovery of novel (Fig 7.1b) secondary metabolites.

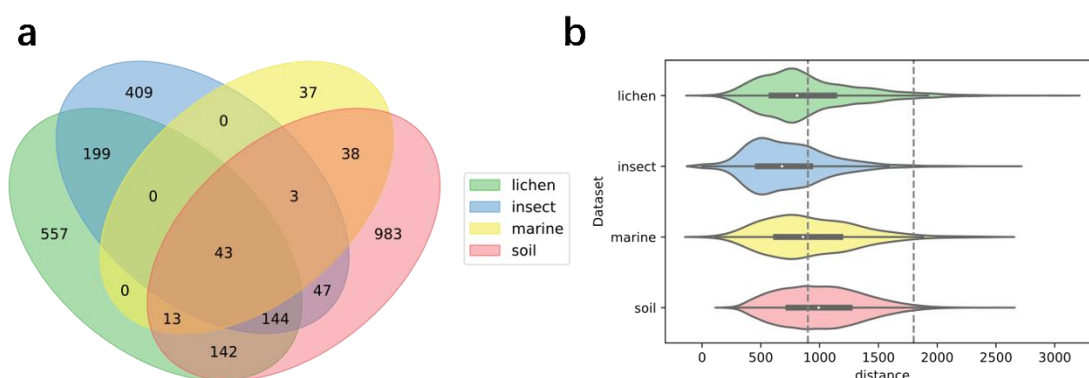


Fig 7.1 BiG-SLiCE comparison analysis of four environmental datasets.

- Venn diagram showing the distribution of all BiG-SLiCE GCFs in the four environmental datasets. 557 GCFs were unique to the lichen dataset.
- BiG-SLiCE membership assignments for four environmental datasets. 3.74% BGCs in the lichen dataset have orphan memberships, while 0.8%, 2.54% and 1.8% of the insect, marine and soil BGCs are 'orphan' BGCs.

7.2 Gene cluster and metabolites from BGC031

This section describes the heterologous expression and pathway refactoring of a puromycin-like BGC isolated from an actinomycete from New Zealand lichens. It was found that the yields of the pathway could be enhanced by overexpressing the PPTases or incorporating specific trace elements into the medium. These facts suggested a different regulation network existed in BGC031 from those previously reported.¹²⁵

This part also demonstrated the feasibility of using GNPS to uncover new variations of compounds and biosynthetic logic from previously investigated pathways. Even though antiSMASH calculated that BGC031 has 100% biosynthetic similarity to a

known puromycin pathway, cutting-edge MS/MS platform GNPS enabled the discovery of novel puromycin congeners (Fig 5.5). We traced part of the congeners back to the well-characterised puromycin pathway, while the remainders suggested a bypass (Fig 5.10a) and divergence (Fig 5.10b) in the biosynthesis of BGC031. This discovery will stimulate the future exploration and validation of the biochemical transformation mechanisms. One potential application of the new knowledge could be generating new puromycin congeners with enhanced bioactivities.

7.3 Gene cluster and metabolites of the BGC005

Here, we have used our in-house GNPS workflow to rapidly identify several pathway-specific molecular families from the complex heterologous expression systems and leading to the successful validation and characterisation of a new compound from one of our *Streptomyces sp.* (strain 438-3) that exhibited significant antibacterial activity when tested against *S. aureus*.

Many intriguing biosynthesis events took place during the biosynthesis of BGC005. An uncommon starter unit generation strategy has been uncovered, and different starter unit choices have also been observed. Sloppy chain length control has been reported previously in heterologous expression systems^{140,141} and engineered strains^{150,151}. The identification of JB1081B (**9**) and homorableomycin (**10**) supported the hypothesis that the promiscuous chain length control also existed in the native/non-manipulated host system.

Additionally, based on the presence of AbXO gene on BGC005, BGC005 is proposed to encode an atypical angucycline. Atypical angucyclines (BGCs) are rare. Based on the sugar biosynthesis genes on BGC005 and phylogenetic placement of AbXO genes, we believe we identified a novel atypical angucycline BGC sourced from New Zealand.^{137–139} Current failure of identifying the BGC005 final products suggested that future work should focus on alternative separation (e.g., HILIC column) and detection mode (e.g., APCI-MS) to isolate and characterise the final products of this BGC.

7.4 MbtH-like protein

Our heterologous expression results indicated that most of the NRPS-related pathways we selected were silent or poorly expressed. Among those NRPS-derived BGCs, the presence of the MbtH-like protein is a conserved feature. It has become clearer that MbtH-like proteins are crucial to bacterial NRPS biosynthesis. In the absence of DptG, an MbtH-like protein on daptomycin BGC, the production of daptomycin is reduced.¹⁵⁴ The knockout of an MbtH-like protein GplH caused the abortion of glycopeptidolipids production.¹⁵⁵ According to a recent study, certain MbtH-like proteins can also function as chaperoning proteins for NRPSs, promoting the folding, stability, and solubility of NRPSs. Driessen's group successfully stimulated filamentous fungi NRPS synthesis by using several bacterial MbtH-like proteins.¹⁵⁶

Reexamination of our antiSMASH results revealed that the ratio of MbtH homologs per genome and the NRPS regions per genome is not always 1:1 (Fig 7.2). The result supported previous findings that one MbtH homolog may crosstalk with more than one NRPS BGCs.^{154,156} Inspired by the Qu group PPTase-based activation strategy¹⁹, future work could aim at constructing an MbtH-like protein-based global activation system.

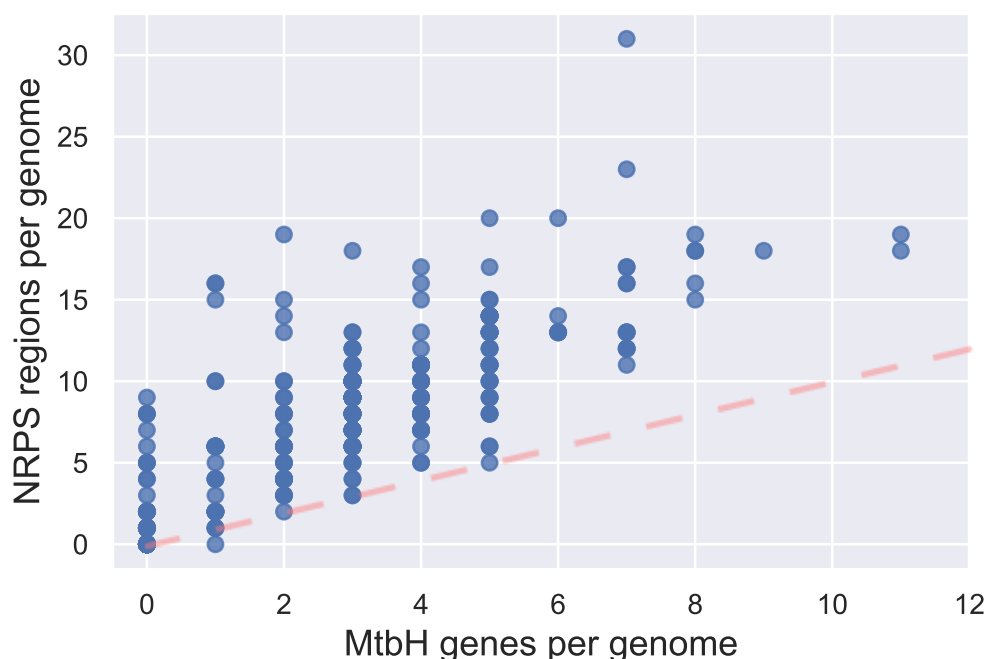


Fig 7.2 Correlation between MbtH-like proteins and NRPS regions of each genome in our datasets

7.5 Culture First vs. Genetic First

In the golden age of antibiotic discovery, bacteria were isolated, cultivated and screened for antibiotic activity. However, only bioactive compounds were characterised through the culture-first approach, and the chemical potential of the bacteria was not fully tapped.

With the advances in genome sequencing and data analysis, sequencing the bacteria genome and genome mining allowed scientists to gain insight into the biosynthetic potential bacteria harboured. Genetic-first approach enabled us to choose the bacteria/pathways we wanted to culture and study in a smarter, more targeted way.

In this thesis project, through genome mining, we did the heterologous expression study of the selected BGCs and fermentation study of the native producing strains of the BGCs in parallel (data not shown).

Genome mining-based study yielded several promising findings:

- BGC005 is proposed to encode a rare atypical angucycline. So far, only less than 10 atypical angucyclines have been reported. Atypical angucyclines are usually good anticancer drug candidates.^{134,137,143,146,149}
- We also identified several BGCs that are proposed to encode antibiotics like daptomycin and vancomycin. These would be promising leads to the fight against the antibiotic (resistance) crisis.

While we did not have much luck in the culture-based study, we still found two interesting, unexpected instances listed below.

7.5.1 New lanthipeptides detected from the BGC009 producing strain

The unsuccessful heterologous expression of one of our pathways (BGC009) led us to go back to its producing strain. While conducting conventional fermentation of BGC009 producing strain, we still failed to link any metabolites to pathway 009 but were able to identify a new molecular family consisting of potential new

lanthipeptides with masses around 3000 Da (Fig 7.3). When revisiting the antiSMASH result of BGC009 producing strain, we found one candidate BGC, and its predicted peptide (Fig 7.4) matched one of the parent ions (Fig 7.3, node highlighted in yellow). This Molecular cluster/BGC could be a good lead for further study.

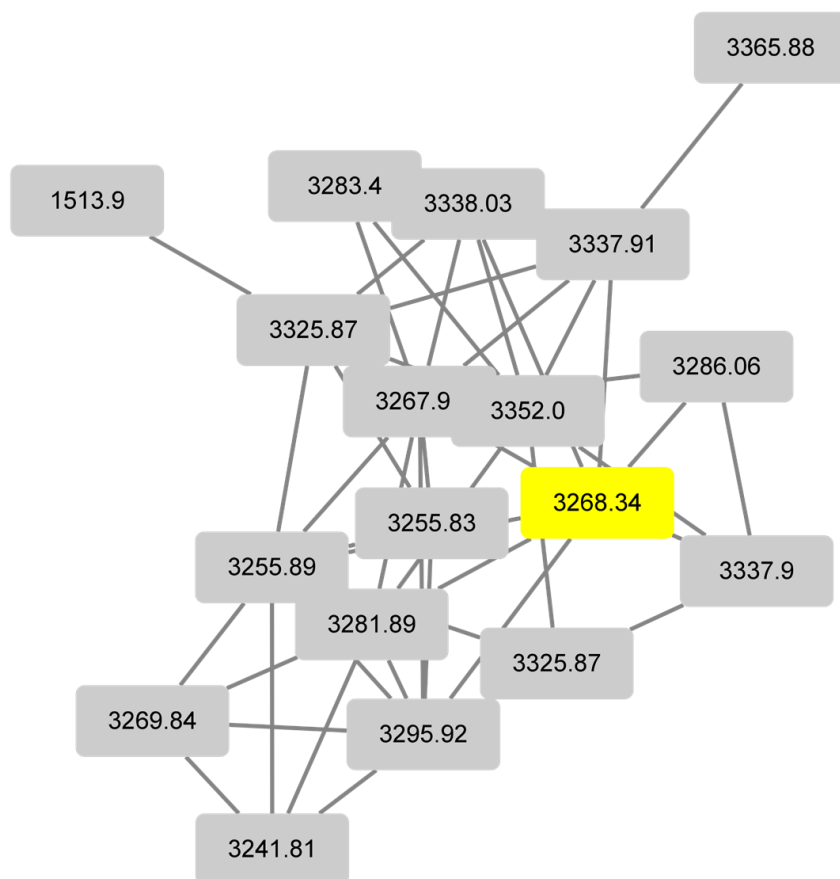


Fig 7.3 A GNPS molecular network identified in the BGC009 native producing strain.

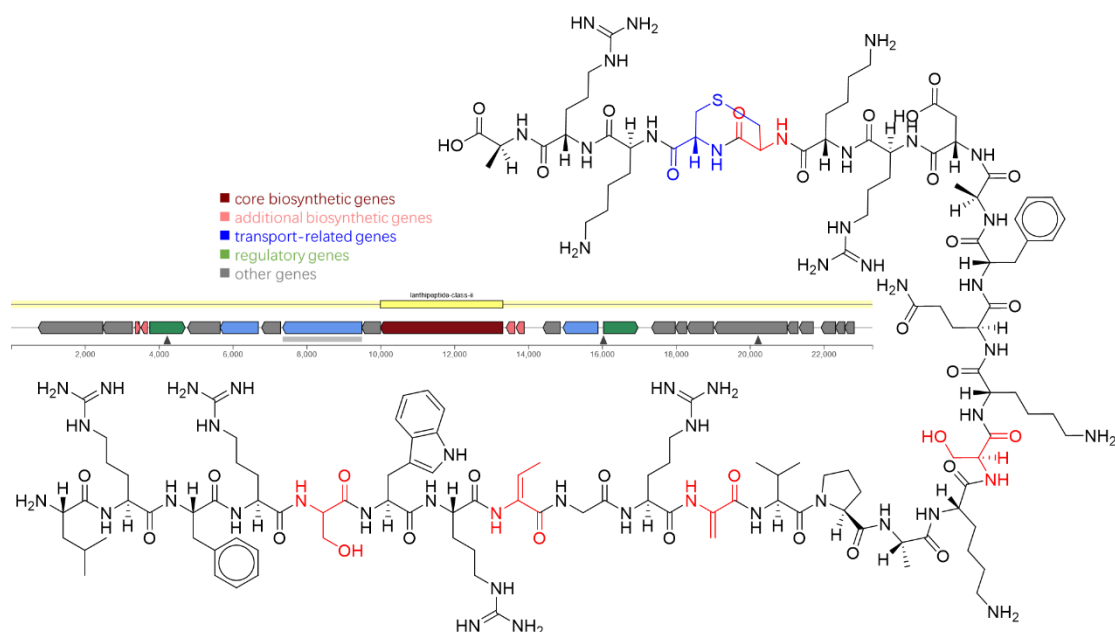


Fig 7.4 A candidate lanthipeptide BGC and its proposed structure.

The high molecular weight in Fig 7.3 molecular networking led us to revisit the antiSMASH result of BGC009 producing strain with a focus on peptide-like BGCs. We found that a core peptide from one candidate lanthipeptide BGC matched the highlighted parent ions in Fig 7.3.

7.5.2 Primary metabolisms of the actinobacteria assemblages in New Zealand lichens

We frequently observed two coloured compounds with the same retention times during cultivation, organic extraction, and fractionation experiments. These compounds were found in the crude extracts of multiple native producing strains and heterologous expression strains, suggesting they are common metabolites among actinomycetes; however, we could not link their presence to any secondary metabolite BGCs. After further investigation using HRESIMS/MS, it was confirmed that they are two known primary metabolites, biliverdin¹⁵⁷ and coproporphyrin III¹⁵⁸ (Fig 7.5). Subsequent analysis using KEGG orthology¹⁵⁹ revealed that the enzymes involved in biliverdin and coproporphyrin III biosynthesis are particularly abundant in actinomycetes (Fig 7.6)¹⁶⁰. It is interesting to explore the reason for this abundance. One hypothesis is that actinomycetes utilised the heme related metabolites to chelate metals presented in the environment to improve their resistance.¹⁶¹

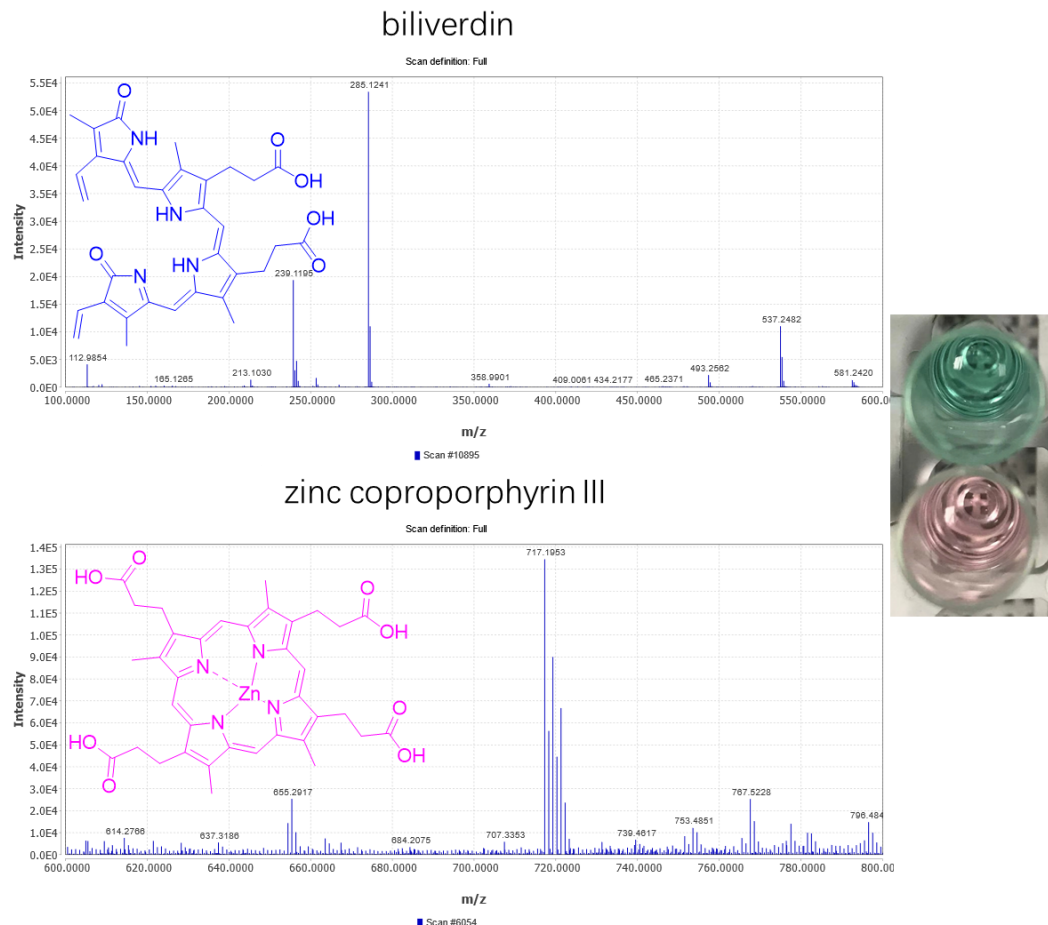
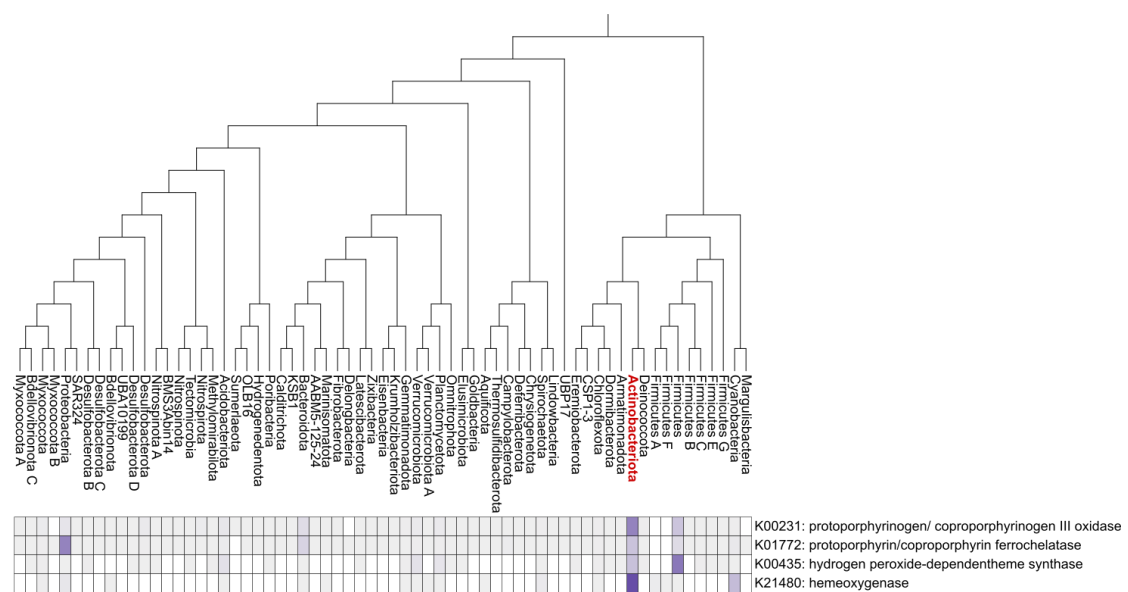


Fig 7.5 Two common metabolites found during fermentation. The structures of these two metabolites were deduced from MS/MS.



Reference

1. Genilloud O. Actinomycetes: still a source of novel antibiotics. *Nat Prod Rep.* 2017;34(10):1203-1232. doi:10.1039/C7NP00026J
2. Newman DJ, Cragg GM. Natural Products as Sources of New Drugs over the Nearly Four Decades from 01/1981 to 09/2019. *J Nat Prod.* 2020;83(3):770-803. doi:10.1021/acs.jnatprod.9b01285
3. Katz L, Baltz RH. Natural product discovery: past, present, and future. *J Ind Microbiol Biotechnol.* 2016;43(2-3):155-176. doi:10.1007/s10295-015-1723-5
4. Azman A-S, Othman I, Velu SS, Chan K-G, Lee L-H. Mangrove rare actinobacteria: taxonomy, natural compound, and discovery of bioactivity. *Front Microbiol.* 2015;6. doi:10.3389/fmicb.2015.00856
5. Kamjam M, Sivalingam P, Deng Z, Hong K. Deep Sea Actinomycetes and Their Secondary Metabolites. *Front Microbiol.* 2017;8. doi:10.3389/fmicb.2017.00760
6. Hamm PS, Caimi NA, Northup DE, et al. Western Bats as a Reservoir of Novel Streptomyces Species with Antifungal Activity. Schloss PD, ed. *Appl Environ Microbiol.* 2017;83(5). doi:10.1128/AEM.03057-16
7. González I, Ayuso-Sacido A, Anderson A, Genilloud O. Actinomycetes isolated from lichens: Evaluation of their diversity and detection of biosynthetic gene sequences. *FEMS Microbiol Ecol.* 2005;54(3):401-415. doi:10.1016/j.femsec.2005.05.004
8. Grimm M, Grube M, Schiefelbein U, Zühlke D, Bernhardt J, Riedel K. The Lichens' Microbiota, Still a Mystery? *Front Microbiol.* 2021;12. doi:10.3389/fmicb.2021.623839
9. Calcott MJ, Ackerley DF, Knight A, Keyzers RA, Owen JG. Secondary metabolism in the lichen symbiosis. *Chem Soc Rev.* 2018;47(5):1730-1760. doi:10.1039/C7CS00431A
10. Parrot D, Antony-Babu S, Intertaglia L, Grube M, Tomasi S, Suzuki MT. Littoral lichens as a novel source of potentially bioactive Actinobacteria. *Sci Rep.* 2015;5(1):15839. doi:10.1038/srep15839
11. Liu C, Jiang Y, Wang X, et al. Diversity, Antimicrobial Activity, and Biosynthetic Potential of Cultivable Actinomycetes Associated with Lichen Symbiosis. *Microb Ecol.* 2017;74(3):570-584. doi:10.1007/s00248-017-0972-4
12. Sánchez-Hidalgo M, González I, Díaz-Muñoz C, Martínez G, Genilloud O. Comparative Genomics and Biosynthetic Potential Analysis of Two Lichen-Isolated Amycolatopsis Strains. *Front Microbiol.* 2018;9.

doi:10.3389/fmicb.2018.00369

13. Bracegirdle J, Hou P, Nowak V V., Ackerley DF, Keyzers RA, Owen JG. Skyllamycins D and E, Non-Ribosomal Cyclic Depsipeptides from Lichen-Sourced *Streptomyces anulatus*. *J Nat Prod*. 2021;84(9):2536-2543. doi:10.1021/acs.jnatprod.1c00547
14. Davies J, Wang H, Taylor T, Warabi K, Huang X-H, Andersen RJ. Uncialamycin, A New Eneidyne Antibiotic. *Org Lett*. 2005;7(23):5233-5236. doi:10.1021/ol052081f
15. Wishart DS. DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res*. 2006;34(90001):D668-D672. doi:10.1093/nar/gkj067
16. Winn M, Fyans JK, Zhuo Y, Micklefield J. Recent advances in engineering nonribosomal peptide assembly lines. *Nat Prod Rep*. 2016;33(2):317-347. doi:10.1039/C5NP00099H
17. Fischbach MA, Walsh CT. Assembly-Line Enzymology for Polyketide and Nonribosomal Peptide Antibiotics: Logic, Machinery, and Mechanisms. *Chem Rev*. 2006;106(8):3468-3496. doi:10.1021/cr0503097
18. Hur GH, Vickery CR, Burkart MD. Explorations of catalytic domains in non-ribosomal peptide synthetase enzymology. *Nat Prod Rep*. 2012;29(10):1074. doi:10.1039/c2np20025b
19. Yan X, Zhang B, Tian W, et al. Puromycin A, B and C, cryptic nucleosides identified from *Streptomyces alboniger* NRRL B-1832 by PPTase-based activation. *Synth Syst Biotechnol*. 2018;3(1):76-80. doi:10.1016/j.synbio.2018.02.001
20. Mollo A, von Krusenstiern AN, Bulos JA, et al. P450 monooxygenase ComJ catalyses side chain phenolic cross-coupling during complestatin biosynthesis. *RSC Adv*. 2017;7(56):35376-35384. doi:10.1039/C7RA06518C
21. Staunton J, Weissman KJ. Polyketide biosynthesis: a millennium review. *Nat Prod Rep*. 2001;18(4):380-416. doi:10.1039/a909079g
22. Shen B. Polyketide biosynthesis beyond the type I, II and III polyketide synthase paradigms. *Curr Opin Chem Biol*. 2003;7(2):285-295. doi:10.1016/S1367-5931(03)00020-6
23. Staunton J, Wilkinson B. Biosynthesis of Erythromycin and Rapamycin. *Chem Rev*. 1997;97(7):2611-2630. doi:10.1021/cr9600316
24. Chen H, Bian Z, Ravichandran V, et al. Biosynthesis of polyketides by trans -AT polyketide synthases in Burkholderiales. *Crit Rev Microbiol*. 2019;45(2):162-181. doi:10.1080/1040841X.2018.1514365

25. Wang B, Guo F, Huang C, Zhao H. Unraveling the iterative type I polyketide synthases hidden in *Streptomyces*. *Proc Natl Acad Sci*. 2020;117(15):8449-8454. doi:10.1073/pnas.1917664117
26. Wang J, Zhang R, Chen X, et al. Biosynthesis of aromatic polyketides in microorganisms using type II polyketide synthases. *Microb Cell Fact*. 2020;19(1):110. doi:10.1186/s12934-020-01367-4
27. Zhang W, Tang Y. Chapter 16 In Vitro Analysis of Type II Polyketide Synthase. In: ; 2009:367-393. doi:10.1016/S0076-6879(09)04616-3
28. Tang Y, Tsai S-C, Khosla C. Polyketide Chain Length Control by Chain Length Factor. *J Am Chem Soc*. 2003;125(42):12708-12709. doi:10.1021/ja0378759
29. Funa N, Ohnishi Y, Fujii I, Shibuya M, Ebizuka Y, Horinouchi S. A new pathway for polyketide synthesis in microorganisms. *Nature*. 1999;400(6747):897-899. doi:10.1038/23748
30. Zha W, Rubin-Pitel SB, Zhao H. Characterization of the Substrate Specificity of PhlD, a Type III Polyketide Synthase from *Pseudomonas fluorescens*. *J Biol Chem*. 2006;281(42):32036-32047. doi:10.1016/S0021-9258(19)84117-0
31. Medema MH, Kottmann R, Yilmaz P, et al. Minimum Information about a Biosynthetic Gene cluster. *Nat Chem Biol*. 2015;11(9):625-631. doi:10.1038/nchembio.1890
32. Albarano L, Esposito R, Ruocco N, Costantini M. Genome Mining as New Challenge in Natural Products Discovery. *Mar Drugs*. 2020;18(4). doi:10.3390/md18040199
33. Syed F, Grunenwald H, Caruccio N. Next-generation sequencing library preparation: simultaneous fragmentation and tagging using in vitro transposition. *Nat Methods*. 2009;6(11):i-ii. doi:10.1038/nmeth.f.272
34. Understanding the NGS workflow. <https://www.illumina.com/science/technology/next-generation-sequencing/beginners/ngs-workflow.html>. Accessed January 14, 2022.
35. Baker M. De novo genome assembly: what every biologist should know. *Nat Methods*. 2012;9(4):333-337. doi:10.1038/nmeth.1935
36. Bankevich A, Nurk S, Antipov D, et al. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *J Comput Biol*. 2012;19(5):455-477. doi:10.1089/cmb.2012.0021
37. Namiki T, Hachiya T, Tanaka H, Sakakibara Y. MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads. *Nucleic Acids Res*. 2012;40(20):e155-e155. doi:10.1093/nar/gks678

38. Rapid Sequencing Kit. <https://store.nanoporetech.com/rapid-sequencing-kit.html>. Accessed January 14, 2022.
39. How nanopore sequencing works. <https://nanoporetech.com/how-it-works>. Accessed January 14, 2022.
40. Chen Z, Erickson DL, Meng J. Benchmarking hybrid assembly approaches for genomic analyses of bacterial pathogens using Illumina and Oxford Nanopore sequencing. *BMC Genomics*. 2020;21(1):631. doi:10.1186/s12864-020-07041-8
41. Ziemert N, Alanjary M, Weber T. The evolution of genome mining in microbes – a review. *Nat Prod Rep*. 2016;33(8):988-1005. doi:10.1039/C6NP00025H
42. Medema MH, Blin K, Cimermancic P, et al. AntiSMASH: Rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. *Nucleic Acids Res*. 2011;39(suppl_2):W339-W346. doi:10.1093/nar/gkr466
43. Blin K, Shaw S, Steinke K, et al. antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline. *Nucleic Acids Res*. 2019;47(W1):W81-W87. doi:10.1093/nar/gkz310
44. Minowa Y, Araki M, Kanehisa M. Comprehensive Analysis of Distinctive Polyketide and Nonribosomal Peptide Structural Motifs Encoded in Microbial Genomes. *J Mol Biol*. 2007;368(5):1500-1517. doi:10.1016/j.jmb.2007.02.099
45. Röttig M, Medema MH, Blin K, Weber T, Rausch C, Kohlbacher O. NRPSpredictor2—a web server for predicting NRPS adenylation domain specificity. *Nucleic Acids Res*. 2011;39(suppl_2):W362-W367. doi:10.1093/nar/gkr323
46. Kautsar SA, Blin K, Shaw S, et al. MIBiG 2.0: a repository for biosynthetic gene clusters of known function. *Nucleic Acids Res*. October 2019. doi:10.1093/nar/gkz882
47. Blin K, Medema MH, Kazempour D, et al. antiSMASH 2.0—a versatile platform for genome mining of secondary metabolite producers. *Nucleic Acids Res*. 2013;41(W1):W204-W212. doi:10.1093/nar/gkt449
48. Kouprina N, Larionov V. TAR cloning: insights into gene function, long-range haplotypes and genome structure and evolution. *Nat Rev Genet*. 2006;7(10):805-812. doi:10.1038/nrg1943
49. Gibson DG, Young L, Chuang R-Y, Venter JC, Hutchison CA, Smith HO. Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat Methods*. 2009;6(5):343-345. doi:10.1038/nmeth.1318
50. Engler C, Kandzia R, Marillonnet S. A One Pot, One Step, Precision Cloning Method with High Throughput Capability. El-Shemy HA, ed. *PLoS One*.

- 2008;3(11):e3647. doi:10.1371/journal.pone.0003647
51. Myronovskyi M, Rosenkränzer B, Nadmid S, Pujic P, Normand P, Luzhetskyy A. Generation of a cluster-free *Streptomyces albus* chassis strains for improved heterologous expression of secondary metabolite clusters. *Metab Eng.* 2018;49:316-324. doi:10.1016/j.ymben.2018.09.004
 52. Zaburannyi N, Rabyk M, Ostash B, Fedorenko V, Luzhetskyy A. Insights into naturally minimised *Streptomyces albus* J1074 genome. *BMC Genomics.* 2014;15(1):97. doi:10.1186/1471-2164-15-97
 53. Rückert C, Albersmeier A, Busche T, et al. Complete genome sequence of *Streptomyces lividans* TK24. *J Biotechnol.* 2015;199:21-22. doi:10.1016/j.jbiotec.2015.02.004
 54. Gomez-Escribano JP, Bibb MJ. Engineering *Streptomyces coelicolor* for heterologous expression of secondary metabolite gene clusters. *Microb Biotechnol.* 2011;4(2):207-215. doi:10.1111/j.1751-7915.2010.00219.x
 55. Kim S-H, Lu W, Ahmadi MK, Montiel D, Ternei MA, Brady SF. Atolypenes, Tricyclic Bacterial Sesterterpenes Discovered Using a Multiplexed In Vitro Cas9-TAR Gene Cluster Refactoring Approach. *ACS Synth Biol.* 2019;8(1):109-118. doi:10.1021/acssynbio.8b00361
 56. Kallifidas D, Kang H-S, Brady SF. Tetarimycin A, an MRSA-Active Antibiotic Identified through Induced Expression of Environmental DNA Gene Clusters. *J Am Chem Soc.* 2012;134(48):19552-19555. doi:10.1021/ja3093828
 57. Yamanaka K, Reynolds KA, Kersten RD, et al. Direct cloning and refactoring of a silent lipopeptide biosynthetic gene cluster yields the antibiotic taromycin A. *Proc Natl Acad Sci.* 2014;111(5):1957-1962. doi:10.1073/pnas.1319584111
 58. Wang B, Guo F, Dong S-H, Zhao H. Activation of silent biosynthetic gene clusters using transcription factor decoys. *Nat Chem Biol.* 2019;15(2):111-114. doi:10.1038/s41589-018-0187-0
 59. Sarker SD, Nahar L. Applications of High Performance Liquid Chromatography in the Analysis of Herbal Products. In: *Evidence-Based Validation of Herbal Medicine.* Elsevier; 2015:405-425. doi:10.1016/B978-0-12-800874-4.00019-2
 60. Overview of Mass Spectrometry for Protein Analysis. <https://www.thermofisher.com/nz/en/home/life-science/protein-biology/protein-biology-learning-center/protein-biology-resource-library/pierce-protein-methods/overview-mass-spectrometry.html>. Accessed January 17, 2022.
 61. Liquid Chromatography Mass Spectrometry (LC-MS) Information. <https://www.thermofisher.com/nz/en/home/industrial/mass-spectrometry/mass-spectrometry-learning-center/liquid-chromatography->

- mass-spectrometry-lc-ms-information.html. Accessed January 17, 2022.
62. WHAT IS MASS SPECTROMETRY? <https://www.broadinstitute.org/technology-areas/what-mass-spectrometry>. Accessed January 17, 2022.
 63. Wang M, Carver JJ, Phelan V V, et al. Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nat Biotechnol.* 2016;34(8):828-837. doi:10.1038/nbt.3597
 64. Larson CB, Crüsemann M, Moore BS. PCR-Independent Method of Transformation-Associated Recombination Reveals the Cosmomyacin Biosynthetic Gene Cluster in an Ocean Streptomyccete. *J Nat Prod.* 2017;80(4):1200-1204. doi:10.1021/acs.jnatprod.6b01121
 65. Bauman KD, Li J, Murata K, et al. Refactoring the Cryptic Streptophenazine Biosynthetic Gene Cluster Unites Phenazine, Polyketide, and Nonribosomal Peptide Biochemistry. *Cell Chem Biol.* 2019;26(5):724-736.e7. doi:10.1016/j.chembiol.2019.02.004
 66. Simon R, Priefer U, Pühler A. A Broad Host Range Mobilization System for In Vivo Genetic Engineering: Transposon Mutagenesis in Gram Negative Bacteria. *Bio/Technology.* 1983;1(9):784-791. doi:10.1038/nbt1183-784
 67. Montiel D, Kang H-S, Chang F-Y, Charlop-Powers Z, Brady SF. Yeast homologous recombination-based promoter engineering for the activation of silent natural product biosynthetic gene clusters. *Proc Natl Acad Sci.* 2015;112(29):8953-8958. doi:10.1073/pnas.1507606112
 68. Spizizen J. TRANSFORMATION OF BIOCHEMICALLY DEFICIENT STRAINS OF BACILLUS SUBTILIS BY DEOXYRIBONUCLEATE. *Proc Natl Acad Sci U S A.* 1958;44(10):1072-1078. doi:10.1073/pnas.44.10.1072
 69. Hong H-J, Hutchings MI, Hill LM, Buttner MJ. The Role of the Novel Fem Protein VanK in Vancomycin Resistance in Streptomyces coelicolor. *J Biol Chem.* 2005;280(13):13055-13061. doi:10.1074/jbc.M413801200
 70. Kieser T, Bibb MJ, Buttner MJ, Chater KF, Hopwood DA, others. *Practical Streptomyces Genetics*. Vol 291. John Innes Foundation Norwich; 2000.
 71. Hennig BP, Velten L, Racke I, et al. Large-Scale Low-Cost NGS Library Preparation Using a Robust Tn5 Purification and Tagmentation Protocol. *G3 Genes/Genomes/Genetics.* 2018;8(1):79-89. doi:10.1534/g3.117.300257
 72. Brady SF. Construction of soil environmental DNA cosmid libraries and screening for clones that produce biologically active small molecules. *Nat Protoc.* 2007;2(5):1297-1305. doi:10.1038/nprot.2007.195
 73. Kallifidas D, Brady SF. Reassembly of Functionally Intact Environmental DNA-Derived Biosynthetic Gene Clusters. In: *Methods in Enzymology*. Vol 517. ;

- 2012;225-239. doi:10.1016/B978-0-12-404634-4.00011-5
74. Lee NCO, Larionov V, Kouprina N. Highly efficient CRISPR/Cas9-mediated TAR cloning of genes and chromosomal loci from complex genomes in yeast. *Nucleic Acids Res.* 2015;43(8):e55-e55. doi:10.1093/nar/gkv112
 75. Chaumeil P-A, Mussig AJ, Hugenholtz P, Parks DH. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. Hancock J, ed. *Bioinformatics*. November 2019. doi:10.1093/bioinformatics/btz848
 76. Navarro-Muñoz JC, Selem-Mojica N, Mullooney MW, et al. A computational framework to explore large-scale biosynthetic diversity. *Nat Chem Biol.* 2020;16(1):60-68. doi:10.1038/s41589-019-0400-9
 77. Kautsar SA, van der Hooft JJJ, de Ridder D, Medema MH. BiG-SLiCE: A highly scalable tool maps the diversity of 1.2 million biosynthetic gene clusters. *Gigascience.* 2021;10(1). doi:10.1093/gigascience/giaa154
 78. assembly_stats 0.1.4. doi:10.5281/zenodo.3968775
 79. Ondov BD, Treangen TJ, Melsted P, et al. Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol.* 2016;17(1):132. doi:10.1186/s13059-016-0997-x
 80. Bateman A, Martin M-J, Orchard S, et al. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.* 2021;49(D1):D480-D489. doi:10.1093/nar/gkaa1100
 81. Tatusov RL. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res.* 2000;28(1):33-36. doi:10.1093/nar/28.1.33
 82. Alcock BP, Raphenya AR, Lau TTY, et al. CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. *Nucleic Acids Res.* October 2019. doi:10.1093/nar/gkz935
 83. Couvin D, Bernheim A, Toffano-Nioche C, et al. CRISPRCasFinder, an update of CRISPRFinder, includes a portable version, enhanced performance and integrates search for Cas proteins. *Nucleic Acids Res.* 2018;46(W1):W246-W251. doi:10.1093/nar/gky425
 84. Makarova KS, Wolf YI, Iranzo J, et al. Evolutionary classification of CRISPR–Cas systems: a burst of class 2 and derived variants. *Nat Rev Microbiol.* 2020;18(2):67-83. doi:10.1038/s41579-019-0299-x
 85. Chylinski K, Makarova KS, Charpentier E, Koonin E V. Classification and evolution of type II CRISPR-Cas systems. *Nucleic Acids Res.* 2014;42(10):6091-6105. doi:10.1093/nar/gku241

86. Richter AA, Mais C-N, Czech L, et al. Biosynthesis of the Stress-Protectant and Chemical Chaperon Ectoine: Biochemistry of the Transaminase EctB. *Front Microbiol.* 2019;10. doi:10.3389/fmicb.2019.02811
87. Tohyama S, Kakinuma K, Eguchi T. The Complete Biosynthetic Gene Cluster of the 28-Membered Polyketide Macrolactones, Halstoctacosanolides, from *Streptomyces halstedii* HC34. *J Antibiot (Tokyo).* 2006;59(1):44-52. doi:10.1038/ja.2006.7
88. Tohyama S, Eguchi T, Dhakal RP, Akashi T, Otsuka M, Kakinuma K. Genome-inspired search for new antibiotics. Isolation and structure determination of new 28-membered polyketide macrolactones, halstoctacosanolides A and B, from *Streptomyces halstedii* HC34. *Tetrahedron.* 2004;60(18):3999-4005. doi:10.1016/j.tet.2004.03.027
89. Gilchrist CLM, Chooi Y-H. clinker & clustermap.js: automatic generation of gene cluster comparison figures. Robinson P, ed. *Bioinformatics.* January 2021. doi:10.1093/bioinformatics/btab007
90. Chen Y, Unger M, Ntai I, et al. Gobichelin A and B: mixed-ligandsiderophores discovered using proteomics. *Medchemcomm.* 2013;4(1):233-238. doi:10.1039/C2MD20232H
91. Landrum G, others. RDKit: Open-source cheminformatics. 2006.
92. Hover BM, Kim S-H, Katz M, et al. Culture-independent discovery of the malacidins as calcium-dependent antibiotics with activity against multidrug-resistant Gram-positive pathogens. *Nat Microbiol.* 2018;3(4):415-422. doi:10.1038/s41564-018-0110-1
93. Wu C, Shang Z, Lemetre C, Ternei MA, Brady SF. Cadasides, Calcium-Dependent Acidic Lipopeptides from the Soil Metagenome That Are Active against Multidrug-Resistant Bacteria. *J Am Chem Soc.* 2019;141(9):3910-3919. doi:10.1021/jacs.8b12087
94. Culp EJ, Waglechner N, Wang W, et al. Evolution-guided discovery of antibiotics that inhibit peptidoglycan remodelling. *Nature.* 2020;578(7796):582-587. doi:10.1038/s41586-020-1990-9
95. Nicolaou KC, Boddy CNC, Bräse S, Winssinger N. Chemistry, Biology, and Medicine of the Glycopeptide Antibiotics. *Angew Chemie Int Ed.* 1999;38(15):2096-2152. doi:10.1002/(SICI)1521-3773(19990802)38:15<2096::AID-ANIE2096>3.0.CO;2-F
96. Haslinger K, Peschke M, Brieke C, Maximowitsch E, Cryle MJ. X-domain of peptide synthetases recruits oxygenases crucial for glycopeptide biosynthesis. *Nature.* 2015;521(7550):105-109. doi:10.1038/nature14141
97. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina

- sequence data. *Bioinformatics*. 2014;30(15):2114-2120. doi:10.1093/bioinformatics/btu170
98. Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics*. 2014;30(14):2068-2069. doi:10.1093/bioinformatics/btu153
 99. Schoch CL, Ciufo S, Domrachev M, et al. NCBI Taxonomy: a comprehensive update on curation, resources and tools. *Database*. 2020;2020. doi:10.1093/database/baaa062
 100. Nothias L-F, Nothias-Esposito M, da Silva R, et al. Bioactivity-Based Molecular Networking for the Discovery of Drug Leads in Natural Product Bioassay-Guided Fractionation. *J Nat Prod*. 2018;81(4):758-767. doi:10.1021/acs.jnatprod.7b00737
 101. Ren H, Wang B, Zhao H. Breaking the silence: new strategies for discovering novel natural products. *Curr Opin Biotechnol*. 2017;48:21-27. doi:10.1016/j.copbio.2017.02.008
 102. Irschik H, Kopp M, Weissman KJ, Buntin K, Piel J, Müller R. Analysis of the Sorangicin Gene Cluster Reinforces the Utility of a Combined Phylogenetic/Retrobiosynthetic Analysis for Deciphering Natural Product Assembly by trans-AT PKS. *ChemBioChem*. 2010;11(13):1840-1849. doi:10.1002/CBIC.201000313
 103. Helfrich EJN, Piel J. Biosynthesis of polyketides by trans-AT polyketide synthases. *Nat Prod Rep*. 2016;33(2):231-316. doi:10.1039/C5NP00125K
 104. Sun Y, Hong H, Gillies F, Spencer JB, Leadlay PF. Glyceryl-S-Acyl Carrier Protein as an Intermediate in the Biosynthesis of Tetrionate Antibiotics. *ChemBioChem*. 2008;9(1):150-156. doi:10.1002/cbic.200700492
 105. Kotowska M, Pawlik K. Roles of type II thioesterases and their application for secondary metabolite yield improvement. *Appl Microbiol Biotechnol*. 2014;98(18):7735-7746. doi:10.1007/s00253-014-5952-8
 106. Schneider T, Gries K, Josten M, et al. The lipopeptide antibiotic Friulimicin B inhibits cell wall biosynthesis through complex formation with bactoprenol phosphate. *Antimicrob Agents Chemother*. 2009;53(4):1610-1618. doi:10.1128/AAC.01040-08
 107. Wang Y, Chen Y, Shen Q, Yin X. Molecular cloning and identification of the laspartomycin biosynthetic gene cluster from *Streptomyces viridochromogenes*. *Gene*. 2011;483(1-2):11-21. doi:10.1016/j.gene.2011.05.005
 108. Ross AC, Xu Y, Lu L, et al. Biosynthetic Multitasking Facilitates Thalassospiramide Structural Diversity in Marine Bacteria. *J Am Chem Soc*. 2013;135(3):1155-1162. doi:10.1021/ja3119674

109. Gu L, Geders TW, Wang B, et al. GNAT-Like Strategy for Polyketide Chain Initiation. *Science* (80-). 2007;318(5852):970-974. doi:10.1126/science.1148790
110. Takano E. γ -Butyrolactones: Streptomyces signalling molecules regulating antibiotic production and differentiation. *Curr Opin Microbiol.* 2006;9(3):287-294. doi:10.1016/j.mib.2006.04.003
111. Ahmed Y, Rebets Y, Tokovenko B, Brötz E, Luzhetskyy A. Identification of butenolide regulatory system controlling secondary metabolism in Streptomyces albus J1074. *Sci Rep.* 2017;7(1):9784. doi:10.1038/s41598-017-10316-y
112. Robbel L, Marahiel MA. Daptomycin, a Bacterial Lipopeptide Synthesized by a Nonribosomal Machinery. *J Biol Chem.* 2010;285(36):27501-27508. doi:10.1074/jbc.R110.128181
113. Miao V, Brost R, Chapple J, She K, Gal M-FC-L, Baltz RH. The lipopeptide antibiotic A54145 biosynthetic gene cluster from Streptomyces fradiae. *J Ind Microbiol Biotechnol.* 2006;33(2):129-140. doi:10.1007/s10295-005-0028-5
114. Baltz RH. Genome mining for drug discovery: cyclic lipopeptides related to daptomycin. *J Ind Microbiol Biotechnol.* 2021;48(3-4). doi:10.1093/jimb/kuab020
115. Sulavik MC, Gambino LF, Miller PF. The MarR repressor of the multiple antibiotic resistance (mar) operon in Escherichia coli: prototypic member of a family of bacterial regulatory proteins involved in sensing phenolic compounds. *Mol Med.* 1995;1(4):436-446. <http://www.ncbi.nlm.nih.gov/pubmed/8521301>.
116. Pompeani AJ, Irgon JJ, Berger MF, Bulyk ML, Wingreen NS, Bassler BL. The Vibrio harveyi master quorum-sensing regulator, LuxR, a TetR-type protein is both an activator and a repressor: DNA recognition and binding specificity at target promoters. *Mol Microbiol.* 2008;70(1):76-88. doi:10.1111/j.1365-2958.2008.06389.x
117. Lee S-K, Kim HR, Jin Y-Y, Yang SH, Suh J-W. Improvement of daptomycin production via increased resistance to decanoic acid in Streptomyces roseosporus. *J Biosci Bioeng.* 2016;122(4):427-433. doi:10.1016/j.jbiosc.2016.03.026
118. Goloborodko AA, Levitsky LI, Ivanov M V., Gorshkov M V. Pyteomics—a Python Framework for Exploratory Data Analysis and Rapid Software Prototyping in Proteomics. *J Am Soc Mass Spectrom.* 2013;24(2):301-304. doi:10.1007/s13361-012-0516-6
119. Nathans D. PUROMYCIN INHIBITION OF PROTEIN SYNTHESIS: INCORPORATION OF PUROMYCIN INTO PEPTIDE CHAINS. *Proc Natl Acad Sci.* 1964;51(4):585-592. doi:10.1073/pnas.51.4.585

120. Enam SU, Zinshteyn B, Goldman DH, et al. Puromycin reactivity does not accurately localize translation at the subcellular level. *Elife*. 2020;9. doi:10.7554/eLife.60303
121. Aviner R. The science of puromycin: From studies of ribosome function to applications in biotechnology. *Comput Struct Biotechnol J*. 2020;18:1074-1083. doi:10.1016/j.csbj.2020.04.014
122. TERCERO JA, LACALLE RA, JIMENEZ A. The pur8 gene from the pur cluster of *Streptomyces alboniger* encodes a highly hydrophobic polypeptide which confers resistance to puromycin. *Eur J Biochem*. 1993;218(3):963-971. doi:10.1111/j.1432-1033.1993.tb18454.x
123. Chen W, Qi J, Wu P, et al. Natural and engineered biosynthesis of nucleoside antibiotics in Actinomycetes. *J Ind Microbiol Biotechnol*. 2016;43(2-3):401-417. doi:10.1007/s10295-015-1636-3
124. Saugar I, Sanz E, Rubio MÁ, Espinosa JC, Jiménez A. Identification of a set of genes involved in the biosynthesis of the aminonucleoside moiety of antibiotic A201A from *Streptomyces capreolus*. *Eur J Biochem*. 2002;269(22):5527-5535. doi:10.1046/j.1432-1033.2002.03258.x
125. Tercero JA, Espinosa JC, Lacalle RA, Jiménez A. The Biosynthetic Pathway of the Aminonucleoside Antibiotic Puromycin, as Deduced from the Molecular Analysis of the pur Cluster of *Streptomyces alboniger*. *J Biol Chem*. 1996;271(3):1579-1590. doi:10.1074/jbc.271.3.1579
126. Abbas M, Elshahawi SI, Wang X, et al. Puromycins B–E, Naturally Occurring Amino-Nucleosides Produced by the Himalayan Isolate *Streptomyces* sp. PU-14G. *J Nat Prod*. 2018;81(11):2560-2566. doi:10.1021/acs.jnatprod.8b00720
127. Bobek J, Šmídová K, Čihák M. A Waking Review: Old and Novel Insights into the Spore Germination in *Streptomyces*. *Front Microbiol*. 2017;8. doi:10.3389/fmicb.2017.02205
128. Hackl S, Bechthold A. The Gene bldA , a Regulator of Morphological Differentiation and Antibiotic Production in *Streptomyces*. *Arch Pharm (Weinheim)*. 2015;348(7):455-462. doi:10.1002/ardp.201500073
129. Espinosa JC, Tercero JA, Rubio MA, Jiménez A. The pur7 Gene from the Puromycin Biosynthetic pur Cluster of *Streptomyces alboniger* Encodes a Nudix Hydrolase. *J Bacteriol*. 1999;181(16):4914-4918. doi:10.1128/JB.181.16.4914-4918.1999
130. Ángel Rubio M, Barrado P, Carlos Espinosa J, Jiménez A, Lobato MF. The pur6 gene of the puromycin biosynthetic gene cluster from *Streptomyces alboniger* encodes a tyrosinyl-aminonucleoside synthetase. *FEBS Lett*. 2004;577(3):371-375. doi:10.1016/j.febslet.2004.09.087

131. Kharel MK, Pahari P, Shepherd MD, et al. Angucyclines: Biosynthesis, mode-of-action, new natural products, and synthesis. *Nat Prod Rep.* 2012;29(2):264-325. doi:10.1039/C1NP00068C
132. Martin GDA, Tan LT, Jensen PR, et al. Marmycins A and B, Cytotoxic Pentacyclic C-Glycosides from a Marine Sediment-Derived Actinomycete Related to the Genus *Streptomyces*. *J Nat Prod.* 2007;70(9):1406-1409. doi:10.1021/np060621r
133. Woo CM, Beizer NE, Janso JE, Herzon SB. Isolation of Lomaiviticins C–E, Transformation of Lomaiviticin C to Lomaiviticin A, Complete Structure Elucidation of Lomaiviticin A, and Structure–Activity Analyses. *J Am Chem Soc.* 2012;134(37):15285-15288. doi:10.1021/ja3074984
134. Colis LC, Woo CM, Hegan DC, Li Z, Glazer PM, Herzon SB. The cytotoxicity of (–)-lomaiviticin A arises from induction of double-strand breaks in DNA. *Nat Chem.* 2014;6(6):504-510. doi:10.1038/nchem.1944
135. Zhang Z, Pan H-X, Tang G-L. New insights into bacterial type II polyketide biosynthesis. *F1000Research.* 2017;6:172. doi:10.12688/f1000research.10466.1
136. Fan K, Zhang Q. The functional differentiation of the post-PKS tailoring oxygenases contributed to the chemical diversities of atypical angucyclines. *Synth Syst Biotechnol.* 2018;3(4):275-282. doi:10.1016/j.synbio.2018.11.001
137. Rix U, Zheng J, Remsing Rix LL, Greenwell L, Yang K, Rohr J. The Dynamic Structure of Jadomycin B and the Amino Acid Incorporation Step of Its Biosynthesis. *J Am Chem Soc.* 2004;126(14):4496-4497. doi:10.1021/ja031724o
138. Tibrewal N, Downey TE, Van Lanen SG, Ul Sharif E, O'Doherty GA, Rohr J. Roles of the Synergistic Reductive O -Methyltransferase GilM and of O -Methyltransferase GilMT in the Gilvocarcin Biosynthetic Pathway. *J Am Chem Soc.* 2012;134(30):12402-12405. doi:10.1021/ja305113d
139. Janso JE, Haltli BA, Eustáquio AS, et al. Discovery of the lomaiviticin biosynthetic gene cluster in *Salinispora pacifica*. *Tetrahedron.* 2014;70(27-28):4156-4164. doi:10.1016/j.tet.2014.03.009
140. Shen Y, Yoon P, Yu T-W, Floss HG, Hopwood D, Moore BS. Ectopic expression of the minimal *whiE* polyketide synthase generates a library of aromatic polyketides of diverse sizes and shapes. *Proc Natl Acad Sci.* 1999;96(7):3622-3627. doi:10.1073/pnas.96.7.3622
141. Huang C, Yang C, Zhu Y, Zhang W, Yuan C, Zhang C. Marine Bacterial Aromatic Polyketides From Host-Dependent Heterologous Expression and Fungal Mode of Cyclization. *Front Chem.* 2018;6. doi:10.3389/fchem.2018.00528
142. Rix U, Wang C, Chen Y, et al. The Oxidative Ring Cleavage in Jadomycin

- Biosynthesis: A Multistep Oxygenation Cascade in a Biosynthetic Black Box. *ChemBioChem*. 2005;6(5):838-845. doi:10.1002/cbic.200400395
143. Liu T, Fischer C, Beninga C, Rohr J. Oxidative Rearrangement Processes in the Biosynthesis of Gilvocarcin V. *J Am Chem Soc*. 2004;126(39):12262-12263. doi:10.1021/ja0467521
 144. Li J, Xie Z, Wang M, Ai G, Chen Y. Identification and Analysis of the Paulomycin Biosynthetic Gene Cluster and Titer Improvement of the Paulomycins in *Streptomyces paulus* NRRL 8115. Virolle M-J, ed. *PLoS One*. 2015;10(3):e0120542. doi:10.1371/journal.pone.0120542
 145. Waldman AJ, Balskus EP. Lomaiviticin biosynthesis employs a new strategy for starter unit generation. *Org Lett*. 2014;16(2):640-643. doi:10.1021/ol403714g
 146. Jiang X, Zhang Q, Zhu Y, et al. Isolation, structure elucidation and biosynthesis of benzo[b]fluorene nenestatin A from deep-sea derived *Micromonospora echinospora* SCSIO 04089. *Tetrahedron*. 2017;73(26):3585-3590. doi:10.1016/j.tet.2017.03.054
 147. Wang B, Ren J, Li L, et al. Kinamycin biosynthesis employs a conserved pair of oxidases for B-ring contraction. *Chem Commun*. 2015;51(42):8845-8848. doi:10.1039/C5CC01986A
 148. Jin J, Yang X, Liu T, et al. Fluostatins M–Q Featuring a 6-5-6-6 Ring Skeleton and High Oxidized A-Rings from Marine *Streptomyces* sp. PKU-MA00045. *Mar Drugs*. 2018;16(3):87. doi:10.3390/md16030087
 149. Kharel MK, Nybo SE, Shepherd MD, Rohr J. Cloning and Characterization of the Ravidomycin and Chrysomycin Biosynthetic Gene Clusters. *ChemBioChem*. 2010;11(4):523-532. doi:10.1002/cbic.200900673
 150. Abdelfattah MS, Rohr J. Premithramycinone G, an Early Shunt Product of the Mithramycin Biosynthetic Pathway Accumulated upon Inactivation of Oxygenase MtmOII. *Angew Chemie Int Ed*. 2006;45(34):5685-5689. doi:10.1002/anie.200600511
 151. Perić-Concha N, Borovička B, Long PF, Hranueli D, Waterman PG, Hunter IS. Ablation of the otcC Gene Encoding a Post-polyketide Hydroxylase from the Oxytetracycline Biosynthetic Pathway in *Streptomyces rimosus* Results in Novel Polyketides with Altered Chain Length. *J Biol Chem*. 2005;280(45):37455-37460. doi:10.1074/jbc.M503191200
 152. Rebets Y, Nadmid S, Paulus C, et al. Perquinolines A–C: Unprecedented Bacterial Tetrahydroisoquinolines Involving an Intriguing Biosynthesis. *Angew Chemie Int Ed*. 2019;58(37):12930-12934. doi:10.1002/anie.201905538
 153. Allen HK, Moe LA, Rodbumrer J, Gaarder A, Handelsman J. Functional metagenomics reveals diverse β -lactamases in a remote Alaskan soil. *ISME J*.

- 2009;3(2):243-251. doi:10.1038/ismej.2008.86
154. Nguyen KT, Kau D, Gu J-Q, et al. A glutamic acid 3-methyltransferase encoded by an accessory gene locus important for daptomycin biosynthesis in *Streptomyces roseosporus*. *Mol Microbiol.* 2006;61(5):1294-1307. doi:10.1111/j.1365-2958.2006.05305.x
 155. Tatham E, sundaram Chavadi S, Mohandas P, et al. Production of mycobacterial cell wall glycopeptidolipids requires a member of the MbtH-like protein family. *BMC Microbiol.* 2012;12(1):118. doi:10.1186/1471-2180-12-118
 156. Zwahlen RD, Pohl C, Bovenberg RAL, Driessen AJM. Bacterial MbtH-like Proteins Stimulate Nonribosomal Peptide Synthetase-Derived Secondary Metabolism in Filamentous Fungi. *ACS Synth Biol.* 2019;8(8):1776-1787. doi:10.1021/acssynbio.9b00106
 157. Frański R, Kozik T. Unexpected interaction between deprotonated biliverdin and alcohols as studied by ESI-MS. *J Mass Spectrom.* 2017;52(2):65-68. doi:10.1002/jms.3900
 158. Cleary JL, Kolachina S, Wolfe BE, Sanchez LM. Coproporphyrin III Produced by the Bacterium *Glutamicibacter arilaitensis* Binds Zinc and Is Upregulated by Fungi in Cheese Rinds. Tullman-Ercek D, ed. *mSystems.* 2018;3(4). doi:10.1128/mSystems.00036-18
 159. Kanehisa M. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* 2000;28(1):27-30. doi:10.1093/nar/28.1.27
 160. Mendler K, Chen H, Parks DH, Lobb B, Hug LA, Doxey AC. AnnoTree: visualization and exploration of a functionally annotated microbial tree of life. *Nucleic Acids Res.* 2019;47(9):4442-4448. doi:10.1093/nar/gkz246
 161. Poole K. At the Nexus of Antibiotics and Metals: The Impact of Cu and Zn on Antibiotic Activity and Resistance. *Trends Microbiol.* 2017;25(10):820-832. doi:10.1016/j.tim.2017.04.010

Appendix 1 Genomic DNA /strains isolated in this study and their corresponding positions on 96-well plates

PLATE: JGO1C	1	2	3	4	5	6	7	8	9	10	11	12
A	187	P3-A6	P2-B6	P1-D2	P1-H7	P1-D4	P1-E6	P1-D5	P1-E5	P2-H11	186-1	186-2
B	184-6	181-2	P2-G12	221-2	P2-F12	P1-G4	181-1	P1-G12	P1-H6	P1-B8	blank	blank
C	P2-E1	P2-B4	P2-D6	P2-A4	P3-G12	P2-E3	P2-B11	P2-B10	P2-A5	P2-E2	P2-A11	P2-E4
D	P2-D3	P2-G11	P3-G3	P3-G10	P3-A4	P3-B7	P3-E9	P2-G9	P3-B6	P3-B10	P3-D8	P2-D2
E	P2-G7	P3-C4	P2-G8	P2-A10	P3-A7	P2-C8	P3-G5	P2-G10	P3-F4	P3-H7	P3-E12	P3-E11
F	P3-E8	P3-E10	P3-E4	P2-B5	P2-H6	P3-D5	P3-G6	P3-F9	P2-D1	193	221-1	P1-D1
G	191	P3-E3	P1-E7	P3-H3	P1-A4	P1-B7	P3-B12	232	P1-A5	P1-A7	P1-C2	186-1
H	P1-B4	229	P1-B6	D1-H2	P3-H12	238	221-2	184	P1-F4	229	P1-E4	186-2

PLATE: JGO2C	1	2	3	4	5	6	7	8	9	10	11	12
A	P1-A8	P2-F10	P3-H11	P3-D1	P3-D2	P3-E3	186	238	234	P1-F5	P2-E10	P1-C6
B	P2-F9	184-4	P3-B1	187-1	P1-F4	221 1-1	P2-H3-1	184-1	184-2	184-3	236-1	236-2
C	233	189-1	189-3	189-2	189-4	184-5	185-1	185-2	200-1	P2-C4	P2-D10	P2-C2
D	P3-C10	P2-C5	P2-D1	P3-C11	P2-G4	P2-C1	293	226	P2-G5	P2-G3	P2-H5	P2-F7
E	P2-H10	P2-H7	P3-F5	262-1	262-2	P3-C1	200-2	200-3	200-4	188-2	200-5	188
F	P1-C8	P1-C7	223	187	185 1-1	185 2-1	P1-F6	134	230-2	230-3	230-1	P2-D8
G	P3-D11	P3-D10	239-1	239-2	239-3	181	P2-D4	P2-B8	P2-F11	P2-B5	P2-B3	P1-A5
H	P2-F1	P3-B4	P2-F2	P1-C4	P3-B11-2	P1-B5	P3-B11	221 1-2	P2-E7	189-1	blank	blank

PLATE: LI1C	1	2	3	4	5	6	7	8	9	10	11	12
A	247-6	464-1	464-2	467-1	467-2	467-4	341-4	341-1	18-8	123-1	341-2	18-7-1
B	464-4	18-10	134-3	18-7-2(9)	463-3	18-3	367-2	367-1	123-7	18-2	123-8	325-2
C	143-2	134-2	217-3	18-6	348-1	217-4	153-2	438-1	157	445	118-2	113-3
D	118-1	113-1	113-2	139-3	139-4	118	476-3	476-5	11-2-1	16-1-1	111-3	457-3
E	111-5	111-4	16-1-2	27-2	22	27-1	31-1	375-1	369-1-1	211-2	210-1	347-3
F	347-5	347-2	465	336-3	336-2	361-1	369-3	441-1	369-1-2	364-2	152-1	22-1
G	11-2-2	23-3	11-1	23-1	23-2	13-9	343-1	343-4	11-4	128-5	148	17-1
H	326	438-7	297-2	174-2	462-1	179-4	179-2	140-1	216-1	254-3	blank	blank

PLATE: LI2C	1	2	3	4	5	6	7	8	9	10	11	12
A	P1-D12	P1-A12	P3-D12	237	P1-E10	221-2-1	P1-E11	347-7	347-6	439-3	439-6	334-6
B	335-1	352-2	471-4	352-1	471-2	304-1	210-1	330-1	113-2	106	113-1	179-1
C	485-1	195	179-2	114-1	343-2	140-2	343-1	128-2	128-1	342-8	128-4	376-2
D	301-3	376-1	313-2	211-1	476-4-a	476-4-b	347-9	300	134-1	325-3	475-2	347-4
E	113-4	247-1	471-1-a	471-1-b-B	471-1-b-A	136-2	471-3	370-2	11-3-1	11-1-a	438-6	450-1
F	247-2	247-3	215-1	215-2	111-1-a	111-1-b	11-3-2	370-1	474-2	474-3	313-1	218-1
G	332-2	19-1	371-1	371-2	476-1	476-3	35-2	438-4	438-5	136-1	133-1	153-1
H	17-3	17-2	457-1	24-1	blank	221-2	P1-E2	P1-C10	P3-G8	P1-G11	blank	P3-F3

PLATE: LI3C(D)	1	2	3	4	5	6	7	8	9	10	11	12
A	P2-C6	237	P1-G7	P1-G9	P2-A9	P1-A2	P1-A1	P2-F5	189-2	16-2	16-4	255-1
B	255-2	245-2	18-3	P2-E6	19-1	18-6	19-2	135-2	135-1	135-4	215-2	140-3
C	140-2	140-1	140-5	140-4	201-2	353-1	201-3	215-1	201-1	13-3	13-2	13-1
D	13-5	125-3	125-4	357-1	125-5	357-2	118-4	118-1	118-2	464-5	140-7	126-2
E	126-1	126-5	126-4	347-3	347-1	347-4	347-2	454-2	339-1	144-2	165-3	165-1
F	254-4	122-4	121-1	122-2	207-3	207-2	209-1	207-5	207-4	297-3	297-2	296-1
G	290-3	290-2	357-3	107-1	256-1	107-3	125-2	129-3	125-1	129-1	129-2	31-3
H	31-4	31-5	31-1	31-2	27-1	18-1	18-2	156-1	307-4	343-5	blank	blank

PLATE: LI4C	1	2	3	4	5	6	7	8	9	10	11	12
A	343-3	171-1	242-1	300-1	300-2	102-5	102-3	438-3	438-4	438-1	439-1	451-1
B	124-5	124-3	124-8	124-7	248-1	219-1	131-8	131-7	219-2	131-9	310-2	219-5
C	122-1-1	217-1	122-2	217-2	342-2	342-1	342-5	342-4	308-1	131-3	131-2	131-1
D	131-6	131-5	131-4	10-4	10-1	248-2	178-2	212-1	218-5	212-2	218-4	166-3
E	166-1	146-1	258-1	146-2	289-2	146-4	146-3	4-1	458-1	173	124-1	124-2
F	10-2	438-10	339-4	219-3	308-4	473-7	109-4	462-4	466-2	466-1	474-4	218-2
G	135-3	102-2	123-5	247-5	128-3	172-2	218-3	166-4	109-1	353-2	296-2	454-1
H	125-6	464-8	464-7	122-1-2	307-3	129-3	5-1	102-8	343-6	343-2	blank	blank

Appendix 2 An Excel workbook containing Sampling_locations, assembly_stats>dbtk.classify, antiSMASH_statistics and BiG-SLiCE_statistics sheets.

<https://github.com/phou0402/Thesis/blob/main/Appendix2.xlsx>

Appendix 3 antiSMASH-json_parser (to generate Fig 3.3).

<https://github.com/phou0402/Thesis/blob/main/Appendix3.py>

```

1  #!/usr/bin/envpython
2  # -*- coding: utf-8 -*-
3  # usage: python Appendix3.5.py --input antismash.json
4  import json
5  import sys, os, argparse
6  import re
7
8  parser = argparse.ArgumentParser()
9  parser.add_argument('--input')
10
11  args = parser.parse_args()
12
13  json_filename = args.input
14  root_name = os.path.splitext(json_filename)[0]
15  stat_filename1 = root_name + "_Psimilarity.csv"
16  stat_filename2 = root_name + "_class.tsv"
17  stat_filename3 = root_name + "_avglen.csv"
18  stat_filename4 = root_name + "_class-count.csv"
19  stat_filename5 = root_name + "_Pedge.csv"
20
21  input_handle = open(json_filename, "r")
22  output_handle1 = open(stat_filename1, "w")
23  output_handle2 = open(stat_filename2, "w")
24  output_handle3 = open(stat_filename3, "w")
25  output_handle4 = open(stat_filename4, "w")
26  output_handle5 = open(stat_filename5, "w")
27
28  with open(json_filename) as json_file:
29      data = json.load(json_file)
30      index=len(data['records'])
31
32  def raw_dict(data):
33      similarity_dict={}
34      chemclass_dict={}
35      bgclen_dict={}
36      edge_dict={}
37      for j in range(index):
38          if 'antismash.modules.clusterblast' in data['records'][j]['modules'].keys():
39              for m in
40                  data['records'][j]['modules']['antismash.modules.clusterblast']['knownclu
41                      ster']['results']:
42                      node1 =
43                          data['records'][j]['modules']['antismash.modules.clusterblast']['know
44                              ncluster']['record_id']
45                          region1=m['region_number']
46                          dict_keys1=str(node1)+'|'+str(region1)
47
48                          rank=m['ranking']
49
50                          if len(rank) ==0:
51                              similarity_dict[dict_keys1]=0
52                          else:
53                              genes=len(m['ranking'][0][0]['tags'])
54                              temp_l1=[]
55                              for item in m['ranking'][0][1]['pairings']:
56                                  temp_l1.append(item[2]['name'])
57                              temp_l2 = list(set(temp_l1))
58                              similarity=int(100*len(temp_l2)/genes)
59                              similarity_dict[dict_keys1]=similarity
60
61                      node2=data['records'][j]['id']
62                      for k in data['records'][j]['features']:
63                          if k['type'] == 'region':
64                              region2=k['qualifiers']['region_number']
65                              dict_keys2 = str(node2) + '|' + str(region2)
66                              location=re.findall(r"\d+\.\d*", k['location'])
67                              bgclen=abs(int(location[0])-int(location[1]))
68                              bgclen_dict[dict_keys2] = bgclen
69                              chemclass_dict[dict_keys2] = k['qualifiers']['product']
70                              edge_dict[dict_keys2]=k['qualifiers']['contig_edge']
71
72      return (similarity_dict,chemclass_dict,bgclen_dict,edge_dict)

```

```

69 similarity_dict=raw_dict(data)[0]
70 chemclass_dict=raw_dict(data)[1]
71 bgclen_dict=raw_dict(data)[2]
72 edge_dict=raw_dict(data)[3]
73
74 def similarity_stat():
75     s20 = 0
76     # '<20'
77     s20_90 = 0
78     # '20-90'
79     s90 = 0
80     # '90-100'
81     for values in similarity_dict.values():
82         if values <= 20:
83             s20 += 1
84         elif 20 < values < 90:
85             s20_90 += 1
86         else:
87             s90+=1
88     try:
89         p_s20=s20*100/len(similarity_dict)
90         p_s20_90=s20_90*100/len(similarity_dict)
91         p_s90=s90*100/len(similarity_dict)
92         output_handle1.write("%s,%s,%s,%s\n" % (root_name,p_s20,p_s20_90,p_s90))
93     except ZeroDivisionError:
94         output_handle1.write("%s,%s,%s,%s\n" % (root_name, 0, 0, 0))
95     output_handle1.close()
96     return
97
98 def bgclen_stat():
99     lbd_1000=len(bgclen_dict.keys())*1000
100     total_length=0
101     for i in bgclen_dict.values():
102         total_length+=i
103     try:
104         lavg=total_length/lbd_1000
105         output_handle3.write("%s,%s\n" % (root_name,lavg))
106     except ZeroDivisionError:
107         output_handle3.write("%s,%s\n" % (root_name,0))
108
109     # output_handle2.close()
110     output_handle3.close()
111     return
112
113 def chemclass_stat():
114     Other=0
115     NRP=0
116     n=["cdps","nrps","nrps-like","thioamide-nrp"]
117     Polyketide=0
118
119     p=["hgls-ks","pks-like","ppys-ks","t1pks","t2pks","t3pks","transat-pks",'transat-
120     pks-like']
121     RiPP=0
122     r=["bottromycin","cyanobactin","fungal-ripp","glycocin","lap","lantipeptide
123     class i","lantipeptide class ii","lantipeptide class iii","lantipeptide class
124     iv","lantipeptide class
125     v","lassopeptide","linaridin","lipolanthine","microviridin","proteusin","ranthipe
126     ptide","ras-ripp","ripp-like","rre-containing","sactipeptide","thioamitides","thi
127     opeptide","bacteriocin","head_to_tail","lanthidin","lanthipeptide","tfua-related"
128     ,"microcin","lanthipeptide"]
129     Saccharide=0
130     s=["amglyccycl","oligosaccharide","saccharide","cf_saccharide"]
131     Terpene=0
132     t=["terpene"]
133     Hybrid=0
134
135     for values in chemclass_dict.values():
136         if len(values)==1:
137             for i in values:
138                 if i.lower() in n:
139                     NRP+=1
140                 elif i.lower() in p:

```

```

133         Polyketide += 1
134     elif i.lower() in r:
135         RiPP+=1
136     elif i.lower() in s:
137         Saccharide+=1
138     elif i.lower() in t:
139         Terpene+=1
140     else:
141         Other+=1
142     if len(values)>1:
143         Hybrid+=1
144     output_handle4.write("%s,%s,%s,%s,%s,%s,%s,%s\n" % (root_name, Other, NRP,
Polyketide, RiPP, Saccharide, Terpene, Hybrid))
145     for k,v in chemclass_dict.items():
146         output_handle2.write("%s\t%s\t%s\n" % (root_name, k,v))
147     output_handle4.close()
148     output_handle2.close()
149     return
150
151 def edge_stat():
152     et=0
153     ef=0
154     for values in edge_dict.values():
155         if values == ['True']:
156             et+=1
157         else:
158             ef+=1
159     try:
160         p_et=et*100/(len(edge_dict))
161         p_ef=ef*100/(len(edge_dict))
162         output_handle5.write("%s,%s,%s\n" % (root_name,p_et,p_ef))
163     except ZeroDivisionError:
164         output_handle5.write("%s,%s,%s\n" % (root_name,0,0))
165
166     output_handle5.close()
167     return
168
169 similarity_stat()
170 bgc_len_stat()
171 chemclass_stat()
172 edge_stat()
173 print(root_name,'Done!')
174

```

Appendix 4 Knowncluster_networking (to generate Fig 3.8b)

<https://github.com/phou0402/Thesis/blob/main/Appendix4.py>

```

1  #!/usr/bin/env python
2  # -*- coding: utf-8 -*-
3  #!/usr/bin/envpython
4  # -*- coding: utf-8 -*-
5  # usage: python Appendix3.6.py --input antismash.json --T 0.2
6
7  import json
8  import pandas as pd
9  import sys, os, argparse
10 import re
11 from collections import defaultdict
12 import numpy as np
13
14 parser = argparse.ArgumentParser()
15 parser.add_argument('--input')
16 parser.add_argument('--T',type=float,default=0.2)
17 args = parser.parse_args()
18
19 json_filename = args.input
20
21 root_name=os.path.splitext(json_filename)[0]
22
23 stat_filename = root_name + "_KC.csv"
24 stat_filename2 = root_name + "_compKC.csv"
25
26 with open(json_filename) as json_file:
27     data = json.load(json_file)
28     index=len(data['records'])
29
30
31
32
33 chemclass_dict={}
34 edge_dict={}
35 for j in range(index):
36     node2=data['records'][j]['id']
37     for k in data['records'][j]['features']:
38         if k['type'] == 'region':
39             region2=k['qualifiers']['region_number'][0]
40             dict_keys2 = root_name+"."+str(node2)+ str(region2)
41             edge_dict[dict_keys2]=k['qualifiers']['contig_edge'][0]
42             chemclass_dict[dict_keys2] = k['qualifiers']['product']
43 # print(chemclass_dict)
44 similarity_list=[]
45 for j in range(index):
46     if 'antismash.modules.clusterblast' in data['records'][j]['modules'].keys():
47         for m in
48             data['records'][j]['modules']['antismash.modules.clusterblast']['knowncluster
49             ']['results']:
50                 node1 =
51                 data['records'][j]['modules']['antismash.modules.clusterblast']['knownclu
52                 ster']['record_id']
53                 region1 = m['region_number']
54                 dict_keys1 = str(node1) + str(region1)
55                 rank = m['ranking']
56                 if len(rank) == 0:
57                     item = (root_name + "." + dict_keys1,root_name + "." + dict_keys1,
58                             0, 0, 0)
59                     similarity_list.append(item)
60                 else:
61                     mibig = m['ranking'][0][0]['accession']
62                     genes = len(m['ranking'][0][0]['tags'])
63                     mibig_types =
64                     m['ranking'][0][0]['description']+"."+m['ranking'][0][0]['cluster_typ
65                     e"]
66                     temp_l1 = []
67                     for item in m['ranking'][0][1]['pairings']:
68                         temp_l1.append(item[2]['name'])
69                     similarity = len(list(set(temp_l1))) / genes
70                     if similarity < args.T or m['ranking'][0][1]['core_gene_hits'] == 0:
71                         item = (root_name + "." + dict_keys1, root_name + "." +
72                                 dict_keys1, 0, 0, 0)

```



```

65         else:
66             item = (root_name + "." + dict_keys1, mibig + ".0",
                    similarity, str(chemclass_dict[root_name + "." + dict_keys1]),
                    str(mibig_types))
67             similarity_list.append(item)
68
69     raw_sim=pd.DataFrame(similarity_list)
70
71     raw_sim.columns=['query','ref','similarity','qtype','rtype']
72
73     edge=pd.DataFrame.from_dict(edge_dict, orient='index').reset_index()
74     edge.columns=['query','on_contig_edge']
75
76     merged=
77     pd.merge(raw_sim[['query',"ref",'similarity','qtype','rtype']],edge[["query",'on_cont
78     ig_edge']],how='left')
79     merged.to_csv(stat_filename,index=False,header=False)
80     merged[merged['on_contig_edge']=='False'].to_csv(stat_filename2,index=False,header=Fa
lse)
81
82     print(root_name,', Done! ')

```

Appendix 5 BiG_SLiCE- SQLite3_parser (to generate Fig 3.4)

<https://github.com/phou0402/Thesis/blob/main/Appendix5.py>

```

1  #!/usr/bin/envpython
2  # -*- coding: utf-8 -*-
3  import numpy as np
4  import pandas as pd
5  import os, argparse,sys
6  import csv, sqlite3
7
8  db_filename = sys.argv[1]
9
10 con = sqlite3.connect(db_filename)
11 cur = con.cursor()
12 cur.execute("DROP TABLE IF EXISTS chem_class;")
13 cur.execute("DROP TABLE IF EXISTS chem_subclass;")
14 cur.execute("DROP TABLE IF EXISTS chem_subclass_map;")
15 cur.execute("CREATE TABLE IF NOT EXISTS chem_class (id, name);")
16 cur.execute("CREATE TABLE IF NOT EXISTS chem_subclass (id,class_id,name);")
17 cur.execute("CREATE TABLE IF NOT EXISTS chem_subclass_map
    (class_source,type_source,subclass_id);")# use your column names here
18 # cur.execute("CREATE TABLE IF NOT EXISTS hmm (id, accession, name, db_id,
    model_length);")
19
20 with open('chem_class.csv','r') as fin: # `with` statement available in 2.5+
21     dr = csv.DictReader(fin) # comma is default delimiter
22     to_db = [(int(i['id']), i['name']) for i in dr]
23 cur.executemany("INSERT INTO chem_class (id, name) VALUES (?, ?);", to_db)
24
25 with open('chem_subclass.csv','r') as fin: # `with` statement available in 2.5+
26     dr = csv.DictReader(fin) # comma is default delimiter
27     to_db = [(int(i['id']),int(i['class_id']),i['name']) for i in dr]
28 cur.executemany("INSERT INTO chem_subclass (id,class_id,name) VALUES (?, ?, ?);",
    to_db)
29
30 with open('chem_subclass_map.csv','r') as fin: # `with` statement available in 2.5+
31     dr = csv.DictReader(fin) # comma is default delimiter
32     to_db = [(i['class_source'], i['type_source'],int(i['subclass_id'])) for i in dr]
33 cur.executemany("INSERT INTO chem_subclass_map
    (class_source,type_source,subclass_id) VALUES (?, ?, ?);", to_db)
34
35 con.commit()
36 con.close()
37
38
39 root_name = os.path.splitext(db_filename)[0]
40 stat_filename = root_name + "_metadata.csv"
41 stat_filename2 = root_name + "_metadata-complete.csv"
42 output_handle = open(stat_filename, "w",newline='')
43 output_handle2 = open(stat_filename2, "w",newline='')
44
45 with sqlite3.connect(db_filename) as con:
46     cur = con.cursor()
47     print("loading clustergbk metadata..")
48     file_names, bgc_ids,gcf_ids,gcf_values,contig_edges, length_nts,folder =
    list(zip(*cur.execute(
49         "select bgc.orig_filename,
        bgc.id,gcf_membership.gcf_id,gcf_membership.membership_value,bgc.on_c
        ontig_edge, bgc.length_nt,bgc.orig_folder"
50         " from bgc,gcf_membership"
51         " where bgc.id=gcf_membership.bgc_id"
52         " order by gcf_membership.gcf_id"
53     ).fetchall()))
54
55     print("loading class information..")
56     class_titles = sorted(set([
57         "{}:{}".format(class_name, subclass_name) \
58         for class_name, subclass_name in cur.execute(
59             "select chem_class.name, chem_subclass.name from chem_subclass,
        chem_class"
60             " where chem_class.id=chem_subclass.class_id"
61         ).fetchall()])))

```

```

62     class_presences = {}
63     for class_title in class_titles:
64         class_name, subclass_name = class_title.split(":")
65         subclass_presences = pd.Series(
66             np.full(len(bgc_ids), False),
67             index=bgc_ids
68         )
69
70     try:
71         subclass_bgc_ids = list(zip(*cur.execute((
72             "select distinct bgc_class.bgc_id from bgc_class, chem_subclass,
73             chem_class "
74             "where chem_subclass.class_id=chem_class.id "
75             "and bgc_class.chem_subclass_id=chem_subclass.id "
76             "and chem_class.name like ? and chem_subclass.name like ?"
77         ), (class_name, subclass_name)).fetchall()))
78         subclass_presences[subclass_bgc_ids] = True
79     except:
80         pass
81     class_presences[class_title] = subclass_presences
82
83     print("merging datasets...")
84     temp_metadata = pd.DataFrame({
85         "bgc": file_names,
86         "bgc_id": bgc_ids,
87         "gcf": gcf_ids,
88         "gcf_value": gcf_values,
89         "contig_edge": contig_edges,
90         "len_nt": length_nts,
91         "folder": folder
92     }, index=gcf_ids)
93
94     for class_title in sorted(class_titles):
95         temp_metadata["class-" + class_title] = class_presences[class_title].values
96
97     temp_metadata.to_csv("temp_query-metadata.csv", sep=",", index=False)
98
99     class_dict={}
100     dict_from_csv = pd.read_csv('temp_query-metadata.csv', header=0,
101                                index_col=0).T.to_dict()
102
103     for k_mibig in dict_from_csv.keys():
104         for k,v in dict_from_csv[k_mibig].items():
105             if 'class' in k and v==True:
106                 class_dict[k_mibig]=k
107
108     temp_class= pd.DataFrame.from_dict(class_dict, orient="index").reset_index()
109     temp_class.columns = ['bgc', 'class']
110     merged =
111     pd.merge(temp_class[['bgc', 'class']], temp_metadata[['bgc', 'bgc_id', 'gcf', 'gcf_value',
112     'contig_edge', 'len_nt', 'folder']], how='left')
113     merged.to_csv(output_handle, index=False)
114     complete = merged[merged["contig_edge"] == 0]
115     complete.to_csv(output_handle2, index=False)
116     os.remove("temp_query-metadata.csv")
117
118     output_handle.close()
119     output_handle2.close()

```

Appendix 6 BiG_SLiCE-scatter_plot (to generate Fig 3.4a)

<https://github.com/phou0402/Thesis/blob/main/Appendix6.py>

```

1  #!/usr/bin/python
2  # -*- coding: utf-8 -*-
3
4  from itertools import groupby
5  import pandas as pd
6  import os, argparse
7  import statistics
8  import matplotlib.pyplot as plt
9
10 parser = argparse.ArgumentParser()
11 parser.add_argument('--input')
12 parser.add_argument('--T', type=int, default=900)
13 args = parser.parse_args()
14 csv_filename = args.input
15 T=args.T
16
17 path = os.path.dirname(csv_filename)
18 stat_filename = path+'GCF0&1T'+str(T) + "(sub).svg"
19 stat_filename2 = path+'GCF0&1T'+str(T) + "(all).svg"
20
21 query_meta = pd.read_csv(csv_filename)
22 n_bgcs=query_meta.shape[0]
23
24 dict_from_csv = pd.read_csv(csv_filename, header=0, index_col=0).T.to_dict()
25 folder_list=[]
26 for k_bgc in dict_from_csv.keys():
27     folder_tuple=(dict_from_csv[k_bgc]['gcf'],dict_from_csv[k_bgc]['folder'].split('/')[
28         '][-1],dict_from_csv[k_bgc]['gcf_value'])
29     folder_list.append(folder_tuple)
30 folder_genome={}
31 for i in range(len(folder_list)):
32     folder_genome.setdefault(folder_list[i][0], []).append(folder_list[i][1])
33 folder_value={}
34 for i in range(len(folder_list)):
35     folder_value.setdefault(folder_list[i][0], []).append(folder_list[i][-1])
36 #####
37
38 value_stat_tmp={}
39 for k,v in folder_value.items():
40     temp_value={}
41     if len(v)>1:
42         value_tuple=(statistics.stdev(v),len(v))
43         temp_value['std']=statistics.stdev(v)
44         temp_value['size']=len(v)
45         value_stat_tmp[k]=temp_value
46 value_stat={k:v for (k,v) in value_stat_tmp.items() if len(v)!=0}
47 std_list=[]
48 for k,v in value_stat.items():
49     std_list.append(value_stat[k]['std'])
50
51 n_singleton=len(folder_value)-len(value_stat)
52
53 #####
54
55 folder_value_list=[]
56 for k,v in folder_genome.items():
57     folder_value_list+=v
58 n_genomes=len(set(folder_value_list))
59
60 n_gcfs=len(folder_genome)
61
62 span_stat={}
63
64 for k,v in folder_genome.items():
65     span_stat[k]={}
66     span_stat[k]['span']=(len(set(v))*100)/n_genomes
67

```

```

68 #####
69
70 def merge(x, y):
71     """Given two dicts, merge them into a new dict as a shallow copy."""
72     z = x.copy()
73     z.update(y)
74     return z
75
76 gcf_span={}
77 for kv,vv in value_stat.items():
78     for kg,vg in span_stat.items():
79         if kv==kg:
80             gcf_span[kv]=merge(vv,vg)
81 #####
82
83 gcf_span_metadata= pd.DataFrame.from_dict(gcf_span,orient='index').reset_index()
84 gcf_span_metadata.columns=['gcf','std','BGC counts','span']
85
86 #####
87 ax3=gcf_span_metadata.plot.scatter(x='span', y='std',c='BGC counts',cmap='viridis_r',alpha=0.8,edgecolors='black')
88 ax3.set_ylim([0, max(std_list)+200])
89 ax3.set_xlim([0, 105])
90 ax3.set_xticklabels([0,'20%','40%','60%','80%','100%'])
91 ax3.spines['top'].set_visible(False)
92 ax3.spines['right'].set_visible(False)
93 ax3.vlines(ymin=0,ymax=max(std_list)+200, x=5, color='black')
94 ax3.hlines(y=max(std_list)+200, xmin=0,xmax=5, color='black')
95
96 xlabel="GCFs shared by % isolates\nnon_contig_edge=0&1, T="+str(T)
97
98 ax3.set_xlabel(xlabel)
99 ax3.set_ylabel("stdev (GCF membership values)")
100 ax3.figure.savefig(stat_filename2)
101
102 ax4=gcf_span_metadata.plot.scatter(x='span', y='std',c='BGC counts',cmap='viridis_r',alpha=0.8,edgecolors='black',colorbar=None)
103 ax4.set_ylim([0, max(std_list)+200])
104 ax4.set_xlim([0, 5.2])
105 ax4.set_xticklabels([0,'1%','2%','3%','4%','5%'])
106 ax4.set_xlabel("")
107 ax4.set_ylabel("")
108 ax4.figure.savefig(stat_filename)
109

```

Appendix 7 BiG_SLiCE-stacked_plot (to generate Fig 3.5b)

<https://github.com/phou0402/Thesis/blob/main/Appendix7.py>

```
1  #!/usr/bin/python
2  # -*- coding: utf-8 -*-
3
4  from itertools import groupby
5  import pandas as pd
6  import os, argparse
7
8  parser = argparse.ArgumentParser()
9  parser.add_argument('--input')
10 args = parser.parse_args()
11 csv_filename = args.input
12 T=900
13
14 path = os.path.dirname(csv_filename)
15 stat_filename = path+'T'+str(T) + "BGC0-1.svg"
16
17 dict_from_csv = pd.read_csv(csv_filename, header=0, index_col=0).T.to_dict()
18 class_list=[]
19 class_dict={}
20 for k_bgc in dict_from_csv.keys():
21     if "class-RiPP" in dict_from_csv[k_bgc]['class']:
22         class_tuple=("RiPP",dict_from_csv[k_bgc]['gcf_value'])
23         class_list.append(class_tuple)
24     if "class-Other" in dict_from_csv[k_bgc]['class']:
25         class_tuple=("Other",dict_from_csv[k_bgc]['gcf_value'])
26         class_list.append(class_tuple)
27     if "class-Polyketide" in dict_from_csv[k_bgc]['class']:
28         class_tuple=("Polyketide",dict_from_csv[k_bgc]['gcf_value'])
29         class_list.append(class_tuple)
30     if "class-Terpene" in dict_from_csv[k_bgc]['class']:
31         class_tuple=("Terpene",dict_from_csv[k_bgc]['gcf_value'])
32         class_list.append(class_tuple)
33     if "class-Saccharide" in dict_from_csv[k_bgc]['class']:
34         class_tuple=("Saccharide",dict_from_csv[k_bgc]['gcf_value'])
35         class_list.append(class_tuple)
36     if "class-NRP" in dict_from_csv[k_bgc]['class']:
37         class_tuple=("NRP",dict_from_csv[k_bgc]['gcf_value'])
38         class_list.append(class_tuple)
39     if "class-Alkaloid" in dict_from_csv[k_bgc]['class']:
40         class_tuple=("Alkaloid",dict_from_csv[k_bgc]['gcf_value'])
41         class_list.append(class_tuple)
42 for i in range(len(class_list)):
43     class_dict.setdefault(class_list[i][0], []).append(class_list[i][1])
44
45 range_list=[]
46 range_dicttemp={}
47 for kl,v in class_dict.items():
48     for k, g in groupby(sorted(v),key=lambda x: x//300):
49         range_key=k*300
50         range_tuple=(kl,(range_key,len(list(g))))
51         print(range_tuple)
52         range_list.append(range_tuple)
53
54 for i in range(len(range_list)):
55     range_dicttemp.setdefault(range_list[i][0], []).append(tuple((range_list[i][1])))
56
57 range_dict={}
58 for k,v in range_dicttemp.items():
59     v2=dict((x, y) for x, y in tuple(v))
60     range_dict[k]=v2
61
62 range_metadata= pd.DataFrame.from_dict(range_dict).reset_index().fillna(0)
63 range_metadata=range_metadata.rename(columns={'index':'membership values'})
64
65 column_list = list(range_metadata)
66 j=len(column_list)
67 column_list.remove("membership values")
68 range_metadata["sum"] = range_metadata[column_list].sum(axis=1)
69
```

```

70 range_metadata.sort_values('membership values',inplace=True)
71
72 df_total_raw = range_metadata[['membership values','sum']]
73 df_total_raw.sort_values('membership values',inplace=True)
74 df_total=df_total_raw.iloc[:,-1]
75 df = range_metadata.iloc[:, 0:j]
76 ax = df.plot.bar(stacked=True,x='membership
values',color={'Terpene':'#6600cc','Other':'#808080',
'RiPP':'#0000ff','Polyketide':'#ff8000','NRP':'#00994c','Saccharide':'#ff3333'})
77 ax.set_xlabel("BGC-to-centroid distance, on_contig_edge=0&l, T="+str(T))
78
79 for i, v in enumerate(df_total):
80     if v < 500:
81         ax.text(i, v+25, str(v).split(".")[0])
82
83 xticklables=[]
84 for item in range_metadata['membership values']:
85     item_new=str(item)+"-"+str(item+300-1)
86     xticklables.append(item_new)
87
88 ax.set_xticklables(xticklables)
89 ax.set_ylabel("BGC counts")
90 ax.spines['top'].set_visible(False)
91 ax.spines['right'].set_visible(False)
92 ax.vlines(ymin=0,ymax=3500, x=2.5, color='gray',linestyles='dotted')
93 ax.vlines(ymin=0,ymax=3500, x=5.5, color='gray',linestyles='dotted')
94 ax.text(0,3000,'core')
95 ax.text(3,3000,'putative')
96 ax.text(6,3000,'orphan')
97 ax.figure.savefig(stat_filename)

```

22

Appendix 8 Gene fragments and Primers used in this study

BGC004		
Fragment name	Sequence	Description
004_TwistL	ATGACCAACATCAACAAAACGTCGGATTGCGCGTTAGATTGCTTCCCGGAGCTAGTTGACGCGTCATC TGCCCGGCACCTAATGTAGGGGATGCGTCAACCAAATGGTGTGGGGCCGCGGGTGGTCAGCCGTAGTTGG CTCACTGTCGGCATCGAGATCGCAGTGCAGTCCGCCGAGGCCGTGGACGCAAGCTCCAGCAAGGGAATC CCATGAGCACTGAGAACAGGCCGGAATTCAGGCACTGCTGACGCCGAGGAGAGCGTTGTCTGACTGAT TGACCACCAGCCGTTTCAGTTCGCCAACCTGCACAGCCACGAACCGACGATGGTCGTCAACAACGTCGTCG GCCTCGCCAAGGCCGCCAAGGATTGACGTTCCGACCATTCTGACGACCGTTCTGGAGGAGCGCGGGCGG CCTTCTCCTCCAGGGTCTGCAGGATGTGTTCCCGGAGCAGAAGCCGATCAACAGGACCTTCATCAACACCT GGGAGG	DNA fragment of homology arms
004_TwistR	AGTTGACGGTGAGTCTGGCCCGTACGGTGTCCGTACCGCGCAGGTGCAAGGTGCTGCTCGCCCGCTGTGTC CCGGAGCCGGATGGTCCGGCTCGTCCGTTCCCGTCCCGGAGGTCAAGTTGGACGCTGAGCCACGCAGGC CGGGCTCGGGTGAGGAATTCGTCCTTCCGCGCCGAGGTCCCGGAGAACACCAGCAGCTTCATACCCGCA CGGGCTCGGCGAAGTGCAGTCCAGGAATCGCCGATGCCGTACCCGGTCTCCTTCGCGGCCAGTAGCG GTTGTTGAAGCCGTCGACCGCGCGGCGGCTGAGCCGACTGCGCGGTGGAGGCGCGGCATTCCGA CGGCGGCACCGACTCGGGTTTCCCGTCTCCTCTCGCCAGGCCGGAACAGACCGCCAGGTGCGGGAGC GCGAACCATGATCCCAACGCCAGCAGGATCA	
Primer name	Sequence	Description
004-up-1	TTCTAATACGACTCACTATAGTTCTGGGCGAAGACGGGAGTTTATAGACTAGA	for synthesising sgRNA, used together with sgRNA-uni-f/r in Table 2.9
004-up-2	TTCTAATACGACTCACTATAGTACCGGGACTTCGTCGAGTTGTTTATAGACTAGA	
004-do-1	TTCTAATACGACTCACTATAGTCCGCCATCCTCGTCGGGGCGTTTATAGACTAGA	
004-do-2	TTCTAATACGACTCACTATAGACGATCGGCAGCCGACCCGCTTTTATAGACTAGA	
004-uptest-f	CTGCTGCTAGTAGCGCAACG	for testing the cutting efficiency of the synthesised sgRNA
004-uptest-r	CGTCAGCAGTGCCGTGAAGTC	
004-dotest-f	TCGGGTTTCCCGTCTCCTC	
004-dotest-r	GGAGTGATCTGCGAAGCGTG	
004_che_0-F	GACGGGTGACTTCTGTGTCA	for screening and checking the cloned pathway
004_che_0-R	GTCTCCCTCTCCCACGTGTA	
004_che_1-F	CCCTGGAGGTGTTCTTCTCA	
004_che_1-R	GACGTCGGTGAATTGGTAGC	
004_che_2-F	CGGTCAGTTCCTCAAGCTC	for knocking out the ctg1_627-628
004_che_2-R	GGGTGCAACGGATTTCTG	
004_HygE_F	CCCAGGAAAGCGTCCGGTCAATTCAGTTTCCGTTAAAGGGGTTGGCGGtcaggcgccggggcggtgtc	
004_HygE_R	AACTCCTGCTCGCCGTGCGTCAAGGTGGTGGGGGGCTCTCCTGGGTCAatggggcctcgttctagacg	
004_KOche_FO	CGACAAGCTTCCGGTTTCGC	for screening and confirming the
KO_HygChe_int	CAACCCCGTACTGGTCGGCG	
BGC005		
Fragment name	Sequence	Description
005_TwistL	ATCATCTGGCCCGCAGCTTGATCAGGGCGTGCATCTCGAACTTGTCTCCAGGACGCCGATCAGCCAGAT CAGCGCGGCCCGGAGAGCAGCGCGCGGGGCTCGTTGGACTGCTCGAAGACCCCGCTGAGGTTCTGCAG GTTGCTGGCGACCAGCAGTCCGGCGCACAGGCCGAAGAACATCGCGATACCGCCGAGACGCGGTGTCGG TTCCGGTGGACGTGCGGGCGCGGATCGCGGGCATCGCGCCGATCGCGATGGCGAACTTCCGACCCGG ACCGGTACAGAAATAAGTACCGCGGCCGTACGCGAGCGTCAAGAGGTAATCACGCACAGGCTGCCCG ACAGGAATCGCCGACCATCTGAGCCACACACTAACTCGGAGAAGGACGGGTGTTGGAATGCGCGTTGC CACATGAGGTAAGGACATACAGCGCGCGCATGTTGTCACAACTCGTCACACCTATGCCCGCCGCG TTTCTACCGTA	DNA fragment of homology arms
005_TwistR	AGTTGACGCCGAGCTTCTGAAGGCCGCGCTGGACATCGTCAAGTTTCATCACCGGCAGGAAGGTCACTTG GCGCCGAGGTCTCTGAAGAAGCGGCTCGGCGCACACTCGGAATGTCGGAGGCGTCTTGAGATCGAAGACG ATGTACGCGGTGCGCATGCCCTCTGCTCCCGAAGTACGCGGCCCTCCGGCTTGATCCGGTCAACACCG ACTTCATGGTCTGGGGCAGCGTGCAGTGTGATCGCCTTGTTCGCTTCGCGGTGTCATCTGAGCCGTCA GCAGACCCGCATGTCGTCGCCCTCTCAGCAGCGCT	
Primer name	Sequence	Description
005-up-1	TTCTAATACGACTCACTATAGTGATCACGACGAGGGCGACGTTTATAGACTAGA	for synthesising sgRNA, used together with sgRNA-uni-f/r in Table 2.9
005-up-2	TTCTAATACGACTCACTATAGTCTGACGATCTGTGGATCCGTTTATAGACTAGA	
005-up-3	TTCTAATACGACTCACTATAGACCTGCAATCAACATCACGCGTTTATAGACTAGA	
005-do-1	TTCTAATACGACTCACTATAGTCACTCCGCTGCCCTTCCGCTTTTATAGACTAGA	
005-do-2	TTCTAATACGACTCACTATAGTCAGTTGTGCGCCTTCATCGGTTTATAGACTAGA	for testing the cutting efficiency of the synthesised sgRNA
005-do-3	TTCTAATACGACTCACTATAGCCGCCTCTCAGCAGCGCTCGTTTATAGACTAGA	
005-uptest-f	CCCGAGTCGCCCATGAAGAT	
005-uptest-r	TCCGAGTTAGTGTGTGGC	
005-dotest-f	GTAGGGATTGACCGTACGGC	for screening and checking the cloned pathway
005-dotest-r	TCGGAATGTCGGAGGCGTCC	
005_0-FORWARD	CATGAACACGACGATTACG	
005_0-REVERSE	TCTGAGCTTCCGTACGACT	
005_1-FORWARD	ACGGCCATCTGAACATTCTC	for knocking out the ctg4_380-382
005_1-REVERSE	ACCAGGTCGTCCACTGGTAG	
005_2-FORWARD	CACAGGCTCGGCTACAC	
005_2-REVERSE	GGTGACGGTCTTGAGTACGG	
005_PKS_KO_F	cgccgatgctgctgagcgagcgctttaccactggcgccgctgctgatcagcgccggggcggtgtcc	for knocking out the ctg4_395
005_PKS_KO_R	cgctgtgtgagcgctgacgagagcgagcgctgtgttttgatgacatGACGTCCCCCTCTCTGGACCTGC	
005_AbxO_KO_F	cgccccgtcgctgccccgagcaaatcgccccgacacacgactctccctatcagcgccggggcggtgtcc	
005_AbxO_KO_R	gcccagaacatggtggcgagctcgccccgcagcaagagcaccctcacatgggcccctctgttctagacg	

* For deletion screening and verification of 005ΔPKS and 005ΔAbxO, using 005_PKS_KO_F and 005_AbxO_KO_F with KO_HygChe_int

BGC009		
Fragment name	Sequence	Description
009_TwistL	ACAGCACC GCCCAGGATCGCTCGGGTCGCGACATCTACGGCGATGTTGTGGCGGTAGTTCTCGACTGACGTTGACCACTGAGGACCTGGAGATGACCAAGGTGGTGCCGGAAGGATCAGAGCCAGTGTGGGTCTCGACATGACAAGTTGGGTGTCTTCTGCAGTTCAAGGTGACGGGTTCGGTCTGCGTGCACCTACAACGCTCCGCCCTGAGCGGTTCCACGGCGGTGGCTAGAATTTCCGCCATGGTGACACGGGGGAACAGGTACGGCGCAGACCCGAGAATCCGGGCTGCCGCGATCAGCATGGGCGCCGCTTGACGGCGCCATTGTGGCGAGTGCAGCATGTGGGCGTGGGGGGTGAAGTTTACATGCGCGGTGACGATCGGGTGGTTCGTGTTCTGTGGAGATTGGCCGCGCTGACCGCGGTCTCTCGGGCCATGTGGCCCGCCGACGGGCCAAGCATCAGAGG	DNA fragment of homology arms
009_TwistR	TGGAAGGCGATCACAAGGTTCTCCGCCATCAGATCGTCCGCCACGTCCGATACCCAGCCCGCCGGCGCGTACCAGTGGATGTCTGTCGGCGTAGCGGTTCGAACAGCAGCGGCAAGGCATCGGGCTCGTCCAGGACCGTTTCGATCACCGAGGCGTCGAGTCCAACCGCTCGTTCGTGCCCTCGGTACGGTGGACTCCGACGCTCGGTGAACGGTCTCGGTCCGCTTAAAGCTCATGTTTTTCCCTCGAAGACGCGGGCATGTGCACCTCTGTTTCGCCATGGCCCCGATTGAGTTCGCCGTTCTCTGT	
Primer name	Sequence	Description
009-up-1	TTCTAATACGACTCACTATAGCATCTCCGCCAGCAACAGGGGTTT TAGAGCTAGA	for synthesising sgRNA, used together with sgRNA-uni-f/r in Table 2.9
009-up-2	TTCTAATACGACTCACTATAGCAGCGACAGGAGCCGATCTG TTTTAGAGCTAGA	
009-do-1	TTCTAATACGACTCACTATAGCATTGGTCCCTCCGTCACCGTTT TAGAGCTAGA	
009-do-2	TTCTAATACGACTCACTATAGCAGGACGTGCGCTTCGAGGA GTTT TAGAGCTAGA	
009-uptest-f	ACCTTCGGCAGCATGGGATG	for testing the cutting efficiency of the synthesised sgRNA
009-uptest-r	TGCTGCCACTCGCCACAATG	
009-doche-f	CGACGCTCGGTGAACGGTC	
009-doche-r	CCCGCTCAGCTTCTGTGAGG	
009_0-FORWARD	GAGTTCCGCTGTACGTCTC	for screening and checking the cloned pathway
009_0-REVERSE	AGTCCAGCCGGAGCAAATAC	
009_1-FORWARD	GGTGGGACACAGGGAATC	
009_1-REVERSE	GCTGGGACCGAGTACATA	
009_2-FORWARD	ATTACGCCATTGTGCGAGAT	
009_2-REVERSE	GAACGGCTTGGAGTACTTG	
BGC014		
Fragment name	Sequence	Description
014_TwistL	atccacatcgagtcgagaaatggcggttgatctcgccagggatcggtggtagcgctgacgctcggaacgctccgctgagcgagcggtgacgcgctgaccagccgagcgccgacactgtcgagctgtccaaggacgctggttcggtatgagcgagcgaggtgagcgatcgagcagccgagcggtgatctctcgagaaatcggtctctctcgctcggtcgagagctgagcgagcgagagaaatcctgatccggtgagcgagagctcgagaaagcgctgctgtgtgacgagagacatggggtcccgagctcgagcaggtcgagaaatgtcgcgagcgagagctgtcatcgagcgagcgcgacgcccggagcagaataccgctcccacggagcggtgctcctgtcggaagcagccacacagggtccgaggggattcagcgctaaa	DNA fragment of homology arms
014_TwistR	aacctgtgtgtcgagcgtgatgtctccgcacatcgacggtctcgaggtgtgtgcgcccgggtgcgagcggcggtcggaacgcggtgatctcctgacgcgctgacgagcgcgccgagacatcggtggcgagactgatctgggtgcgagactacatcagaagccgttcggtcgccgaggtgcgcccgggtgctgtccgttctgcccgttcgcccgcggggagagcgggcgctcgacccaactcctgtctgcgggagatcgagatgaacacgagactcgaagatcgccggtgctgcccgtgtgcgggtggaactctcccaacgagatccgggtgtctcacatctgctgtaacgggaacgggtgtgacgagatcaactcttggaacgggtgtgggaactcggtgagggcgatcagcggtgtgtaagacatcatctctatctgcgcccgaactcgacgcgtggtggcccggtgatgcgaactcgg	
Primer name	Sequence	Description
014-up-1	TTCTAATACGACTCACTATAGCTGTTCGGTGATGCCGCAT TTTTAGAGCTAGA	for synthesising sgRNA, used together with sgRNA-uni-f/r in Table 2.9
014-up-2	TTCTAATACGACTCACTATAGAAATGGCGTTTGATCTCGCCGTTT TAGAGCTAGA	
014-do-1	TTCTAATACGACTCACTATAGCGGGTAGATACTCCGCGAGCGTTT TAGAGCTAGA	
014-do-2	TTCTAATACGACTCACTATAGTCCCTCCGACCTCATCTCGCTTT TAGAGCTAGA	
014-uptest-f	GGTCGGAGTTTCCGAGCAAG	for testing the cutting efficiency of the synthesised sgRNA
014-uptest-r	CTCTCCGCGGCGACATTTCCG	
014-doche-f	CCGGTTGCTCCACCATCTGC	
014-doche-r	TGCACCAGCAGGCGGTAGTC	
TAR 014_Chk1_F	CTGCCAGCAGATGTTGTAGC	for screening and checking the cloned pathway
TAR 014_Chk1_R	ATCTGACCGGAGCCAACTC	
TAR 014_Chk2_F	CAGAAGGTGTGCGGGTTCT	
TAR 014_Chk2_R	TTCTCCCTCATGAAGCTCGT	
TAR 014_Chk3_F	GAGACGGTCTGCAAAACCAT	
TAR 014_Chk3_R	ATACTCGTCAACGGCACGAT	
BGC027		
Fragment name	Sequence	Description
027_TwistL	GTCTTGAAGTTGCGGACCGGTGGGCGATCTCGCTGCCCTTCATGGAGCAGTTGTAGGTGTGACGGGGGGGAAGTGCCTGAGATCCGGGTGCCGTCGACGCCCTCCCACAGGAATGTGTGGTGGGGGAAGGTGTTGGTCTGCGACGAGGATCTTGTGGTGAGCAGCGCTTGGAGCCGGCCGCTTTGATGATCTCGCGCAGGCCGGCGGCAACCGAAGGTGTCCGGCAGCCAGGCCCTCTGCTTCTCATGCCGAACCTCGTCGAGGAAGAAGCGTTTGGGTGCACGAACCTCCGGGCCATCGCTCCGAGCCGGGCATGTTGGTGTGCGACTCCACCCACATCCCGCCGGGGGACGAACCCGCTCGCGCAGCGCCCTTTCACCTT	DNA fragment of homology arms
027_TwistR	AAAAGTGCGGTTCATCGTTCCGGCAGCTCACTGTGCGACGAGGGGAGGTTCTCCGGGATCCAAACCTTCGCGGAGCCAGTGTCAATATCTTGTCTGACATCTCTCATGCCCTCCCGCAGCGCAGTGGGAATGCAGTACAAAACATTGACAGTGTGCCGTGCGGCTCTCACTCTGTGGAGGCGGACGACCGGGCATCGCGTCTCACCTGTGACAGGCTCCCGGGGTCTGTGACCCGAGCCGAGGAGCGTCCATGCCGAGTCCGTCCC GTCTTTCGACCTGATCAGATGGGCGGTATGCGAGTGCATCTCTACCCCTCCAGTCCGGGTGATCCCTGGAGCAGGTGGAGTCTTCGGCAAGTTCTCGGGGGTTCGGCCGCCAATG	
Primer name	Sequence	Description
027-up-1	TTCTAATACGACTCACTATAGTCAAGGACAAGGCGCTGCGCCTTTT TAGAGCTAGA	for synthesising sgRNA, used together with sgRNA-uni-f/r in Table 2.9
027-up-2	TTCTAATACGACTCACTATAGCTCGGCCCTGTGCGAAGAACGGT TTTTAGAGCTAGA	
027-up-3	TTCTAATACGACTCACTATAGCTTGGCGACCATCTCGCGGGGTTT TAGAGCTAGA	
027-do-1	TTCTAATACGACTCACTATAGCTTCGGCAGTATCTGCACCGTTT TAGAGCTAGA	
027-do-2	TTCTAATACGACTCACTATAGTTCGGTGTGGACGACCGCT TTTTAGAGCTAGA	for testing the cutting efficiency of the synthesised sgRNA
027-do-3	TTCTAATACGACTCACTATAGAAGATCTCGCAGAAGGTGAT TTTTAGAGCTAGA	
027-uptest-f	TCCAGATGCGGTCCAGTTCC	
027-uptest-r	CCTGCCGCAGATCATCAAG	
027-dotest-f	GCAGGTCCGAATGTGAGTAC	for screening and checking the cloned pathway
027-dotest-r	GGTCAAGACGGGTGATGCCG	
027_0-FORWARD	GTAGACCTCGGCCTTCCAGT	
027_0-REVERSE	CAGTGGCAGAACCTCAACG	
027_1-FORWARD	TCTCAGGAGTGTGTCCACGA	for screening and checking the cloned pathway
027_1-REVERSE	ACGTATCATGAGGTCAAGAGC	
027_2-FORWARD	ACCTCGTCGATCAGGGTGAAG	
027_2-REVERSE	CGACGTCTCATGTCTCAAC	

BGC218-3		
Primer name	Sequence	Description
218-3_CAP_L_fw	CCTAGCGTAACATATCGATCTCGAGGCCGGACAGTACGAATCCTTC	for amplifying homology arms
218-3_CAP_L_rev	TTGATACCTCCTCAGCCGTACGGATGTTTAAACGTCGAACACGGGGAAGTG	
218-3_CAP_R_fw	ATCCTGACGGCTGAGGAGTA	
218-3_CAP_R_rev	CTGCAGGAGCTCGCATGCTCTAGAGGACCAGCGGGAAGATCACTA	
218-3_lipo_L_PAM	GAAATTAATACGACTCACTATAGG GGGCTGCGTGCAGTCATGTGTTTAGAGCTAGAAATAGC	for synthesising sgRNA, used together with sgRNA-uni-f/r in Table 2.9
218-3_lipo_L_PAM	GAAATTAATACGACTCACTATAGG GTCATGTGCGGCTCTCAGGAGTTTAGAGCTAGAAATAGC	
218-3_lipo_L_PAM	GAAATTAATACGACTCACTATAGG GTCGTGGCGGCCCGGACCTGTTTTAGAGCTAGAAATAGC	
218-3_lipo_R_PAM	GAAATTAATACGACTCACTATAGG GTCACGCCACACGGTGCAGGGGTTTTAGAGCTAGAAATAGC	
218-3_lipo_R_PAM	GAAATTAATACGACTCACTATAGG GTTCCGCGGCTTTTAGAGCTAGAAATAGC	
218-3_L_PAM_chk	GATGCGACTGGCCTTGCT	for testing the cutting efficiency of the synthesised sgRNA
218-3_L_PAM_chk	AGATCGGTGTGGTCCACGTA	
218-3_R_PAM_chk	ATGATGATGGGCTTCGACTT	
218-3_R_PAM_chk	ACTCCGCGGTGAGAAGTG	
3_9-3-3_F	TCTTCGTCAACACCCCTTGTG	for screening and checking the cloned pathway
3_9-3-3_R	GTTCCGTGAAGGTGAAGCTC	
3_9-3-M_F	CGGCTCGCGCAACAGGAGCA	
3_9-3-M_R	GTGGCTGGAGACCGCGGG	
3_9-3-2_F	GATTCGAGTTCTCCGACGAC	
3_9-3-2_R	TGAGCATGGACTTGTGTCC	
3_9-3-4_F	AATTCGACACCGACGACAAC	
3_9-3-4_R	ACGGTCGACGACACTGTGT	
218_3_HygE_F	cgtggaattcgtcggtgagaagccacggaacacgagagaggtcgccctcagccgcccggggcggtgtcc	for HygE218-3 construction
218_3_HygE_R	gtacggcgacgagcgctcacagaggtcgggggcgagatctgtggggggccatatggggctctgttctagacg	
218_ins_chef	cgccatccgggtctcatcgc	*
dptEF_F	TCGTGCCGTTGGTAGGATCGTCTAGAACAGGAGGCCCAatggccccccacagatccgcc	for pJdptEF218-3 construction
dptEF_R	GGATCCAAGCTTAGATCTATGCAGGTGCAGCTCTAGTTAATctattccgtttccgtcgacg	
* for screening and confirming the HygE218-3 HygE knockin, used together with KO_HygChe_int		
BGC031		
Fragment name	Sequence	Description
031_TwistL	aacccccacgtaccgagggcgaccatcacgggaagagtgctggattccgcttcggtagaggtgtaggtctcttctggccgcgagtgccgt cgtgcgctgtgagagacgtgctctgccacgcttcgcgagcgctcgcgagggagcttggaacgcatgcccctcgttccgggccatcag atcgcggaaagtgaacgcccacggaagatcacgtcacggagacgatcagagttccacgaggtctccacgacctccacgcgcgctgtatc gcacaaccgaacgcggtccctcgtgcacatcacgcacctctgactgattgcgactcgaacccctcatctgagggtgcgtaacagttgcagt tcgaatacactcttggaggtgagtcgcatgaccaaggtccgcggtgcacaactccatgtctctccgacggtctacgcgacgagcgagacgat cacgatcagcacaccatcgggggcgccgagaggtgttc	DNA fragment of homology arms
031_TwistR	agttccgcgagttgcagctgtgacaggaacgcgcacctgtctccaccgccgcaccatgcagtgccgcatccgcccgtgggtcggggtgagc gctgcgcccctcgaacgagctggagccgggctgcggtctgtcgtctatgcattcaagtaagcagagttcgtgatgagatgtatcgagat gcgcagagggagagagaggtgtcccccggcgccgggggaaattgccgtgtgtccaggagattggaacggcccggaagcgggcagg gggtcgaagcggcgaagaattacaaacggcgaacaaatccgcactctcgcagcatcagggctgacggaaggaagcagcagcatgtctcctgaagc gacgcgcgaagaagaagcgccgcgcgaagaacgcgcgcacacacg	
Primer name	Sequence	Description
031-up-1	TTCTAATACGACTCACTATAG ATGAGTTCGTTACGAGGTTGGTTTAGAGCTAGA	for synthesising sgRNA, used together with sgRNA-uni-f/r in Table 2.9
031-up-2	TTCTAATACGACTCACTATAG GAGTTCCTCCGCGAGGATGCGTTTATAGAGCTAGA	
031-do-1	TTCTAATACGACTCACTATAG GTCTGAGCGAAGGCCCTGACGTTTATAGAGCTAGA	for testing the cutting efficiency of the synthesised sgRNA
031-do-2	TTCTAATACGACTCACTATAG TAGCGCTAGCGCTAGCGGGAGTTTATAGAGCTAGA	
031-uptest-f	AGTCTCCTGAACGCTCACGG	for screening and checking the cloned pathway
031-uptest-r	CGCGTTCGGGTTGTGCGATC	
031-doche-f	CAGTACGGGCTGACGGAAGG	
031-doche-r	ACGACGGTCAATGGCGACCAAC	
031_0-FORWARD	AGGAGACCTGAGAAGGACGAG	for constructing pJ2449
031_0-REVERSE	TTGTCCACCACTCCAGTGTC	
031_1-FORWARD	GCGAGATCGAGGAGGAGAC	
031_1-REVERSE	GAAAGGAGGACCGCTCAG	
031_2-FORWARD	GCGATCTGTACATGCTGGTG	
031_2-REVERSE	CGGTGAACGTGCCAACAAATG	
sfp-GbA-fw	GTGGGCACAATCGTCCGGTTGGTAGGATCGTCTAGAACAGGAGGCCCCATATGAAAGGAGG	for constructing pJ2449
svp-GbA-rv	TGAGAACTAGGATCCAACTTAGATCTATGCAGGTGCAGCTCTAGTTAATCTTACGGGACGCGCGG	

Appendix 9 GNPS_network analysis python script for positive mode

<https://github.com/phou0402/Thesis/blob/main/Appendix9.py>

```
1  #!/usr/bin/envpython
2  # -*- coding: utf-8 -*-
3  import sys
4  pathway=sys.argv[1]
5  medium=sys.argv[2]
6  group=sys.argv[3]
7  clustersummary=medium+"-pos.clustersummary"
8
9  f_1=pathway+"_"+medium+"_1-pos.mzXML"
10 f_3=pathway+"_"+medium+"_2-pos.mzXML"
11 f_5=pathway+"_"+medium+"_3-pos.mzXML"
12
13 f_2='pTAR_'+medium+"_1-pos.mzXML"
14 f_4='pTAR_'+medium+"_2-pos.mzXML"
15 f_6='pTAR_'+medium+"_3-pos.mzXML"
16
17 # get cluster composition info
18 import pandas as pd
19 df =
20 pd.read_csv(clustersummary,sep='\t')[['DefaultGroups','componentindex','RTMean_min','
21 precursor charge','precursor mass']]
22
23 # analysis clusters contain mutiple nodes
24 df_dict = df.to_dict('records')
25
26 grouped_tupleinfo=[]
27 for element in df_dict:
28     if element['componentindex'] !=-1:
29         grouped_tupleinfo.append((element['componentindex'],element['DefaultGroups']))
30 # print(grouped_tupleinfo)
31 grouped_dictinfo={}
32 for index,member in grouped_tupleinfo:
33     grouped_dictinfo.setdefault(index, []).append(member)
34 print(grouped_dictinfo)
35 grouped_component_list=[]
36 for key,val in grouped_dictinfo.items():
37     if set(val)=={group}:
38         grouped_component_list.append(key)
39 print(grouped_component_list)
40
41 grouped_tuple=[]
42 for element in df_dict:
43     if element['componentindex'] in grouped_component_list:
44         grouped_tuple.append((element['precursor
45 mass'],(round(element['RTMean_min'],2),element['precursor charge'])))
46 #print(grouped_tuple)
47
48 result_dict2={}
49 for mass,value in grouped_tuple:
50     result_dict2.setdefault(mass, []).append(value)
51 print(result_dict2)
52
53 from pyteomics import mzxml
54
55 import pylab
56 import matplotlib
57 matplotlib.use('Agg')
58
59 tol=0.03
60 img_list = []
61 for k, v in result_dict2.items():
62     mz = k
63     left, right = mz - tol, mz + tol
64
65     rt_1, intens_1 = [], []
66
```

```

67 with mzxml.MzXML(f_1) as reader_1:
68     for scan_1 in reader_1:
69         if scan_1['msLevel'] == 1:
70             i_1 = scan_1['m/z array'].searchsorted(left)
71             j_1 = scan_1['m/z array'].searchsorted(right)
72             rt_1.append(scan_1['retentionTime'])
73             # integrate, since the scans are in profile mode
74             intens_1.append(scan_1['intensity array'][i_1:j_1].sum())
75
76 rt_2, intens_2 = [], []
77 with mzxml.MzXML(f_2) as reader_2:
78     for scan_2 in reader_2:
79         if scan_2['msLevel'] == 1:
80             i_2 = scan_2['m/z array'].searchsorted(left)
81             j_2 = scan_2['m/z array'].searchsorted(right)
82             rt_2.append(scan_2['retentionTime'])
83             # integrate, since the scans are in profile mode
84             intens_2.append(scan_2['intensity array'][i_2:j_2].sum())
85
86 rt_3, intens_3 = [], []
87 with mzxml.MzXML(f_3) as reader_3:
88     for scan_3 in reader_3:
89         if scan_3['msLevel'] == 1:
90             i_3 = scan_3['m/z array'].searchsorted(left)
91             j_3 = scan_3['m/z array'].searchsorted(right)
92             rt_3.append(scan_3['retentionTime'])
93             # integrate, since the scans are in profile mode
94             intens_3.append(scan_3['intensity array'][i_3:j_3].sum())
95
96 rt_4, intens_4 = [], []
97 with mzxml.MzXML(f_4) as reader_4:
98     for scan_4 in reader_4:
99         if scan_4['msLevel'] == 1:
100             i_4 = scan_4['m/z array'].searchsorted(left)
101             j_4 = scan_4['m/z array'].searchsorted(right)
102             rt_4.append(scan_4['retentionTime'])
103             # integrate, since the scans are in profile mode
104             intens_4.append(scan_4['intensity array'][i_4:j_4].sum())
105
106 rt_5, intens_5 = [], []
107 with mzxml.MzXML(f_5) as reader_5:
108     for scan_5 in reader_5:
109         if scan_5['msLevel'] == 1:
110             i_5 = scan_5['m/z array'].searchsorted(left)
111             j_5 = scan_5['m/z array'].searchsorted(right)
112             rt_5.append(scan_5['retentionTime'])
113             # integrate, since the scans are in profile mode
114             intens_5.append(scan_5['intensity array'][i_5:j_5].sum())
115
116 rt_6, intens_6 = [], []
117 with mzxml.MzXML(f_6) as reader_6:
118     for scan_6 in reader_6:
119         if scan_6['msLevel'] == 1:
120             i_6 = scan_6['m/z array'].searchsorted(left)
121             j_6 = scan_6['m/z array'].searchsorted(right)
122             rt_6.append(scan_6['retentionTime'])
123             # integrate, since the scans are in profile mode
124             intens_6.append(scan_6['intensity array'][i_6:j_6].sum())
125
126 # y = max(intens_1 + intens_2 + intens_3 + intens_4 + intens_5 + intens_6) * 1.1
127
128 ax1 = pylab.subplot(321)
129 # ax1.set_ylim(bottom=0, top=y)
130 pylab.plot(rt_1, intens_1, label='m/z, (rt,charge) = {:.3f}, {{\n{}}}'.format(mz,v,
131 f_1))
132 pylab.xlabel('RT, ' + rt_1[0].unit_info)
133 pylab.legend(loc='upper center', bbox_to_anchor=(0, 1.75, 0.5, 0.5),
134 edgecolor='1')
135 # pylab.legend()

```

```

134
135 ax2 = pylab.subplot(322)
136 # ax2.set_ylim(bottom=0, top=y)
137 pylab.plot(rt_2, intens_2, label='m/z = {:.3f}\n{}'.format(mz, f_2))
138 pylab.xlabel('RT, ' + rt_2[0].unit_info)
139 pylab.legend(loc='upper center', bbox_to_anchor=(0, 1.75, 0.5, 0.5),
140 edgecolor='1')
141
142 ax3 = pylab.subplot(323)
143 # ax3.set_ylim(bottom=0, top=y)
144 pylab.plot(rt_3, intens_3, label='m/z = {:.3f}\n{}'.format(mz, f_3))
145 pylab.xlabel('RT, ' + rt_3[0].unit_info)
146 pylab.legend(loc='upper center', bbox_to_anchor=(0, 1.75, 0.5, 0.5),
147 edgecolor='1')
148 # pylab.legend()
149
150 ax4 = pylab.subplot(324)
151 # ax4.set_ylim(bottom=0, top=y)
152 pylab.plot(rt_4, intens_4, label='m/z = {:.3f}\n{}'.format(mz, f_4))
153 pylab.xlabel('RT, ' + rt_4[0].unit_info)
154 pylab.legend(loc='upper center', bbox_to_anchor=(0, 1.75, 0.5, 0.5),
155 edgecolor='1')
156
157 ax5 = pylab.subplot(325)
158 # ax5.set_ylim(bottom=0, top=y)
159 pylab.plot(rt_5, intens_5, label='m/z = {:.3f}\n{}'.format(mz, f_5))
160 pylab.xlabel('RT, ' + rt_5[0].unit_info)
161 pylab.legend(loc='upper center', bbox_to_anchor=(0, 1.75, 0.5, 0.5),
162 edgecolor='1')
163 # pylab.legend()
164
165 ax6 = pylab.subplot(326)
166 # ax6.set_ylim(bottom=0, top=y)
167 pylab.plot(rt_6, intens_6, label='m/z = {:.3f}\n{}'.format(mz, f_6))
168 pylab.xlabel('RT, ' + rt_6[0].unit_info)
169 pylab.legend(loc='upper center', bbox_to_anchor=(0, 1.75, 0.5, 0.5),
170 edgecolor='1')
171 # pylab.legend()
172
173 pylab.subplots_adjust(wspace=0.8, hspace=2)
174 # pylab.xlabel('RTMean: '+str(v))
175 pylab.savefig(pathway + '_' + medium + '_' + str(mz) + '.png', bbox_inches='tight')
176 img_list.append('pathway-' + pathway + '_' + str(mz) + '.png')
177 pylab.close()
178 print(img_list)
179

```

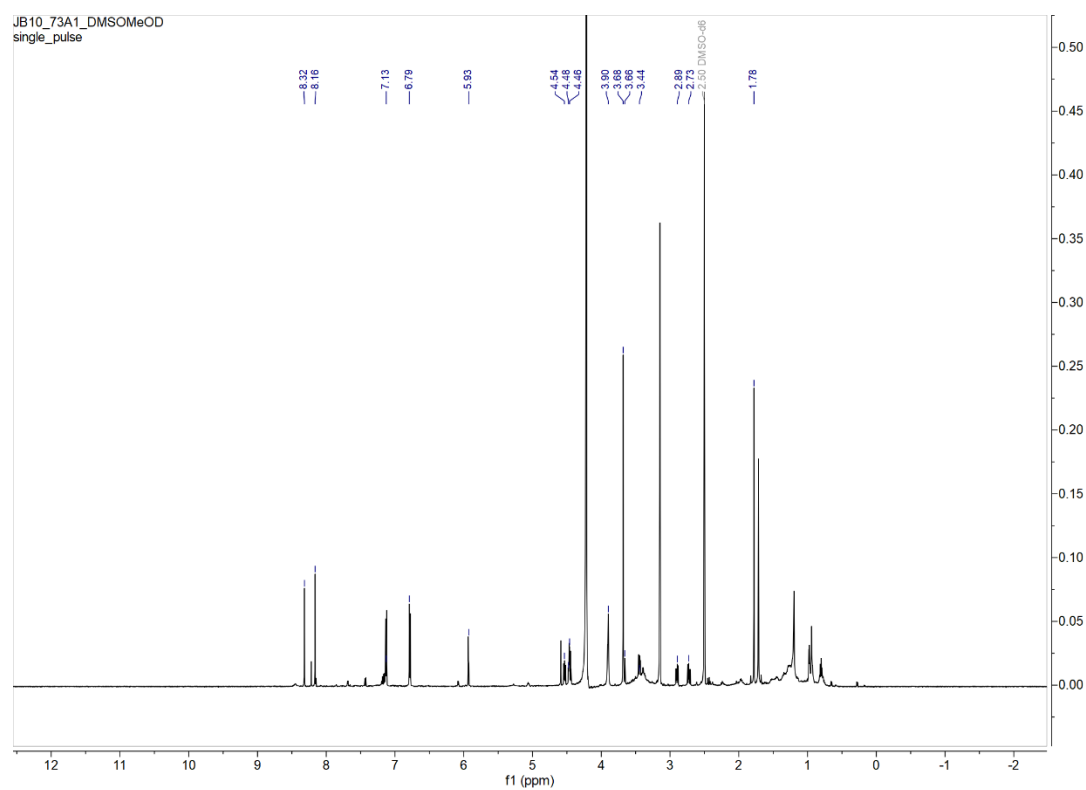
Appendix 10 GNPS_network analysis python script for negative mode

<https://github.com/phou0402/Thesis/blob/main/Appendix10.py>

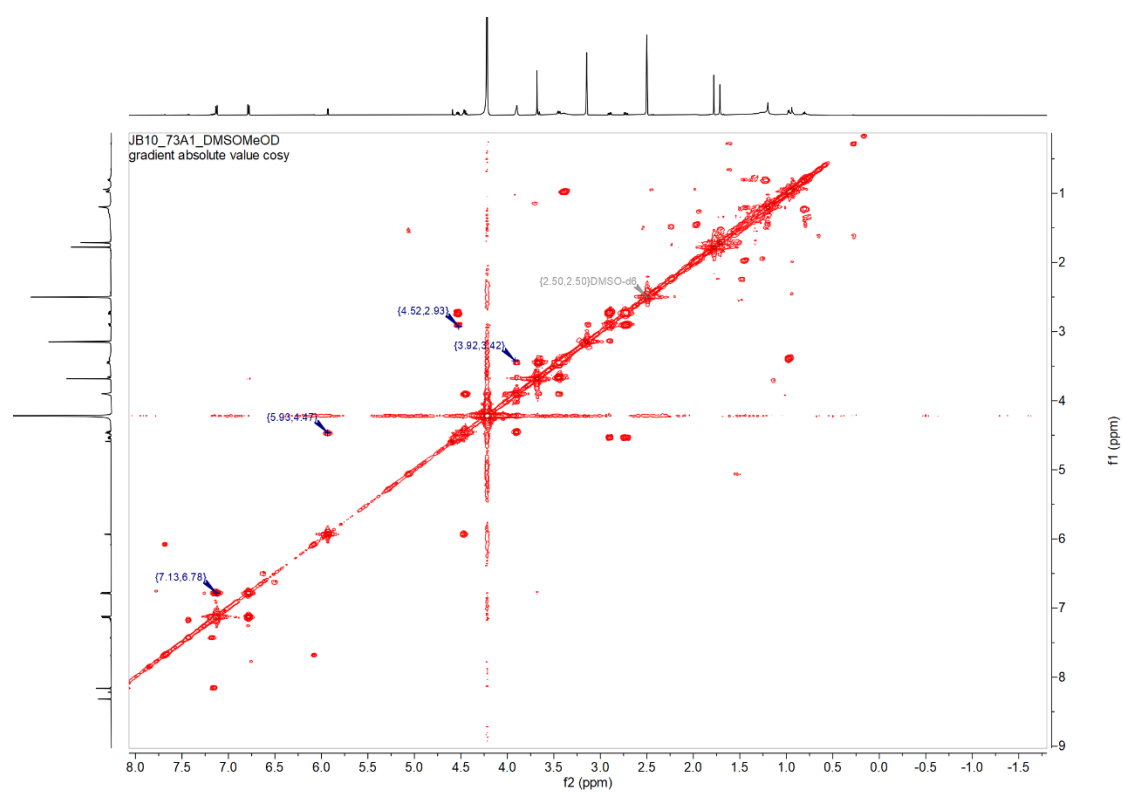
Appendix 11 BGC218-3 and BGC016 validated ions generated by the python analysis workflow

the positive ion mode			
pathway	medium	ions	retention time
del14_HygE218-3	ISP4	519.781	16.79, 16.77
		528.786	16.74
		1056.57	16.69
		1057.57	16.83
		1109.57	16.04
		1097.6	16.14
		1110.57	16.04
	R5a	1056.63	15.54, 15.59
	SMM	669.322	14.46
		1056.64	15.61
the negative ion mode			
pathway	medium	ions	retention time
del14_HygE218-3	R5a	1117.61	17.59
	SMM	1109.55	15.74
		1110.56	15.77
del14_BGC016	SMM	493.003	18.14
	R5a	453.179	15.08,15.11

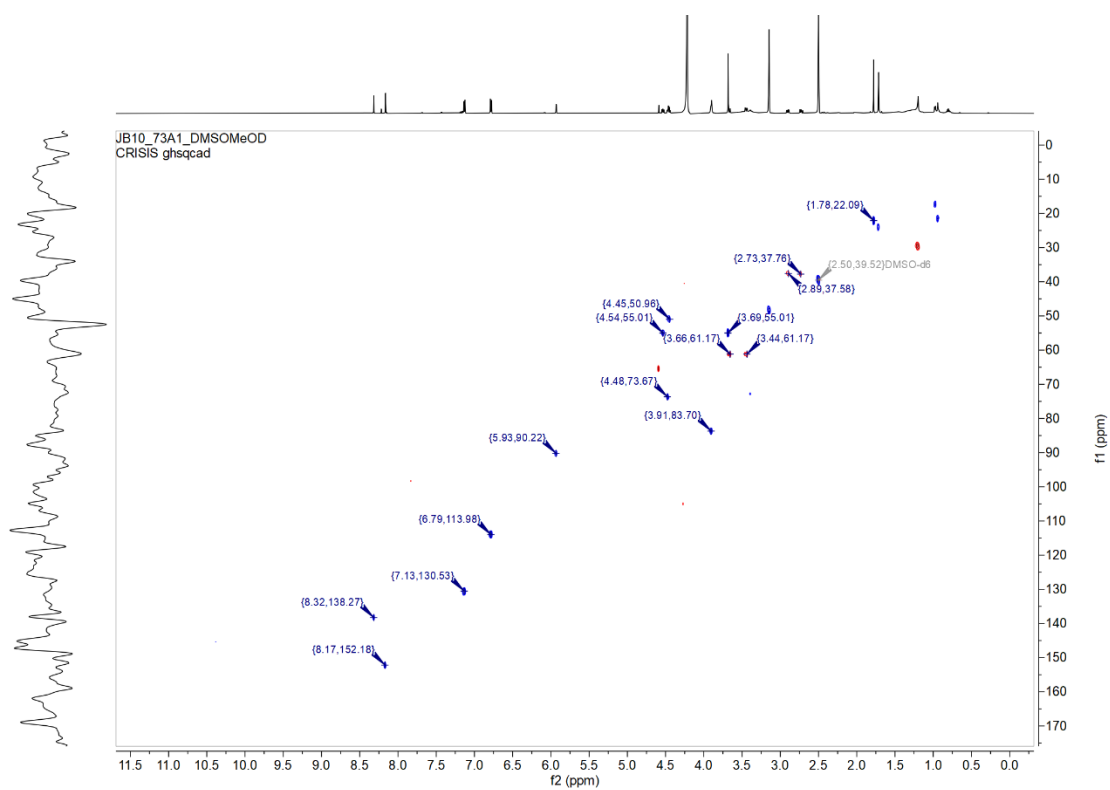
Appendix 12 ^1H NMR of puromycin B (compound **1**)



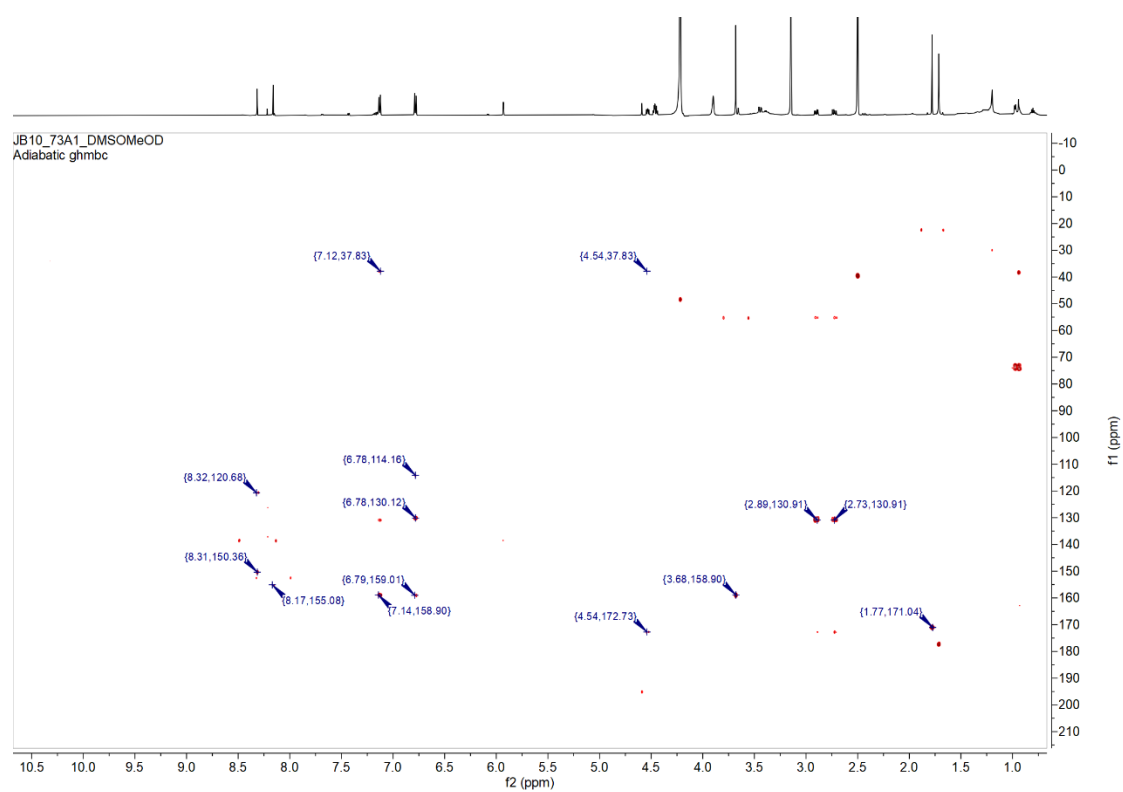
Appendix 13 COSY of puromycin B (compound **1**)



Appendix 14 HMQC of puromycin B (compound 1)

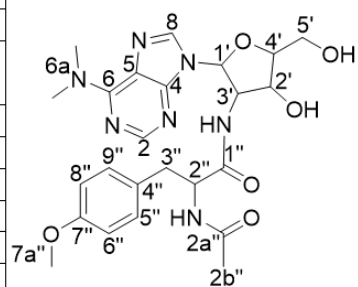


Appendix 15 HMBC of puromycin B (compound 1)



Appendix 16 ^{13}C (150 MHz) and ^1H (600 MHz) NMR Data for puromycin B (compound 1, DMSO- d_6 :MeOD= 1:1)

Position	δ_c	δ_H	multi	$J(1\text{H}-1\text{H})$ in Hz	COSY	HMBC
2	152.18	8.16	s			6,4
4	150.36					
5	120.68					
6	155.3					
6a	ND					
8	138.27	8.32	s			5, 4
1'	90.22	5.93	d	3.2	3'	8, 4
2'	73.67	4.48	m			
3'	50.96	4.46	m		1'	
4'	83.7	3.9	s		5'	
5'	61.17	3.66	d	2.1	4'	
		3.44	dd	12.6,3.5		
1''	172.73					
2''	55.01	4.54	dd	8.5,6.2	3''	1'', 3''
3''	37.67	2.89	dd	13.3,6.3	2''	5'',9'',1'',2''
		2.73	dd	13.9,8.5		
5'',9''	130.53	7.13	d	8.4	6'',8''	7'',3'',5'',9''
6'',8''	113.98	6.79	d	8.4	5'',9''	7'',5'',9'',6'',8''
7''	158.9					
7a''	55.01	3.69	s			7''
2a''	171.04		s			
2b''	22.09	1.78				2a''



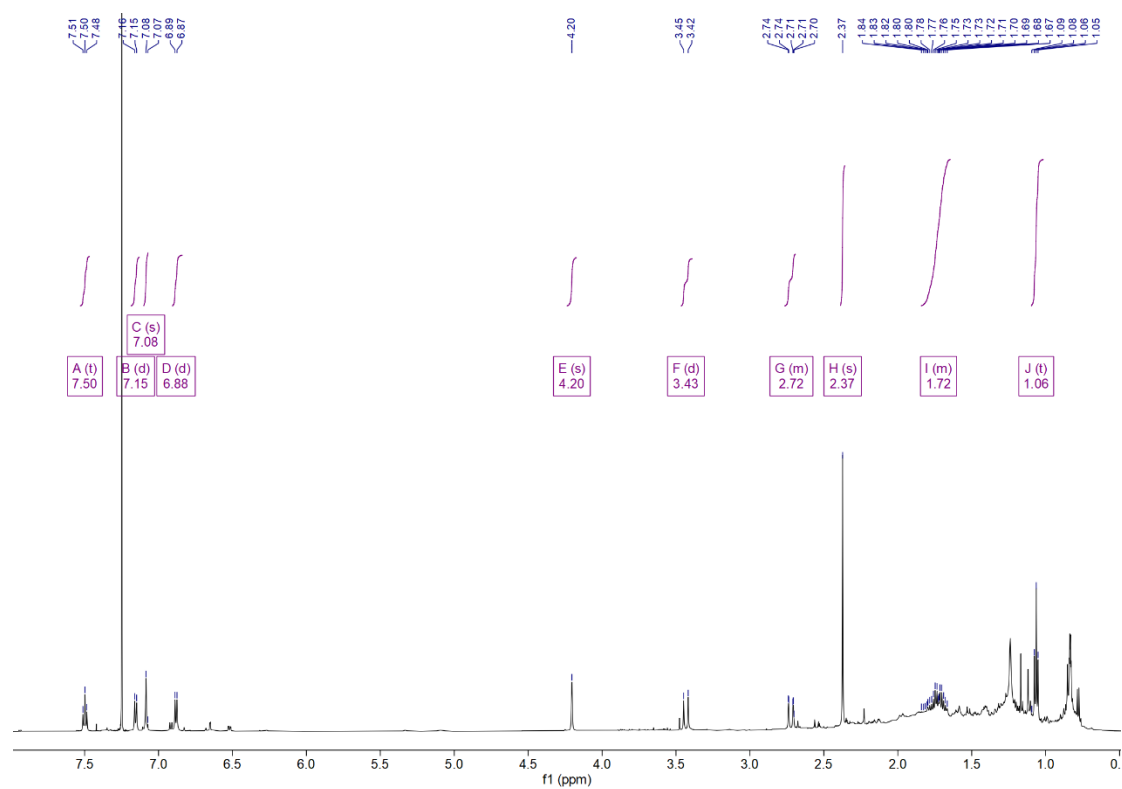
Appendix 17 Characterisation of compound **9**

Compound **9** (JB1081B) was isolated as a yellow film. A deprotonated molecule was observed in negative-ion mode HRESIMS at m/z 313.1090, which indicated the molecular formula $C_{18}H_{18}O_5$ and requiring 10 degrees of unsaturation. A multiplicity-edited HSQC experiment revealed nine carbon centres connected to 15 protons, which indicated the presence of three exchangeable protons. Sample paucity hindered the direct acquisition of a ^{13}C NMR spectrum and accordingly, structure elucidation was achieved using 2D NMR data.

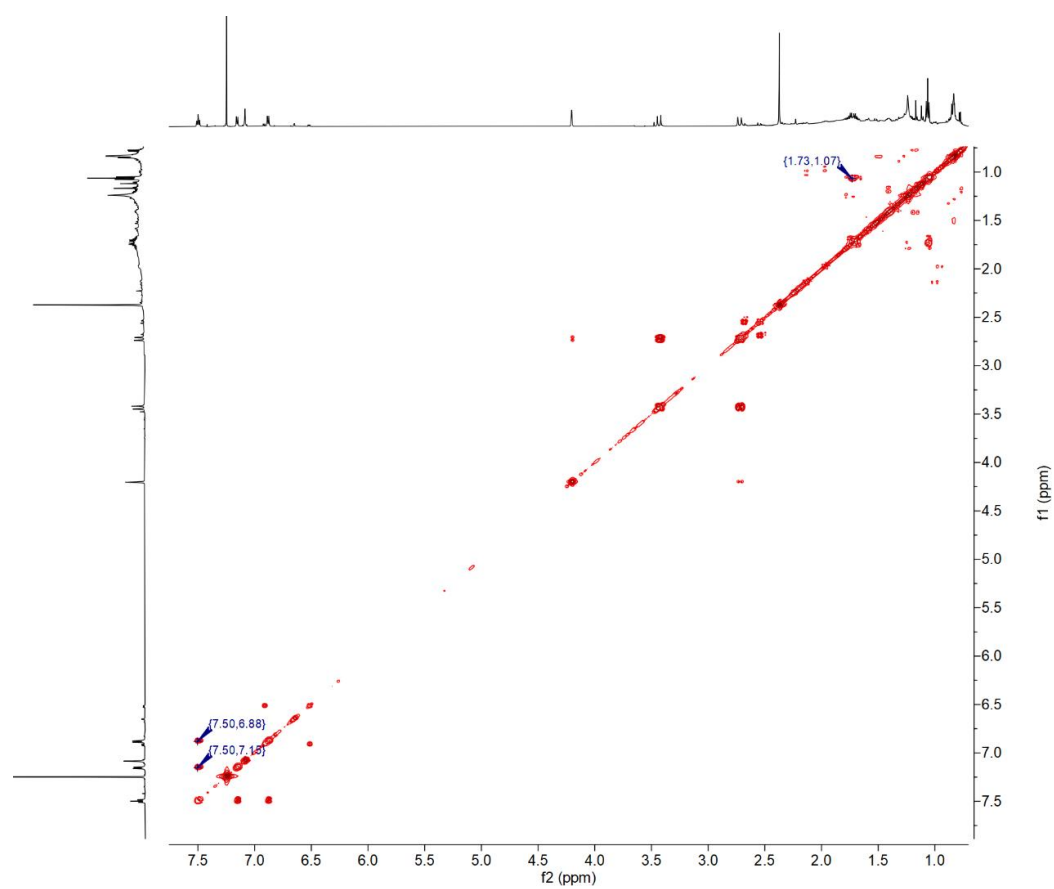
A 1,2,3-trisubstituted aromatic ring was indicated by diagnostic scalar coupling constants and the observation of COSY correlations from doublets H-5 (δ_H 7.17, J = 8.1 Hz) and H-7 (δ_H 6.90, J = 8.1 Hz) to the H-6 triplet (δ_H 7.51, J = 8.0 Hz). Correlations in the HMBC spectrum established the relative positions of the non-protonated carbons with three-bond correlations appearing more strongly than two bond correlations in aromatic rings. Strong correlations were observed from H-6 to non-protonated carbons C-5a (δ_C 139.0) and C-8 (δ_C 158.3), with the latter established to be *ortho*- to CH-7 due to a weak correlation from H-7 to C-8. The 1H NMR resonances for H-5 and H-7 each exhibited reciprocal correlations to each other's carbon (δ_C 118.2 and 111.8, respectively) and a shared correlation to C-8a (δ_C 113.4), with an extra correlation observed from H-5 to singlet aromatic methine C-10 (δ_C 119.1). In addition to C-5 and C-8a, H-10 (δ_H 7.10) displayed three-bond HMBC correlations to non-protonated carbon C-9a (δ_C 108.4) and the singlet methine carbon C-4 (δ_C 61.2). Substitution by a methyl ketone at C-4 was established through HMBC correlations from the methyl CH_3 -12 (δ_H 2.39, δ_C 31.6) proton resonances to C-4 and C-11 (δ_C 205.4). In addition to C-9a, C-10 and C-11, H-4 (δ_H 4.22) correlated to methylene C-2 (δ_C 45.7), non-protonated aromatic C-4a (δ_C 131.9) and oxygenated quaternary carbon C-3 (δ_C 73.8), thereby establishing C-3 and C-4a as the remaining neighbouring carbons either side of C-4. An ethyl branch chain connected to C-3 was indicated through COSY correlations between protons on CH_3 -14 (δ_H 1.08, δ_C 6.6) and CH_2 -13 (δ_H 1.74, δ_C 32.1), in conjunction with HMBC correlations from H-13 and H-14 to C-3. The diastereotopic methylene protons of CH_2 -2 (δ_H 3.45, 2.74, δ_C 45.7) correlated to C-3 and C-4 in addition to ketone carbon C-1 (δ_C 202.6) and a very weak correlation to C-9a. With

one carbon, three oxygens, three exchangeable protons and one degree of unsaturation left to be assigned, these could be accounted for by an additional ring with C-9 bridging between C-8a and C-9a, and with three hydroxyls situated at positions C-3, C-8, and C-9 to give the planar structure of JB10_81B. The relative configuration of C-3 and C-4 was determined from a ROSEY correlation between H-4 and H₃-14, indicating both to be on the same face of the ring (Appendix 22). ¹H NMR, COSY, HSQC, HMBC of compound **9** was provided in Appendices 18-21.

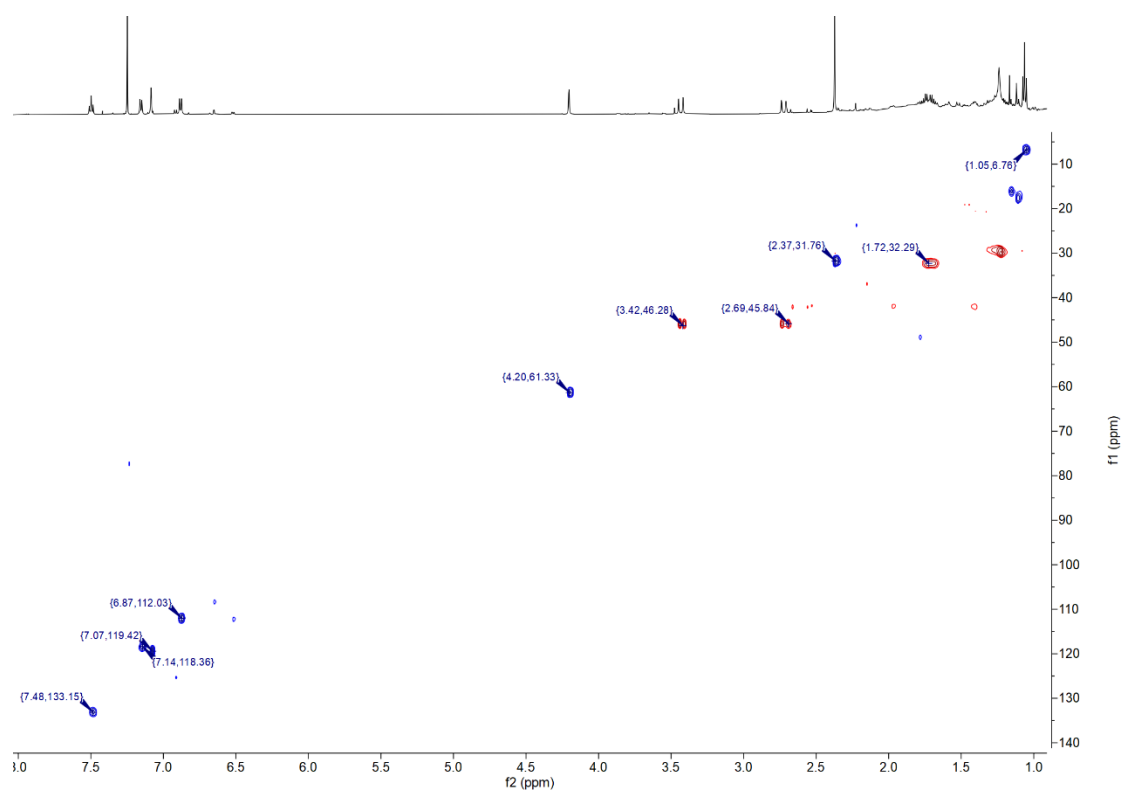
Appendix 18 ^1H NMR of compound JB1081B (compound **9**)



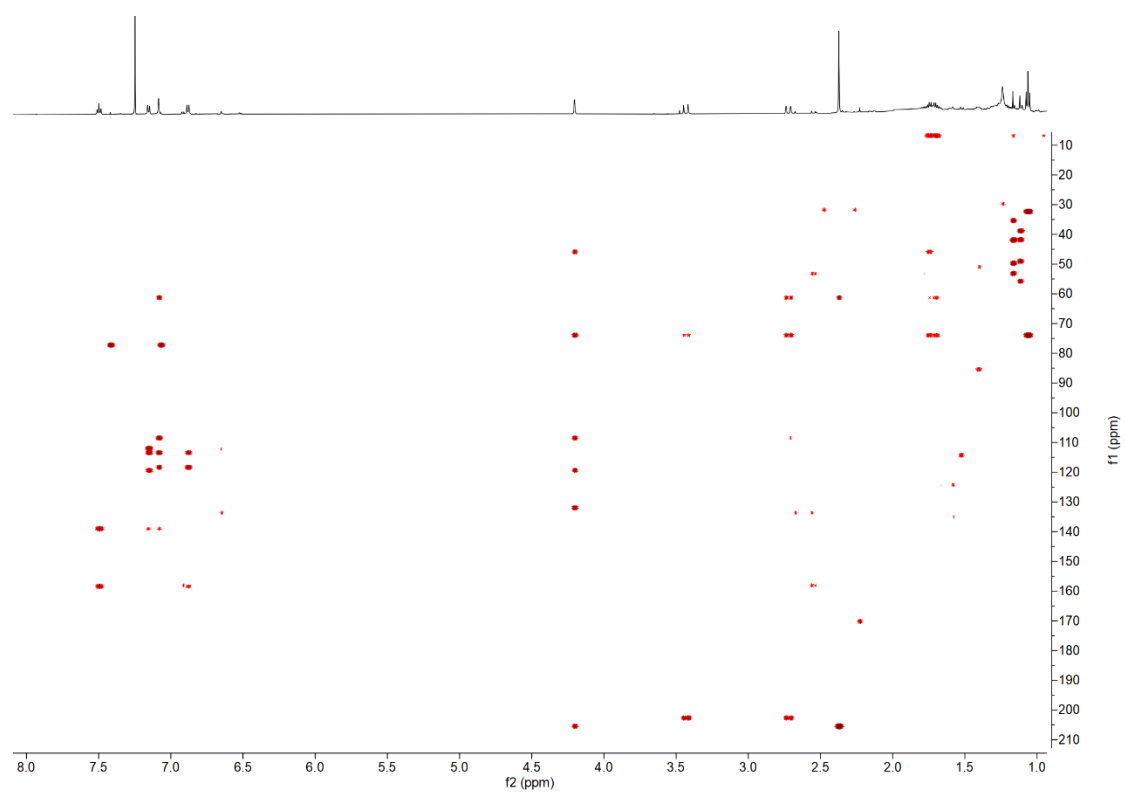
Appendix 19 COSY of JB1081B (compound **9**)



Appendix 20 HSQC of JB1081B (compound 9)



Appendix 21 HMBC of JB1081B (compound **9**)

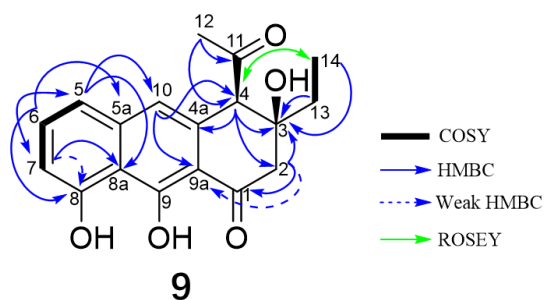


Appendix 22 ^{13}C (150 MHz) and ^1H (600 MHz) NMR Data for JB1081B (compound **9**, CDCl_3)

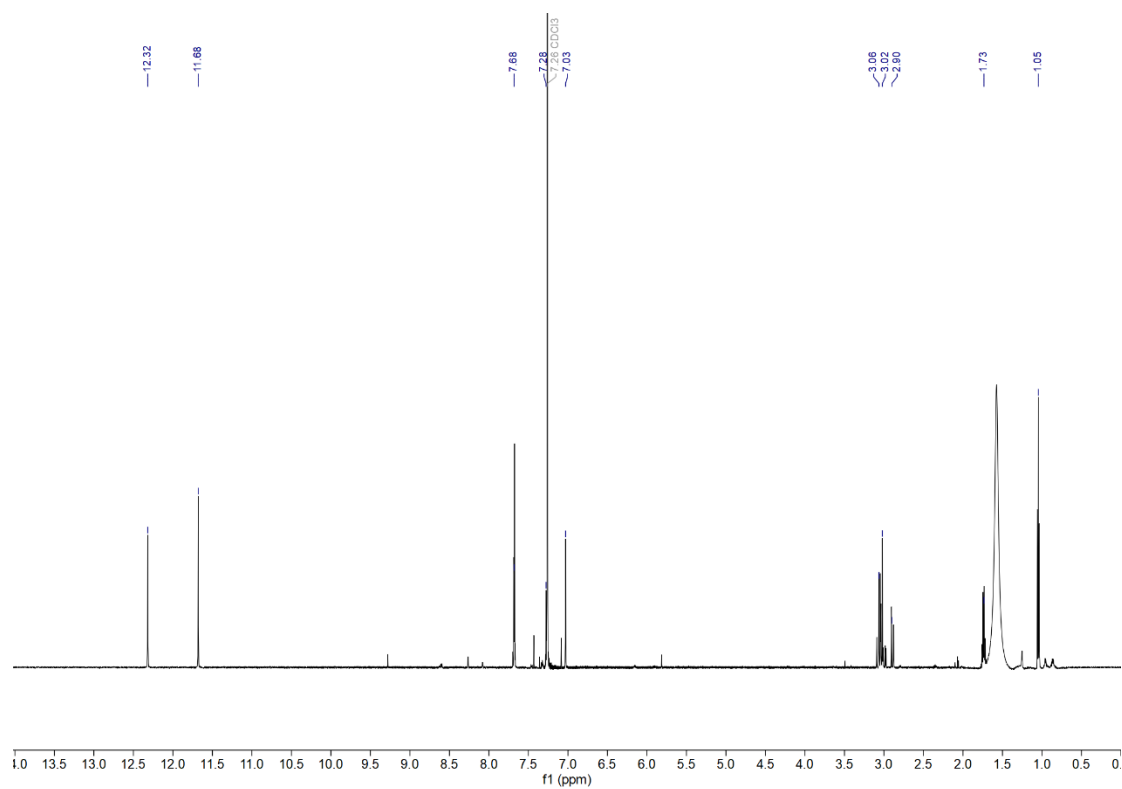
pos.	^{13}C		^1H			COSY	HMBC (^1H to ^{13}C)	ROSEY (^1H to ^1H)
	δ (ppm)	mult	δ (ppm)	mult	J (Hz)			
1	202.6	C						
2a	45.7	CH_2	3.45	d	18.4	2b	1, 3	
2b			2.74	d	18.4	2a	1, 3, 4, 9a(w)	
3	73.8	C						
4	61.2	CH	4.22	s			2, 3, 4a, 9a, 10, 11	14
4a	131.9	C						
5	118.2	CH	7.17	d	8.1	6	5(w) 7, 8a, 10	
5a	139.0	C						
6	132.9	CH	7.51	t	8.0	5,7	5a, 8	
7	111.8	CH	6.90	d	8.1	6	5, 8(w), 8a	
8	158.3	C						
8a	113.4	C						
9	ND	C						
9a	108.4	C						
10	119.1	CH	7.10	s			4, 5, 5a(w) 8a, 9a	
11	205.4	C						
12	31.6	CH_3	2.39	s			4, 11	
13	32.1	CH_2	1.74	mult		14	2, 3, 14	
14	6.6	CH_3	1.08	t	7.5	13	3, 13	4

ND Not Detected

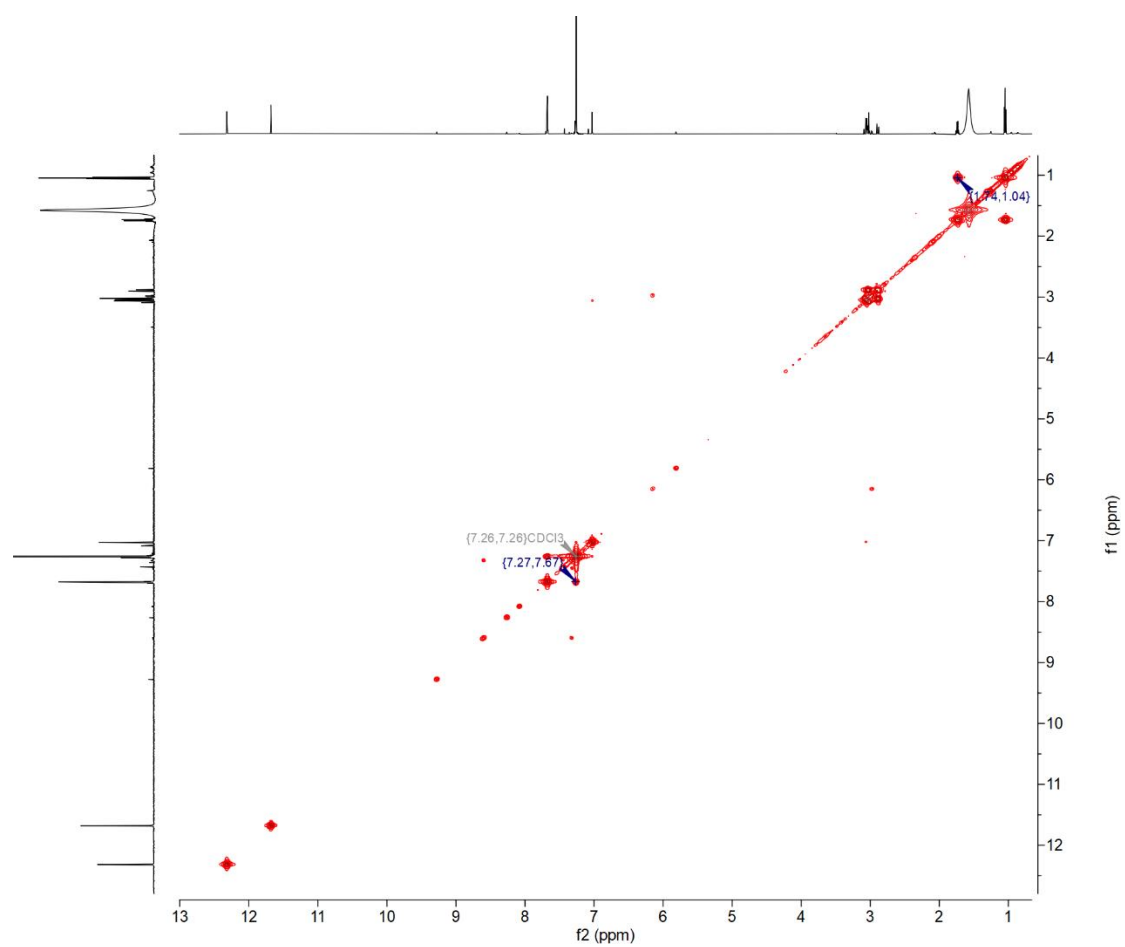
(w) Weak



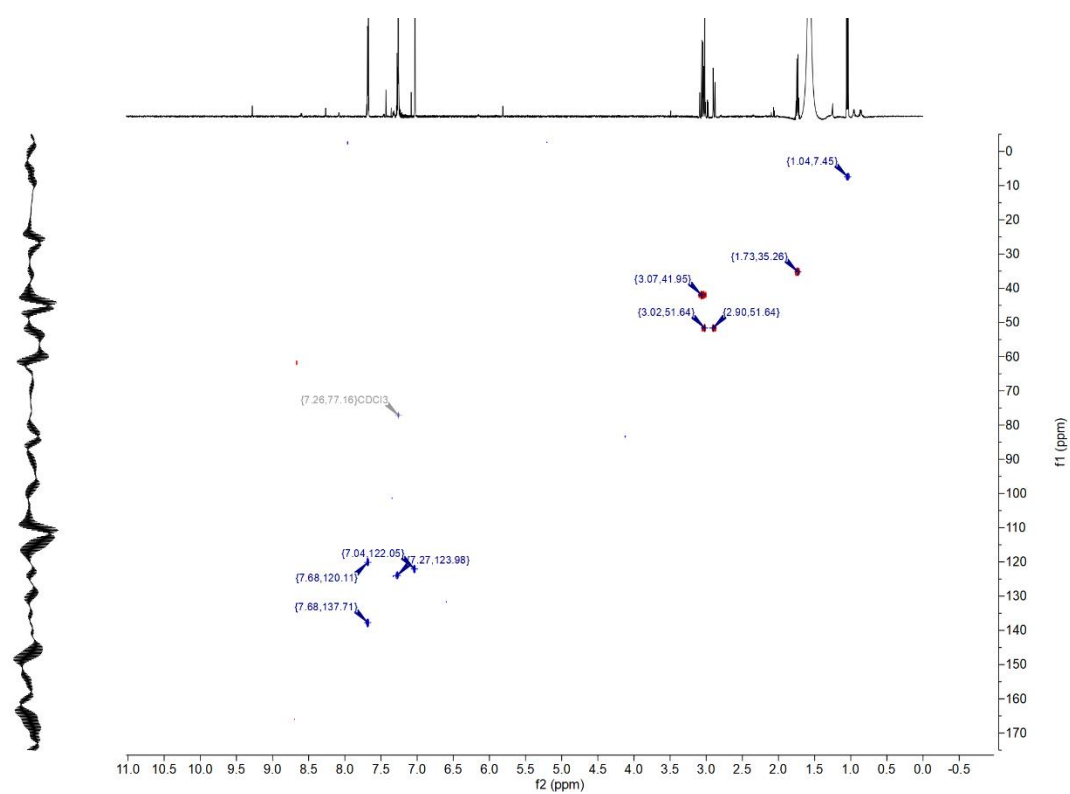
Appendix 23 ^1H NMR of homorableomycin (compound **10**)



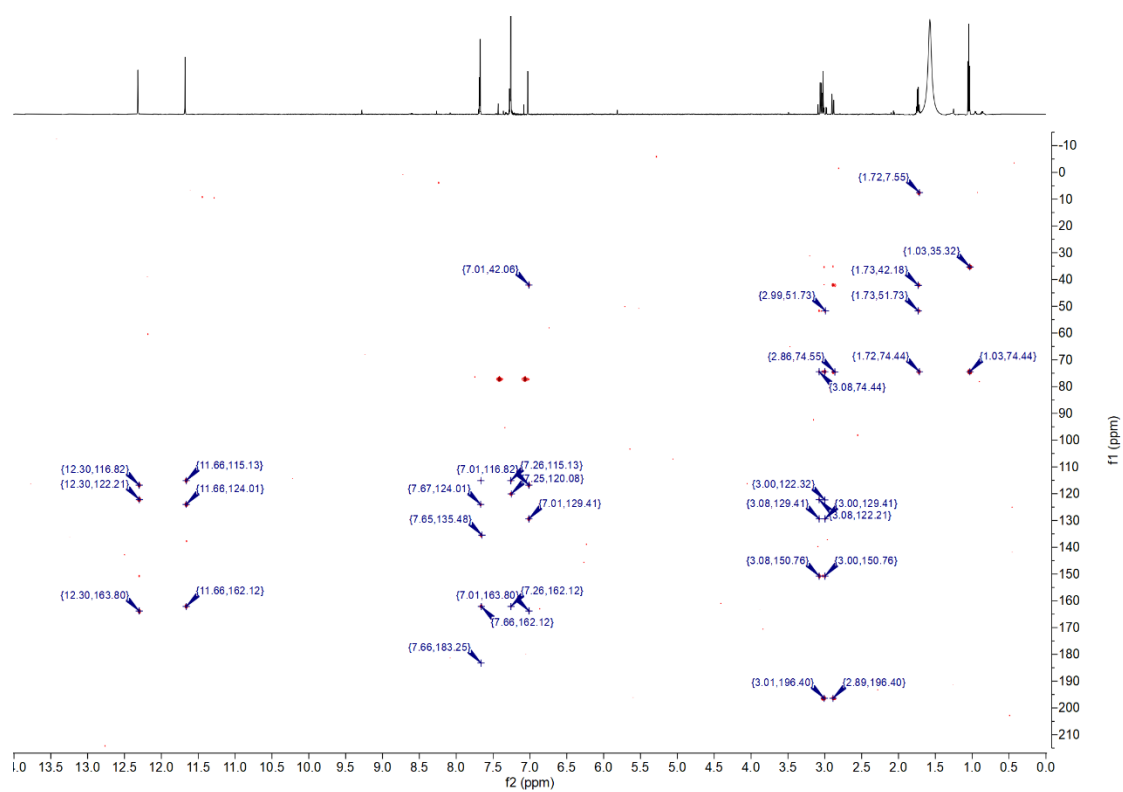
Appendix 24 COSY of homorableomycin (compound **10**)



Appendix 25 HSQC of homorableomycin (compound **10**)



Appendix 26 HMBC of homorableomycin (compound **10**)



Appendix 27 ^{13}C (150 MHz) and ^1H (600 MHz) NMR Data for homoreableomycin (compound **10**, CDCl_3)

Position	δ_c	δ_H	multi	$J(1\text{H}-1\text{H})$ in Hz	COSY	HMBC
1	196.4					
2	51.7	3.02 2.9	m dd	14.8, 1.9		1,3,4, 4a(w),12b
3	74.4					
3-OH	ND					
4	42.0	3.07	m			4a,5,12b,3
4a	150.8					
5	122.1	7.04	s			4,6a,12b,6
6	163.8					
6-OH		12.32	s			6,5,6a,4a(w)
6a	116.7					
7	ND					
7a	115.1					
8	162.1					
8-OH		11.68	s			7a,8,9
9	124.0	7.27	m		10 or 11	7a,11,8
10	137.7	7.68	m		9	8,11a,9,7a,12
11	120.1	7.68	m		9	
11a	135.5					
12	183.1					
12a	ND					
12b	129.4					
13	35.3	1.73	t	7.6	14	4,2,3,14
14	7.5	1.04			13	13,3

