EVOLUTION OF THE GLYCOPEPTIDE BIOSYNTHETIC GENE CLUSTER FAMILY

EVOLUTION OF THE GLYCOPEPTIDE BIOSYNTHETIC GENE CLUSTER FAMILY

By NICHOLAS A. WAGLECHNER, B.Sc, M.Sc

A thesis submitted to the School of Graduate Studies in partial fulfilment of requirements for the degree Doctor of Philosophy

DESCRIPTIVE NOTE

McMaster University, DOCTOR OF PHILOSOPHY (2020), Hamilton, Ontario (Health Sciences)

TITLE: Evolution of the Glycopeptide biosynthetic gene cluster family

AUTHOR: Nicholas A. Waglechner, B.Sc., M.Sc. (McMaster University)

SUPERVISOR: Dr. Gerard D. Wright

NUMBER OF PAGES: xx, 292

LAY ABSTRACT

The discovery of antibiotics is a triumph of the 20th century. Most antibiotics are derived from microorganisms – they are natural products. The biosynthetic gene clusters (BGCs) encoding their production are being revealed through an unprecedented genomic sequencing. My work considers the way BGCs are represented and ways of comparing the biosynthetic potential of different strains. Glycopeptide antibiotics (GPAs) are chemically diverse, made by several genera of Actinobacteria. Using BGCs derived from our in-house and public databases I build a natural history of these molecules, dating their emergence to 150-400 million years ago and connect it with GPA resistance. I further explore their evolution by considering the wider set of related BGCs to propose an expanded classification system that naturally points the way to the discovery of novel molecules. This culminates in discovery of corbomycin, shown to possess a new mechanism of action shared by other molecules in my classification.

ABSTRACT

The serendipitous discovery of antibiotics in the 20th century paved the way for safer, modern medical interventions. Bacteria in the phylum Actinobacteria are the most prolific producers of natural products, including antibiotics. The availability of low-cost, high-throughput generation of bacterial genome sequence data transforms natural product discovery, making it possible to judge the biosynthetic capacity of a strain based on its genome sequence.

I developed a software tool to compare biosynthetic gene clusters (BGCs) to show that streptothricin production is distributed among *Streptomyces* and that the capacity to produce common natural products does not predict the remaining potential of *Streptomyces* species. This approach provides a way to consider the rarity of particular natural products and grounded a biotechnological approach using CRISPR/Cas9 engineering to facilitates the identification of rare natural products in these strains.

Glycopeptide antibiotics (GPAs) are encoded by BGCs in several genera of Actinobacteria. Their diversity is the product of an intricate evolutionary history. We show that GPA biosynthesis and resistance maps to approximately 150-400 million years ago, from an older, pre-existing pool of components. We find that resistance appeared contemporaneously with biosynthetic genes, raising the possibility that the mechanism of action of glycopeptides was a driver of diversification in these gene clusters.

In a set of GPA BGCs, we identify several scaffolds distinct from the traditional D-Ala-D-Ala binding antibiotics. while complestatin, kistamicin, and longer peptides like enduracidin and ramoplanin are known, others are uncharacterized. Through phylogenetic analysis of these BGCs we develop a new classification scheme to organize these BGCs into four major classes. Structural predictions led us to purify complestatin and a novel compound we named corbomycin. Both possess antibacterial activity. Mutations conferring decreased susceptibility to these compounds suggest a novel mechanism of action distinct from known compounds in the GPA family.

ACKNOWLEDGEMENTS

The love and support of many people made this work possible. To whoever is reading this: if you think this could possibly mean *you*, it probably does. To all the people with which I have been fortunate enough to collaborate, thank you for letting me be a part of your science.

Many thanks to my supervisor, Gerry Wright. Your laboratory has been the nucleus of my professional life for many years. Too few people tell you how remarkable it is. Even more remarkable is the guidance and support you have provided me during this time – it has changed my life, thank you for letting me a part of it. I would like to thank my supervisory committee members, Dr. Andrew McArthur and Dr. Paul Higgs. Your mentorship, in and out of the classroom and laboratory, has made all the difference.

The personnel of the Wright lab – you are too numerous to list. A tide of people passed through and I have learned something from all of you, thank you. In particular, I owe much to the people who have made glycopeptide antibiotics into such an interesting story in the Wright lab: Vanessa, Lindsay, Andrew K., Maulik, Wen, Grace, Beth, Peter, Ricardo, and Kalinka, This is all your fault.

Thank you to my parents, especially my late father. At first, I wrote '*I wish you were here'*, but you are here in everything that I do.

Lastly, to my wife Susan. You have been my constant and enduring champion for as long as I have known you. This could not have happened without you at my side. You have/are my love.

TABLE OF CONTENTS

DESCRIPTIVE NOTE ii
LAY ABSTRACT iii
ABSTRACT iv
ACKNOWLEDGEMENTS v
TABLE OF CONTENTS vi
LIST OF FIGURES x
LIST OF TABLES xiv
LIST OF ABBREVIATIONS xv
DECLARATION OF ACADEMIC ACHIEVEMENT xx

CHAPTER ONE:
Introduction 1
INFECTIOUS DISEASE IN THE PRE-ANTIBIOTIC ERA 2
NATURAL PRODUCT ANTIBIOTICS 3
ANTIBIOTIC RESISTANCE 6
GENOMICS OF ANTIBIOTICS AND RESISTANCE 11
Genome Sequencing and Annotation 11
Representations of Resistance 14
Methods for Annotating Antibacterial Resistance 21
Representations of Antibiotic Biosynthesis 26
Methods for identifying Biosynthetic Gene Clusters 28
RESEARCH GOALS 34

CHAPTER TWO:
Representing and comparing biosynthetic gene clusters 38
CHAPTER TWO PREFACE 39
ABSTRACT 40
INTRODUCTION 41
METHODS 45
Strain Selection 45

Genome Sequencing of Wildtype Streptothricin Producing Strains 45
Phylogenomic Analysis 46
Identification of BGCs and Assessment of diversity 46
Development of the evoc Software 48
RESULTS 49
Phylogeny of Streptothricin Producing Strains 49
The evoc Software and Representation of BGCs 54
Measures of Strain-wise and BGC-wise Similarity 59
DISCUSSION 66

CHAPTER THREE:
Phylogenetic reconciliation reveals the natural history of glycopeptide antibiotic biosynthesis and resistance 70
CHAPTER THREE PREFACE 71
ABSTRACT 72
INTRODUCTION 73
METHODS 77
Growth of organisms, DNA sequencing and assembly 77
Selection of BGCs 78
Annotation of BGCs and ORF sequences 79
Construction of species trees 80
Construction of domain/gene trees and obtaining BGC-related sequences 83
Phylogenetic reconciliation 83
Data Availability 84
Code availability 85
RESULTS 86
GPA production is distributed in Actinobacteria 86
GPA biosynthesis is accomplished by at least 90 component families 89
GPA precursor biosynthesis is over 1 billion years old 93
The GPA scaffold is three to five hundred million years old 98
Reconciliation of GPA tailoring strategies 100

Efflux and regulatory elements reveal possible multiple lateral gene transfer events 104
Resistance is contemporary with GPA biosynthesis dating to one to four hundred million years 105
DISCUSSION 106
SUPPLEMENTARY MATERIAL 109
Supplementary Figures 110
Supplementary Tables 121

CHAPTER FOUR:
Phylogenetics predict diverse members of the glycopeptide biosynthetic gene cluster family 136
CHAPTER FOUR PREFACE 137
ABSTRACT 138
INTRODUCTION 139
METHODS 143
GPA-like BGC sequences 143
GPA BGC analysis 143
Phylogenetic trees 144
Growth of organisms 144
Fermentation and purification of complestatin 145
Raising resistance mutants through serial passage in the presence of antibiotic 146
RESULTS 148
Several distinct scaffold types are encoded by GPA-like BGCs in Actinobacteria 148
The kistamicin scaffold is a hybrid Class III and IV 167
Analysis of class II scaffolds 169
Analysis of class IV scaffolds 175
Streptomyces sp. WAC01325 produces the class IVa scaffold complestatin 181
Streptomyces sp. WAC01529 produces corbomycin 186
Corbomycin and complestatin have antibiotic activity 191
DISCUSSION 200
SUPPLEMENTARY MATERIAL 207

Supplementary Figures 208

CHAPTER FIVE:
Discussion and future directions 249
Discussion 250
Future Directions 254
Concluding Remarks 255

REFERENCES 256

APPENDICES 271
Appendix 1: Culp, E. and Yim, G. *et al.* (2019) Hidden antibiotics in actinomycetes can be identified by inactivation of gene clusters for common antibiotics. *Nat Biotechnol* 37(10):1149-54. 271
Appendix 2: Culp, E. *et al.* (2020). Evolution-guided discovery of antibiotics that inhibit peptidoglycan remodelling. *Nature* 578:582-87. 281

LIST OF FIGURES

Chapter One

Figure 1-1: Mechanisms of antibiotic resistance....9
Figure 1-2: Overview of DNA sequencing and assembly....12

Chapter Two

Figure 2-1: Maximum-likelihood *Streptomyces* phylogeny....51
Figure 2-2: Entity relation diagram of the evoc schema....55
Figure 2-3: Composition and annotation of a BGC....58
Figure 2-4: Pairwise genome vs genome sum BGC distance in 42 streptothricin producers....61
Figure 2-5: Pairwise genome vs genome sum BGC distance versus phylogenetic distance....62
Figure 2-6: Pairwise genome vs genome BGC sum Goodman-Kruskal gamma index....64

Chapter Three

Figure 3-1: GPAs, BGC evolution and precursor biosynthesis....74
Figure 3-2: Species phylogeny and reconciliation....87
Figure 3-3: GPA precursor biosynthesis reconciliation dates....95
Figure 3-4: GA scaffold A-domain phylogeny and reconciliation dates....99
Figure 3-5: GPA tailoring, resistance and regulation reconciliation dates....102
Figure 3-6: Summary of major events in GPA BGC evolution inferred from reconciliation....107
Supplementary Figure 3-1: GPAs and related structures....110
Supplementary Figure 3-2: GPA BGC synteny....112
Supplementary Figure 3-3: Workflow overview....114
Supplementary Figure 3-4: 16S rRNA species tree....116
Supplementary Figure 3-5: Dated amino acid species trees....117
Supplementary Figure 3-6: densiTree plot of species trees....119
Supplementary Figure 3-7: Amino acid precursor biosynthesis....120

Chapter Four

Figure 4-1: Feglymycin structure, BGC, and NRPS configuration 151
Figure 4-2: Ramoplanin and enduracidin structures, BGCs, and NRPS configurations 153
Figure 4-3: Vancomycin structure, BGC, and NRPS configuration 155
Figure 4-4: Kistamicin structure, BGC, and NRPS configuration 156
Figure 4-5: Approximate maximum likelihood phylogeny of GPA BGC adenylation domains 158
Figure 4-6: Approximate maximum likelihood phylogeny of GPA BGC condensation domains 164
Figure 4-7: Approximate maximum likelihood phylogeny of GPA BGC peptidyl-carrier domains 166
Figure 4-8: Approximate maximum likelihood phylogeny of GPA BGC X domains.169
Figure 4-9: GP1438 predicted structure, BGC, and NRPS configuration 172
Figure 4-10: Approximate maximum likelihood phylogeny of GPA BGC thioesterase domains 174
Figure 4-11: GP6369 predicted structure, BGC, and NRPS configuration 175
Figure 4-12: Approximate maximum likelihood phylogeny of GPA BGC cytochrome P450 monooxygenases 178
Figure 4-13: GP6738 predicted structure, BGC, and NRPS configuration 180
Figure 4-14: Complestatin structure, BGC, and NRPS configuration 182
Figure 4-15: Variant class IVa BGC comparison 184
Figure 4-16: Approximate maximum likelihood phylogeny of GPA BGC halogenases 185
Figure 4-17: GP1529 prediction and structure of corbomycin, the corbomycin BGC and NRPS configuration.....187
Figure 4-18: Approximate maximum likelihood phylogeny of GPA BGC sensor histidine kinases 190
Figure 4-19: Approximate maximum likelihood phylogeny of GPA BGC response regulators 191
Figure 4-20: GPA scaffold class relationships 204
Supplementary Figure 4-1: A-domain phylogeny subset A 208
Supplementary Figure 4-2: A-domain phylogeny subset B 209

Supplementary Figure 4-3: A-domain phylogeny subset C 210
Supplementary Figure 4-4: A-domain phylogeny subset D 211
Supplementary Figure 4-5: A-domain phylogeny subset E 212
Supplementary Figure 4-6: A-domain phylogeny subset F 213
Supplementary Figure 4-7: A-domain phylogeny subset G 214
Supplementary Figure 4-8: A-domain phylogeny subset H 215
Supplementary Figure 4-9: A-domain phylogeny subset I 216
Supplementary Figure 4-10: A-domain phylogeny subset J 217
Supplementary Figure 4-11: A-domain phylogeny subset K 218
Supplementary Figure 4-12: A-domain phylogeny subset L 219
Supplementary Figure 4-13: Condensation domain phylogeny subset A 220
Supplementary Figure 4-14: Condensation domain phylogeny subset B 221
Supplementary Figure 4-15: Condensation domain phylogeny subset C 222
Supplementary Figure 4-16: Condensation domain phylogeny subset D 223
Supplementary Figure 4-17: Condensation domain phylogeny subset E 224
Supplementary Figure 4-18: Condensation domain phylogeny subset F 225
Supplementary Figure 4-19: Condensation domain phylogeny subset G 226
Supplementary Figure 4-20: Condensation domain phylogeny subset H 227
Supplementary Figure 4-21: Condensation domain phylogeny subset I 228
Supplementary Figure 4-22: Condensation domain phylogeny subset J 229
Supplementary Figure 4-23: Condensation domain phylogeny subset K 230
Supplementary Figure 4-24: PCP domain phylogeny subset A 231
Supplementary Figure 4-25: PCP domain phylogeny subset B 232
Supplementary Figure 4-26: PCP domain phylogeny subset C 233
Supplementary Figure 4-27: PCP domain phylogeny subset D 234
Supplementary Figure 4-28: PCP domain phylogeny subset E 235
Supplementary Figure 4-29: PCP domain phylogeny subset F 236
Supplementary Figure 4-30 PCP domain phylogeny subset G 237
Supplementary Figure 4-31: PCP domain phylogeny subset H 238
Supplementary Figure 4-32: PCP domain phylogeny subset I 239
Supplementary Figure 4-33: PCP domain phylogeny subset J 240

Supplementary Figure 4-34: PCP domain phylogeny subset K.....241
Supplementary Figure 4-35: PCP domain phylogeny subset L 242
Supplementary Figure 4-36: PCP domain phylogeny subset M 243
Supplementary Figure 4-37: Cytochrome P450 monooxygenase phylogeny subset A 244
Supplementary Figure 4-38: Cytochrome P450 monooxygenase phylogeny subset B 245
Supplementary Figure 4-39: Cytochrome P450 monooxygenase phylogeny subset C 246
Supplementary Figure 4-40: Cytochrome P450 monooxygenase phylogeny subset D 247
Supplementary Figure 4-41: Cytochrome P450 monooxygenase phylogeny subset E 248

LIST OF TABLES

Chapter One

Table 1-1: Summary of antimicrobial resistance bioinformatics resources 20

Table 1-2: A selection of core sequences used to identify BGCs 30

Chapter Two

Table 2-1: Basic types loaded on type table creation 57

Table 2-2: Basic relationships loaded during type_relationship table creation 57

Chapter Three

Table 3-1: Component families involved in GPA biosynthesis 91

Supplementary Table 3-1: Description of organisms, clusters, and sequences used in this study 121

Supplementary Table 3-2: Time tree node details 127

Supplementary Table 3-3: Kendall-Colijn distance between species trees.....131

Supplementary Table 3-4: Normalized Robinson-Foulds distance between species trees 132

Supplementary Table 3-5: Node reconciliation details 133

Chapter Four

Table 4-1: GPA BGC classification criteria.....149

Table 4-2: Scaffold peptides and classes.....160

Table 4-3: Initial broth dilution MIC values 193

Table 4-4: Genotypes of *B. subtilis* 168 strains raised to be resistant to complestatin and corbomycin compared to parental and NCBI reference genome 194

Table 4-5: Genotypes of *S. aureus* ATCC 29213 strains raised to be resistant to complestatin and corbomycin compared to parental and NCBI reference genome.....197

LIST OF ABBREVIATIONS

16S rRNA	16S ribosomal RNA
46DH	NDP-sugar 4,6 dehydratase
4KR	4-keto reductase
95%HPD	95% highest posterior distribution
ABC-ATP	ABC transporter ATPase
abhyd	alpha/beta hydrolase
Ac-CoA	acyl co-enzyme A
ACP	acyl carrier protein
Actm.	*Actinomadura*
Actp.	*Actinoplanes*
acyltf	acyltransferase
A-domain	adenylation domain
Am.	*Amycolatopsis*
AMP	adenosine monophosphate
AMR	antimicrobial resistance
antiSMASH	antibiotics & secondary metabolites analysis shell
antiSMASHdb	antiSMASH database
APH	aminoglycoside phosphotransferase
ARDB	Antibiotic Resistance Database
aSDomain	antiSMASH domain
AT	acyltransferase
ATCC	American Type Culture Collection
BD	birth-death process
BEAST	Bayesian evolution analysis sampling trees
BGC	biosynthetic gene cluster
BHI	brain heart infusion medium
Bht	β-hydroxytyrosine

BiGSCAPE	biosynthetic gene cluster similarity cluster and prospecting engine
BLAST	basic local alignment search tool
βOHase	amino acid betahydroylase
CAMHB	cation-adjusted Mueller-Hinton broth
CARD	Comprehensive Antibiotic Resistance Database
CAT	Bayesian mixture model for categorizaion of sites
cat_antiporter	cationic antiporter
catAcSymporter	cation/acetate symporter
C-domain	condensation domain
CDS	coding sequence
Cglyc	glycopeptide-type condensation domain
Cit	citrulline
ClusterFinder	hidden Markov model trained on BGCs
CM	chorismate mutase
coal	coalescent process
CORASON	core analysis of syntenic orthologs to prioritize natural product gene clusters
CRISPR	clustered regularly interspersed short palindromic repeats
DAHPS	7-deoxy-arabino-heptulosonate 7-phosphate synthase
dalbaheptide	D-Ala-D-Ala binding glycopeptide
DH	dehydratase
DMSO	dimethylsulfoxide
DNA	3'-deoxyribonucleic acid
DPG	3,5-dihydroxy phenylglycine
DSM	Deutsche Sammlung von Mikroorganismen und Zellkulturen GmbH
DTL	duplication-transfer-loss
eDNA	environmental DNA
EDTA	ethylene diamine tetra-acetic acid

.embl	European molecular biology laboratory file format
End	enduracididine
ER	enoyl-reductase
evoc	evolution of clusters
fungiSMASH	fungal antiSMASH
.gbk, .gb	genbank flat-file format
GFF3	general feature format version 3
GK-γ	Goodman-Kruskal gamma
GPA	glycopeptide antibiotic
gtf	glycosyltransferase
GTR	general time-reversible
Hal	halogenase
Hb.	*Herbidospora*
HGT	horizontal gene transfer
HisK	sensor histidine kinase
HMM	hidden Markov model
HPG	4-hydroxy phenylglycine
HPLC	high-performance liquid chromatography
HR-ESI-MS	high resolution electrospray ionisation mass spectrometry
HSP	high-scoring segment pair
Kb.	*Kibdellosporangium*
k-mer	oligonucleotide sequence of length k bases
KS	ketosynthase
LC-MS	liquid chromatography-mass spectrometry
Ma	millions of years ago
mbgtf	membrane-bound glycosyltransferase
Mc.	*Micromonospora*
MCC	maximum clade credibility

MCMC	Markov chain Monte Carlo
MHB	Mueller-Hinton broth
MIBiG	minimum information about a biosynthetic gene cluster
MIC	minimum inhibitory concentration
Micros.	Micromonosporaceae
ML	machine learning
MP	most parsimonious
MPR	most parsimonious reconciliation
MRCA	most recent commo ancestor
MT	methyltransferase
mtf	methyltransferase
MultiGeneBlast	modified BLAST procedure for BGCs
NCBI	National Center for Biotechnology Information
NDP	nucleoside diphosphate
NMR	nuclear magnetic resonance
NMT	N-methyltransferase
nmtf	N-methyltransferase
No.	*Nonomuraea*
nr	NCBI non-redundant BLAST database
NRBC	National Institute of Technology and Evaluation Biological Resource Center
NRPS	non-ribosomal polypeptide synthase
NRRL	Agricultural Research Service Culture Collection
omtf	O-methyltransferase
ORF	open reading frame
Orn	ornithine
pantiSMASH	plant antiSMASH
PCP	peptidyl carrier protein
pdh	prephenate dehydrogenase

PFAM	protein family
PGb	peptidoglycan binding protein
pHMM	profile hidden Markov model
PKS	polyketide synthase
QC	quality control
ResFam	resistance family
RespReg	DNA-binding response regulator
RF	Robinson-Foulds
RNA	ribonucleic acid
SDS	sodium dodecyl sulfate
SH-like	Shimodaira-Hasegawa like support values
smCOGS	secondary metabolite cluster of orthologous groups
St.	*Streptomyces*
stf	sulfotransferase
Streptosp.	Streptosporangiales
strict	strict molecular clock
TDP	thymidine diphosphate
TE	thioesterase
TERA	tree estimation using reconciliation algorithm
TIGRFAM	The Institute for Genomic Research families
TSB	typtone soy broth
unc	log-normal uncorrelated relaxed molecular clock
UV	ultraviolet
WAC	Wright Actinomycete Collection
WAG	Whelan and Goldman substitution model
WHO	World Health Organization

DECLARATION OF ACADEMIC ACHIEVEMENT

I have performed all of the research in this body of work except where indicated in the preface of each chapter.

CHAPTER ONE: Introduction

INFECTIOUS DISEASE IN THE PRE-ANTIBIOTIC ERA

Infectious disease is among the leading causes of death and disease burden worldwide (*Antibiotic Resistance Threats in the United States, 2013*). Three major developments were required to form a modern understanding this topic in the 19th century. First, the hygiene theory of Ignaz Semmelweis, showing that simple hand washing was sufficient to reduce the occurrence of fever in maternity wards. Pasteur's germ theory of disease showed that the presence of microbes, invisible to the naked eye and ubiquitous in the world, connected hygiene to spoilage and infection. Lastly, Koch and his postulates demonstrated that the presence of specific organisms related to the presence of specific disease. These ideas firmly connected the mostly unseen microbial world to the material reality of health and disease in a way that provides a rational foundation for the study of infectious diseases. This suggested that these diseases may be treated by inhibiting the organisms responsible, accomplished originally by reducing their transmission from wherever they reside to healthy people to cause disease. More importantly, these ideas raised the possibility that pathogenic organisms may be inhibited directly, either before or after disease is established.

Realizing this concept began with the developments of synthetic compounds having narrow spectra of inhibition against a specific set of organisms. Activity here means that exposure of microbes to this material results in toxic impairment of the microbe with the necessary specificity so this toxicity should not also impair the host infected by these organisms. Ehrlich's 'magic bullet' encapsulated the idea that specific toxin could be 'aimed chemically' at a particular microbial target (Gensini, Conti, &

Lippi, 2007). The historical development of organic chemistry on continental Europe was fueled by the quest for alternatives to the costly import and refinement of materials for the production of dyes and other industrial chemicals (Brock, 2000). From dyes containing toxic arsenic, the first agents with specific activity against organisms causing infectious disease were put forward (Sneader, 2005).

Two decades later, the serendipitous discovery of penicillin proved to be pivotal for a number of reasons. First, unlike early chemotherapy, penicillin had both a much broader range of activity against a wider array of organisms and a wider margin of safety owing to decreased toxicity in humans (Sneader, 2005). Second, it was produced by one microbe, a *Penicillium* fungus, when incubated on the same plate of media as a *Staphylococcus* pathogen. It appeared that one organism growing in competition with another for the same physical space and resources on solid media could produce and secrete a compound with the ability to inhibit its competitor. This competition repeats itself across the microbial world (Cornforth & Foster, 2015; Gottlieb, 1976). Generally, organisms possess a repertoire of small molecules that could be harvested by humans to fight infectious disease.

NATURAL PRODUCT ANTIBIOTICS

Antibiotics, as they came to be known, are among the most important discoveries of the 20th century (Bud, 2007). Antibiotics have been transformative in modern medicine, making procedures such as surgery, organ transplantation, other invasive procedures, and even cuts and scrapes no longer subject to simple or complicated infections. Antibiotics are defined as any compound with the specific ability to impair the

growth of bacteria (Waksman, 1947). They are members of the larger class of compounds called antimicrobials which include compounds with activity against viruses, fungi, parasites, and bacteria.

These small molecules are known as natural products, the products of secondary (or non-essential) metabolism of microbes, plants and animals. After the first successful discovery of an antibiotic, research programs were initiated to replicate the discovery of penicillin. These early efforts entailed systematic replication of the discovery of penicillin on a larger scale – more organisms, grown on more media types were screened to determine if they produced compounds with antimicrobial properties (Demain, 2014). This was known as the golden age of antibiotic discovery, lasting several decades from the 1940s to the 1970s, and it produced examples of essentially every major class of antibiotic known to medicine, as well as countless molecules unsuitable for development by the pharmaceutical industry (Demain, 2014). Natural products are an important source of compounds that have been ultimately developed into pharmaceuticals, antibiotics being the largest and most successful examples. It is estimated that nearly 80% of antibiotics currently in use are themselves or have been derived from natural products (Kieser, Bibb, Buttner, Chater, & Hopwood, 2000).

One of the lessons learned from the screening efforts used to identify natural product production is that some bacterial taxa are more prolific than others in terms of the number and diversity of natural products they produce (Baltz, 2005). In particular, the phylum Actinobacteria have been identified as being particularly good sources of natural products. In Actinobacteria, the genus *Streptomyces* has contributed nearly 80% of the

natural product antibiotics. *Streptomyces coelicolor* A(32), established as a model organism of the genus by David Hopwood, produces several pigmented antibiotics at different times during its life cycle (Kieser et al., 2000). The phenotype of pigment production was a useful genetic tool to study the biosynthesis of these natural products. Significantly, the genes encoding for the biosynthesis of antibiotics are clustered together (Martin, 1992). These clusters of genes are presumably linked because together they encode the capacity to convert primary metabolic products, like sugars, lipid precursors, and amino acids, through a pathway that produces a new compound termed a secondary metabolite, that provides a benefit to the host. In the case of antibiotics, it is believed that the ability to inhibit the growth of susceptible neighbouring organisms provides a competitive advantage to the producer, and so there is positive selection to keep these biosynthetic gene clusters (BCGs) intact (Gregory L Challis & Hopwood, 2003). Before and during the early genome era, as BCGs were studied many whole and partial sequences were deposited in the public sequence databases confirming the idea that genes for natural product biosynthesis were indeed clustered, greatly expanding the genetics and biochemistry of antibiotic production.

A major revelation occurred when the genome of *Streptomyces coelicolor* A(32) was sequenced in 2002 (Bentley et al., 2002). Extending the study of previous BGCs, it was predicted that this well-studied organism encoded the potential biosynthesis of 18 other natural products that were never before observed under any growth conditions (G. L. Challis, 2014). The subsequent publication of other model *Streptomyces* spp.(Ikeda et al., 2003; Ohnishi et al., 2008) and other Actinobacterial, and bacterial genomes, in

general has demonstrated that there is a wealth of undiscovered genetic potential for natural products.

ANTIBIOTIC RESISTANCE

Antibiotics became victims of their own success. Soon after the introduction of every new class of antibiotic, treatments began to be less effective. This takes the form of longer necessary duration of treatment, higher or more frequent dosing to achieve the same effect, and ultimately treatment failure requiring substitution to a different drug entirely. This phenomenon became known as antimicrobial resistance (AMR), and it was initially described empirically. As the discovery of antibiotics progressed, old compounds were replaced with newer, less toxic, cheaper, or more effective versions. Medicinal chemists had a hand in modifying natural products to have more desirable properties such as increased potency, broader spectra of activity, increased stability, and decreased toxicity. With many options available, it was initially easy to switch from one compound to another. The concept of antibiotic resistance was known from the time of the discovery of penicillin, namely that sub-lethal doses select for resistant members of a population of bacteria (Demain, 1974). Antibiotics by definition are selective agents, preferentially killing susceptible members of a population upon exposure, leaving the resistant individuals behind. Susceptibility and resistance are variable traits within and between populations of bacteria.

Every exposure of organisms to a growth inhibiting compound results in an increased relative frequency of resistant bacteria. The inverse of a compound having specific activity against a particular taxon of bacteria is that there are other bacteria who

are resistant to that compound. For a variety of reasons that will be discussed, the rate at which new compounds were discovered did not match the rate at which organisms became resistant. Over the past decade, multiple national and international organizations took concrete steps to recognize the urgent risk posed by antimicrobial resistance. In 2015, the World Health Organization began to develop a Global Action Plan to address the new reality that antimicrobial resistance is an existential threat to modern medicine (World Health Organization, 2015). In the United States, the Center for Disease Control estimates that over 2 million Americans become infected by resistant organisms, resulting in over 23,000 deaths per year (*Antibiotic Resistance Threats in the United States, 2013*). The Public Health Agency of Canada has developed their own Federal Action Plan on Antimicrobial Resistance in 2015 including building a Canadian Antibiotic Resistance Surveillance System to monitor the spread of resistance between humans, animals and the environment (*Federal Action Plan on Antimicrobial Resistance and Use in Canada*, 2015). In 2019, a report by the United Nations Ad hoc Interagency Coordinating Group on Antimicrobial Resistance estimates that AMR will result in over 10 million deaths per year worldwide and cause catastrophic damage to the global economy by 2050 (*No Time To Wait: Securing The Future From Drug-Resistant Infections*, 2019).

The inhibitory actions of antibiotics against bacteria come as a result of inhibiting essential cellular functions, which are constrained to several major classes. Major targets of antibiotics are the bacterial cell wall and membrane, including peptidoglycan biosynthesis, that help bacterial cells maintain structural integrity. Additionally, key cellular processes such as DNA replication, translation and transcription are all targets for

multiple classes of antibiotics. Similarly, the ability for bacteria to ameliorate the toxic effects of antibiotics are constrained to several major strategies. Antibiotic targets are located on or within bacterial cells, which antibiotics have to reach at a high enough concentration to inhibit. For targets within the cell, an obvious strategy is to reduce permeability of the antibiotic, frequently achieved via modification of the cell wall and/or membrane or altered porin expression, or to actively lower the intracellular concentration via efflux. Protein targets may evolve lower intrinsic affinity for antibiotics, or they may be enzymatically modified such that this affinity is lowered or blocked altogether. Where this modification takes place inside the cell, the available donors of these chemical groups tend to be sourced from primary metabolism. Bacteria may harbor both susceptible and non-susceptible targets and may be able to alter the relative proportions of each through regulation of gene expression. Protection proteins may evolve that interact with targets to prevent access to the antibiotic. A major class of resistance comes from modification of antibiotics. Antibiotics may be modified by group transfer enzymes where chemical groups are added to the antibiotic that prevent interaction with the target. Antibiotics may be subject to redox reactions which can lead to partial or total metabolism of the compound. Another important strategy is enzymatic hydrolysis of labile bonds leading to the inactivation of antibiotics. Lastly, non-protein targets can bypass antibiotic binding through enzymatic modification. Examples of each mechanism are depicted in Figure 1-1.

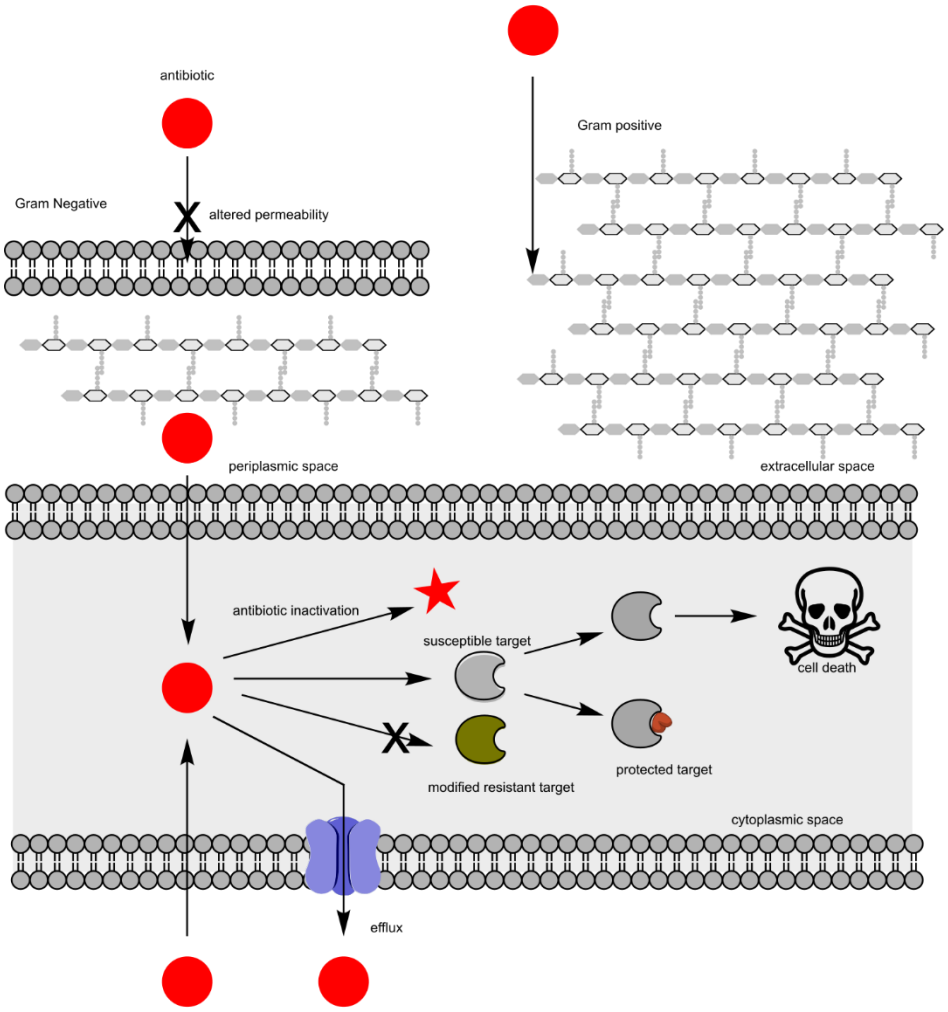


Figure 1-1: Mechanisms of antibiotic resistance. Only antibiotic interaction with a susceptible target can result in cell death. The effective concentration of an antibiotic inside the cell is reduced by preventing accumulation inside the cell, by transporting it back out once it is inside, or eliminating it by modification or inactivation. The other component of this interaction is the susceptible target, which may be modified to reduce interaction with the antibiotic or be protected from the antibiotic. These are the general ways cells become resistant to antibiotics.

Various behavioural strategies for increasing resistance are known. Bacteria may form biofilms, which are a structured growth of cells, strengthened by secretion of a proteinaceous and/or carbohydrate matrix that decreases permeability and increases adhesion of a population of bacteria to organic and inorganic surfaces. Biofilms also

increase resistance to physical forces such as fluidic flow and shearing. A large enough biofilm creates different microenvironments, such as a surface and interior, that are free to vary in behavior (such as gene expression) and experience external stimuli (such as antibiotic concentrations) depending on their level of exposure. The phenomenon of persistence, a state of lowered metabolic activity, is also known to protect bacterial cells from antibiotics. The reasons small numbers of a bacterial population enter and exit these states of dormancy are not fully understood but may involve responses to stimuli. Evidence exists that persister cell formation may be a stochastic event, and therefore the increased resilience of persisters might be a general strategy for survival and resilience (Harms, Maisonneuve, & Gerdes, 2016).

All of the elements embodying these strategies considered together are termed the resistome (G. D. Wright, 2010). This concept can be broadened to all bacteria or narrowed to a single species or even a single organism. An underlying assumption of the resistome concept is that every source of resistance, whether a protein, a behavior, gene expression or regulation, is encoded in the genome of an organism. This is important because it means that the resistome is heritable and therefore subject to evolution, most obviously by selection with antibiotics.

GENOMICS OF ANTIBIOTICS AND RESISTANCE

Genome Sequencing and Annotation

The phenomena of antibiotic biosynthesis, sensitivity and resistance is written in the DNA of bacteria. These heritable sequences are passed down vertically to cellular descendants and horizontally between organisms. Genomics is the collective study of these sequences on a large scale. The development of genomics in general goes hand in hand with the development of sequencing technology and the computational techniques used to organize these data.

Sequencing reads are the direct product of DNA sequencing and are typically around a few hundred base pairs long for second generation sequencing and several thousand base pairs long for third generation sequencing. Regardless of the method used to produce small fragments of DNA, assembly techniques are used to computationally produce larger and more contiguous sequences, known as contigs, by oversampling fragments from many copies of bacterial genomes (Figure 1-2). The expectation is that each position of a bacterial genome sequence will be sampled on many fragments and that these fragments may be overlapped and combined to reproduce the original sequence. These assembled contigs represent whole or partial bacterial chromosomes and/or plasmids ranging from tens of thousands to millions of base pairs long. In practice, while generating these longer sequences it is often difficult to distinguish between chromosomal and plasmid sequences without additional information such as depth of coverage or comparing the assembled sequences against a reference database of known sequences. It is also difficult to distinguish erroneous contigs produced from

contaminating reads not originating from the genome being sequenced, or contigs produced by combining these reads with correct reads, or contigs produced via misassembly of real reads. Genome sequences produced without careful, expensive and time-consuming experimental validation should be considered draft quality but can still contain much useful information for downstream experiments.

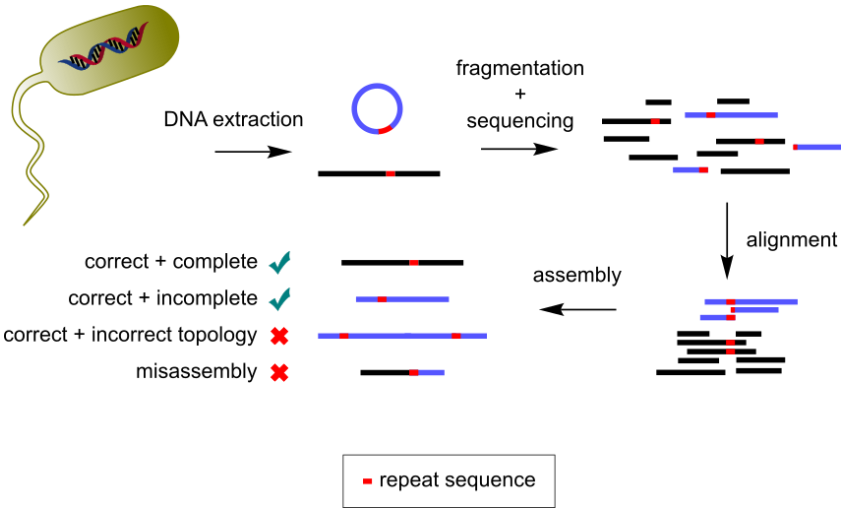


Figure 1-2: Overview of DNA sequencing and assembly. The general procedure of DNA sequencing and assembly consists of several steps common to most sequencing technologies. Often, the number, ploidy, and topology of the molecules being sequenced is unknown *a priori*. Whole molecules are not amenable for sequencing, so a fragmentation step is necessary in addition to the technology specific library preparation steps. The actual sequencing will produce a large number of digital fragments of the original sequences. These fragments are typically aligned with one another, either directly (all versus all alignment) or indirectly via decomposition into subsequence of length *k* (k-mers) into an intermediate form. Consensus sequences are assembled from this intermediate form using heuristics with the goal of reproducing the sequence of the original DNA molecule but may also produce sequences with various kinds of errors. These errors can include, but are not limited to, sequences with incorrect topologies, and incorrect joining of sequences owing to sequencing mistakes and the presence of repeated sequences (red).

Genome sequences themselves are not useful without annotation. There are several formats for annotating nucleotide sequences, but what they have in common are

meta-data and a coordinate system that refer to positions in the nucleotide sequence of the nucleotide sequences in a genome assembly. Metadata can in principle be used to track any desired information for each sequence and minimally should include an identifier that can be used to label the nucleotide sequence. Beyond this, the data that is tracked in metadata can be anything and may be format specific. The different formats in standard use will have defined support for some specific fields built into the specification, see for example the Genbank flat-file format at the National Center of Biotechnology Information (NCBI) (gbk or gb, https://www.ncbi.nlm.nih.gov/Sitemap/samplerecord.html, accessed Sept. 2019) which stores fields related to the metadata stored and referenced by the GenBank databases. Other common formats are the European Molecular Biology Laboratory (embl) flat file format, and General Feature Format version 3 (GFF3). These formats are just specifications. In daily practice, parsing software is used to read and write these formats and this software may only implement a subset of the features and fields of these specifications, or may support reading but not writing (so-called 'round trip' parsing). For example, the SeqIO module of BioPython, a bioinformatics software library written for the Python language, can read a wider variety of data from a GenBank format file than it can write so if a file needs to be read then written as part of a larger analysis pipeline than metadata other than the basic fields is often lost.

Some of this lost flexibility can be regained through feature annotations on nucleotide sequences. A feature is represented by a coordinate interval and a type that describes the feature. One such common feature is the coding sequence, or CDS, that

represents an interval of nucleotide sequence derived from the parent nucleotide record than can be conceptually translated (via a specified translation table) to yield a protein sequence. While the actual GenBank format lists the various types of features supported by the specification, parsing software can read almost any type of feature that can be associated with a coordinate interval without needing the feature type to be hardcoded into the parser. Depending on the format, these custom types remain human-readable when opening the file in a text editor. In a similar way that arbitrary features can be associated with arbitrary nucleotide sequences, features can have arbitrary fields associated with them that can be read and written in an application-specific manner as opposed to a format-specific manner enforced via the parsing tool.

The discussion of these formats is necessary because genomics research is built upon a shifting foundation of standards, formats, and implementation of software tools that make carefully planning a project and data storage important considerations. The software combined with data define how knowledge is or can be represented and impacts many decisions throughout the research described by this thesis.

Representations of Resistance

Resistance is a problem-oriented phenomenon. The decreasing effectiveness of antibiotics after their introduction prompted investigations into how and why antibiotic resistance seemed to grow, leading the identification of specific genetic determinants of resistance for many antibiotics. This led ultimately to speculation regarding the origin of resistance (Davies, 1990). These activities require methods to acquire and store molecular resistance data.

Quantifying resistance is an empirical activity. It is impossible to track whether individual molecules of an antibiotic are capable of inhibiting individual bacterial cells, so inhibition is commonly tracked by measuring the proxy of cell growth. Cellular growth depends on many other factors such as strain properties, inoculum density, medium, temperature, and pressure in addition to the presence or absence of antibiotics. When coupled with the ability of bacteria to sense and respond to stimuli, growth becomes a complicated function of these variables, sensitive to small changes. The antibiotic concentration at which no visible growth is observed is known as the minimum inhibitory concentration (MIC), typically measured over a series of concentration doublings. This MIC value is specific to the strain, inoculum concentration, and growth conditions and is considered accurate to within one doubling of concentration due to biological and technical noise. Because planktonic growth in liquid is different than colony formation on solid media, depending on the organism, the respective MIC values are typically different but correlated.

A similar empirical method is the disk diffusion assay. Paper disks are impregnated with an antibiotic compound and deposited on the surface of solid media inoculated with a culture of an indicator organism. After a period of incubation, the growth of the indicator strain around the disk is assessed with resistance measured as the radius of the zone of inhibition. In addition to the sources of variability above, the solubility of the compound can influence diffusion of the compound into the surrounding media and therefore the size of the inhibitory zone around the disk.

These methods, with their shortcomings, have been used to map out the relative MIC values for various strains to various antibiotics. A shift to a higher MIC implies that a change occurred in the genome of an organism correlating to the presence of a determinant of resistance, whether this change arose *de novo* or was acquired. The scale of these changes varies with the mechanism of resistance and can involve a single nucleotide polymorphism, or the loss or acquisition of one or more genes. The simplest way to verify the effect of a determinant of resistance is compare the MIC of a strains with an identical genetic background with and without the putative determinant. The early techniques of molecular biology allowed for correlating changes in MIC with the gain or loss of a plasmid harbouring a resistance determinant. Now it has become straightforward to deliberately engineer strains with resistance determinants expressed at one or more standardized levels using characterized expression systems at high copy number, low copy number or integrated into a chromosome (Cox et al., 2017). While these systems are effective tools to study resistance, a caveat is that they divorce the resistance determinant from its wild-type context which can obscure the understanding its role in nature. Great care must be used when designing and interpreting the results of comparative experiments.

Most characterized resistance determinants come in the form of genes and the proteins they encode. Since it is still difficult to consider the DNA context of any gene and understand if and under what conditions it is expressed, whether it is constitutively expressed or repressed—particularly for non-model organisms—the presence of a resistance determinant is only necessary but not sufficient to infer the capacity for

resistance. Comparative sequence analysis is used to identify putative resistance determinants. This depends entirely on prior knowledge of resistance determinants and procedures for performing comparison of a hypothetical set of putative sequences against that database. Because they were originally characterized one by one, over the course of decades during and prior to the genome era, the public nucleotide databases were the original repository for every sequence linked to a publication describing that sequence. Annotation was initially community driven, following on from the exemplar sequence usually labelled by its gene symbol. Different systems of nomenclature for each gene are in use, for example the Ambler classification of beta-lactamases (Classes A, B, C and D) (Hall & Barlow, 2005) and the aminoglycoside resistance determinants (mechanism, regiospecificity, compound profile, and ordinal number) (Ramirez & Tolmasky, 2010). In a few cases, the name of a gene is revised after antibiotic resistance activity is discovered, however the gene may still be labelled with its original name in the primary database in which it was originally deposited. Because antibiotic resistance research has been ongoing for five decades, these problems are becoming more common and are related to the phenomenon of annotation poisoning where an old, or erroneous annotation is used to annotate a new sequence thereby perpetuating the error.

Lists of resistance determinants are curated informally and formally. An example of an informal collection would be PFAM, a database of amino acid sequence profiles, where sequence families are annotated as being composed of, or containing, examples of resistance determinants. PFAM PF00144 beta-lactamase is a superfamily consisting of sequence profiles captured by the PFAM was initially associated with the sequences of

beta-lactamases. The superfamily includes many other sequences that do not possess beta-lactamase functionality yet do possess one or more of the sequence profiles. Another example is PFAM PF01636 APH, or aminoglycoside phosphotransferase, family. This family is named after resistance determinants but belongs to a larger clan of phosphotransferase sequences. It was initially closely associated with APHs, and related MPH or macrolide phosphotransferases, but has since grown to include other protein kinases that share the sequence profile of this family. Informal annotation efforts exist to organize sequences in an empirical *post hoc* manner. This is effective because the number of new sequences generated by the combined worldwide sequencing effort greatly outstrips the capacity to characterize these sequences. Annotation via 'guilt by association' in this case means that a sufficient level of similarity allows large numbers of uncharacterized sequences to coalesce around a few characterized sequences. An unintended consequence of guilt by association annotation is that sequences can become labelled by their association with a resistance determinant without being verified. If, in turn, the incorrectly labelled sequence is used to annotate a second uncharacterized sequence the incorrect annotation spreads beyond the first error. This is called annotation poisoning and is a known phenomenon (Radivojac et al., 2013; Richardson & Watson, 2013). This can happen with sequence search tools like the Basic Local Alignment Search Tool (BLAST) and variants, or with clustering procedures like USEARCH, and can happen with protein family grouping using models of families like PFAM (Altschul et al., 1997; Edgar, 2010; Finn et al., 2016).

Formal curation efforts follow a dedicated program of collecting sequences of resistance determinants through searching literature and sequence databases for records of protein and nucleotide sequences. Depending on the criteria being used, this can be very labour intensive. There are several examples of curated antibiotic resistance databases, most notably the Antibiotic Resistance Database (ARDB) and the Comprehensive Antibiotic Resistance Database (CARD) (B. Liu & Pop, 2009; McArthur et al., 2013). Databases may be distinguished based on scope, annotation methods, accompanying tools, how they are maintained and updated, and their stated goals (Table 1-1) and see recent reviews (Boolchandani, D'Souza, & Dantas, 2019; Xavier et al., 2016).

Table 1-1: Summary of antimicrobial resistance bioinformatics resources. (Updated November 2019)

Name	Full Name	Scope	Data Source	Features	Last Update	Reference
CARD	Comprehensive Antibiotic Resistance Database	general AMR	public sequence databases ardb Lahey-Clinic	curated BLASTp cut-offs, detection models, resistance gene identifier, antibiotic resistance ontology	2020	(Alcock et al., 2020)
ardb	Antibiotic Resistance Database	general AMR	public sequence databases	curated BLAST cut-offs	2009	(B. Liu & Pop, 2009)
ARG-ANNOT	Antibiotic Resistance Genes Annotation	general AMR	ardb, CARD, ResFinder, Lahey Clinic, public sequence databases	BLAST searching	2018	(Gupta et al., 2014)
Lahey Clinic		beta-lactams	public sequence databases, manual addition	sequence variants, resistance phenotype	2015	unpublished
BLDB	Beta-Lactamse DataBase	beta-lactams	PDB, public sequence databases	kinetics, protein structures, sequence variants, BLAST	2019	(Naas et al., 2017)
ARTS	Antibiotic Resistance Targets Seeker	AMR in BGCs	TIGRFAM core bacterial genes, CARD, LacED, Lahey Clinic, ResFams	Horizontally transferred determinants, duplicated housekeeping genes, known resistance associated with BGCs	2017	(Alanjary et al., 2017)
PARFuMs	Parallel Annotation and Re-assembly of Functional Metagenomic Selections	beta-lactams, aminoglycosides, amphenicols, sulfonamides, and tetracyclines	manual BLAST curation	functional detection of multidrug resistant gene cassettes	2015	(Forsberg et al., 2012)
NDARO	National Database of Antimicrobial Resistant Organisms	general AMR, resistant organisms	CARD, ResFinder, Lahey Clinic, Pasteur Institute Beta Lactamases	AMRfinder plus, HMMs, curated hierarchy of protein families, Antibiotic suceptibility data, genome browser	2020	(Feldgarden et al., 2019)
MEGARes	Metagenomic Antibiotic Resistance	metagenomic AMR, biocide and metal resistance	ResFinder, ARG-ANNOT, CARD, Lahey Clinic	AMR++ metagenome resistance identification pipeline, acyclic hierarchy of annotations	2020	(Doster et al., 2020)
RESfams	Resistance Families	general AMR	CARD, LacED, Lahey Clinic	HMMs	2015	(Gibson, Forsberg, & Dantas, 2015)
PATRIC	Pathosystems Resource Integration Center	ARM phenotype from genotype for four pathogens	resistant pathogen phenotypes, CARD, NDARO	Machine learning classifiers for several classes of drugs for four pathogens	2016	(Davis et al., 2016)
ARGs-OAP	Antibiotic Resistance Genes Online Analysis Platform	general AMR	CARD, ARDB, public sequence databases	SARG database, SARGfams HMMs, metagenomic profilig	2019	(Yin et al., 2018)
ResFinder	Resistance Finder	general AMR	Marilyn Roberts tetracycline resistance database, ARDB, public sequence databases	BLAST searches	2020	(Zankari et al., 2012)
RED-DB	Resistance Determinant Database	general AMR	Public sequence databases, manual curation	BLAST searches	N/A	unpublished
LacED	The Lactamase Engineering Database	beta-lactamases	Public sequence databases, PDB	sequence variants	2012	(Thai, Bos, & Pleiss, 2009)
DeepARG	Deep Learning Antibiotic Resistance Genes	general AMR	CARD, ARDB, UNIPROT	short sequence (DeepARG-SS) and long sequence (DeepARG-LS) detection models, ARGminer crowdsourced annotation inspection	2017	(Arango-Argoty et al., 2018)
MARA	Multiple Antibiotic Resistance Annotator	Gram-negative general AMR	public sequence databases, Lahey Clinic, ISfinder, Marilyn Roberts tetracycline resistance database, INTEGRALL, manual literature searches	Repository of antibiotic resistance cassettes (RAC), context based annotation, mobile element annotation	2017	(Partridge & Tsafnat, 2018)

Because antibiotic resistance can encompass variants of protein sequences that are otherwise functionally identical to sensitive versions, determinants of resistance can include nucleotide and amino acid variants. CARD is the only database that explicitly stores such variant information. It is important to note that these variants may be taxon-specific. In addition to the practical reality of it being easier to identify variants in closely related sequences, it also can be biologically unrealistic to expect to find taxon-specific resistance variants in distantly related species—for example, it might not make sense to find a *Mycobacterium* specific protein variant of the gene *RpoB* in Enterobacteria. CARD tracks such information, but this is complicated by the fact that variants are not protein or gene sequences in themselves and exist with respect to the wild-type sequence (Alcock et al., 2020; Jia et al., 2017).

Methods for Annotating Antibacterial Resistance

Purely bioinformatic methods for identifying and describing resistance relies primarily on sequence comparison. Search tools such as BLAST were designed to take a query sequence as input, either nucleotide or amino acid, and a database, and return records from the database that match the query sequence according to some filtering criteria. These criteria typically include percent identity over the length of the match, percent of coverage (either percent length of query covered by subject, percent subject covered by query, or both with the same or different cutoff values), alignment score, or expectation value (E-value) which is the expected number of sequences having an equal or greater score than the subject in a database of a given size and composition. The result of a BLAST search is a list of alignments to database records, termed High-scoring

Segment Pairs (HSPs), or colloquially ‘hits’, meeting the search criteria in ranked order of score, the assumption being that hits at the top of the result are more similar to the query than hits at the bottom of the result.

A procedure such as BLAST is useful because while DNA and protein sequences are straightforward objects that computers can easily manipulate, two sequences sharing a common ancestor are assumed to have been identical at some point and are theoretically expected to acquire differences relative to one another over time. These differences can include insertions and deletions of sequence resulting in the relative gain and loss of alignment gaps in addition to substitutions of one character in the string for another. Not all of these differences are equally significant, and so merely counting differences between two sequences leads to incorrectly scoring and ranking how similar they are. These underlying evolutionary assumptions are accounted for in the scoring system for the search procedure which determines how much matches are worth, and how gaps in the sequence alignment penalize the score for each hit.

Given a list of putative protein sequences, predicted in a bacterial genome for example, and a database of resistance determinants, it is straightforward to search each of these proteins against the database and return results where hits meet the search criteria according to the search procedure being used. Without any further information, each query sequence is evaluated against every subject in the database.

BLAST is not the only available search procedure. Profile Hidden Markov Models (pHMMs) are a type of probabilistic model trained on an alignment of related

sequences. The most widely used software implementation of these models is the HMMer package which provides utilities for training HMMs, aligning sequences against a profile, searching sequences against one or multiple HMMs and performing profile-profile matches of nucleotide and amino acid sequences (Eddy, 2011). The scoring system for evaluating expectation values of sequences aligned against profiles was shown to be nearly equivalent to the Lipman-Altschul statistics used to generate e-values via BLAST, meaning that the statistical underpinning of the distribution of scores for sequencing matching might be independent of the search procedure used (Eddy, 2009).

Profile HMMs consist of an abstracted ordered sequence of states, beginning with a start state and finishing with an end state. The sequence transitions through a series of match, insertion, or deletion states. Each column of a sequence alignment is mapped to a match state. Each match state has a probability of transitioning to the following match state, skipping the following match state by transitioning to a deletion state, or transitioning to an insertion state. Each match or insertion state is associated with probabilities of emitting a character from either a nucleotide or amino acid alphabet depending on what is being modelled. These emission probabilities are learned from the distribution of characters from the corresponding column of the alignment, while the emission probabilities of insertion states are learned from the overall character distribution of the alignment. Exiting from deletion states result in transitions to downstream matching states, so potential gaps in an alignment are modelled as skipping over match states in the alignment, so transitions to and from deletion states are learned by the frequency and position of gaps in the alignment. Overall, the profile HMM is an

abstraction of the alignment used to train the parameters of the HMM. A profile HMM is a generative model, meaning any sequence can be evaluated in terms of the probability of being generated by the model in the sense that every valid path through the states of a profile HMM is a product of a sequence of transition probabilities, and the observed characters of the sequence are product of the character emission probabilities of each state in the state path. The label 'hidden' refers to the fact that *a priori* the path through the states that best models any sequence is not known, however for any given model and sequence the Forward algorithm is capable of summing over all possible paths to determine the probability of the input sequence (Krogh, Brown, Mian, Sjolander, & Haussler, 1994). The hmm-build command from HMMer uses the information in the alignment to curate internal cutoffs that are subsequently embedded in the model, so sequences scoring higher than the 'trusted cutoff' field stored in the model may be considered a hit to the model. Similarly, the Viterbi algorithm can compute the most-probable state path for a given model-sequence pair which is useful for aligning a sequence to the profile (Krogh et al., 1994). Being probabilistic, local alignment probability for each character can be used to determine the goodness of fit of a sequence aligned to a model. These two features make pHMMs a valuable tool for annotation.

Provided that a user can supply an appropriate alignment of related sequences, the features in this alignment can be probabilistically abstracted into a representative profile model. Query sequences can be both scored against this model and aligned to this model. The downside to this approach is that there has to be enough information in terms of sequence number and diversity in each alignment in order to appropriately train each

model, otherwise the increased search and alignment sensitivity gained through a probabilistic approach will not perform better than simpler methods like BLAST (Altschul et al., 1997).

Profile HMMs are the central components of approaches like RESFams, which utilize the knowledge and data collected in other databases to produce HMMs that can be used for downstream annotation applications (Gibson et al., 2015). Attempts to validate the accuracy of RESFams predictions against sequence similarity searching with BLAST were aided by construction of a 'gold-standard' manually curated dataset as well as functional screening with a metagenomics approach (Gibson et al., 2015; Gibson et al., 2016). Because the manual gold-standard validation only used 18 antibiotics, only 54 of the 166 RESFams families were able to be validated, demonstrating high performance (Gibson et al., 2015). Sequences scoring above the trusted cutoff for each model can be annotated with the label associated with the model. This is approach is borrowed from other HMM databases like PFAM, where proteins are assigned into families and alignments by human curators, then HMMs are built for each family and used for downstream annotation (Finn et al., 2016). An alternative approach for curating family-specific information is to use curated score cutoffs for annotation, and form the model-centric curation efforts for antibiotic resistance databases like CARD.

Annotation is both incredibly important and notoriously difficult to automate. Data, particularly molecular sequence data, is produced faster than knowledge. The transfer of knowledge into data is one of the fundamental objectives of bioinformatics (Stevens, 2013). Much like the efforts of overall bacterial genome annotation, antibiotic

resistance annotation efforts are perpetually catching up with literature and is always biased towards what is currently known. Something that sets antibiotic resistance apart is the usefulness of the Antibiotic Resistance Ontology, a hierarchical controlled vocabulary that structures and classifies the knowledge of antibiotic resistance (McArthur et al., 2013). This enables a shared vocabulary, and a consistent set of annotation targets that encompasses nomenclature, publication, sequence and variant data into a common structure. True *de novo* resistance prediction may not ever be achievable; however, some progress is being made with machine learning (ML) models applied to growing datasets (Arango-Argoty et al., 2018; Davis et al., 2016; Rahman, Olm, Morowitz, & Banfield, 2018). Using the smaller world of resistance serves as a gentle introduction to annotation concepts and highlights the approaches and challenges currently in wide use.

Representations of Antibiotic Biosynthesis

The genetic structure of natural product biosynthesis took longer to be recognized than antimicrobial resistance (L. F. Wright & Hopwood, 1976). The best and most comprehensive introduction comes through the Minimum Information about a Biosynthetic Gene (MIBiG) cluster database, a project of the Genomics Standards Consortium and an attempt to standardize the output of several decades of research on BGCs (Medema et al., 2015). Painstaking biochemical characterization coupled with the elaboration of BGC sequences has resulted in a human curated set of 1727 whole or partial BGCs. Each BGC is linked to a chemical structure, a GenBank entry for the primary nucleotide sequence, and a reference publication though this publication may be

a reference for the sequence, the annotation of the cluster, the taxonomy of the producing organism, or a combination of one or more of these items.

Each cluster entry consists of the metadata available for that cluster, centered around the BGC nucleotide sequence. Since these sequences all refer to GenBank records, the BGC annotation data is overlaid on top of the data available for GenBank records. Each record is available for download in Javascript-Object Notation (JSON) format, or as a GenBank flat format file. GenBank taxonomic information, accession number, as well as predicted CDS features and annotations are available. MIBiG specific information is presented as additional features, and feature annotations including MIBiG version history, changelog, reference publication, name and structure and external database reference for compounds produced by this BGC, the BGC type for this record, and email address for the human curator.

BGCs are presented as an interval of a nucleotide sequence, either from a whole or partial genome, but possibly as specifically sequenced piece of DNA such as a cosmid. CDS features are predicted, though prediction pipelines usually do not record which software was used in the BGC itself. Beyond the DNA level, it is useful to think of a BGC as a collection of genes, or a collection of proteins encoded by those genes. At a lower level, individual genes may have functional domains annotated on them. This is particularly true for BGC types that rely on large multi-modular synthases such as polyketide synthases (PKS) and non-ribosomal peptide synthases (NRPS) (Fischbach & Walsh, 2006). These systems can be composed of multiple polypeptides each thousands of amino acids long consisting of functional modules each responsible for synthesizing a

single ketide or peptide unit, respectively, that is incorporated into the resulting product. Each module within these proteins consists of functional domains that act in concert to synthesize and join in assembly-line fashion, sometimes shuttling the partially synthesized natural product between multiple polypeptides during its biosynthesis. Beyond those two specific examples, BGCs are classified by the types of enzymes they possess, and it is possible for more than one type of enzyme to be present in hybrid clusters.

The borders of these clusters were initially determined by the judgement of the researchers who generated and analyzed each sequence. Most of these BGCs are not fully characterized in their expression or in terms of conditions under which the genes are expressed, and their biosynthetic potential in terms of the number of molecular species including precursors, side-products, and whole or partial products synthesized by the BGC, or under which growth conditions these products are produced. In general, these details are not entirely predicted solely from the BGC sequence and this is essentially the modern challenge of identifying and annotating these sequences for downstream applications.

Methods for identifying Biosynthetic Gene Clusters

Characterizing putative BGCs from DNA sequence along is a type of specialized annotation. This characterization makes use of the techniques mentioned previously: sequence comparison, compilation of databases and searching, but also rule-based approaches, machine learning, and cluster analysis.

The identification and annotation of BGCs followed the identification and annotation of the various different kinds of BGC components as they became recognized. Table 1-2 describes a non-exhaustive list of some core sequences used to identify a selection of BGC types (Medema et al., 2011). A brief discussion of these tools as they relate to modular NRPS and PKS BGCs serves to illustrate this process. A key development in the analysis of the modular biosynthesis of NRPSs was the first crystal structure of the phenylalanine-adenylating domain of gramicidin S synthase 1 bound to phenylalanine and adenosine monophosphate (AMP) (Osterlund, Nookaew, Bordel, & Nielsen, 2013). It was recognized that the physical chemical properties of the amino acid binding site of each A-domain determined which amino acid was activated an incorporated by each NPRS module. The motif of these sites could be read as a specificity-conferring code, in conjunction with the co-linearity rule, allowed the prediction of the complete peptide produced by the modules of an NRPS BGC (Osterlund et al., 2013). A similar analogy, using phylogenetic sequence clusters instead of protein crystal structures, was applied to modular polyketide synthases to predict the identify of each acyl-Co-enzyme A monomer, and along with the presence of various redox domains (enoyl-reductase, dehydratase domains), the oxidation state of each ketide unit synthesized by each module (Haydock et al., 1995; Khosla, Gokhale, Jacobsen, & Cane, 1999).

Table 1-2: A selection of core sequences used to identify BGCs.

Abbreviation	Name	Description	BGC type
A	adenylation	activation and loading of amino acids onto cognate PCP	NRPS
C	condensation	peptide bond formation between adjacent PCP-bound amino acids	NRPS
PCP	peptidyl carrier protein	phosphopantithienylated carrier protein	NRPS
TE	thioesterase	cleavage and/or cyclization of completed scaffold	NRPS, PKS
ACP	acyl carrier protein	phosphopantihienylated carrier protein	PKS
KS	ketosynthase	claisen condensation of adjacent ACP-bound CoA monomers	PKS
AT	acyltransferase	activation and loading of acyl-CoA monomers onto cognate ACP	PKS
ER	enoyl-reductase	ketide reduction	PKS
DH	dehydratase	ketide dehydration	PKS
	terpene or lycopene synthase	N- and C-term domains, terpene biosynthesis	Terpene
	terpene or lycopene cyclase	many classes of cyclase enzymes for terpenoid molecules	Terpene
	beta-lactam or clavulanic acid synthetase	formation of clavam ring	Beta-lactam
	lantibiotic synthase	formation of lanthionine ring in immature scaffold	Lantibiotic
	lantibiotic dehydratase	dehydration of the immature lantibiotic	Lantibiotic
	lantibiotic peptide	the immature ribosomally encoded lantibiotic peptide, may have a leader peptide	Lantibiotic
	bacteriocin peptide	immature ribosomally encoded peptide scaffold	Bacteriocin
	SpcD/SpcK-like thymidyl transferase	activation of sugar subunits	Aminoglycoside
	aminoglycoside glycosyltransferase	addition of sugar subunits to aminoglycoside precursor	Aminoglycoside
	siderophore synthase	formation of aerobactin-like siderophore from N^6-acetyl-N^6-hydroxylysine subunits	Siderophore
	ectoine synthase	hydro-lyase catalyzing the cyclization of ectoine	Ectoine
	AfsA-like butyrolactone synthase	produces the starting linear substrate for butyrolactone formation	Butyrolactone
	StaD-like chromopyrrolic acid synthase	formation of the dichromopyrrolic acid precursor substrate	Indole
	LipM-like nucleotidyltransferase	formation of the nucleoside precursor scaffold	Nucleoside
	MelC-like melanin synthase	Bacterial tyrosinase, oxidation of L-Trp to form melanin	Melanin

All-purpose BGC identification pipelines began by combining smaller, more focused tools into larger pipelines, typified by the antiSMASH software, a redevelopment of the CLUSEAN project, and the PRISM software (Medema et al., 2011; Skinnider, Merwin, Johnston, & Magarvey, 2017; Weber et al., 2009). These are rule-based approaches that depend on identifying specific features, along with a series of rules that then categorize putative BGCs with into different classes. The original rules used by antiSMASH are included in Medema *et al* supplementary table II (Medema et al., 2011) and generally consist of identifying one or more core domains (via profile matching using HMMs, curated from the PFAM database and primary BGC publications) that meet or exceed threshold scores for positive matches and do not meet or exceed threshold scores for negative matches. Cases where a putative BGC sequence meets one or more rulesets for different BGC types are considered hybrid types. To solve the problem of localizing the borders of each putative BGC, the authors use a proximity cutoff of 15kbp upstream and downstream from an identified core domain. The PRISM software follows a similar rule-based approach to identifying BGCs, however the antiSMASH pipeline has become the most widely used tool to identify natural product BGCs, including variants for identifying BGCs in fungi (fungiSMASH) and plants (plantiSMASH) (Kautsar, Suarez Duran, Blin, Osbourn, & Medema, 2017). As of September 2019, the website http://antismash.secondarymetabolites.org provides a web interface for running the pipeline and has tallied over 577,000 jobs processed by the latest antiSMASH version 5 software (Blin et al., 2019).

Several additions to the basic antiSMASH annotation have been included in the software by the original antiSMASH authors and others. Following the creation of a curated set of known BGC sequences, a sequence analysis and clustering step was applied to the entire set of amino acid sequences to build the secondary metabolite clusters of orthologous groups (smCOGS). The 301 smCOG families are divided into four categories, 'transporter', 'regulator', 'biosynthetic_smcog' and 'other', which are mainly used to colour-code the various CDSs of a BGC that are not considered 'core' components of a BGC in the antiSMASH output. A list of the available smCOGs is available hard-coded into the antiSMASH database schema (https://github.com/antismash/db-schema/blob/master/smcogs.sql, accessed Sept. 2019). The particular smCOG type, for example "SMCOG1045 glycosyl transferase group 1", provides a basic functional annotation for a CDS but is not very specific since this SMCOG does not identify which sugar moiety is being transferred (substrate specificity), how the substrate is being activated (nucleoside diphosphate linked, or polyprenyl phosphate linked), where it is being transferred on the scaffold molecule (regiospecificity) and how the sugar is to be linked (O-, N-, or C- glycosylation). Each smCOG is used to generate an alignment and a corresponding profile HMM that is then included with an antiSMASH analysis along with the option to perform a full PFAM analysis. The smCOGS annotation acts as a subset of the PFAM annotation, so in principle it may be possible to map each smCOG onto a parent PFAM but this has not been formally attempted. In practice, while the smCOGS were derived only from sequences present in BGCs while PFAMs are derived from whole-genome sequences, the

PFAMs might be more useful since they are equally non-specific when it comes to putative function prediction, but are able to annotate a greater proportion of protein sequences found in BGCs.

PFAM annotations were used in the ClusterFinder model to define putative BGCs based on their composition (Cimermancic et al., 2014). ClusterFinder is a simple two-state HMM consisting of the 'BGC' and 'non-BGC' state. These states have emission probability matrices that are trained on the sequence of PFAM annotations of a set of curated BGCs, and once trained is able to identify BGCs in a whole genome sequence by exploiting the difference in the frequency of PFAM domains in BGCs versus the background frequencies of a bacterial genome overall. Because PFAM domain composition does not entirely classify BGC sequences, ClusterFinder has a recognized propensity to produce false positives which may outweigh its ability to identify BGCs lacking a well-defined rule for identification (Baltz, 2018).

The curated set of BGCs was formalized into the MiBIG project which attempted to establish a standard annotation procedure for BGC sequences (Medema et al., 2015). The most significant result of this project is a human-curated database of BGC sequences linked to known products and stored in the MiBIG repository. Because only a few BGCs have been extensively characterized, the level of annotation for the approximately 1800 entries of this repository varies considerably. More positively, MiBIG serves as a database which new BGCs can be compared to see if a known BGC has been rediscovered, provided a procedure exists which can compare a query BGC against a subject BGC.

ClusterBLAST is such a procedure, based on the MultiGeneBlast tool (Medema, Takano, & Breitling, 2013). ClusterBLAST queries a set of input protein sequences against a subject of BGC sequences where the scores are weighted depending on their classification in the BGC. For example, core protein sequences are weighted higher than other sequences. In practice, ClusterBLAST is not useful for comparing the large multi-modular sequences found in BGCs, but can be helpful in quickly visualizing similarity of the smaller, single-domain sequences shared among one or more BGCs. This ecosystem of bioinformatics tools is rich with possibility. The antiSMASH pipeline, across its various iterations, provides a more or less consistent and stable base from which to build new tools that can answer new questions regarding natural product biosynthesis in general, and antibiotics in particular.

RESEARCH GOALS

The background information in this introduction frames research questions in this thesis. The history of antibiotics and resistance is described alternately by serendipity and inevitability. After facile discovery of the 'low hanging fruit' antibiotics was exhausted during the golden age of discovery, investigators turned to increasingly exotic discovery techniques, hopefully equating exotic sources and methods with exotic results. Confronted with the growing problem of antibiotic resistance, traditional Waksman-like screening that relied on vast numbers was soon abandoned by the pharmaceutical industry *en masse*. Synthetic chemistry was unsuccessful in its attempt to replace natural products as a source of new compounds with new activities but was marginally useful to modify the properties of known antibiotics.

The genome era presents new challenges for both natural product discovery and the study of antibiotic resistance. I take the point of view that evolution is the linking concept between these two phenomena, and accordingly have chosen to focus on origins rather than outcomes to address the problem with antibiotics and resistance. The general view of antibiotic resistance is that it is a response to the anthropogenic use of antibiotics, but it has been well established that antibiotic resistance widespread and predates the modern use of antibiotics (Bhullar et al., 2012; D'Costa et al., 2011). The label 'pre-antibiotic' era is actually a misnomer, since it is believed bacteria have been producing antibiotics long before humans discovered and re-purposed them. Few lines of investigation into the age of antibiotics have been followed. Recognizing that the human use of antibiotics has resulted in selection that has changed the frequency and distribution of antibiotic resistance of some bacteria, namely pathogens, raises the question of what the background frequency and distribution of resistance determinants are in bacteria globally. Several lines of reasoning suggest that antibiotics are also ancient features of the microbial lifestyle.

My first goal is to harness the flood of genomic data that can now be cheaply and easily produced by even small laboratories. This entails the production of new tools, and application of existing tools to enable the conversion of genomic data into new biological knowledge. For natural product biosynthesis, this means developing ways of organizing genomes and the BGC data produced by ourselves and others and move beyond counting to categorization and comparison. I describe features implemented in the python module evoc (evolution of clusters) and demonstrate its ability to compare and categorize

streptothricin-producing Actinobacteria by their biosynthetic potential. Streptomycin is an aminoglycoside natural product reported to be commonly found when screening extracts of *Streptomyces* species. Currently, software makes it possible to identify if a particular strain harbours a putative streptomycin or related BGC. No method exists that can characterize and compare the remaining BGC complement or assess how similar the biosynthetic potential of two strains are. This is important because it motivates prospective silencing of streptomycin (or other common antibiotic) BGCs to reveal the production of other, potentially rarer, natural products obscured by streptomycin activity in extracts. My work highlights the distribution of streptomycin BGCs in *Streptomyces* genomes, and in the wider context of a larger number of bacteria taxa.

Next, my research focused on the evolution of a family of BGCs that encode glycopeptide antibiotics (GPAs). These clinically important compounds were once considered antibiotics of last resort for serious Gram-positive infections. GPAs are an attractive model because their biosynthesis is well-studied, consisting of the synthesis of rigid peptide scaffold incorporating both proteinogenic and non-proteinogenic amino acids, a diverse set of tailoring reactions producing a wide assorting of final structures, and a well-characterized mechanism of resistance that is shared between producers and pathogens. I have mined the genome sequences of an in-house strain collection and public databases to generate a comprehensive list of known and uncharacterized potential GPA producers and investigate the evolutionary history of the biosynthesis of these important compounds. I consider the different components of these biosynthetic gene clusters and leverage a phylogenetic technique of reconciliation to estimate the times and

locations within a dated GPA producer genome phylogeny to understand the evolutionary history of these antibiotics. I also show that the canonical resistance operon, consisting of the *vanH*, *vanA*, and *vanX* genes can be reconciled to the time that the GPA clusters appeared, suggesting that possibility that there is a dual relationship between resistance and antibiotic biosynthesis.

Lastly, investigating the more diverse BGCs of the GPA and GPA-like producer strains, I have used my phylogenetic analysis to predict where novel members of this larger family may be discovered. Looking at the components of these clusters and what is known about the wider family suggests there may be natural products with different mechanisms of action to be mined in these strains. This culminates in the discovery of corbomycin, a molecule that shares an evolutionary history and features of GPAs but appears to act via a new mechanism on the Gram-positive cell wall.

CHAPTER TWO: Representing and comparing biosynthetic gene clusters

CHAPTER TWO PREFACE

Portions of the work presented in this chapter have been published as:

Culp, E, Yim G, Waglechner N, Wang W, Pawlowski AC, Wright GD. (2019) Hidden antibiotics in actinomycetes can be identified by inactivation of gene clusters for common antibiotics. *Nature Biotechnology* **37(10)**:1149-54.

Author Contributions

EC, GY and GDW conceived of the broader study, and designed experiments. NW conceived of the experiments in this chapter. GY initially identified the list of streptothricin producers. NW assembled whole genome sequences and performed phylogenetic and BGC content analysis, and wrote this chapter. EC, GY and GDW wrote the published manuscript, see Appendix 1.

Acknowledgements

Computational resources for genome assembly and analysis were provided by Andrew G. McArthur at McMaster University. This research was funded by a Canadian Institutes of Health Research grant (MT-14981), the Ontario Research Fund, and by a Canada Research Chair (to G.D.W.). E.C. was supported by a CIHR Vanier Canada Graduate Scholarship. G.Y. was supported by a M.G. DeGroote Fellowship Award and a CIHR postdoctoral fellowship. N.W. was supported by a CIHR Canada Graduate Scholarship Doctoral Award. We thank Christy Groves for graphical edits to figures.

ABSTRACT

Streptomyces, specifically, and bacteria in the phylum Actinobacteria in general, are among the most prolific producers of natural products, including antibiotics. A traditional Waksman-like antibiotic screening platform will reject a strain once any antimicrobial activity derived from this strain is believed to come from a known molecule. The ability to judge the biosynthetic capacity of a strain based on its genome sequence is valuable as a way to direct the construction and screening of a strain library. We implemented the evoc software as a fine-grained approach to represent and group the biosynthetic gene clusters of streptothricin- and streptomycin producing bacteria and show that streptothricin production is distributed among *Streptomyces* and that the capacity to produce common natural products does not predict the remaining potential of a *Streptomyces* sp. This approach provides a way to consider the rareness of particular natural products and grounds a biotechnological approach using CRISPR/Cas9 engineering that facilitates the identification of rare products from these strains.

INTRODUCTION

Sequencing bacterial genomes is now a common starting point for natural product discovery, alongside more traditional activity-based screening (Adamek et al., 2018; Adamek, Alanjary, & Ziemert, 2019; Nadine; Ziemert, Weber, & Medema, 2019). An ecosystem of tools has been developed to identify, annotate and classify biosynthetic gene clusters (BGCs) in these genome sequences with the goal of facilitating novel natural product discovery, particularly antibiotics (Weber & Kim, 2016). Initial surveys of the diversity of BGCs among sequenced genomes used a variety of similarity measures to map BGC space (Doroghazi et al., 2014). Within the genus *Salinispora*, MultiGeneBlast (a tool that processes BLAST results for multiple sequences into a single score) was used to explore BGC diversity in way that that is intuitive and familiar to many researchers (N. Ziemert et al., 2014). While pairwise similarity measures have been shown to be useful, few tools exist to evaluate the biosynthetic potential of strains, where analysis is mainly restricted to the BGC level. For example, MultiGeneBlast was developed into ClusterBlast, and integrated into the antiSMASH pipeline as a weighting scheme for BLAST results that performs BGC vs BGC comparisons (Cimermancic et al., 2014), MiBIG is a repository that stores curated BGCs, BiGSCAPE generates BGC comparison networks and CORASON visualizes dendrograms of these comparisons (Medema et al., 2015; Navarro-Muñoz et al., 2018). There is no general consensus on the best way to compare BGCs with one another, and only *ad hoc* ways to assess diversity of BGCs in a set of strains. As most BGCs are uncharacterized, finding novel natural products is not difficult since novel clusters abound, consistent with the conclusions of

analyses of BGC marker sequences that sampling of BGCs has not been saturated (Charlop-Powers, Owen, Reddy, Ternei, & Brady, 2014; Charlop-Powers et al., 2015; Crits-Christoph, Diamond, Butterfield, Thomas, & Banfield, 2018). The knowledge of the general distribution of BGCs is not currently exploited to streamline the screening process, or to specifically target unusual or rare biosynthetic potential. Novelty is not the major criteria for which to characterize BGCs. Rather, once a target BGC is chosen- perhaps because its product was a hit in a screen, for example a comparison is made with both known and uncharacterized BGCs residing in databases like MiBIG and antiSMASHdb to retroactively establish biosynthetic novelty (Blin, Medema, Kottmann, Lee, & Weber, 2017). Firsthand measures of BGC diversity can be used to evaluate the diversity in a set of organisms, such as a culture collection or a set of genomes, and can drive downstream applications such as strain engineering or development of novelty-enriched extract libraries for use in standard forward-screens of compound activity.

Bacteria of the genus *Streptomyces* are especially prolific with respect to BGC diversity, however it is known that observed diversity is not equally distributed among strains (Baltz, 2006; Cimermancic et al., 2014; Kieser et al., 2000; Lewis, 2012). Streptothricin is one of the most common antibiotics found in extracts of *Streptomyces,* and a determinant of resistance to this antibiotic is known (Baltz, 2006; Cox et al., 2017). Screens of natural product extracts with antimicrobial activity from environmentally derived strains historically suffer from the ‘dereplication problem’ as the same compounds are rediscovered repeatedly, such as streptothricin (G. D. Wright, 2014). This occurs a direct result of the largely unknown structure of BGC diversity in bacteria in

conjunction with the ability to bring only a limited number of environmental strains into laboratory culture.

It is unclear how to reliably access the novel chemical matter that surveys of biosynthetic environmental DNA predict is present (Charlop-Powers et al., 2015; Charlop-Powers et al., 2016). One strategy involves counter-screening extracts from these strains against a panel of organisms harbouring resistance determinants to common antibiotics, reserving those extracts retaining activity for costly and time-consuming characterization (Cox et al., 2017).

It has been suggested that taxonomically diverse organisms will produce diverse natural products. Phylogenetic analysis of natural product producers might be a way to access diverse natural products. McDonald and Currie selected a set of sequenced *Streptomyces* to produce a robust phylogeny that represents the overall diversity in the genus (McDonald & Currie, 2017). In this work, we combine this set with the identified streptothricin producers in our strain collection to explore the taxonomic relationship and correlation of streptothricin production in the wider background of the genus *Streptomyces*. To accomplish this, we develop software to facilitate this exploration using the base annotation provided by the ubiquitous antiSMASH pipeline that identifies BGC sequences in microbial genomes (Blin, Wolf, et al., 2017). An optional component of this annotation includes a comparison with known BGC sequences in MIBiG, however antiSMASH does not provide a simple way to compare sequences with each other. We address by implementing the python 3 module evoc, which post-processes antiSMASH

annotations and provides a platform fine-grained expert annotation and comparison of BGCs.

METHODS

Strain Selection

Two sets of genome sequences were identified in the work of (Culp et al., 2019) (See Appendix 1). One set was identified as putative streptothricin producers from our in-house strain collection by virtue being positive hits against a counter screen of streptothricin resistant reporter strains. Others were identified as putative producers if they possessed a streptothricin BGC (MIBiG accession BGC0000432) after BLAST and manual examination of their genome sequence (See Appendix 1, Methods). A larger set of 122 strains was previously published (McDonald & Currie, 2017), which represents a selection of organisms sampled mainly from the genus *Streptomyces*, and other organisms in the phylum Actinobacteria. These strains were selected to provide a robust comparison of organisms both within and without the genus *Streptomyces*.

Genome Sequencing of Wildtype Streptothricin Producing Strains

Genome sequences for these strains were produced for (Culp et al., 2019) (see Appendix 1). Strains were grown in TSB at 30°C, 250 rpm to midlog phase, pelleted for lysis with standard lysozyme, proteinase K and SDS treatment, followed by phenol/chloroform cleanup and ethanol precipitation or column purification. Illumina MiSeq sequencing (300 bp, paired end reads) was performed by the Farncombe Genomics Facility (McMaster University). Raw reads were assembled following QC trimming and filtering using Skewer 0.2.2 (Jiang, Lei, Ding, & Zhu, 2014) with "-q 30 -Q 30" options, followed by read merging with Flash v1.2.11 (Magoc & Salzberg, 2011) using "-M 300" parameter to produce files with paired unmerged reads, and a file with

unpaired merged reads. These data were assembled using SPAdes v3.12 (Bankevich et al., 2012) using default parameters.

Phylogenomic Analysis

The hidden Markov models (HMMs) corresponding to every family listed under TIGRFAM genome property 0799 'bacterial core gene set, exactly 1 per genome' (https://genome-properties.jcvi.org/cgi-bin/Listing.cgi, accessed Sept. 12th 2018) were collected. HMMER3 was used to analyze every genome using each model's trusted cutoff (Eddy, 2011). The top hits for each model were retained and aligned as a group against the model HMM. If a genome lacked a hit for a model, gaps equal to the length of the missing sequence were added to the alignment for that genome. These aligned model families were subsequently concatenated into an overall alignment which was inspected manually. This alignment was used for phylogenetic analysis using fasttree2 using the WAG substitution model and otherwise default parameters (Price, Dehal, & Arkin, 2010).

Identification of BGCs and Assessment of diversity

The entire set of streptothricin BGC containing genome sequences were subjected to analysis by antiSMASH v4 using the '--smcogs' '--knownclusterblast' and '--full-hmmer' in addition to default options (Blin, Wolf, et al., 2017). The genomes derived from public databases and McDonald and Currie already had coding sequence (CDS) features predicted, while the genomes produced from our strain library had CDSs predicted using Prodigal, the default tool provided with antiSMASH (Hyatt et al., 2010).

Each BGC in each genome was extracted from the overall genome annotation and examined individually by a custom Python script using the BioPython library (Cock et al., 2009). Translated coding sequences (CDSs) in each cluster were extracted whole if they contained zero or one 'aSdomain' features. However, if they contained two or more 'aSdomain' features the coding sequence was divided at the borders of each 'aSdomain' and extracted separately, including the sequences, if any, at the beginning and end of each CDS and between labelled domains. The entire set of extracted CDS sequences were then output to a fasta file and subjected to clustering using USEARCH v8.1.181 (Edgar, 2010) using the 'cluster_fast' mode with a 60% identity cutoff and a minimum 75% length filter. After clustering, each translated CDS and CDS fragment was labelled with the cluster number it was assigned, so each BGC can be described as a sequence of labelled translated CDS fragments.

We compared the overall biosynthetic capacity pairwise between each genome by implementing the Jaccard Index component of the Lin index (Cimermancic et al., 2014; Lin, Zhu, & Zhang, 2006) using the unique elements (genes/domains) of the set of combined genome BGC CDS fragment labels from the whole genome. In addition to the Jaccard Index component, the Goodman-Kruskal gamma function was implemented as described by (Lin et al., 2006). This function, as previously implemented, counts the set of paired domains, and reverse paired domains between two proteins but we apply it to the representation of a BGC as a sequence of domains. The difference in these counts is scaled to the interval [-1, 1] by adding 1 and dividing by the two times sum total of forward and reverse pairs. The possible values indicate 100% set coverage in reverse

order, and 100% set coverage in forward order, in other words having the same domains occurring in the same order either forward or reverse.

Development of the evoc Software

To streamline this analysis, the steps of post-processing genomic antiSMASH annotations was combined into a python 3 module called evoc, which stores all the genome annotations and processed sequences in a single sqlite3 database file. HMM-based core gene phylogenomics was incorporated into evoc as the HMMERprint module, packaged along with the TIGRFAM hmm profiles used for this analysis. The ability to provide a custom set of hmm profiles was included, in case a user wishes to use search, extract, and produce a concatenated amino acid sequence phylogeny for custom markers. We designed this package to be suitable as a platform for further sequence analysis of BGCs, coupled to a phylogenomic analysis of the strains from which these BGCs are derived.

RESULTS

Phylogeny of Streptothricin Producing Strains

To determine the potential chemical diversity available in streptothricin producers chosen by Culp and Yim *et al*, we performed a phylogenetic analysis to examine the distribution of BGCs present in known streptothricin producers (Culp et al., 2019) (see Appendix 1). TIGRFAM genome property 0799 defines a set of 111 core bacterial genes existing as exactly one copy in at least 95% of bacterial genomes. A robust phylogeny is necessary to establish any patterns observed for some trait in a set of related organisms. In this case, we want to determine if the genetic potential to produce streptothricin in *Streptomyces* depends on how closely related the producers are in the species tree. The associated HMM for each bacterial single-copy TIGRFAM gene was used to both search and produce a profile alignment for each family. We chose concatenated multi-locus sequence analysis of 156 *Streptomyces* strains, including 11 genomes from streptothricin producers from our in-house strain collection to represent diversity across the genus and find streptothricin production is distributed across this phylogenetic tree (Fig 2-1). We included an additional 40 bacterial genomes as outgroups to *Streptomyces* for a total of 197 strains in a final alignment consisting of 44012 amino acid sites.

The current best molecular phylogeny established for the genus *Streptomyces* defined an overall structure of sister monophyletic Clade I and Clade II strains, as well as an outgroup of strains that do not appear monophyletic. Our phylogeny is consistent with the general structure of the tree produced by McDonald and Currie (McDonald & Currie, 2017). Streptothricin biosynthesis appears to be a rare trait in the outgroup *Streptomyces*

(labelled Other Lineages, Fig 2-1) while it is associated with specific groups of taxa in Clades I and II.

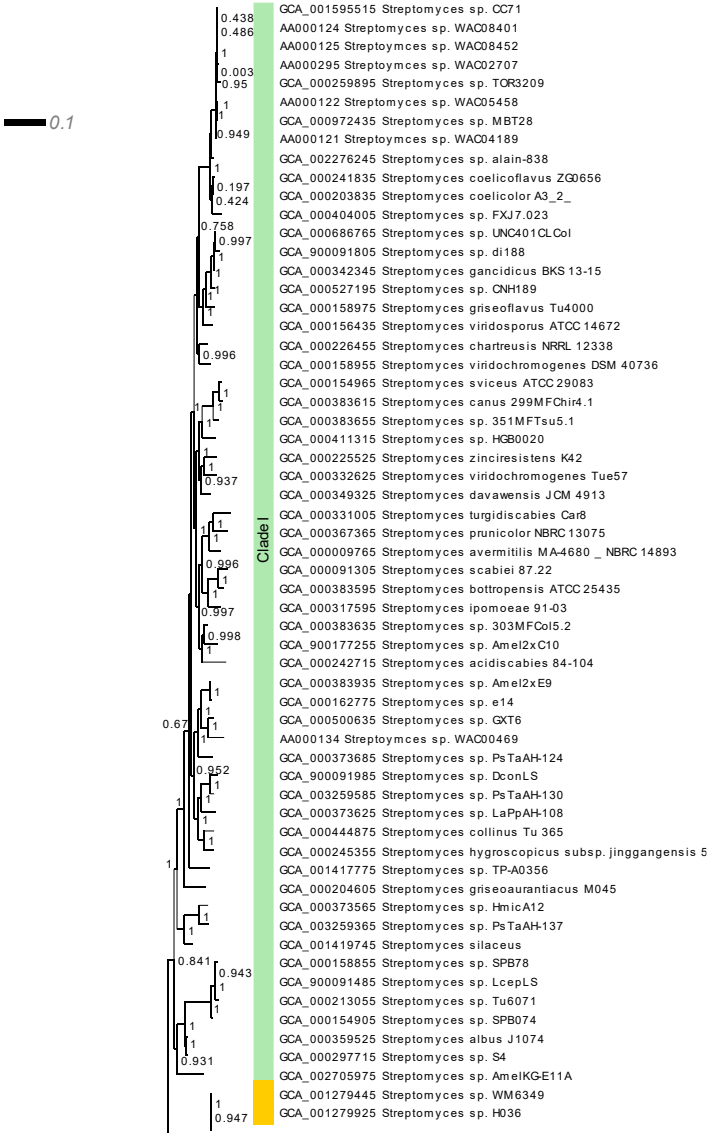


Figure 2-1: Maximum-likelihood *Streptomyces* phylogeny. (Continued).

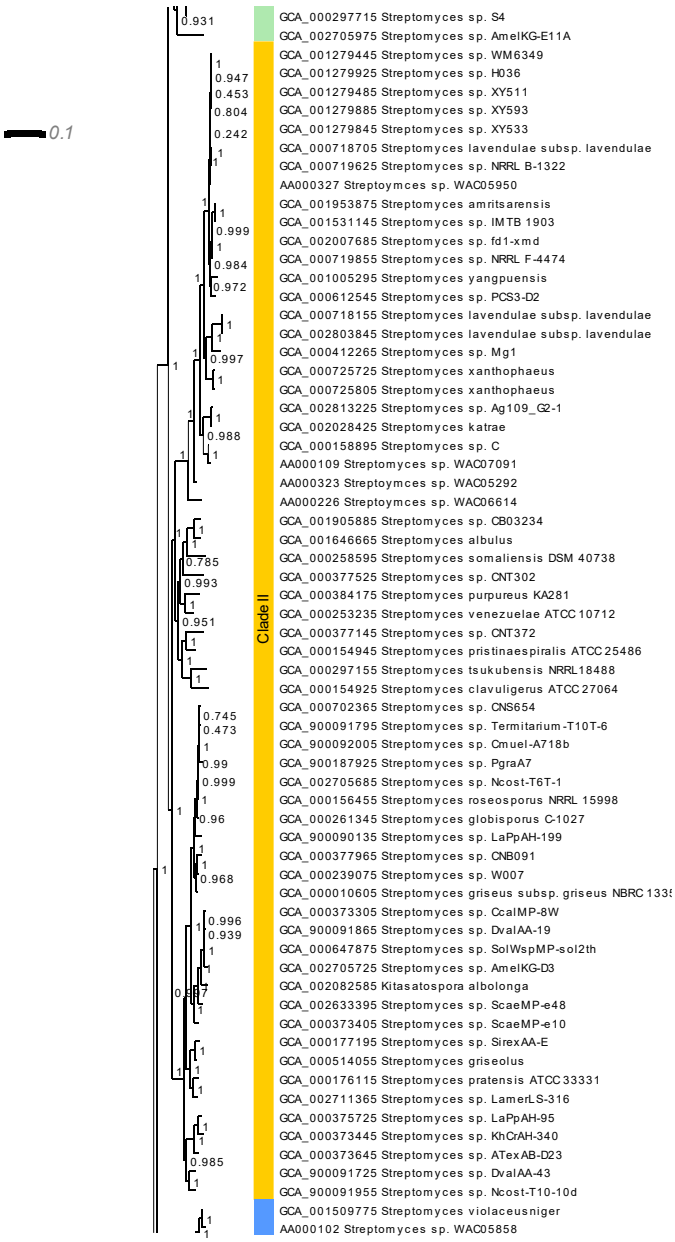


Figure 2-1:Maximum-likelihod *Streptomyces* phylogeny. (Continued)

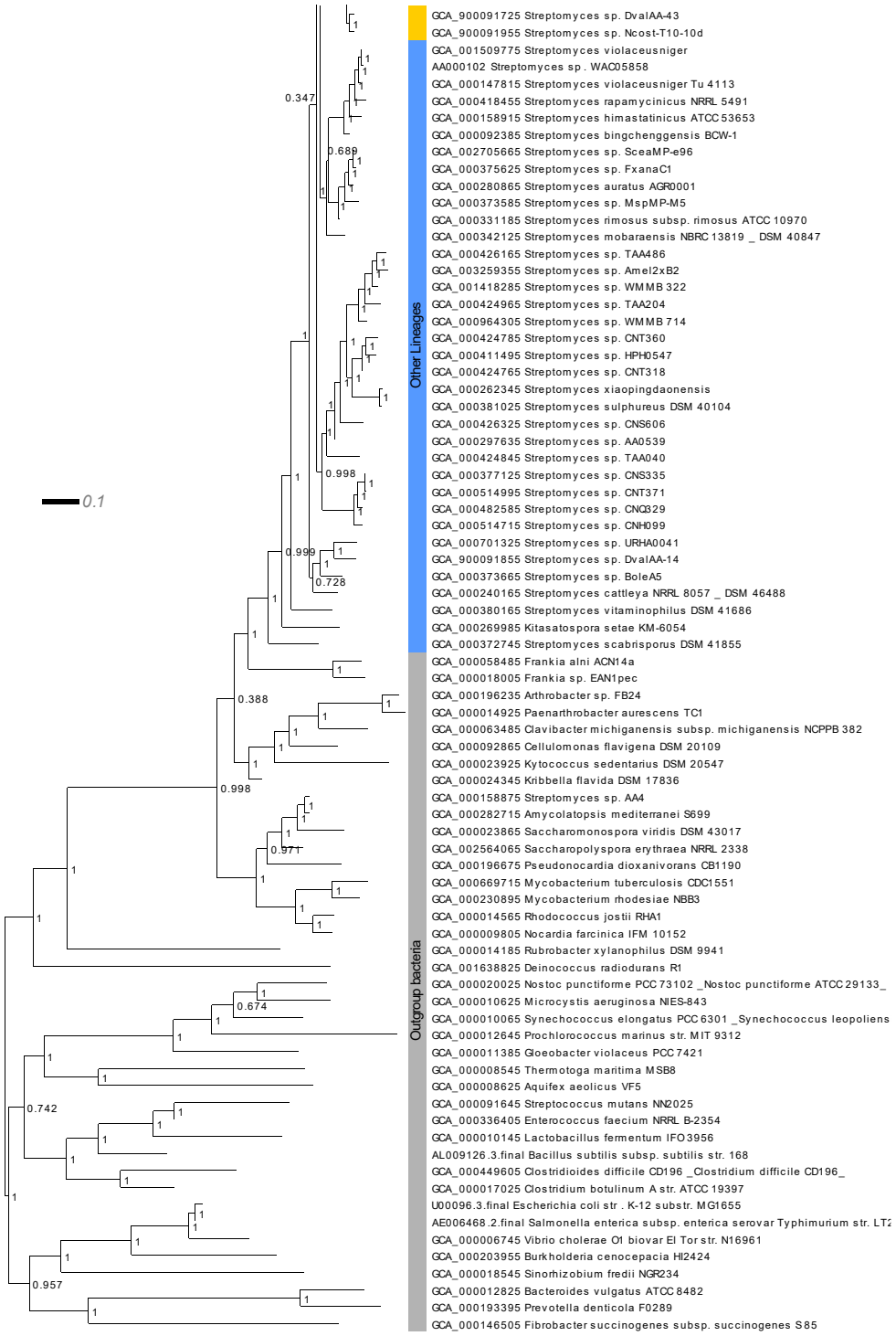


Figure 2-1:Maximum-likelihod *Streptomyces* phylogeny. Core, single copy genes defined in the TIGRFAM 0799 genome property chosen from 196 bacterial genomes

were aligned concatenated, and analyzed under the WAG model with the CAT approximation in FastTree 2 (see methods). The overall structure of the genus *Streptomyces* consists of an outgroup of other lineages (blue), and two monophyletic clades, Clade I (green) and Clade II (gold).

The evoc Software and Representation of BGCs

The output of antiSMASH software consists of annotated Genbank flat files along with an HTML visualization. BGCs are annotated separately for visualization without preserving the original genomic coordinates, but are conveniently also provided as in single genome file containing CDS features with optional genome-wide annotations (the '--full-hmmer' option) (Blin, Wolf, et al., 2017). Each strain is processed separately in a typical antiSMASH installation which does not facilitate comparison of strains or BGCs without reference to a central database. This is an ongoing area of development for other groups resulting in an expansion of the BGC tool ecosystem (Blin, Medema, et al., 2017; Hadjithomas et al., 2015). The evoc software was developed to organize and facilitate strain vs strain and BGC vs BGC comparisons starting from common antiSMASH version 4 annotation. BioPython is used to parse each flat file, track taxonomy where available from each genome record, extract and organize sequences and annotations (Cock et al., 2009). This information was stored and indexed in a single relational database file having a schema depicted in Figure 2-2. The organizing principle of this database is stored in the type and type_relationship tables, intended to be hierarchical and ontology-like, modeled after the cvterm and cvterm_relationship table in the chado schema and the original version of CARD (McArthur et al., 2013; Mungall, Emmert, & FlyBase, 2007).

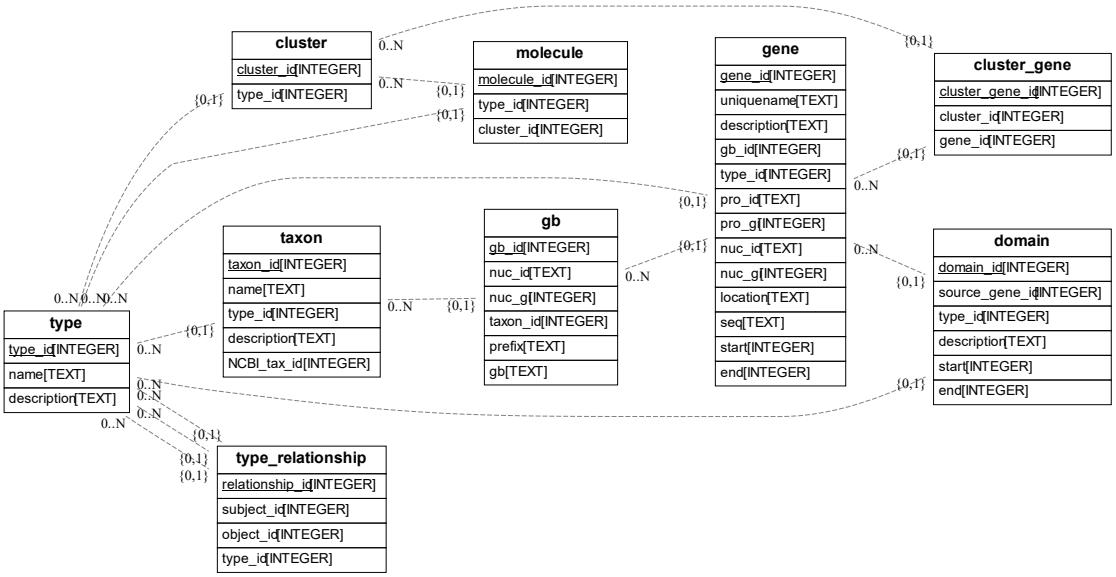


Figure 2-2: Entity relation diagram of the evoc schema. The type and type_relationships are central to every record stored in evoc except those in the gb table. Taxon provides a field for storing an NCBI taxonomy id, where available, and is used to provide a taxon type for each gb record which contains the BGC annotation. Gene records are derived from these gb records, and domain records are derived from each gene record. The cluster_gene table links specific genes to specific BGCs, since gb records can contain more than just BGC entries. Entries in the taxon, cluster, molecule, gene, and domain can be considered children of top-level entries in the type table. This schema was derived from https://bitbucket.org/waglecn/evoc commit 36239b637.

The links from the gene, domain, taxon, cluster and molecule tables to the type table indicate that the annotation of each of these elements may be stored as a type entry with few constraints, for example no duplicate names are allowed, while the relationships between these types are stored in the type_relationship tables. For simplicity, the only type used currently is the ‘is_a’ relationship, though more basic and complex relationships such as ‘part_of’ and ‘participates_in’ may be modelled if necessary (Smith et al., 2005). As a result, there are top-level entities modelled in the type table, in the sense that they have no parent entities, for each of the described tables (Tables 2-1, 2-2).

When loading a genome annotation possessing an assigned NCBI taxonomy identifier, each taxonomy term is added via ‘is_a’ relationships to an existing parent term, which should lead to precise interoperability with the NCBI taxonomy. In the same way, the precise definitions for which sequences, and therefore clusters, match antiSMASH identification rules are documented via types and relationships in these tables. Where no such annotation is available, catch-all terms such as ‘unknown_gene’ and ‘unknown_domain’ may be used, however once the usearch clustering step is applied to group sequences together in families by similarity, these families may be represented as hierarchical groups of ‘unknown_gene’ types, which can potentially simplify downstream annotation. Without anything but the most basic antiSMASH annotation, given specific clustering evoc makes it possible to organize components in a label-free manner, meaning it should be possible to compare BGCs detected by antiSMASH without knowing anything else about these features.

Table 2-1: Basic types loaded on type table creation.

type name	description
none	nothing
parent	a parent
child	a child
is_a	a relationship
organism	an organism
unidentified_microorganism	an unknown organism (NCBITax 81726)
cluster	a cluster
unknown_cluster	an unknown cluster
gene	a gene
unknown_gene	an unknown gene
domain	a domain
unknown_domain	an unknown domain
molecule	a product of a cluster
unknown_molecule	an unknown molecule

Table 2-2: Basic relationships loaded during type_relationship table creation.

subject	relationship	object
child	is_a	parent
unidentified_microorganism	is_a	organism
unknown_cluster	is_a	cluster
unknown_gene	is_a	gene
unknown_domain	is_a	domain
unknown_molecule	is_a	molecule

A limitation of using a defined set of annotations to label the parts of a BGC comes from the limited number of profile models available for annotation. Neither the smCOGs nor the PFAMs used by antiSMASH are able to fully annotate every CDS in a BGC. They do offer a finite set of possible families, which eases the creation of probabilistic models such as ClusterFinder where the emission probability vector for the BGC and non-BGC states has a fixed dimension (Cimermancic et al., 2014). We consider

a BGC to be composed of a sequence of translated CDSs. The antiSMASH rules for detecting a BGC uses profile matches to a small number of fixed domains characteristic of each BGC type, therefore each CDS will be annotated with 0, 1 or more of these basic 'aSDomain' types. We decomposed each CDS into 1 or more subsequences, which we refer to as a general 'domain', using the combined start and end coordinates of its CDS and any internal 'aSDomain' features. For example, a CDS with a single 'aSDomain' type annotated as a sub-sequence of its CDS will be decomposed into the following three intervals: [cds_start, aSDomain_start -1], [aSDomain_start, aSDomain_end], and [aSDomain_end + 1, cds_end]. Only 'domains' with nonzero lengths are retained. This converts a BGC into a sequence of 'domains', where each 'domain' either represents a single CDS, or a subsequence of a CDS (Fig 2-3).

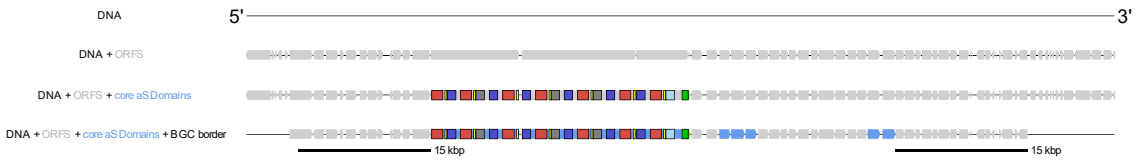


Figure 2-3: Composition and annotation of a BGC. Proceeding in steps, first CDS are predicted from open reading frames (ORFs, grey). Each ORF is scanned to detect core antiSMASH domains (aSDomains). The presence of these domains are compared with detection rules to infer the type of BGC and core ORFs are labelled (blue). The furthest upstream and downstream core ORF plus 15 kbp are used to define the cluster border.

This permits us to the use nearly all amino acid sequence information in a BGC for comparison rather than just the subset that match the predefined PFAM or smCOG profiles included in standard antiSMASH annotation. This also prevents us from assuming, *a priori*, how many different possible 'domains' can be observed in any set of BGCs.

Measures of Strain-wise and BGC-wise Similarity

The post-processing antiSMASH annotation of 42 streptothricin producers motivates a technique for evaluation of their biosynthetic potential. Culp and Yim *et al* showed these strains to have the capacity to produce a wide variety of specialized metabolites that rarely overlap in identity (Culp et al., 2019) (see Appendix 1). After each BGC in each genome is decomposed into ordered sets of 'domains', these domains are subjected to a cluster analysis using the greedy uclust algorithm implemented in usearch (see methods). Each 'domain' will belong to a single family (these could also be described as clusters, but for clarity we use the term family to distinguish between these clusters of 'domains' and BGCs). Each family consists of one or more members which is represented by a centroid member, that is, a representative member of the entire sequence family such that every other family's centroid sequence has a degree of similarity below the clustering threshold, while every family member has a measure of similarity to the family centroid above the clustering threshold (Edgar, 2010). Once assigned into unique families, 'domain' types in the evoc sqlite3 database are organized hierarchically using the type_relationship table. Each BGC may be represented as a set of domain types, and the Jaccard index as described by Lin *et al* for comparing two BGCs is trivially implemented using these domain sets (Lin et al., 2006).

The union of the BGC domain sets from an entire genome may then be used as a measure of the biosynthetic potential for that genome and compared in the same way as for individual BGCs. This is shown for each of the streptothricin producers in Fig 2-4. When compared to the phylogeny of these producers, it becomes possible to see how

clades close to each other in the tree are more likely to share more of their biosynthetic potential than clades more distantly related in the tree. Distantly related streptothricin producers appear to not share the majority of their biosynthetic potential.

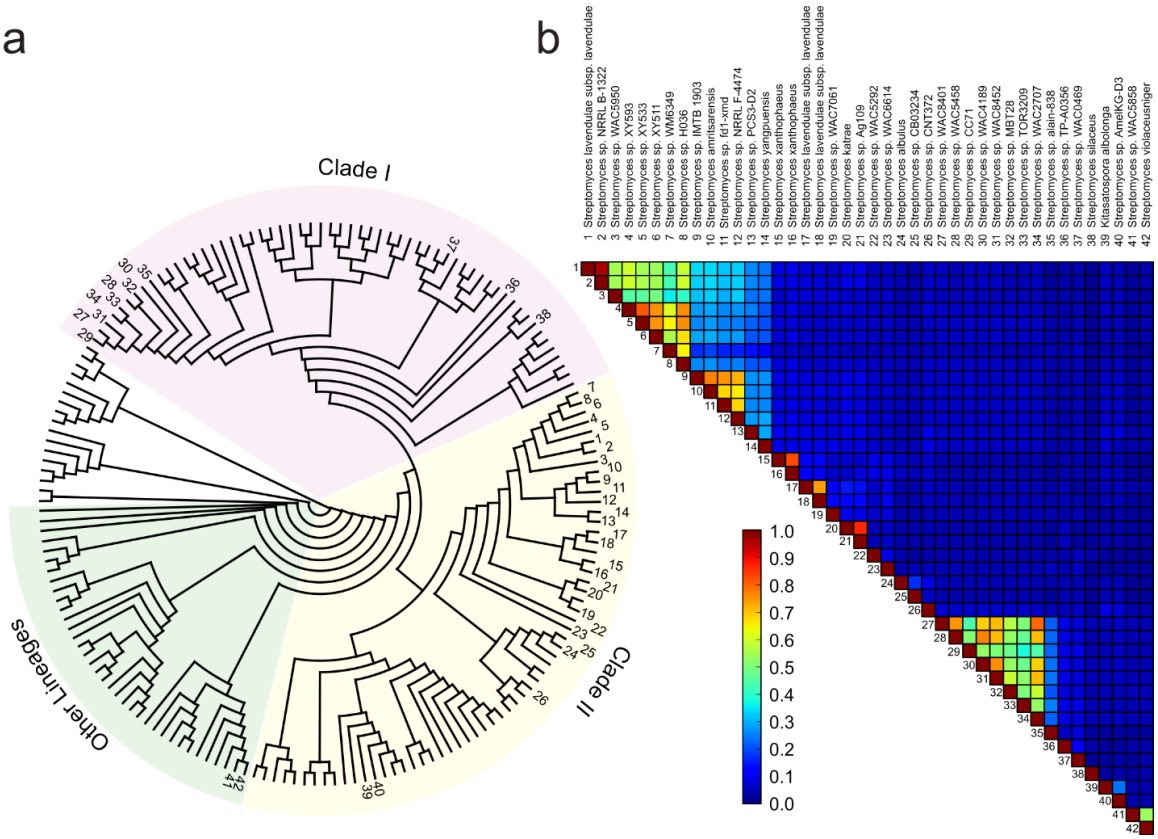


Figure 2-4: Pairwise genome vs. genome sum BGC distance in 42 streptothricin producers. a The bacterial core-gene phylogeny from Figure 2-1 represented as a circular cladogram. The streptothricin producing strains are indicated with numbers, as well as their position in the genus *Streptomyces* clades as organized by McDonald and Currie. **b** The Jaccard index applied to the genome-wise union of BGC sequences in streptothricin producers shows that shared streptothricin production between all but the closest neighbours on the tree does not predict other biosynthetic capacity.

When the MRCA distance of two leaves from the phylogeny are compared with the genome-vs-genome sum BGC Jaccard distance, less than half of biosynthetic potential is lost after two units on the tree. Alternatively, 808 out of 922 or 87.6% of species pairs share less than half biosynthetic potential in this set of organisms (Figure 2-5).

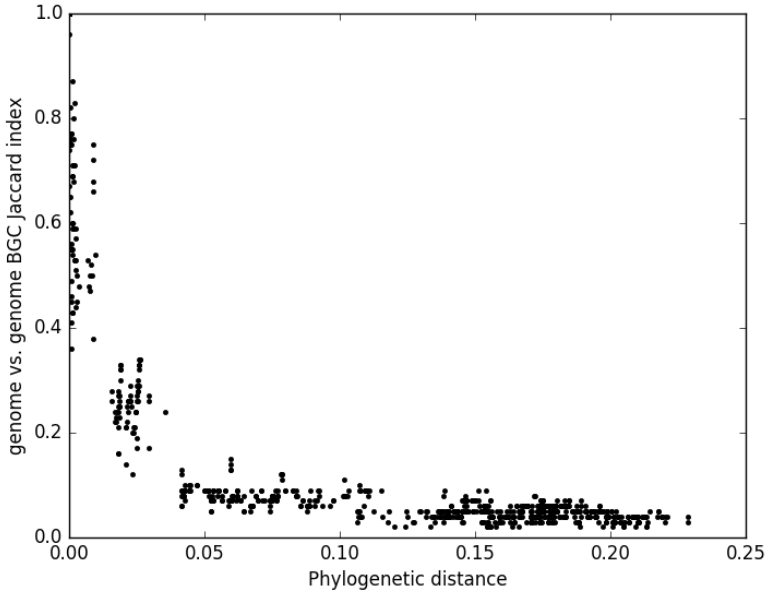


Figure 2-5: Pairwise genome vs. genome sum BGC distance versus phylogenetic distance. The shared biosynthetic content between genome pairs rapidly drops below 50% within 0.02 expected substitutions per site for strains with the streptothricin BGC on the phylogeny depicted in Figure 2-1.

In contrast, the Goodman-Kruskal γ index provides a poor measure of genome-vs-genome BGC similarity. Figure 2-6 shows the same comparison among strains using this measure. This may be attributed to the representation of BGCs as they are predicted from draft genomes, where incomplete contigs may produce multiple partial BGC predictions across two or more contigs which is likely to occur when genomes sequences are produced using short read sequencing technology like Illumina MiSeq or HiSeq. This is known to occur for large and/or highly repetitive cluster types, like non-ribosomal polypeptide synthase and polyketide synthase type BGCs, which require additional techniques to reconstruct (Meleshko et al., 2019). The implementation of Lin *et al.* was intended to compare two protein sequences consisting of a series of domains (Lin et al.,

2006), but protein sequences are intrinsically ordered from N-terminus to C-terminus, so ordered pairs of domains derived from such a protein are meaningful. Two BGCs may be equivalent even though their component genes and domains derived from these gene sequences may occur in a rearranged order. Similarly, two draft genome assemblies may be equivalent even if they possess randomly reverse-complemented contig sequences, however the GK-γ index applied to the evoc-processed BGC annotations from these genomes will produce scores randomly multiplied by -1 or 1 depending on the relative orientation of each BGC sequences.

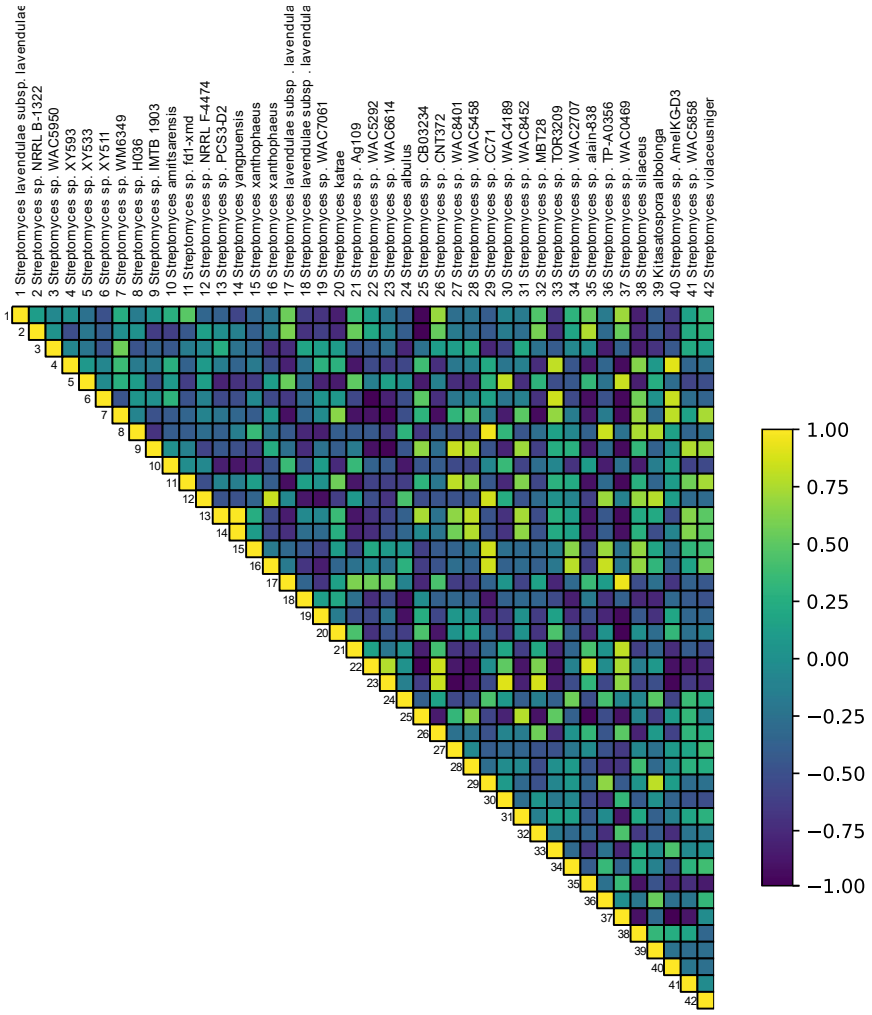


Figure 2-6: Pairwise genome vs genome BGC sum Goodman-Kruskal gamma index. Unlike the Jaccard index, this measure relies on pairs of domains that may be randomly in the forward or reverse orientation due to orientation of sequences in the genome assembly. For genomes that share less overall BGC content, it is not informative about the genome in general if that shared fraction lies in the same orientation or not.

The representation of BGCs used by evoc bins different components by sequence similarity into families. In the BGC vs BGC case, an ordered sequence of domains for each BGC is converted into a set, which is an unordered datatype in Python. The Jaccard index in a sense compares the capacity of these domains, and is more like a ‘bag of

sequences' that encodes the production a respective natural product and is more robust to the types of draft sequences used here than the GK-γ index.

DISCUSSION

High throughput microbial sequencing has enabled even small laboratories to quickly produce many complete and draft genomes, often with the purpose of studying a specific set of genes relating to a particular function. It is unusual for annotation in public databases to achieve the same level of accuracy for these sequences that a lab with specific expertise can produce, however lack of significant bioinformatics resources can result in the use of *ad hoc* and unreproducible methods. Spreadsheets and unorganized computer files used to represent this expertise goes against ideal practices, and makes it difficult to share back to the community (Hao et al., 2016).

Evoc is a package of command-line tools that aids the organization, annotation and phylogenomic analysis of conserved genomic regions. Starting with a preliminary set of records in common formats, evoc can help users perform iterations of extracting, organizing, phylogenetic analysis, visualization and annotation. evoc can potentially help subject-matter experts to rapidly curate minimally annotated sequences into a single file suitable for downstream applications such as biochemical investigation, ontology building and phylogenomics.

The subdivision of protein sequences into smaller fine-grained components naturally defines a hierarchy of families. By using sequence similarity and length of alignment as the clustering metric it is expected that domains will be organized into alignable families. These families, aided by the way they are stored by evoc, can be quickly converted to multiple sequence alignments for downstream analysis. The

structure of these clusters is used to produce a loose hierarchy, even without manual curation, which may be considered as a rudimentary ontology.

The clustered domain family alignments may be inspected as part of a larger curation effort, to be used by subject experts to iteratively refine the domain boundaries to produce more biologically relevant subdivisions of CDS sequences. This can be used to rapidly investigate CDS sequences that may be composed of previously unrecognized functional domains (Stogios et al., 2016; Truman et al., 2009). Additionally, the family alignments may be used to quickly produce profile models for additional searching to obtain new sequences, contributing to the iterative building of a larger, more accurate, dataset.

Annotations of the 'domain' families may be propagated back to the original records to rapidly classify and organize the components of the input 'clusters'. Instead of annotating the individual members of each cluster, representative sequences may be used to easily annotate the family as a whole. This may be especially useful when only few members of the family have been biochemically characterized.

Refined high-quality 'domain' alignments may be used to produce phylogenetic trees that represent the evolution of the components of these 'clusters'. A framework like phylogenetic reconciliation, or supertree phylogenetic methods can be used to study the evolution of these component trees (Jacox, Chauve, Szöllősi, Ponty, & Scornavacca, 2016; Stolzer, Siewert, Lai, Xu, & Durand, 2015). Even at the 'gene'-level, traditional phylogenetic techniques cannot easily be applied to multi-domain CDS sequences like

large non-ribosomal polypeptide megasynthases because each functional domain in these proteins might have a different history (Page & Charleston, 1997). The history of an entire cluster of genes, and of the multi-domain gene, might include duplication, horizontal transfer and loss events that can only be identified by studying the evolution of the aligned sub-sequences of these objects. The evoc framework is designed towards organizing and facilitating this type of analysis.

The structure of the type and type_relationship table is especially useful for rapid ontology development. A user can view the alignments and phylogenetic trees produced by these tools and quickly determine if the structure of the rudimentary ontology agrees with the sequence annotations, and correct annotations or the structure when there is disagreement. By providing rapid turnaround and directly linking to sequence data, an ontology that is used to organize and classify sequences can become a living document that reflects up-to-date understanding of a collection of sequences, as has been successfully used elsewhere (McArthur et al., 2013).

The application to streptothricin production in *Streptomyces* illustrates other uses of evoc. Blind extract screening, relying on large numbers of strains, has been the traditional source for natural product discovery. Substantial genetic capacity for natural product biosynthesis has been documented for several industrially important Genera of bacteria (Doroghazi et al., 2014). Large databases of genome and predicted BGC sequences are now easily and routinely constructed, and need to be rationalized against the phenomena of rediscovery of common natural products in these screens. Effective measures of strain vs strain biosynthetic potential are needed, as well as methods for performing BGC vs

BGC searches that can leverage these strains and sequence collections. This knowledge directly translates into greater understanding of how BGC diversity is distributed among bacteria. The evaluation of BGC potential and phylogenetic analysis provides a strong rational motivation for investigating the antimicrobial extracts of streptothricin CRISPR-inactivated strains that would otherwise be discarded in a traditional screening approach (Brown & Wright, 2016).

CHAPTER THREE: Phylogenetic reconciliation reveals the natural history of glycopeptide antibiotic biosynthesis and resistance

CHAPTER THREE PREFACE

This work was previously published as:

Waglechner N, McArthur AG, Wright GD. (2019) Phylogenetic reconciliation reveals the natural history of glycopeptide antibiotic biosynthesis and resistance. *Nature Microbiology* **4(11)**:1862-71.

Author Contributions

NW and GDW conceived the project. NW, GDW and AGM designed experiments. NW performed the analysis. NW, AGM and GDW wrote the manuscript. NW wrote this chapter.

Acknowledgements

Christy Groves provided valuable input on the figures. This research was funded by the Canadian Institutes of Health Research (MT-14981) and by a Canada Research Chair (to G.D.W.). N.W. was supported by a Canadian Institutes of Health Research graduate scholarship. A.G.M. holds a Cisco Research Chair in Bioinformatics, supported by Cisco Systems Canada, Inc. Some computer resources were provided by the McMaster Service Lab and Repository computing cluster, funded in part by grants from the Canadian Foundation for Innovation (34531 to A.G.M.)

ABSTRACT

Glycopeptide antibiotics are produced by Actinobacteria through biosynthetic gene clusters that include genes supporting their regulation, synthesis, export, and resistance. The chemical and biosynthetic diversity of glycopeptides are the product of an intricate evolutionary history. Extracting this history from genome sequences is difficult since conservation of the individual components of these gene clusters is variable and each component can have a different trajectory. We show that glycopeptide biosynthesis and resistance maps to approximately 150-400 million years ago in Actinobacteria. Phylogenetic reconciliation reveals that the precursors of glycopeptide biosynthesis are far older than other components implying that these clusters arose from a pre-existing pool of genes. We find that resistance appeared contemporaneously with biosynthetic genes, raising the possibility that the mechanism of action of glycopeptides was a driver of diversification in these gene clusters. Our results put antibiotic biosynthesis and resistance into an evolutionary context and can guide future discovery of compounds possessing new mechanisms of action. Furthermore, this work emphasizes the uniqueness of antibiotics as ancient natural wonders, coevolved with resistance, that are imperiled by human activity.

INTRODUCTION

The availability of microbial genome sequences is transforming the study of antibiotic biosynthesis and resistance, information that can be leveraged both in understanding and predicting resistance emergence and in the discovery of new antibiotics (Truman et al., 2014). Various informatic tools (Skinnider et al., 2017; Weber et al., 2015) facilitate the mining and discovery of biosynthetic gene clusters (BGCs), the sets of genes from genome sequences that encode all the necessary elements for production of natural products, including antibiotics. Only a few studies have explored BGCs as co-evolved units (Cruz-Morales et al., 2016). Here we focus on the BGCs that encode the glycopeptide antibiotic (GPA) class to demonstrate the utility of a phylogenetic reconciliation approach in understanding the origins and evolution of these units and their products.

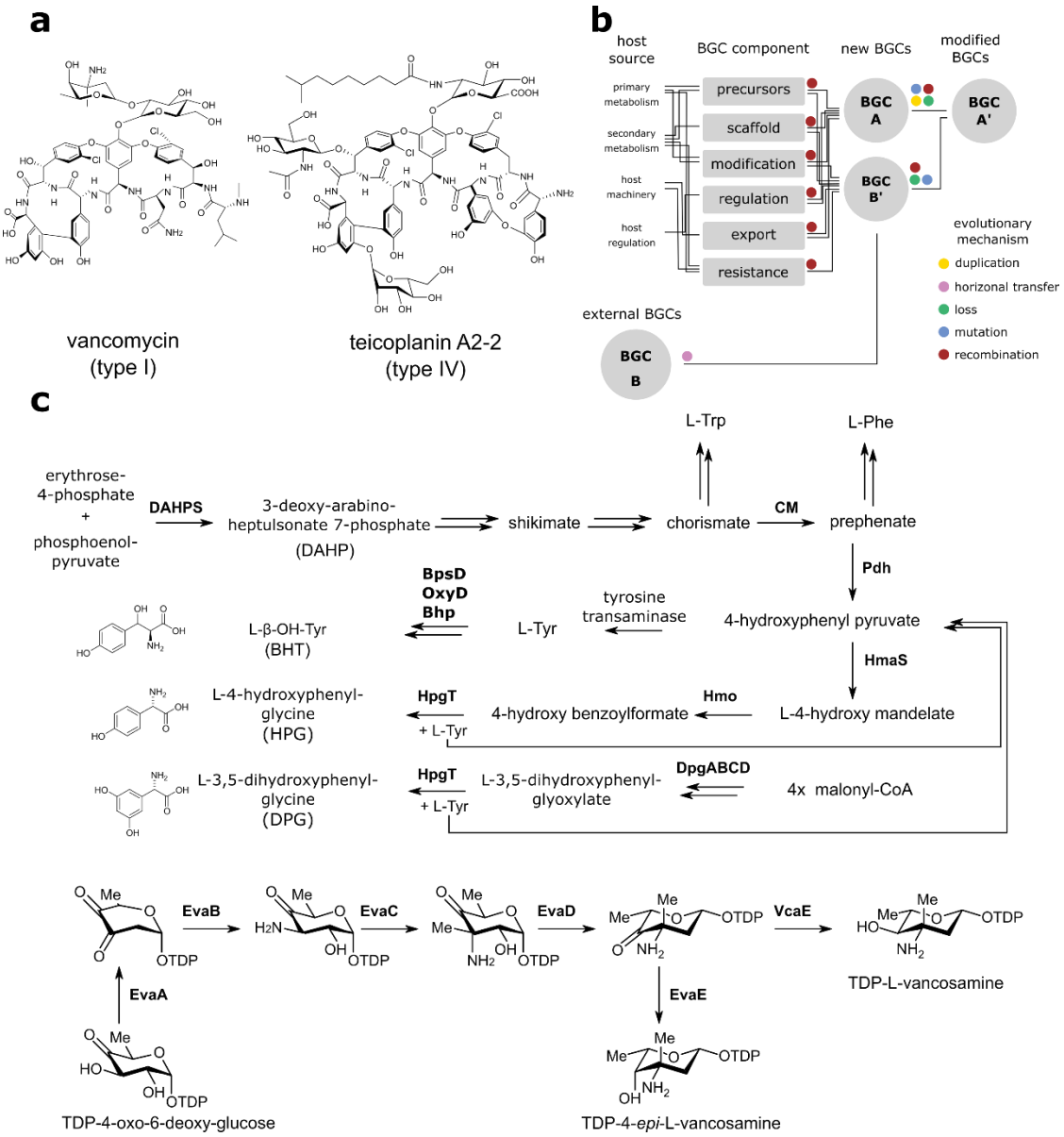


Figure 3-1: GPAs, BGC evolution and precursor biosynthesis. a Structure of vancomycin and teicoplanin A2-2. **b** BGC components are a pool of sequences derived from various aspects of host biology. Different combinations of these components make up various BGCs like BGC A. Whole or partial BGCs originating outside the host, such as BGC B, can be acquired via horizontal transfer events and become part of the host repertoire like BGC B'. The host repertoire is subject to evolutionary mechanisms like duplication, HGT, loss, mutation and recombination that can lead to BGCs like A'. **c** GPA Precursor biosynthetic pathways. While HPG originates from the shikimate pathway, DPG is synthesized from malonyl co-enzyme A units. The final reaction in both

HPG and DPG biosynthesis shares a common aminotransferase which regenerates an intermediate precursor for more HPG biosynthesis. The starting substrate for vancosamine biosynthesis is a thymidine diphosphate (TDP) activated glucose. The specific combination of enzymes present in each BGC result in different end products like the two epimers of vancosamine depicted. DAHPS – 3-deoxy-arabino-heptulosonate 7-phosphate synthase, CM – chorismate mutase.

GPAs such as vancomycin and teicoplanin (the two archetypal peptide scaffolds of this class) (Figure 3-1a, Supplementary Figure 3-1) are WHO essential medicines, used for the treatment of infections caused by Gram-positive bacteria. By leveraging the sequence and chemical diversity of GPA biosynthesis we plot a natural history that reveals the logic of GPA BGC assembly over time. As a general principle, the assembly of BGCs includes the co-opting or divergent evolution of genes required for the production of precursor components (amino acids, acyl-CoAs, sugars, etc.), scaffold assembly (e.g. peptide or polyketide synthetases), tailoring enzymes (halogenation, glycosylation, acylation, etc.), transport, regulation, and resistance in the case of antibiotics (Figure 3-1b). The importance of GPAs, the number of GPA BGC sequences, and the association between clinical and environmental resistance through a common mechanism of bacterial cell wall biosynthesis (Marshall, Broadhead, Leskiw, & Wright, 1997) make these antibiotics an excellent model for evolutionary studies.

GPAs share several notable features beginning with a non-ribosomally synthesized peptide scaffold (Nicolaou, Boddy, Brase, & Winssinger, 1999), assembled via non-ribosomal peptide synthetases (NRPSs). These NRPSs incorporate the constituent standard (Asn, Leu) and non-proteinogenic amino (dihydroxyphenylglycine (DPG), hydroxyphenylglycine (HPG), and β-hydroxytyrosine (BHT)) in a modular assembly-line fashion. The genes required to synthesize these amino acids (Chen, Tseng, Hubbard, &

Walsh, 2001) and the specialized aminosugars are also present in BGCs (Figure 3-1c). During synthesis, peptide scaffolds undergo intra-molecular cyclization by oxidative cross-linking and are further modified during and after release from the NRPS machinery by enzymatic tailoring reactions including N-, O-, and C-methylation, acylation, glycosylation, sulfation, and halogenation all encoded by clustered genes (Supplementary Figure 3-2) (Yim, Thaker, Koteva, & Wright, 2014). Lastly, genes responsible for GPA resistance along with regulatory and export mechanisms are required (Kilian, Frasch, Kulik, Wohlleben, & Stegmann, 2016; Lo Grasso et al., 2015). With few exceptions, these features are not universally conserved in all GPA BGCs, making their roles in the origin and evolution of GPAs challenging to interpret in isolation, but a natural history of GPAs can be reconstructed through comparison of several BGCs. Since the components of BCGs are not universally conserved across producers (which leads to chemical diversity in the class), the homology relationships between duplicated components are not obvious and obtaining robust estimates of rate parameters for each of these sequences is difficult. We compare the phylogeny of each gene or domain component separately with a dated species tree using reconciliation to identify milestones in the evolution of GPAs (Supplementary Figure 3-3). Rather than an ancestral reconstruction, using this species tree, we can infer which nodes representing taxonomy and time were most important during the evolution of GPAs (Chevrette & Currie, 2018).

METHODS

Growth of organisms, DNA sequencing and assembly

Strains from our in-house collection were grown in Tryptone Soy Broth (BD Biosciences, TSB) for 3 to 7 days at 30°C shaking at 250 RPM. For short read sequencing, pelleted cells resuspended in 180 mM monobasic sodium phosphate, pH 8, with GES (0.51 M guanidine thiocyanate, 10 mM EDTA, 3.4 mM N-lauroyl sarcosine) were subjected to mechanical lysis with 0.1 mm glass beads followed by treatment with lysozyme (5 mg/mL) and RNase A (0.1 mg/mL) at 37°C for 1 hour. A second enzymatic lysis was performed by addition of 0.625% SDS, Proteinase K (0.5 mg/mL) and 0.3 M NaCl with gentle mixing and incubation at 65°C. Cell lysate was pelleted by centrifugation, and the supernatant removed and extracted with an equal volume of freshly prepared 25:24:1 phenol-chloroform-isoamyl alcohol. Following centrifugation, the organic layer was removed and combined with 200 uL of DNA binding buffer (Thermo Scientific) then bound to GeneJET spin columns (Thermo Scientific). Bound samples were washed twice with wash buffer (Thermo Scientific), dried, then DNA was eluted with warmed 10mM TRIS-HCl buffer. For long read sequencing, high molecular weight DNA was extracted from pelleted cells grown to mid-log phase treated with lysozyme (30 mg/mL) followed by lysis with Proteinase K (5 mg/mL) and 1% SDS. Cell lysate was extracted with phenol/chloroform and DNA precipitated from the aqueous phase using cold isopropanol with 0.3 M sodium acetate, pH 5.2, avoiding shearing. DNA was washed with cold 70% ethanol then resuspended in 10 mM TRIS-HCl (pH 8) and treated with RNAse A (0.1 mg/mL). The resulting DNA was again cleaned by isopropanol precipitation and

dissolved in 10mM Tris-HCl buffer. DNA was submitted for sequencing on the Illumina platform at the McMaster Genomics Facility at the Farncombe Family Digestive Health Research Institute at McMaster University (Hamilton, Ontario, Canada). Libraries were constructed using either 250 bp or 300 bp paired end protocols, using Nextera kits or 650 bp insert size selection TruSeq kits according to manufacturer's instructions (Illumina, Inc.). High molecular weight DNA was submitted for PacBio sequencing at the Génome Québec Innovation Center at McGill University (Montréal, Québec, Canada), and sequenced on the Pacific Biosciences RS-II. Illumina data were quality filtered and trimmed using skewer v0.2.2 using -q 25 and -Q 25 quality filters, Nextera adapter sequences were used for trimming using the -x option on Nextera-prepared libraries, while non-Nextera prepared samples were trimmed using the default Illumina universal adapter sequences (Jiang et al., 2014). These trimmed and filtered read pairs were then overlapped and merged using FLASH v1.2.11 using default parameters prior to assembly (Magoc & Salzberg, 2011). Illumina reads were assembled using SPAdes v3.10.0 using default parameters (Bankevich et al., 2012). PacBio data were assembled at Génome Québec using the HGAP pipeline (Chin et al., 2013). Where available, short read data were downloaded from the SRA and reassembled using this assembly protocol to help complete BGCs that were incomplete in the publicly available genome assemblies. The source of all sequences used in this study are listed in Supplementary Table 3-1.

Selection of BGCs

Strain descriptions and BGC accessions are included in Supplementary Table 3-1. The BGCs of known glycopeptides and related molecules reported in the literature were

collected from public databases (31 BGCs, Supplementary Table 3-1). BLASTp searches were conducted using GPA BGC fingerprint sequences (Thaker et al., 2013) as queries to identify suspected producers having hits of 85% length to at least three of the query sequences (Altschul et al., 1997) (18 BGCs). Where available, whole genome sequences for known or predicted GPA producers not in our strain collection were downloaded from Genbank, otherwise genome sequences were produced as described for strains from our strain collection (22 BGCs). In total, 56 of the 71 strains harbouring a GPA-like BGC had a genome sequence available for phylogenetic analysis.

Annotation of BGCs and ORF sequences

The antiSMASH software consists of a pipeline using hidden Markov models (HMMs) and BGC-specific rules to identify regions of a genome sequence that are likely to encode the production of a natural product of an expected type. Since many of the known GPA BGCs were first identified long before antiSMASH was available, detection of BGCs was performed using antiSMASH v4.2.0, for both identification of BGCs from newly available sequences and consistent first-pass annotation of known BGCs (Blin, Wolf, et al., 2017). Genome sequences lacking open reading frame predictions were run through Prodigal v2.6.2 (part of the antiSMASH pipeline) using default parameters (Hyatt et al., 2010). Amino acid sequences for all ORFs in the set of predicted biosynthetic clusters were divided into two sets based on the features present in the antiSMASH Genbank format output; those that were predicted by antiSMASH to contain the ‘aSDomain’ feature type and those that did not. The ‘aSDomain’ features are domains identified using antiSMASH-specific HMMs that are used to predict the BGC type. Since these domains

are frequently found as part of multi-domain megasynthase proteins, the sequences of these domains were separately extracted and organized hierarchically by type (as identified by antiSMASH, according to the HMM that identified each domain), along with the amino acid sequences between these identified domains. Whole ORFs were then further binned into families using the cluster_fast option implemented in usearch v8.1 with 65% identity and minimum 70% query coverage cutoffs (Edgar, 2010). To generate consistent annotations for each sequence family, the centroid sequence (a central representative sequence, as defined and output by usearch) of each family was manually annotated using BLAST hits to the nr database at NCBI. The annotation of this centroid sequence was then used to label the entire family. This process ensured that each family consisted of sequences that were more alike each other than they were to sequences in any other family, that these sequences have consistent labels, and ensured that sequences from GPA BGCs were more similar to the other BGC sequences than they were to other sequences in the public database.

Construction of species trees

Two approaches were used to construct a species level phylogeny for the strains considered in this work. 16S ribosomal DNA sequences were obtained, where available, for known GPA producers from NCBI (Supplementary Table 3-1). For organisms with available genome sequences, ribosomal DNA sequences were extracted using rnammer v1.2 (Lagesen et al., 2007). These 16S sequences were aligned using the aligner tool at the ribosomal database project (Cole et al., 2014). These alignments were subjected to phylogenetic analysis using GTR model with default rate categories as implemented by

fasttree2 (Supplementary Figure 3-4). Following McDonald and Currie (2017), the set of HMMs included in TIGRFAM (release 15.0) 'bacterial core gene set' genome property 0799, a set of genes expected to be present in a single copy in every bacterial genome (Selengut et al., 2007) (http://genome-properties.jcvi.org/cgi-bin/GenomePropDefinition.cgi?prop_acc=GenProp0799) were used to analyze the available genome sequences used in this study, using trusted cutoffs for each model. Amino acid sequences were extracted and aligned for each TIGRFAM using HMMER 3.1b2 (Eddy, 2011). Missing sequences were replaced with gap symbols, any missing families, if present, were replaced with a single column of gaps, and finally each TIGRFAM alignment was concatenated into an overall alignment. The final alignment was used for phylogenetic analysis using fasttree2 v2.1.8 using the WAG substitution model with the default number of rate categories for preliminary comparison with the 16S tree (Price et al., 2010). We compared the topology of the 16S tree and the TIGRFAM core-genome trees using normalized Robinson-Foulds difference (implemented in ete3 (Huerta-Cepas, Serra, & Bork, 2016)) and found they were similar, and proceeded with the amino acid sequences for the rest of the analysis.

The concatenated TIGRFAM core genome sequence alignment was analyzed using BEAST v1.84 and libhmsbeagle on a NVIDIA GTX 1080 Ti GPU using the WAG substitution model, and either relaxed uncorrelated lognormal distributed or strict clock priors, and either a constant-size population coalescent or a birth-death tree prior over 2-4 runs for 50 to 250 million generations each depending on model complexity (Ayres et al., 2012; Drummond & Rambaut, 2007). Up to 60% of each MCMC run was discarded as

burn-in when assessed with Tracer before results were combined using logcombiner and treeannotator in the BEAST package to generate the maximum clade credibility (MCC) trees, retaining mean node heights, for each combination of priors. The priors for the age of the root of the BEAST trees and most recent common ancestor (MRCA) nodes for Proteobacteria, Terrabacteria and Actinobacteria were calibrated as uniform distributions bounded by previously published estimates for comparison with splits older than ancestors the GPA BGC harboring organisms (Battistuzzi, Feijao, & Hedges, 2004; McDonald & Currie, 2017). Specifically, the tree root was uniformly distributed from 3500-4000 MA. The MRCA of *E. coli* and *Salmonella* uniformly distributed from 51-176 Ma, initial value 102 Ma. The MRCA of the Terrabacteria (*B. subitilis* and Actinobacteria) uniformly distributed from 2734-3434 Ma, initial value 3051 Ma. The MRCA of Actinobacteria uniformly distributed from 2512-3076 Ma, initial 2734 Ma. The oldest nodes in the species tree have the longest branches and consequently the largest uncertainty in the ages of the major taxonomic groups of GPA producers.

Dated species trees using different clock models and tree priors agreed well with each other (Supplementary Table 3-2). Topologies produced by the strict clock using either tree prior were identical. Small differences in topology, particularly within *Amycolatopsis*, between the strict and relaxed clock models, and between the two tree priors used with a relaxed clock in BEAST, produced a maximum normalized Robinson-Foulds distance of 0.09 as computed by ete3 and a maximum distance of 11.62 using the Kendall-Colijn metric as implemented in the R package treespace (Drummond & Rambaut, 2007; Huerta-Cepas et al., 2016; Jombart, Kendall, Almagro-Garcia, & Colijn,

2017) (Supplementary Figure 3-5, 3-6, Supplementary Table 3-3, 3-4). The eight genera represented in these trees are organized into four groups by Family and Order in Actinobacteria: *Amycolatopsis* and *Micromonospora* (Family Micromonosporaceae), *Kibdellosporangium* and *Amycolatopsis* (Family Pseudonocardiaceae), *Actinomadura, Nonomuraea* and *Herbidospora* (Order Streptosporangiales), and finally *Streptomyces* (Family Streptomycetaceae). All species trees displayed high interior node support values, providing consistent and robust topologies for further analysis of the individual cluster components.

Construction of domain/gene trees and obtaining BGC-related sequences

Phylogenetic trees were constructed for each domain/gene family with at least three sequences (since rooted trees with two or fewer members are trivial), aligned with MUSCLE, inspected manually for presence of gaps and misalignment, then analyzed with fasttree2 to produce unrooted phylogenetic trees using the WAG model suitable for phylogenetic reconciliation (Edgar, 2004). We also performed the reconciliation of inter-domain sequences by including the sequence between the functional domains in NRPS sequences (i.e. the Condensation-Adenylation linker domain as reported elsewhere (Bozhuyuk et al., 2018; Bozhuyuk et al., 2019).

Phylogenetic reconciliation

Reconciliations take the form of a mapping between gene/domain tree nodes, labelled with one of several events, and species tree nodes. Given a set of event costs, an optimal reconciliation under the parsimony criterion is the one with minimal cost. Typically, multiple optimal reconciliations share a minimal cost therefore support for an event can

be computed as the fraction of reconciliations with the same event mapping among the set of optimal reconciliations. Using a dated species tree, estimates can be made for various features of GPA biosynthesis represented by the domain trees. We chose to use the node numbers from the relaxed clock, birth-death prior tree for comparison in this work and used the most conservative node age intervals for the analysis. Time estimates for the major events in GPA evolution were produced by summarizing the most parsimonious (MP) reconciliations of the domain family trees against the time-calibrated species tree using ecceTERA v1.24, including the estimation of MPR root positions, using the following sets of DTL (δ, τ, λ) event cost vectors: (1, 1, 1) = A, (1, 3, 1) = B, without (A00, B00) and with (A10, B10) transfers to/from a dead lineage, or the Pareto-optimal strategies 1 = C01 and 3 = C03 as previously published (Jacox et al., 2016; Libeskind-Hadas, Wu, Bansal, & Kellis, 2014; To, Jacox, Ranwez, & Scornavacca, 2015). For each domain family, the set of MP reconciliations were used to estimate the frequency-weighted time range for the corresponding events discussed in the text. All events and time estimates are included in Supplementary Table 3-5.

Data Availability

All genome sequences produced for this study (22 organisms) are deposited to GenBank under the BioProject accession PRJNA472056. The source of every BGC is listed in Supplementary Table 3-1. Dates for all nodes in all dated trees are provided in Supplementary Table 3-2. Input BGC sequences, 16S rRNA sequences, 16S rRNA alignment, 16S rRNA tree, concatenated TIGRFAM core genome sequence alignment, all dated BEAST species trees, extracted gene/domain family sequences, annotated

gene/domain families, gene/domain family alignments, gene/domain family trees and all reconciliations (scheme A00, A10, B00, B10, C01 and C03) are available at http://github.com/waglecn/GPA_evolution.

Code availability

Reconciliations were visualized by overlaying the reconciled nodes of each gene tree to the species tree using a custom Python script, available at http://github.com/waglecn/GPA_evolution.

RESULTS

GPA production is distributed in Actinobacteria

We sourced 71 GPA BGC clusters consisting of previously published sequences (31 BGCs), and used GPA fingerprint sequences (Thaker et al., 2013) and BLASTp followed by antiSMASH (Weber et al., 2015) (software consisting of rules and sequence models that detect regions of genomes likely to be BGCs) to identify BGCs from Genbank (18 BGCs) and our in-house collection (22 BGCs) (Methods, Supplementary Table 3-1). The GPA BGCs studied here are derived from eight genera from the phylum Actinobacteria, and we constructed multiple species phylogenies using available 16S rRNA sequences (58 strains) and concatenated TIGRFAM single-copy bacterial core genome protein sequences (56 strains) (Methods, Supplementary Figure 3-4) (Selengut et al., 2007). The 16S rRNA (fasttree2) (Price et al., 2010) and core genome phylogeny under various combinations of clock model and tree prior (implemented in BEAST) (Drummond & Rambaut, 2007) produced similar topologies as measured by normalized Robinson-Foulds distance (implemented in ete3 (Huerta-Cepas et al., 2016)) and the Kendall-Colijn metric (Jombart et al., 2017) (Methods, Supplementary Figure 3-5, 3-6, Supplementary Table 3-2, 3-3). A *Bacillus subtilis* and two proteobacteria genomes were used to calibrate the clocks for these trees (Supplementary Table 3-4) and comparison with previously published dates for splits outside of the GPA producing organisms (Battistuzzi et al., 2004; McDonald & Currie, 2017). We chose to use the relaxed clock, birth-death prior tree for reconciliation and present the most conservative node age intervals for the analysis.

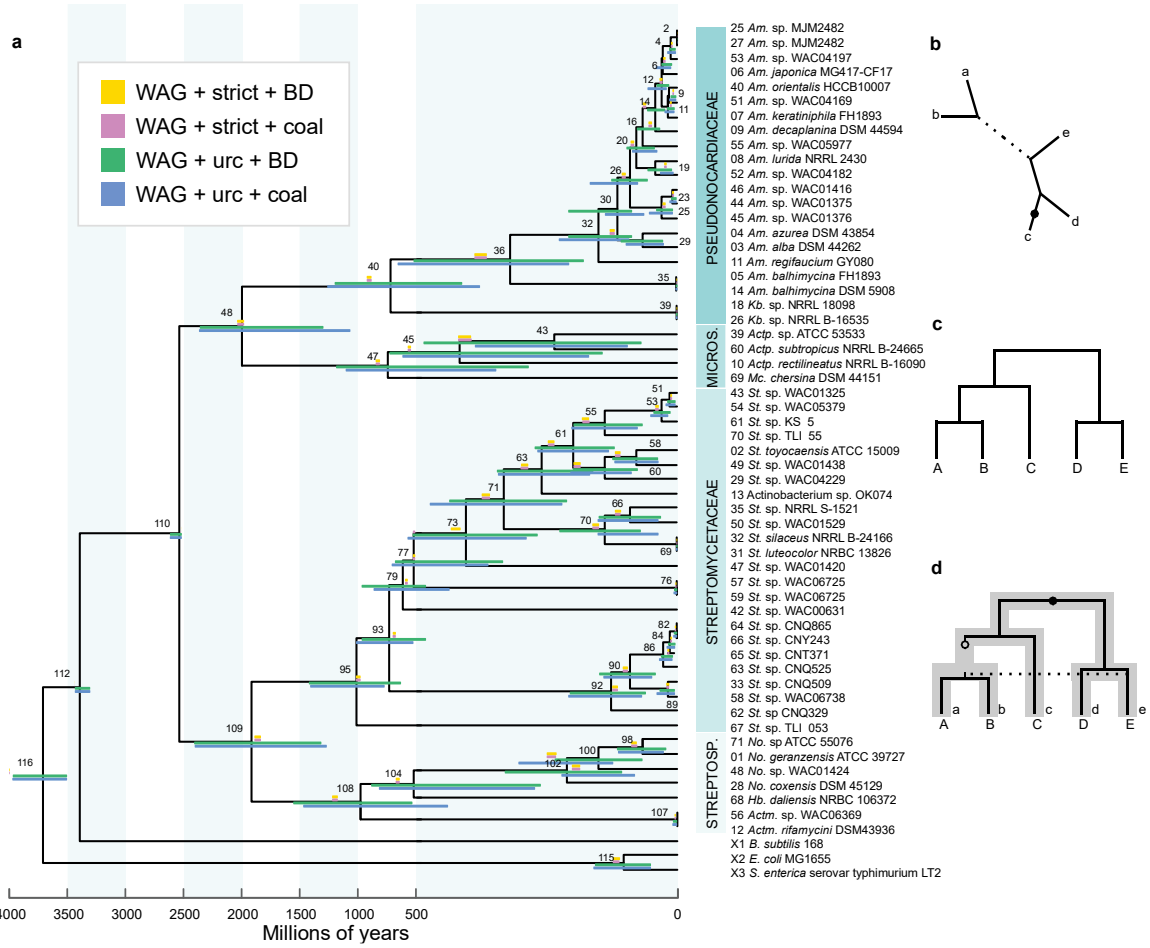


Figure 3-2: Species phylogeny and reconciliation. a GPA-producing species phylogeny. The topology and node heights as drawn come from the tree using a birth-death tree prior and the log-normally distributed uncorrelated relaxed clock implemented in BEAST. Node numbers are used throughout the manuscript to refer to time and place in this phylogeny. Overlaid on each node are the 95% highest posterior distribution (HPD) values for the median height of each comparable node from the trees computed using other models. WAG – WAG substitution model, strict – strict molecular clock, BD – birth-death tree prior, urc – log-normal distributed uncorrelated relaxed clock, coal – fixed population size coalescent tree prior. Genus abbreviations: *Am.* – *Amycolatopsis*; *Kb.* – *Kibdellosporangium*; Micros. – Micromonosporaceae; *Actp.* – *Actinoplanes; Mc.* – *Micromonospora*; *St.* – *Streptomyces*; Streptosp. - Streptosporangiales; *No.* – *Nonomuraea*; *Hb.* – *Herbidospora*; *Actm.* – *Actinomadura*. Note that nodes in this tree only refer to the ancestor of the GPA BGC strains included here wh ich are not the ancestors of the entire group of taxa of which these strains are a subset. We use the combined largest interval of the 95% highest posterior distribution (HPD) of species nodes defining a branch as an estimate for the uncertainty when mapping an event to that branch (Supplementary Table 3-5). **b** An unrooted cartoon gene tree depicting five genes using lower case names a-e. **c** A cartoon species tree depicting five species using upper

case labels A-E. **d** A cartoon showing the gene tree from B reconciled to species tree C showing inferred position of the root (solid circle), a loss event (open circle) in the ancestor leading to species A and B, and a horizontal gene transfer event (dashed line) from the branching leading to species E to the ancestor of species A and B. This is one possible reconciliation of these trees. If the interior nodes of this cartoon species tree were dated, then the inferred events in the mapped gene tree can be dated.

The diversity of their products (Cimermancic et al., 2014) mirrors the evolutionary complexity of BGCs. Domains evolve within genes, genes evolve within BGCs, and BGCs evolve within the genomes of producers. In our analysis, the simplest type of biosynthetic unit is that of a protein domain. For single domain proteins the domain tree is equivalent to the gene tree but for higher levels of organization (multiple domains in a gene, e.g. NRPSs; multiple genes in a BGC, e.g., cytochrome P450s) evolutionary history encompasses more than molecular sequence evolution and may not be adequately represented by trees at all. For the sake of clarity, we refer to the trees of single-domain proteins along with the trees using domains from multi-domain protein as domain trees and eschew the explicit construction of BGCs trees entirely. We used a clustering procedure to gather each of these domain sequences into families to generate annotation, alignment, and phylogenetic trees (see Methods). To avoid the dual meaning of ‘cluster’, we refer to clustered biosynthetic genes as BGCs and clusters of aligned domain sequences as families. To construct a model of GPA evolution we compare the set of domain family trees to the dated species phylogeny using reconciliation (Goodman, Moore, Romero-Herrera, & Matsuda, 1979), a technique that recognizes phylogeny of one set of sequences can differ from another for reasons other than speciation and vertical inheritance (Libeskind-Hadas et al., 2014; Stolzer et al., 2015), identifying well-

supported speciation, duplication, transfer and loss events during the evolution of the GPA BGCs using ecceTERA v1.2.4 (Jacox et al., 2016) (Figure 3-2b, c, d). Dates of mapped events are presented as millions of years ago (Ma) based on median node ages within the species tree.

GPA biosynthesis is accomplished by at least 90 component families

The clustering procedure employed on the protein sequences of the entire set of GPA BGCs was limited by two factors. First, because some clusters were predicted *de novo* antiSMASH analysis, their borders may be inaccurate either because of the arbitrary limits of a window of 15kbp up- and downstream of an identified core biosynthetic domain, or because of the presence of a sequence gap from the originating genome assembly of the source organism. The danger of the first problem is that sequences that are non-biosynthetic may become inadvertently grouped with the BGC sequences, while the danger of the latter is that sequences may be missed. Since the construction of a rooted phylogenetic tree requires at least three leaf sequences, at least three copies of a false positive sequence must be included at the border of a BGC in order to affect downstream analysis, which might be expected to occur if a set of related organisms all have a GPA BGC at the same chromosomal locus. This has occurred, demonstrated by the work of Adamek et al. (2018) showing that *Amycolatopsis* have several common loci occupied GPA BGCs. Re-assembly was used for strains with available read data to improve the contiguity of genome sequences and reduce the number of gaps and identify whole genome BGC sequences for *Kibdellosporangium* sp. NRRL B-16545, *Streptomyces luteocolor* NRBC 13826, *Streptomyces* sp. CNQ329, *Streptomyces* sp.

CNQ-525, *Streptomyces* sp. CNQ826, *Streptomyces* sp. CNT371, *Streptomyces* sp. CNY243, *Herbidospora daliensis* NRBC 106372, *Micromonospora chersina* DSM44151, and *Streptomyces* sp. TLI_55 (Supplementary Table 3-1).

A total of 91 families were produced by the sequence clustering procedure. They are summarized in Table 3-1. 8 families were rejected because they either had too few members for phylogenetic analysis and reconciliation, because they were found in strains missing from the species tree, or because manual curation of the sequences and alignment suggested they were not involved in GPA biosynthesis. The remainder of this chapter will consist of analysis of reconciliation data of these families and discussion of where they fit into overall GPA BGC evolution.

Table 3-1: Component Families Involved in GPA Biosynthesis.

Class	Family Name	Description	# of Sequences	Reason for exclusion
Precursors	DAHPS_I	deoxy-6-arabino-hexulosonate phosphate synthase I	63	
	DAHPS_II	deoxy-6-arabino-hexulosonate phosphate synthase II	25	
	CM_I	chorismate mutase I	27	
	CM_II	chorismate mutase II	123	
	pdh	prephenate dehydrogenase	130	
	HmaS	hydroxymandelate synthase	64	
	Hmo	hydroxymandelate oxidase	64	
	HpgT	4-hydroxy phenylglycine amino transferase	67	
	DpgA	3,5-dihydroxy phenylglycine biosynthesis	57	
	DpgB	3,5-dihydroxy phenylglycine biosynthesis	57	
	DpgC	3,5-dihydroxy phenylglycine biosynthesis	58	
	DpgD	3,5-dihydroxy phenylglycine biosynthesis	43	
	OxyD	β-OH tyrosine monooxygenase	26	
	Bhp	PCP-bound β-OH tyrosine hydrolase	27	
	BpsD	minimal tyrosine loading NRPS	27	
	evaA	NDP-4-oxo-6-deoxy-glucose 3-dehydratase	26	
	evaB	NDP-3,4-dioxo-2,6-dideoxy glucose 4-transaminase	25	
	evaC	NDP-3-amino-2,3,6-trideoxy glucose 3-methyltransferase	12	
	evaD	NDP-3-amino-2,3,6-trideoxy 3-methyl glucose 3-epimerase	25	
	4KR	NDP-aminosugar 4-ketoreductase	20	
	G1PTtf	hexose 1-phosphate thymidyltransferase	4	No genome sequences with gene leaf sequences
	bOHase	amino acid β-hydroxylase	7	
	46DH	NDP-sugar 4,6 dehydratase	4	No genome sequences with gene leaf sequences
NRPS	Adomain	NRPS amino acid adenylation	664	
	Condensation	NRPS condensation	561	
	Epimerization	NRPS epimerization	253	
	PCP	peptidyl carrier protein	648	
	Thioesterase	scaffold peptide release	78	
	Xdomain	Recruits P450s for crosslinking	61	
	NtermCOM	NRPS interaction	134	
	CtermCOM	NRPS interaction	53	
	AMTFill	between A and MT domains	6	
	APCPFill	between A and PCP domains	633	
	ATEFill	between A and TE domains	19	
	CondAFill	between C and A domains	544	
	CondPCPFill	between C and PCP domains	4	No genome sequences with gene leaf sequences
	ECondFill	between E and C domains	147	

	ECtermCOMFill	between E and C-term COM domains	47	
	MT	Methyltransferase	10	
	MTPCPFill	between MT and PCP domains	6	
	PCPCondFill	between PCP and C domains	203	
	PCPEFill	between PCP and E domains	251	
	PCPTEFill	between PCP and TE domains	9	
	PCPXFill	between PCP and X domains	61	
	XAFill	between X and A domains	20	
	XTEFill	between A and TE domains	40	
Tailoring	MbtH	NRPS accessory protein	70	
	Hal	halogenase	50	
	deacetylase	N-acetyl glucosamine deacetylase	32	
	hyd_AcyltfI	hydrolase/acyltransferase I	13	
	hyd_AcyltfII	hydrolase/acyltransferase II	9	
	P450	P450 monoxygenase for scaffold crosslinking	214	
	gtfI	glycosyltransferase I	107	
	gtfII	glycosyltransferase II	8	
	mbgtfI	membrane-bound glycosyl transferase I	28	
	mbgtfII	membrane-bound glycosyl transferase II	7	
	stf	sulfotransferase	12	
	mtfI	methyltransferase I	14	
	mtfII	methyltransferase II	3	
	mtfIII	methyltransferase III	3	No genome sequences with gene leaf sequences
	mtfIV	methyltransferase IV	4	
	omtfI	O-methyltransferase I	11	
	omtfII	O-methyltransferase II	3	
	omtfIII-IV	O-methyltransferase III and IV	3	
	nmtf	N-mehtyltransferase	16	
Other	ABC-ATP	ABC transporter ATPase component	118	
	ABC-perm	ABC transporter permease	21	
	ABC-sol	ABC transporter solute binding	8	
	cat_antiporter	cation antiporter	40	
	catAcSymporterI	cation/acetate symporter I	7	
	catAcSymporterII	cation/acetate symporter II	6	
	AI23trans	AI-23 transporter	6	
	MFS	MFS family transporter	21	Not a coherent family, no individual set of sequences with > 3 members, unlikely related to GPA BGCs
	abhyd	α,β hydrolase family	49	
	ferredoxin	P450 accessory	22	
	vanH	D-lactate dehydrogenase	31	

vanA	D-ala-D-lac ligase	28	
vanX	D-ala-D-ala carboxypeptidase	29	
vanY	D,D carboxypeptidase	5	
PGbI	peptidoglycan binding protein I	14	
PGbII	petptidoglycan binding protein II	7	
PgBIII	peptidogylcan binding protein III	4	
MurF	MurF UDP-N-acetylmuramoyl-tripeptide D-alanyl-D-alanine ligase	11	
MurG	MurG UDP-N-acetylglucosamine-N-acetylmuramyl-(pentapeptide) pyrophosphoryl-undecaprenol N-acetylglucosamine transferase	2	Not enough sequences
FemAB	VanJ-like peptidoglycan biosynthesis protein	4	
DDcarboxypeptidase	cell wall D,D-carboxypeptidase	8	
DalaDala-ligase	Cell wall D-alanyl-D-alanine ligase	5	No genome sequences with gene leaf sequences
StrR	StrR-like regulator	62	
RespReg	Response regulator, part of two-component regulator system	62	
HisK family I	Sensor histidine kinase I, part of two-component regulator system	55	
HisK family II	Sensor histidine kinase II, part of two-component regulator system	4	No genome sequences with gene leaf sequences

GPA precursor biosynthesis is over 1 billion years old

The precursor amino acids HPG, DPG, and BHT share common aromatic amino acid biosynthetic pathways including the enzymes 3-deoxy-D-arabinoheptulosonate 7-phosphate synthase (DAHPS) and chorismate mutase (CM). Two types of DAHPS and CM, as well as a single type of prephenate dehydrogenase (Pdh), are found among GPA BGCs, and additional copies of each are found elsewhere in these genome sequences. Phylogenetic reconciliation hypothesizes these enzymes are ancient, predating their incorporation into GPA biosynthesis (Pdh - node 112, 3401 Ma; DAHPS I and CM II node 110, 2537 Ma) summarized in Figure 3-3a. The most recent common ancestor (MRCA) of the DAHPS type I in GPA BGCs maps to the root of *Amycolatopsis* (node 40, 703 Ma), while the MRCA of DAHPS type II found in GPA BGCs maps to the root of *Streptomyces* (node 77, 503 Ma). The BGC copies of both types of CM could not be dated since they are primarily found in organisms lacking sequenced genomes. These

results suggest that additional copies of at least some of these enzymes were captured in ancestral GPA BGCs in a taxon-specific manner but do not suggest which version, if any, is ancestral in GPA BGCs.

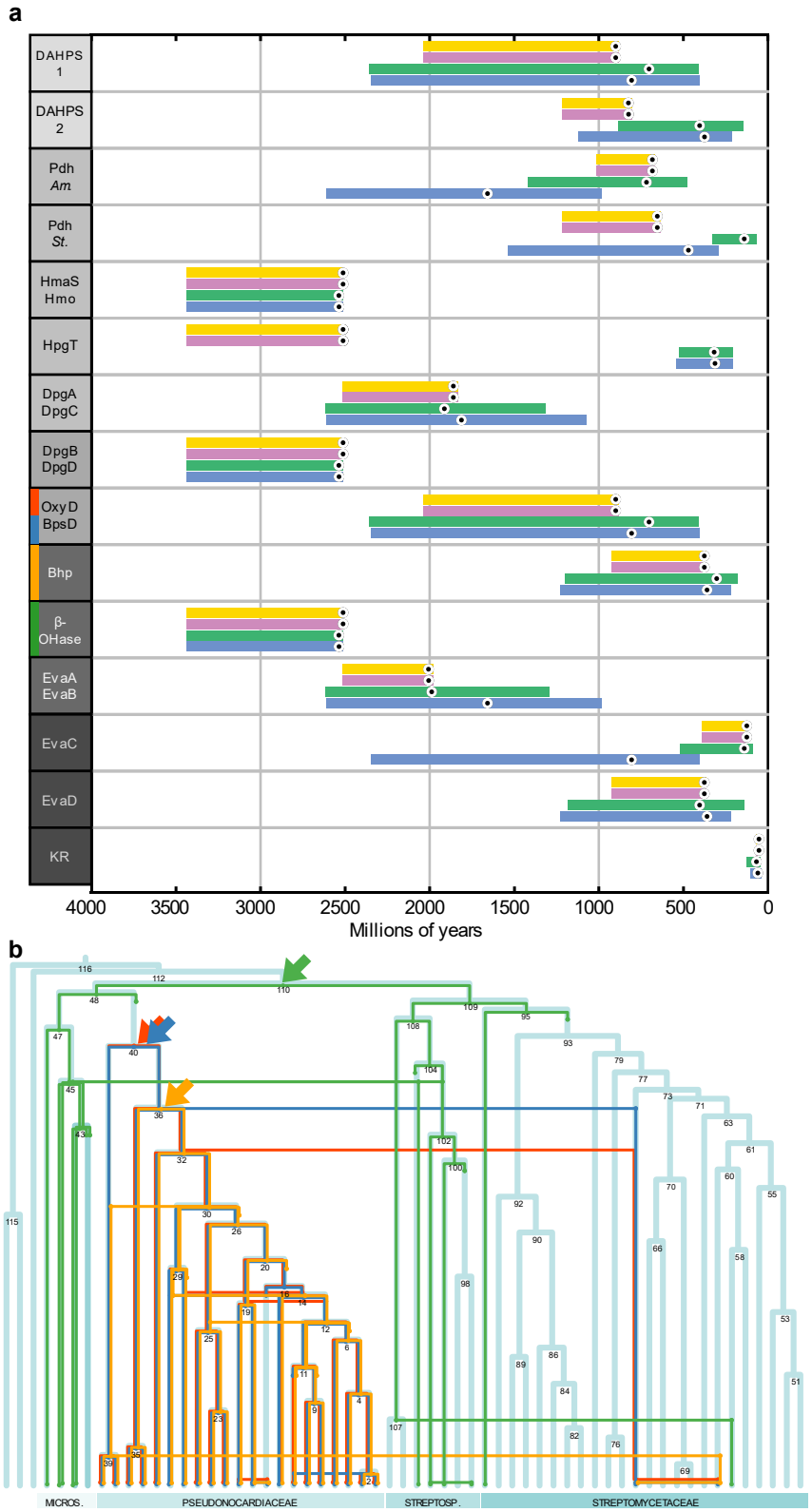

a
DAHPS 1
DAHPS 2
Pdh Am
Pdh St.
HmaS Hmo
HpgT
DpgA DpgC
DpgB DpgD
OxyD BpsD
Bhp
β-OHase
EvaA EvaB
EvaC
EvaD
KR
4000
3500
3000
2500
2000
1500
1000
500
0
Millions of years
b
MICROS.
PSEUDONOCARDIACEAE
STREPTOSP.
STREPTOMYCETACEAE

Figure 3-3: GPA precursor biosynthesis reconciliation dates. a Dates of reconciled nodes of GPA precursor biosynthesis genes versus the four dated species trees. HPD colors for tree models follow Figure 2. **b** Visualized reconciliation for the β-OH tyrosine (BHT) biosynthesis. Root of the ancestral β-hydroxylase from lipodepsipeptide and GPA BGCs (green arrow) maps to the ancient node 110. The MRCA of the GPA β-hydroxylase that converts Tyr in the GPA scaffold to BHT is inferred to transfer from *Nonomuraea* in the branch leading to node 102 to *Actinoplanes* (node 45). The roots of the trees representing the alternate pathway for generating BHT (blue arrow – BpsD, red arrow – OxyD, yellow arrow – Bph) map to early nodes in *Amycolatopsis* (nodes 45 and 36). KR – sugar 4-ketoreductase. Taxa abbreviations as in Figure 3-2.

The HPG biosynthetic enzymes are not always found in these BGCs, perhaps because Pdh is not conserved and the function of the aminotransferase HpgT is not specific to HPG synthesis. The BGC pdh sequences are divided between those found in Streptomyces (MRCA node 93, 718 Ma) and those found in *Nonomuraea* and *Amycolatopsis* (MRCA node 104, 504 Ma). The BGC and genome copies are not strictly monophyletic, suggesting that other non-cluster copies or vice-versa may replace BGC genes. In some *Streptomyces* BGCs Hmo is fused with HpgT. These sequences were split in our construction of their respective trees but genome sequences for the BGCs containing these fused genes are lacking therefore this feature cannot be dated. The MRCA of both Hmo and HmaS suggest HPG biosynthesis has ancient origins (node 110, similar to DAHPS I and CM II).

In contrast, the Dpg genes are usually in an operon; however, DpgD is typically absent in the *Streptomyces* BGCs and is non-essential for DPG biosynthesis (Chen et al., 2001). The reconciled roots of DpgB and DpgD also map to node 110, while those for DpgA and DpgC map to node 109 (1913 Ma). The topology of the aminotransferase HpgT tree differs from the other HPG and DPG enzymes, its reconciliation to the relaxed clock trees

suggest multiple transfers from *Streptomyces* to other producers, with the root mapping to the relatively recent node 71 within *Streptomyces* (318 Ma).

Two distinct routes for BHT biosynthesis are known from tyrosine. In pathway 1, tyrosine is activated and loaded onto a minimal NRPS BpsD where it is hydroxylated to BHT by the P450 enzyme OxyD and then released by the Bhp hydrolase (Figure 3-1c, Supplementary Figure 3-7). In pathway 2, tyrosine is incorporated into the scaffold and subsequently hydroxylated by a β-hydroxylase (Stinchi et al., 2006). The adenylation domain (A-domain) specificity for the GPA NRPS module 6 incorporating either Tyr or BHT agrees with the observed genes in each BGC. A related β-hydroxylase is also predicted to be present in the ramoplanin (Supplementary Figure 3-1, 3-2) BGC as well as in several other related lipodepsipeptide BGCs in which a β-OH group is known or predicted to participate in cyclization of the scaffold (Hoertz, Hamburger, Gooden, Bednar, & McCafferty, 2012). The root of the reconciled β-hydroxylase tree maps to ancient nodes 109-110, compared to the roots of the BpsD/OxyD/Bhp trees, which all map near the root of *Amycolatopsis* (node 40, 703 Ma or node 36, 303 Ma), implying that pathway 2 reflects the ancestral biosynthetic pathway still in use by some BGCs (Figure 3-3b).

GPAs are frequently tailored by aminosugars, glucose, arabinose, and mannose. Aminosugar biosynthesis is initiated with an activated hexose and elaborated via a series of Eva enzymes (Chen et al., 2000) (Figure 3-1c). Phylogenetic roots of EvaA and EvaB map to the MRCA of *Amycolatopsis* & *Actinoplanes* (node 48, 1988 Ma) while the EvaC

root is reconciled to node 32 within *Amycolatopsis* (140 Ma), EvaD is reconciled to node 45 within *Actinoplanes* (404 Ma), the root of the ketoreductase (KR) tree (EvaE or VcaE) is reconciled to node 20 within *Amycolatopsis* (68 Ma). The comparatively older roots for EvaA and EvaB may reflect their pathway function leading to several related aminosugar products, known among a wider set of producing organisms (Thibodeaux, Melancon, & Liu, 2008). In contrast, the GPAs with C3 methylated aminosugars appear to be an innovation originating in or acquired by *Amycolatopsis* (EvaC, Figure 3-3a).

The GPA scaffold is three to five hundred million years old

The GPA scaffold is a heptapeptide where each amino acid incorporated by an NRPS module is encoded on three or four ORFs found in each BGC. The multimodular organization of NRPSs is a recognized source of chemical diversity in GPAs and other natural products (Medema, Cimermancic, Sali, Takano, & Fischbach, 2014) (Figure 3-4a). An analysis of the adenylation (A-, where amino acid recognition occurs) and condensation (C-, site of peptide synthesis) domain phylogenies reveals a MRCA with high support for each A-domain in the GPA NRPSs, rationalizing the pattern of amino acids they are either known, or predicted, to activate (Figure 3-4b). Similarly, the C-domains responsible for amide bond formation between these amino acids in the GPA family also have common ancestors in the domain tree. Together these results suggest that the GPA scaffold has a single origin which can be dated in the species tree. Despite sharing several general molecular features according to the classification of Nicolaou *et al*. (1999), the A- and C-domains encoding the GPA-like compounds complestatin and kistamicin (Nicolaou classification type V) (Supplementary Figure 3-1) scaffolds are not

shared with GPA scaffolds (types I-IV) indicating they evolved independently from the 'true GPAs' (defined as acyl-D-Ala-D-Ala binders).

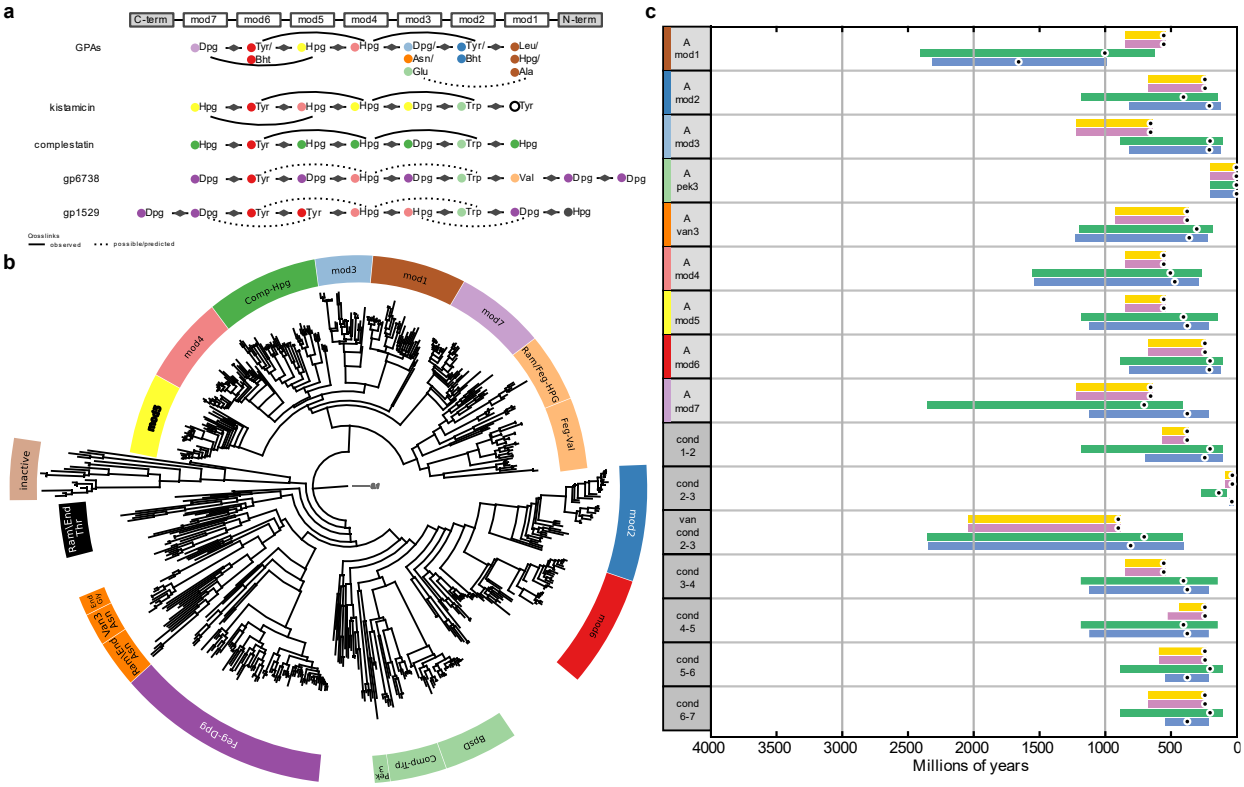


Figure 3-4: GPA scaffold A-domain phylogeny and reconciliation dates. a Linear peptide scaffolds of known and predicted in the GPA BGC family. Circles represent the A-domains activating the amino acid in each position. Crosslinked residues are shown with arcs. Abbreviations: gp1529 – the predicted scaffold of the BGC from Streptomyces sp. WAC01529; gp6738 – the predicted scaffold of the BGC from Streptomyces sp. WAC06738. **b** GPA BGC A-domain phylogeny. Clades representing major divisions in the scaffolds are indicated. Colors for individual A-domains follow panel a. **c** Summary dates of the reconciled roots for GPA BGC A- and C-domains versus the four species trees. HPD colors for phylogenetic models follow Figure 3-2.

Reconciliations of the MRCAs for these important nodes in the A-domain tree shows that modules 1, 2, 3, and 5 of the GPAs map to node 45 within *Actinoplanes* (404 Ma), while modules 6 and 7 map to a contemporary node in *Nonomuraea* (node 102, 202 Ma) (Figure 3-4c). The MRCA of the central HPG (module 4 in GPAs, Figure 3-4a) maps to

node 104 (504 Ma) within *Nonomuraea* followed by a later horizontal gene transfer event into node 45 of *Actinoplanes* (404 Ma). Reconciliation of the C-domain phylogeny tells a similar tale with domains catalyzing amide bond formation in the GPAs sharing common ancestors that map to sub-500 Ma nodes in the species tree. C-domains catalyzing condensations at homologous positions relative to the central HPG4 in known and predicted molecules are ancestral in each C-domain clade. One exception to this is position 3 in the GPAs. The MRCA of A_{Asn3} in the chloroeremomycin/balhimycin/vancomycin group, mapping to node 36 in *Amycolatopsis* (303 Ma), is more closely related to A_{Asp} domains found in lipodespipeptides, while the MRCA of A_{Leu3} from pekiskomycin (Supplementary Figure 3-3) is more closely related to the A-domain in BpsD than to any A-domain from the GPA scaffold NRPS sequences. Likewise, the C-domain responsible for the peptide bond between position 2 and 3 in these BGCs is unrelated to the C-domain catalyzing this bond in the other GPAs. We postulate an NRPS recombination event that reorganized the seven GPA NRPS modules from four genes with a 2-1-3-1 pattern into three genes with a 3-3-1 pattern sometime after the BGCs arrived into *Amycolatopsis* and a second event that converted the A-domain specificity of the third module (Mod3) in pekiskomycin.

Reconciliation of GPA tailoring strategies

Enzyme-catalyzed tailoring of GPAs is a significant source of chemical diversity (Banik, Craig, Calle, & Brady, 2010; Yim, Thaker, et al., 2014; Yim, Wang, Thaker, Tan, & Wright, 2016). A hallmark of the GPA family is a series of intramolecular crosslinks through aromatic C-O and C-C bonds, a feature linked to antibiotic activity. Known and

predicted crosslinks are represented by distinct clades in the phylogeny of the P450s responsible for these modifications. While the root of the reconciled P450 tree maps to ancient node 110, P450s OxyA, OxyC and OxyE have origins at node 102 *Nonomuraea* (202 Ma), similar to the NRPS module 6 and 7 A-domains. The OxyB MRCA maps to node 47 within *Actinoplanes* (731 Ma) (Figure 3-5b). The early branching of the GPA crosslinking enzymes is challenging to interpret due to conflicting branching order of sequences mapping within *Actinoplanes* and *Nonomuraea*; however, we postulate a lower bound for a complete set of P450 enzymes to be present within *Amycolatopsis* (node 30, 106 Ma).

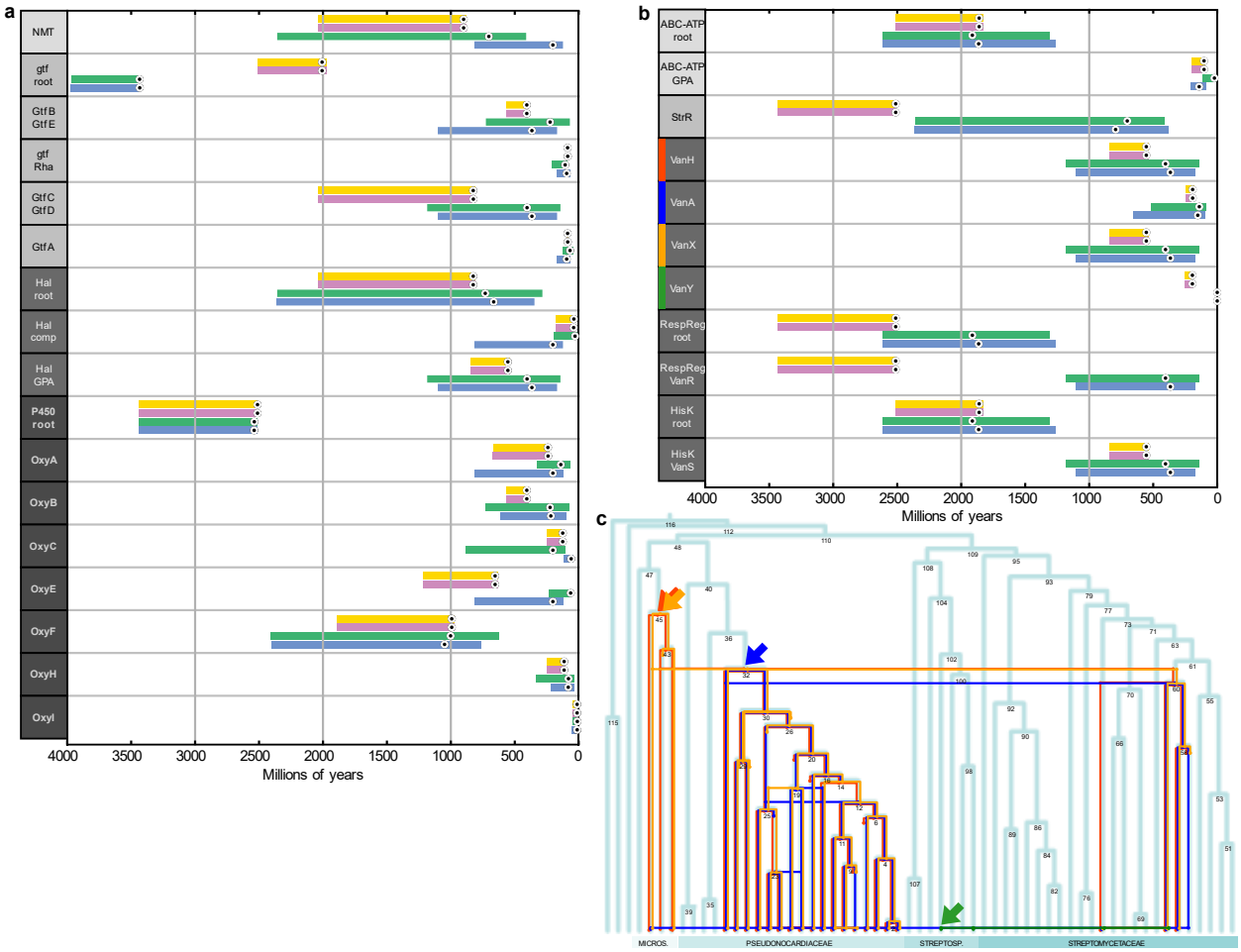


Figure 3-5: GPA tailoring, resistance and regulation reconciliation dates. a Summary dates of the reconciled nodes from GPA BGC modification enzyme phylogenies versus the four species phylogenies. HPD colors for phylogenetic models as in Figure 2. **b** Summary dates of the reconciled GPA BGC export, regulation and resistance versus the four species trees. **c** Reconciliation of GPA BGC resistance genes. Roots of *vanH* (red arrow) and *vanX* (orange arrow) map to *Actinoplanes* node 45, while *vanA* (blue arrow) maps to *Amycolatopsis* node 32. *vanY* (green arrow) maps to *Nonomuraea* and is narrowly distributed in BGC clusters, requiring multiple transfer events. Taxa abbreviations as in Figure 3-2.

The glycosyltransferases (Gtfs) that modify GPAs with sugars all group into a single family. The reconciled root for this family maps to an ancient node in the species tree, but this is possibly an example of poor fitting since it splits a clade of aminosugar glycosyltransferases expected to be monophyletic (Truman et al., 2009) which includes

GtfC from the chloroeremomycin BGC and GtfD from the vancomycin BGC. The MRCA of the clade containing GtfE maps within *Amycolatopsis* (node 43, 224 Ma) and is expected to predate GtfC and GtfD since it acts upstream in the GPA biosynthetic pathway. While there are clear subfamilies in the Gtf phylogeny, outside of the named enzymes here there are few biochemical studies that demonstrate their substrate and regiospecificities. Two additional types of glycosyltransferases are found among GPA BGCs, the first having a high-probability match to pfam02366, and the second to the related pfam13231, both related to the transfer of a sugar from a polyprenylphosphate carrier at or outside the membrane. The first, and more numerous, type in the GPA BGCs is consistent with the prediction of a conserved mannosyltransferase among the GPAs and ramoplanin and related BGCs, while the second is only found in ristocetin BGCs (Supplementary Figures 3-1, 3-2) in *Amycolatopsis* and may represent an unidentified arabinosyl transferase.

GPAs are chlorinated at several aromatic amino acid positions by halogenases. The root of the reconciled halogenase tree maps within *Amycolatopsis* (node 47, 731 Ma). Halogenation is a feature that distinguishes the vancomycin-type GPAs from the teicoplanin-type within *Amycolatopsis*, with the latter BGCs likely losing the ability since halogenation is widespread among other BGCs. Another distinguishing feature in GPAs is the fatty acid acylation of a sugar moiety attached to the central Hpg4 in molecules like A40926 (Supplementary Figure 3-1) and teicoplanin. Precursors of these products have an N-acetylglucosamine in this position that requires deacetylation before addition of the fatty acyl chain (Truman, Robinson, & Spencer, 2006). A putative deacetylase is found in

these BGCs but surprisingly also in every BGC from *Amycolatopsis*, even where glucose is known to be added at position 4 without being ultimately acylated. The root of this tree is reconciled to node 102 within *Nonomuraea* (202 Ma). The acyltransferase tree (hydrolase-acyltransferase II) is also reconciled at its root to node 102, with an HGT event predicted to introduce it into the *Amycolatopsis coloradensis* lineage at node 25 (21.9 Ma).

Efflux and regulatory elements reveal possible multiple lateral gene transfer events

ABC transporters required for export of the synthesized antibiotic with sequence similarity to StaU from the teicoplanin BGC are found in every BGC. Reconciliation of StaU phylogenetic history is complex or may merely reflect poor performance of the ecceTERA algorithm for this aspect of GPA biosynthesis. The reconciled root of the ABC transporter ATP-binding domain phylogeny maps to ancient node 109 (1913 Ma), yet the StaU phylogenetic pattern fits poorly within the species phylogeny requiring late HGTs from *Nonomuraea* after the split at node 98 (60 Ma) to *Amycolatopsis* in the branch leading to node 12 (27 Ma) and *Actinoplanes* at the same time.

An StrR-like regulator is also found in most BGCs and shows a similarly conflicting reconciliation pattern compared to other BGC components. The root of the StrR family maps to node 40 within *Amycolatopsis*, violating the expected monophyly of the GPA StrR family and requiring an HGT into *Nonomuraea* (node 100, 140 Ma).

Resistance is contemporary with GPA biosynthesis dating to one to four hundred million years

Most GPA BGCs include the *vanHAX* genes responsible for self-resistance, with sequence and mechanistic similarity to the van genes found in human pathogens (Marshall et al., 1997). Reconciled roots for VanH and VanX map to node 45 within *Actinoplanes* (404 Ma), while the root of VanA maps slightly later to node 32 in *Amycolatopsis* (140 Ma) (Figures 3-5a, 3-5c). In addition to the vanHAX genes, one or more two-component systems consisting of a paired sensor and response regulator are found in every BGC outside of *Amycolatopsis* (excluding *A. balhimycina*). Such two-component systems are known to be present in some BGCs, with or without accompanying resistance genes (Stegmann, Frasch, Kilian, & Pozzi, 2015) and may play many roles. Yet, while the roots of both components map to ancient node 109, robust clades of sensors and regulators both map to node 45 within *Actinoplanes* (404 Ma), both arriving via lateral transfer events contemporaneous with origin of the BGCs as measured by the NRPS A- and C-domains, that appear to be associated with a resistance gene operon in their BGCs.

DISCUSSION

To understand the evolutionary history of the GPA antibiotics, we have reconciled the phylogeny of gene families having varying levels of conservation in BGCs. Other attempts to date the origin of natural products have focused on phylogenies of a restricted number of marker genes (Baltz, 2005; Joynt & Seipke, 2018; Medema et al., 2014), but our approach can use any sequences for which a robust tree can be estimated. Precursor synthesis genes have some of the oldest reconciled dates, while specific tailoring genes have some of the youngest. The age of the GPAs is best described by comparing the dates of the common ancestors of the NRPS domains responsible for the peptide scaffold making the GPA backbone, approximately 150-400 Ma, which agrees well with other gene and domain families from these BGCs. A timeline of the significant events in GPA evolution according to reconciliation is summarized in Figure 3-6. Our results show that the different features of these molecules have complex evolutionary histories and that the GPA BGCs producing molecules with the well-known ability to bind D-Ala-D-Ala of peptidoglycan precursors arose from a larger pool of biosynthetic components.

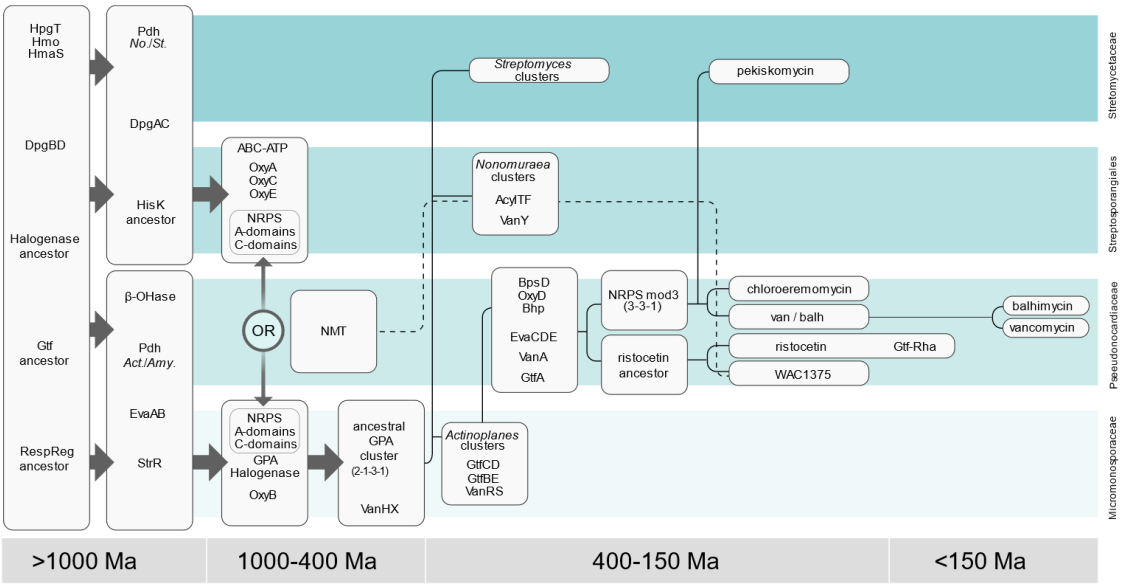


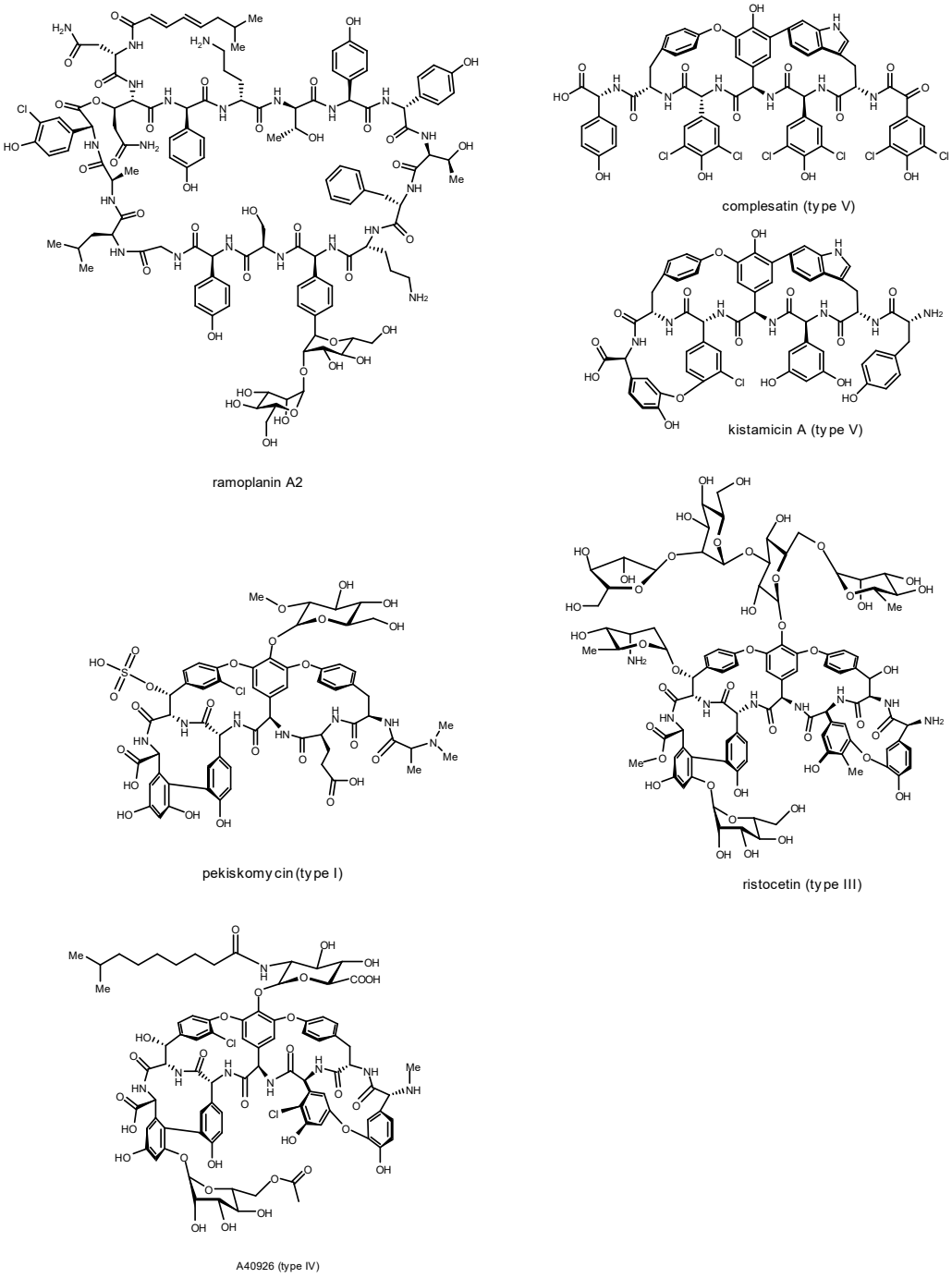
Figure 3-6: Summary of major events in GPA BGC evolution inferred from reconciliation. BGC components responsible for precursor biosynthesis and ancestors of common functions like export and regulation are the most ancient components of BGCs. Ancestral NRPS components of the true GPAs emerge in *Actinoplanes* and *Nonomuraea* around the 1000-400 Ma range and form the GPA BGCs around 400 Ma in *Actinoplanes.* This BGC subsequently spreads to *Nonomuraea, Streptomyces* and *Amycolatopsis* gaining characteristic features like novel scaffold modifications, *vanY* resistance in *Nonomuraea*, a 3-3-1 NRPS module arrangement and alternate BHT biosynthesis in *Amycolatopsis*. Dashed line indicates the transfer of the N-methyltransferase from *Amycolatopsis* to *Nonomuraea*, and the movement of an acyltransferase from *Nonomuraea* to the *Amycolatopsis coloradensis* lineage. β-OHase, β-hydroxylase; Gtf, glycosyltransferase; RespReg, two component system response regulator; HisK, two component system sensor histidine kinase; AcylTF, acyltransferase; Gtf-Rha, rhamnosyltransferase.

Our results offer a lens in which to consider the current antibiotic crisis. The GPAs represent a chemically complex and remarkably tuned family of antibiotics assembled by natural selection over the last 2 billion years. They have been encoded by BGCs in several genera for at least 150 million years. The first GPA antibiotic used in the clinic, vancomycin, was discovered in the 1950s and entered into significant clinical use following the emergence of methicillin-resistant *Staphylococcus aureus* (MRSA) in the late 1970s (Kirst, Thompson, & Nicas, 1998). Clinical resistance in pathogenic

enterococci was reported a decade later (Leclercq, Derlot, Duval, & Courvalin, 1988). When compared to the natural history of GPAs occurring on geologic time scales, the emergence of resistance in the clinic 10 years after their introduction speaks dramatically to the unique fragility of antibiotics as drugs and the great need for careful stewardship in their deployment. More positively, this study strengthens the role for phylogenetics in the discovery of new antibiotics by careful analysis of BGCs and their evolutionary history and the leveraging of this information to direct genome-first discovery efforts towards new molecules with new modes of action (Thaker et al., 2013).

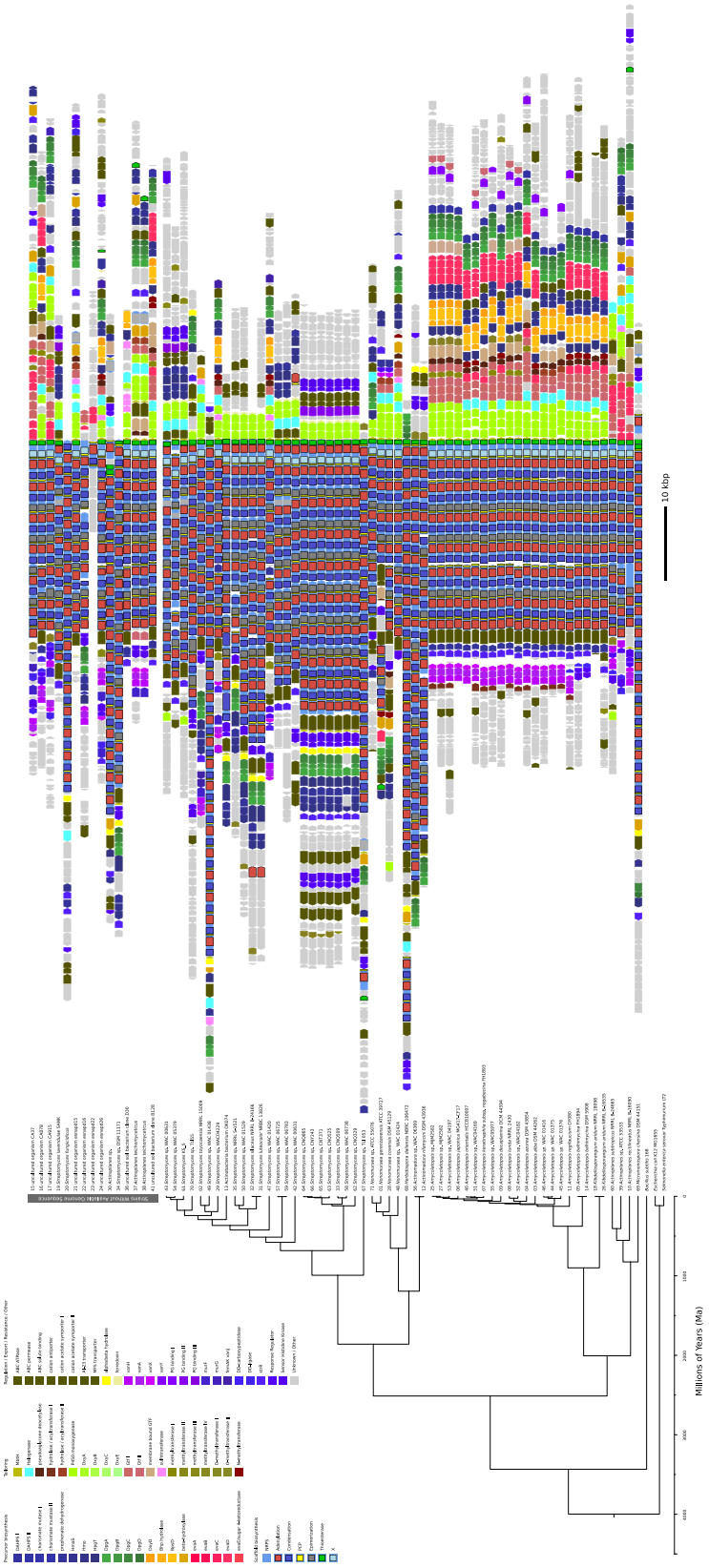
SUPPLEMENTARY MATERIAL

Supplementary Figures

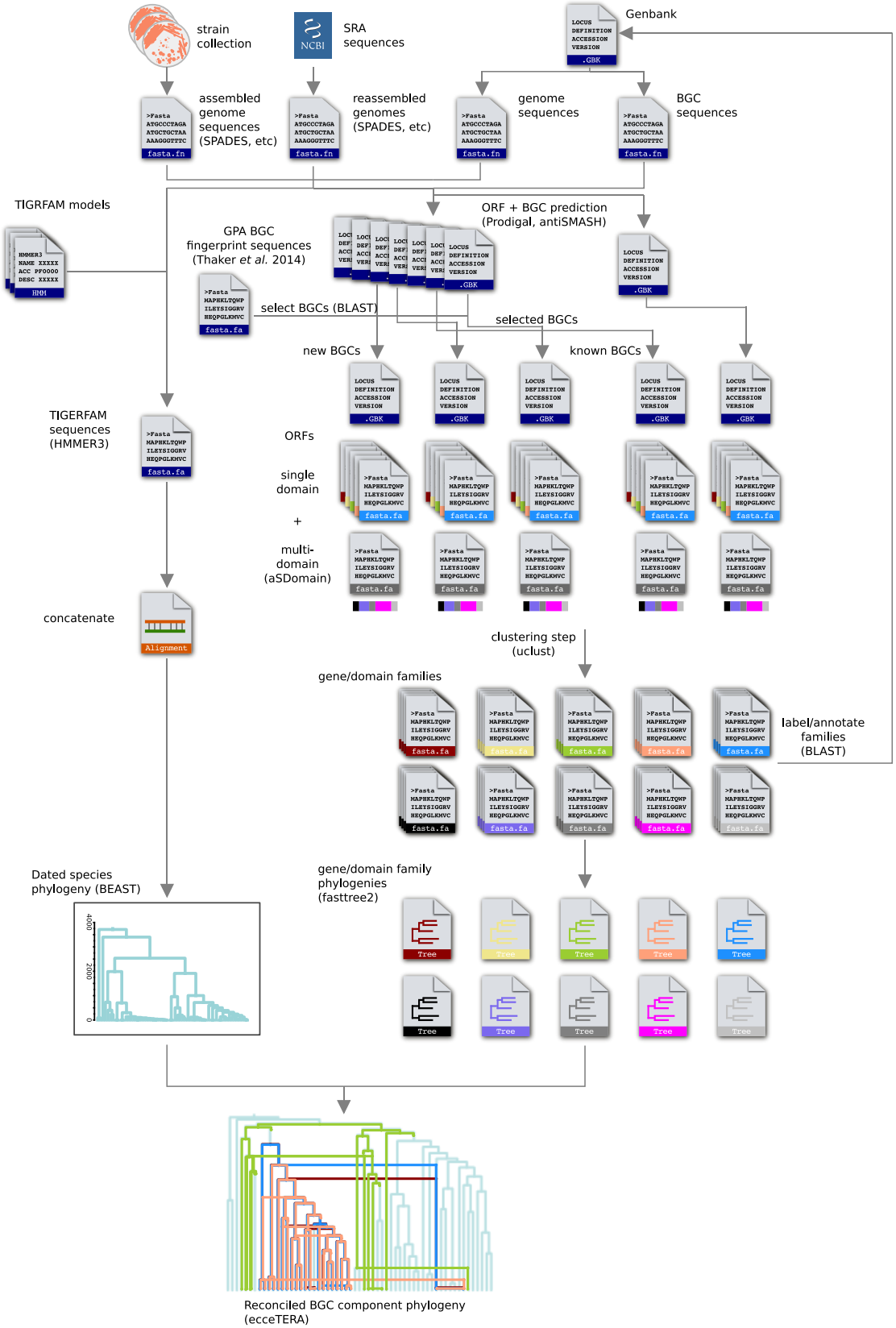


Supplementary Figure 3-1: GPAs and related structures. Type designations for GPAs follow Nicolau *et al*. (1999). Ramoplanin is not a GPA, however it shares several modular features with the GPAs like non-proteinogenic amino acid precursors in its scaffold and tailoring reactions like acylation, glycosylation and halogenation. The type

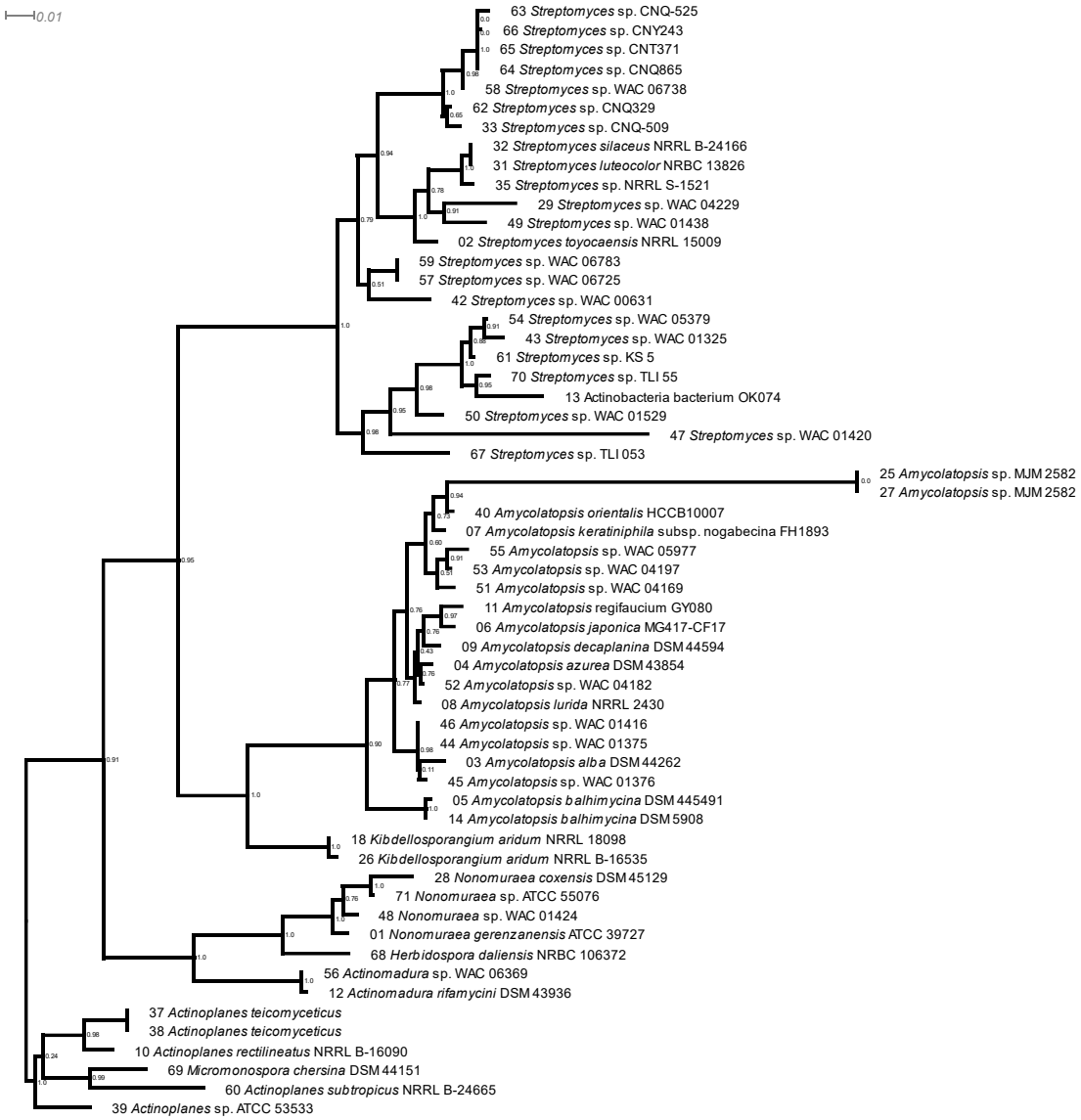
V GPAs, containing crosslinked Trp residues, are more structurally related but evolved independently from the true GPAs.


10 kbp
Millions of Years (Ma)

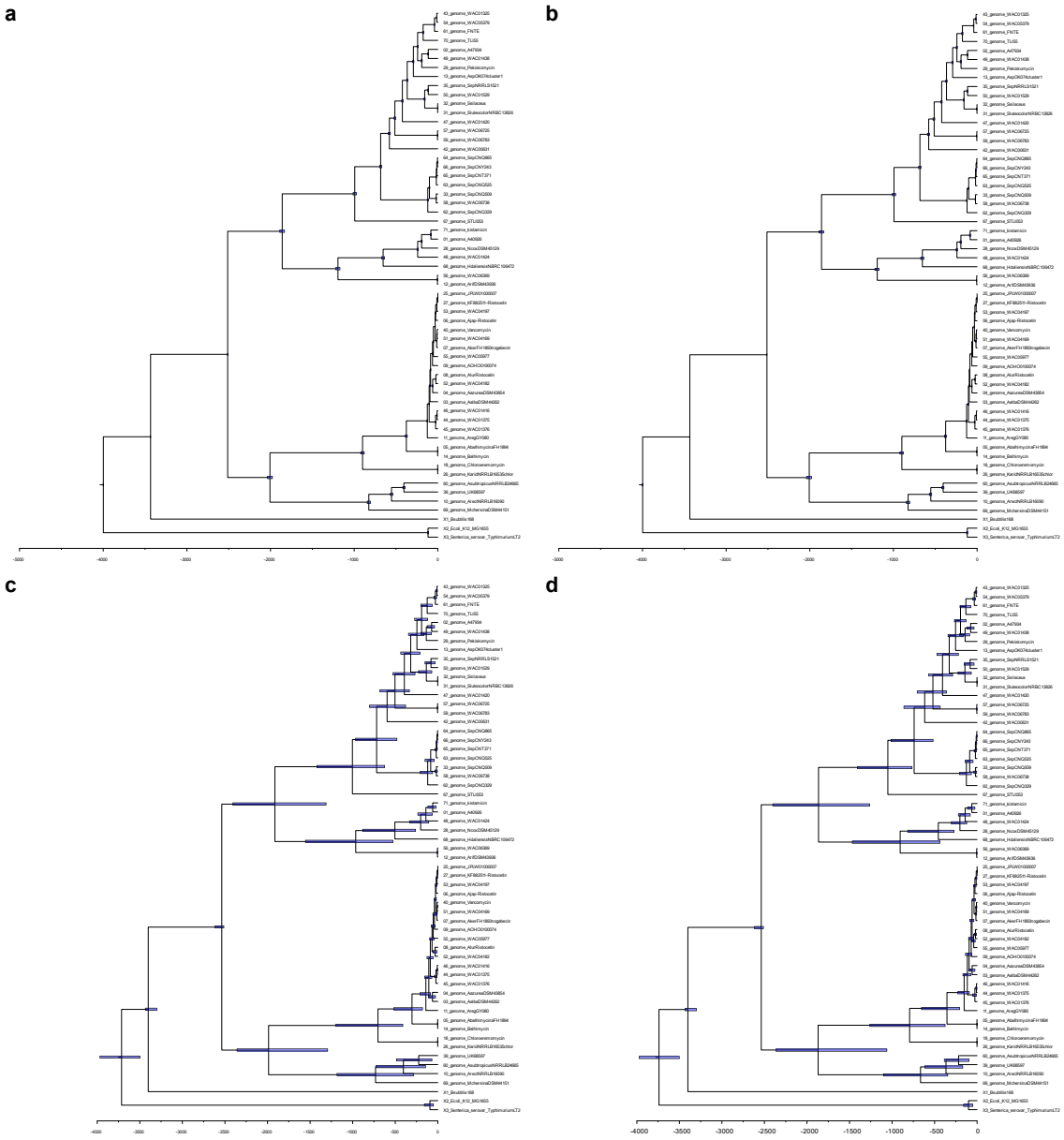
Supplementary Figure 3-2: GPA BGC synteny. 71 GPA BGCs included in this work. Environmental strains derived from uncultured organisms and those lacking sequenced genomes are depicted above, followed by the remaining BGCs originating from strains in the TIGRFAM core-genome species tree estimated in BEAST using the WAG substitution model, a strict clock, and a Birth-Death tree prior. Genes and domains are coloured according to their assigned family.


strain collection
NCBI
SRA sequences
Genbank
assembled genome sequences (SPADES, etc)
reassembled genomes (SPADES, etc)
genome sequences
BGC sequences
TIGRFAM models
GPA BGC fingerprint sequences (Thaker et al. 2014)
ORF + BGC prediction (Prodigal, antiSMASH)
select BGCs (BLAST)
selected BGCs
new BGCs
known BGCs
TIGERFAM sequences (HMMER3)
ORFs
single domain
+
multi-domain (aSDomain)
concatenate
clustering step (uclust)
gene/domain families
label/annotate families (BLAST)
Dated species phylogeny (BEAST)
gene/domain family phylogenies (fasttree2)
Reconciled BGC component phylogeny (ecceTERA)

Supplementary Figure 3-3: Workflow overview. Heterogenous sources of sequence data were used for this workflow. Either previously published GPA BGC sequences or whole genome sequences with known clusters are obtained. Available whole genome sequences were used to extract TIGRFAM single-copy core bacterial genome sequences, aligned, concatenated, and used to produce species trees. ORFs from each identified BGC were taken either whole (single), or broken down (multi-domain), depending on their antiSMASH annotation. These sequences were subjected to a clustering step (UCLUST), to form domain families. Representative members of this families were queried against genbank (BLASTp) to annotate each family with a common label. Families with at least 3 member sequences were subject to alignment, editing if necessary, and unrooted phylogenetic tree estimation. These domain family trees were then reconciled against the species tree (ecceTERA). See Methods.

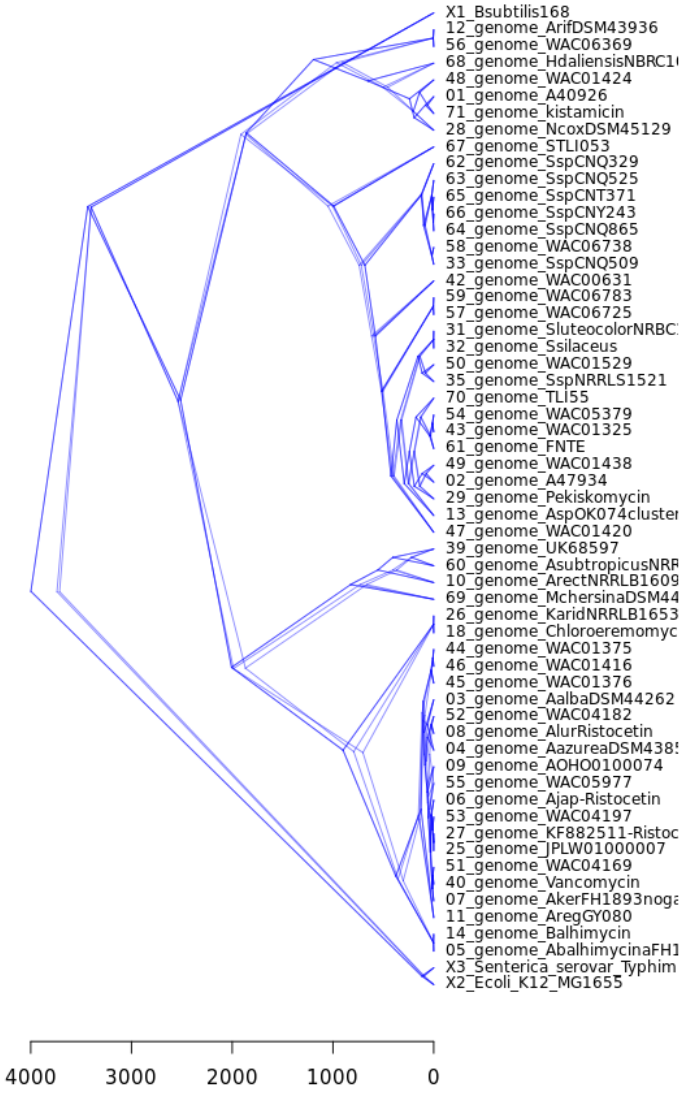


Supplementary Figure 3-4: 16S rRNA species tree. Sequence sources are listed in Supplementary Table 3-1.

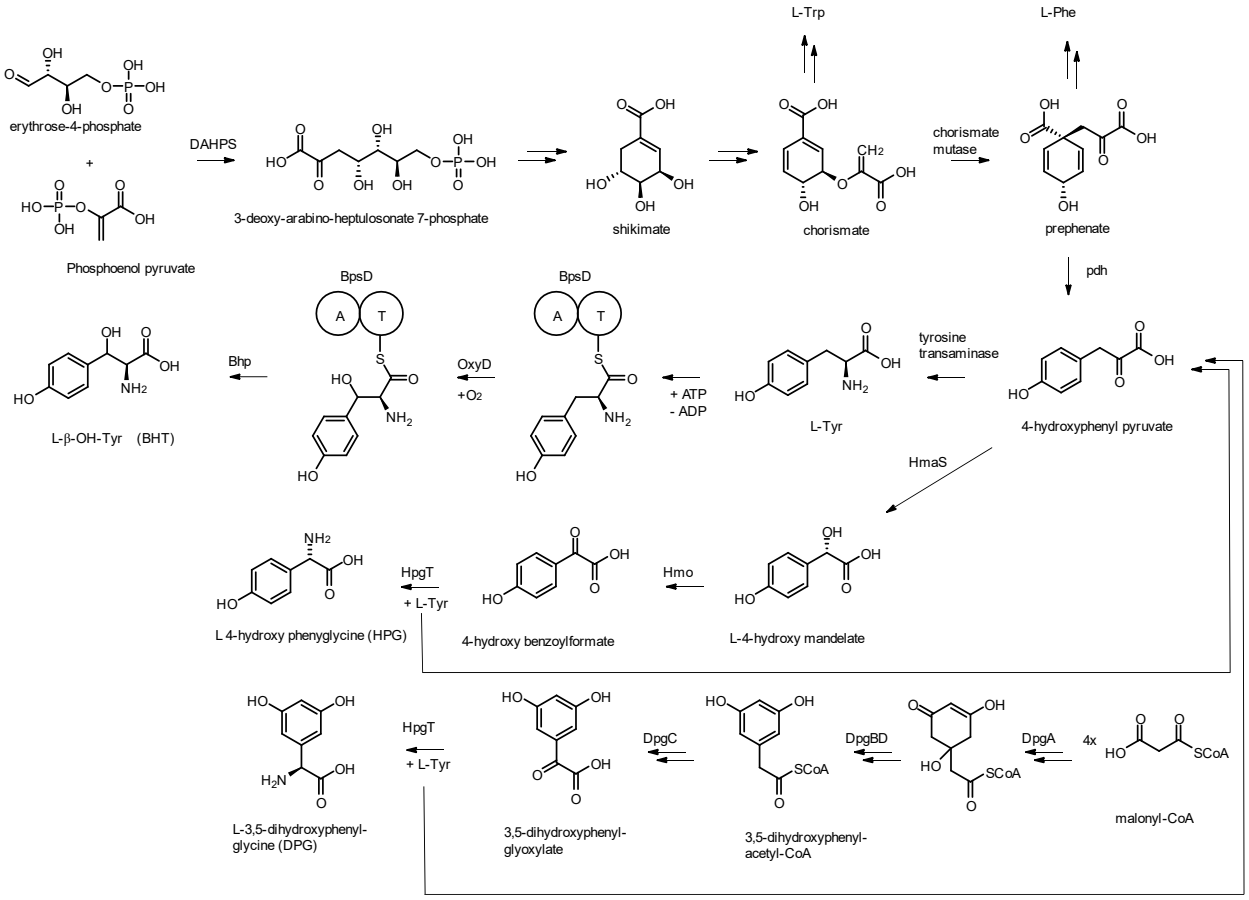


Supplementary Figure 3-5: Dated amino acid species trees. Trees were computed using the following models, node heights are drawn using the posterior median node heights estimated in BEAST (see methods). Blue bars over each node show 95% highest posterior distribution (HPD) of node heights. All trees were estimated using the WAG substitution mode. **a** –strict molecular clock, Birth-Death tree prior. **b** –strict molecular clock, fixed population size coalescent tree prior. **c** – uncorrelated, log-normal distributed

relaxed clock, Birth-Death tree prior. **d** – uncorrelated, log-normal distributed relaxed clock, fixed population size coalescent tree prior.



Supplementary Figure 3-6: densiTree plot of species trees. Overlay of dated species trees computed by densiTree as implemented in the treespace R package. See Methods.



Supplementary Figure 3-7: Amino acid precursor biosynthesis. Nonproteinogenic amino acids used in GPA biosynthesis originate either from the shikimate pathway as in aromatic amino acid biosynthesis or from malonyl-CoA. Both the HPG and DPG pathways require an aminotransferase reaction, provided by HpgT, which uses L-Tyr as an amino donor and regenerates 4-hydroxyphenyl pyruvate that can feedback into BHT or HPG production. The pathway generating BHT via BpsD and OxyD shown here corresponds to pathway 1 in the main text.

Supplementary Tables

Supplementary Table 3-1: Description of organisms, clusters, and sequences used in this study. All genome, BGC, and 16S rRNA sequences attributed to this study are available under BioProject PRJNA472056.

BGC	Genome	Genome accession	Genome reference	Genome reassembled	Taxonomy source	Genus	Species	Strain
1	Y	LT559118.1	N/A	No	as reported	*Nonomuraea*	*gerenzanensis*	ATCC 39727
2	Y	JFCB01000000	(Kwun & Hong, 2014b)	No	as reported	*Streptomyces*	*toyocaensis*	NRRL 15009
3	Y	NZ_ARAF01000000	N/A	No	as reported	*Amycolatopsis*	*alba*	DSM 44262
4	Y	NZ_ANMG01000000	N/A	No	as reported	*Amycolatopsis*	*azurea*	DSM 43854
5	Y	NZ_ARBH01000000	N/A	No	as reported	*Amycolatopsis*	*balhimycina*	DSM 445491, FH1894
6	Y	NZ_CP008953.1	(Stegmann et al., 2014)	No	as reported	*Amycolatopsis*	*japonica*	MG417-CF17
7	Y	MQUP01000000	N/A	No	as reported	*Amycolatopsis*	*keratiniphila*	subsp nogabecina FH1893
8	Y	CP007219.1	(Kwun & Hong, 2014a) (Kaur, Kumar, Bala, Raghava, & Mayilraj, 2013)	No	as reported	*Amycolatopsis*	*lurida*	NRRL 2430
9	Y	AOHO01000000	No	as reported	*Amycolatopsis*	*decaplanina*	DSM 44594	
10	Y	JZKF01000000	N/A	No	as reported	*Actinoplanes*	*rectilineatus*	NRRL B-16090
11	Y	LQCI01000000	N/A	No	as reported	*Amycolatopsis*	*regifaucium*	GY080
12	Y	AULB01000000	N/A	No	as reported	*Actinomadura*	*rifamycini*	DSM 43936
13	Y	LJCV01000000	N/A	No	as reported	*Streptomyces*	sp.	Actinobacteria bacterium OK074
14	Y	QHHU00000000	this study	N/A	as reported	*Amycolatopsis*	*balhimycina*	DSM 5908
15	N	N/A	N/A	N/A	unknown			uncultured organism CA37
16	N	N/A	N/A	N/A	unknown			uncultured organism CA878
17	N	N/A	N/A	N/A	unknown			uncultured organism CA915
18	Y	QHKI00000000	this study	N/A	by 16S	*Kibdellosporangium*	*aridum*	NRRL18098
19	N	N/A	N/A	N/A	as reported	*Streptomyces*	*lavendulae*	SANK
20	N	N/A	N/A	N/A	as reported	*Streptomyces*	*fungicidicus*	
21	N	N/A	N/A	N/A	as reported			uncultured organism esnapd15
22	N	N/A	N/A	N/A	as reported			uncultured organism esnapd16
23	N	N/A	N/A	N/A	as reported			uncultured organism esnapd22
24	N	N/A	N/A	N/A	as reported			uncultured organism esnapd26

25	Y	JPLW01000000	(Kwun et al., 2014)	No	as reported	*Amycolatopsis*	sp.	MJM2582
26	Y	JOAA01000000	N/A	Yes	by 16S	*Kibdellosporangium*	*aridum*	NRRL B-16535 (labelled Rhodococcus rhodnii)
27	Y	JPLW01000000	(Truman et al., 2014)	No	as reported	*Amycolatopsis*	sp.	MJM2582
28	Y	ARBV01000000	N/A	No	as reported	*Nonomuraea*	*coxensis*	DSM 45129
29	Y	QHJH00000000	this study	N/A	as reported	*Streptomyces*	sp.	WAC 04229
30	N	N/A	N/A	N/A	as reported	*Actinoplanes*	sp.	
31	Y	BDGW01000000	this study	Yes	as reported	*Streptomyces*	*luteocolor*	NRBC 13826
32	Y	LIRJ01000000	N/A	No	as reported	*Streptomyces*	*silaceus*	NRRL B-24166
33	Y	CP011492.1	N/A	No	as reported	*Streptomyces*	sp.	CNQ-509
34	N	N/A	N/A	N/A	as reported	*Streptomyces*	sp.	DSM 11171
35	Y	LLZK01000000	N/A	No	as reported	*Streptomyces*	sp.	NRRL S-1521
36	N	N/A	N/A	N/A				D30
37	N	N/A	N/A	N/A	as reported	*Actinoplanes*	*teichomyceticus*	
38	N	N/A	N/A	N/A	as reported	*Actinoplanes*	*teichomyceticus*	
39	Y	QHHV00000000	this study	No	as reported	*Actinoplanes*	sp.	ATCC 53533
40	Y	CP003410.1	N/A	No	as reported	*Amycolatopsis*	*orientalis*	HCCB10007
41	N	N/A	N/A	N/A				Uncultured soil bacterium clone B128
42	Y	QHKJ00000000	this study	No	by 16S	*Streptomyces*	sp.	WAC 00631
43	Y	QHKK00000000	this study	No	by 16S	*Streptomyces*	sp.	WAC 01325
44	Y	QHHW00000000	this study	No	by 16S	*Amycolatopsis*	sp.	WAC 01375
45	Y	QHKL00000000	this study	No	by 16S	*Amycolatopsis*	sp.	WAC 01376
46	Y	QHHX00000000	this study	No	by 16S	*Amycolatopsis*	sp.	WAC 01416
47	Y	QHHY00000000	this study	No	by 16S	*Streptomyces*	sp.	WAC 01420
48	Y	QHHZ00000000	this study	No	by 16S	*Nonomuraea*	sp.	WAC 01424
49	Y	CP029601.1	this study	No	by 16S	*Streptomyces*	sp.	WAC 01438
50	Y	CP029617.1	this study	No	by 16S	*Streptomyces*	sp.	WAC 01529
51	Y	QHJI00000000	this study	No	by 16S	*Amycolatopsis*	sp.	WAC 04169
52	Y	QHJJ00000000	this study	No	by 16S	*Amycolatopsis*	sp.	WAC 04182
53	Y	QHJK00000000	this study	No	by 16S	*Amycolatopsis*	sp.	WAC 04197
54	Y	QHJL00000000	this study	No	by 16S	*Streptomyces*	sp.	WAC 05379
55	Y	QHJM00000000	this study	No	by 16S	*Amycolatopsis*	sp.	WAC 05977
56	Y	QHJN00000000	this study	No	by 16S	*Actinomadura*	sp.	WAC 06369
57	Y	QHJO00000000	this study	No	by 16S	*Streptomyces*	sp.	WAC 06725
58	Y	CP029618.1	this study	No	by 16S	*Streptomyces*	sp.	WAC 06738
59	Y	QHJP00000000	this study	No	by 16S	*Streptomyces*	sp.	WAC 06783
60	Y	JOJL01000000	N/A	No	as reported	*Actinoplanes*	*subtropicus*	NRRL B-24665
61	Y	FNTE01000000	N/A	No	as reported	*Streptomyces*	sp.	KS_5
62	Y	AXVU01000000	N/A	Yes	as reported	*Streptomyces*	sp.	CNQ329

63	Y	JNID01000000	N/A	Yes	as reported	*Streptomyces*	sp.	CNQ-525
64	Y	AUKP01000000	N/A	Yes	as reported	*Streptomyces*	sp.	CNQ865
65	Y	AZWZ01000000	N/A	Yes	as reported	*Streptomyces*	sp.	CNT371
66	Y	ARHU01000000	N/A	Yes	as reported	*Streptomyces*	sp.	CNY243
67	Y	LT629775.1	N/A	No	as reported	*Streptomyces*	sp.	TLI_053
68	Y	BBXF01000000	(Komaki et al., 2015)	Yes	as reported	*Herbidospora*	*daliensis*	NRBC 106372
69	Y	FMIB01000000	N/A	Yes	as reported	*Micromonospora*	*chersina*	DSM 44151
70	Y	OAOJ01000000	N/A	Yes	as reported	*Streptomyces*	sp.	TLI_55
71	Y	CP017717.1	(Nazari et al., 2017)	No	as reported	*Nonomuraea*	sp.	ATCC 55076

Supplementary Table 3-1 (Continued): Description of organisms, clusters, and sequences used in this study. All genome, BGC, and 16S rRNA sequences attributed to this study are available under BioProject PRJNA472056.

BGC	16S?	16S accession	BGC accession	Product	BGC family	Peptide length	BGC reference
1	Y	NR_148586.1	AJ561198.1	A40926	GPA	7	(Margherita Sosio, Stinchi, Beltrametti, Lazzarini, & Donadio, 2003)
2	Y	JFCB01000068.1_105-1619	U82965	A47934	GPA	7	(Pootoolal et al., 2002)
3	Y	NR_024888.1	KB913032.1:5306247-5399940	unknown	GPA	7	this study
4	Y	NZ_ANMG01000133.1_63-1567	ANMG01000033.1:125666-193411 ANMG01000086.1:1-15420 ANMG01000086.1:51876-32518	unknown	GPA	7	(Omura et al., 1979)
5	Y	NR_025564.1	KB913037.1:1-104787	balhimycin	GPA	7	this study
6	Y	NR_025561.1	CP008953.1:6885328-6816619	ristocetin	GPA	7	this study
7	Y	NR_025563.1	MQUP01000022.1:85511-181481	unknown	GPA	7	this study
8	Y	NR_114905.1	CP007219.1:4230016-4310208	ristocetin	GPA	7	(Kwun & Hong, 2014a)
9	Y	NR_025562.1	AOHO01000074.1:355443-260799	unknown	GPA	7	this study
10	Y	JZKF01000194.1_63-1568	JZKF01000002.1:148291-254753	unknown	GPA	7	this study
11	Y	NR_042747.1	LQCI01000003.1:229395-317819	unknown	GPA	7	this study
12	Y	NR_113155.1	AULB01000007.1:8849-16513 AULB01000053.1:1-7382 AULB01000063.1:1-6453 AULB01000066.1:1-5192 AULB01000065.1:1-3684 AULB01000064.1:5221-3 AULB01000007.1:8766-1 AULB01000053.1:19713-11874 AULB01000074.1:1534-1 AULB001000044.1:1-15549 LJCV01000261.1:275-16839 LJCV01000108.1:11021-54 LJCV01000261.1:16798-55068	unknown	N/A	?	this study
13	Y	LJCV01000030.1_33598-35113		unknown	GPA		this study
14	Y	QHHU01000165.1:124-1632	Y16952	balhimycin	GPA	7	(Pelzer et al., 1999)
15	N	N/A	HM486074.1	CA37	N/A		(Banik et al., 2010)
16	N	N/A	HM486075.1	CA878	N/A		(Banik et al., 2010)
17	N	N/A	HM486076.1	CA915	N/A		(Banik et al., 2010)
18	Y	QHKI01000152.1:3401-4907	AJ223998.1 AJ223999.1	chloroeremomycin	GPA	7	(van Wageningen et al., 1998)
19	N	N/A	AF386507.1	complestatin	complestatin	7	(Chiu et al., 2001)
20	N	N/A	DQ403252.1	enduracidin	enduracidin, ramoplanin-like		(Yin & Zabriskie, 2006)
21	N	N/A	KF264554.1	esnapd15	GPA	7	(Owen et al., 2013)
22	N	N/A	KF264555.1 KF264556.1	esnapd16	N/A		(Owen et al., 2013)
23	N	N/A	KF264562.1	esnapd22	N/A		(Owen et al., 2013)
24	N	N/A	KF264565.1	esnapd26	N/A		(Owen et al., 2013)
25	Y	JPLW01000013.1_7-2100	JPLW01000007.1:282114-369000	ristocetin	GPA	7	(Kwun et al., 2014)
26	Y	JOAA01000108.1_1-1320	JOAA01000006.1:260395-353547	chloroeremomycin	GPA	7	this study
27	Y	JPLW01000013.1_7-2100	KF882511.1	ristocetin	GPA	7	(Truman et al., 2014)

28	Y	KB904034.1_12-1456	KB903969.1:41266-1 KB904024.1:4190-1 KB903969.1:632-1 KB903995.1:1-30779	unknown	GPA	7	this study
29	Y	JX440422.1	KC688274.1	pekiskomycin	GPA	7	(Thaker et al., 2013)
30	N	N/A	DD382878	ramoplanin	enduracidin, ramoplanin-like		(Farnet, 2002)
31	Y	N/A	BDGW01000009.1:76837-124328 BDGW01000054.1:1-46946	unknown	WAC01529-like	10	this study
32	Y	LIRJ01000287.1_37462-38974	LIRJ01000175.1:47401-1 LIRJ01000084.1:1-13480 LIRJ01000083.1:1-18145	unknown	WAC01529-like	?	this study
33	Y	CP011492.1_6129288-6130805	CP011492.1:7442803-7539290	unknown	WAC06738-like	9	this study
34	N	N/A	KT809366.1	feglymycin	feglymycin	14	(Gonsior et al., 2015)
35	Y	LLZK01000053.1_18-1530	LLZK01000255.1:1-29980 LLZK01000034.1:1-4905 LLZK01000121.1:247639-289616	unknown	WAC01529-like		this study
36	N	N/A	EU874253.1	TEG	GPA	7	(Banik & Brady, 2008)
37	Y	AB047513.1	AJ632270.1	teicoplanin	GPA	7	(M. Sosio et al., 2004)
38	Y	AB047513.1	AJ605139.1	teicoplanin	GPA	7	(Li et al., 2004)
39	Y	QHHV01000131.1:1676-170	KF192710.1	UK-68,597	GPA	7	(Yim, Kalan, et al., 2014)
40	Y	CP003410.1_1243911-1245417	HQ679900.1	vancomycin	GPA	7	(Xu et al., 2014)
41	N	N/A	EU874252.1	VEG	GPA	7	(Banik & Brady, 2008)
42	Y	QHJK01000340.1:29-1542	QHKJ01000024.1:16380-45867 QHKJ01000550.1:1-624 QHKJ01000229.1:1-8795 QHKJ0101000530.1:1-624 QHKJ01000044.1:1-34189	unknown	complestatin-like	7	this study
43	Y	QHKK01000359.1:1530-17	QHKK01000070.1:32012-1 QHKK01000185.1:1-18540 QHKK01000026.1:62423-19352	complestatin	complestatin	7	this study
44	Y	JX440413.1	JX576190.1	unknown	GPA	7	(Thaker et al., 2013)
45	Y	JX440414.1	QHJK01000008.1:153149-308339	unknown	GPA	7	this study
46	Y	QHHX01000022.1:12-1514	QHHX01000008.1:93295-236922	unknown	GPA	7	this study
47	Y	JX440416.1	JX026280.1	pekiskomycin	GPA	7	(Thaker et al., 2013)
48	Y	QHHZ01000046.1:169-1675	QHHZ01000020.1:1-68658	unknown	GPA	7	this study
49	Y	CP029601.1:2145995-2144480	CP029601.1:7342589-7472015	unknown	enduracidin, ramoplanin-like	23	this study
50	Y	CP029617.1:1153283-1151770	CP029617.1:7746089-7835715	unknown	WAC01529-like	9	this study
51	Y	QHJI01000026.1:5177-3671	QHJI01000006.1:373300-466777	ristocetin	GPA	7	(Thaker et al., 2013)
52	Y	QHJJ01000014.1:29-1533	QHJJ01000006.1:380178-255381	ristocetin	GPA	7	this study
53	Y	QHJK01000023.1:69-1575	QHJK01000007.1:424423-319976	ristocetin	GPA	7	this study
54	Y	QHJL01000470.1:2462-949	QHJL01000251.1:14357-1 QHJL01000627.1:2276-1 QHJL01000163.1:21516-1 QHJL01000064.1:1-32323	complestatin	complestatin	7	this study
55	Y	QHJM01000061.1:5005-3671	QHJM01000012.1:87170-1	ristocetin	GPA	7	this study
56	Y	QHJN01000296.1:1537-31	QHJN01000002.1:1-91218	unknown	N/A	21	this study
57	Y	QHJO01000475.1:48-1562	QHJO01000010.1:63466-1	unknown	complestatin-like	7	this study
58	Y	CP029618.1:6348193-6349708	CP029618.1:7765392-7852876	unknown	WAC6738-like	9	this study
59	Y	QHJP01000053.1:13-1527	QHJP01000004.1:143104-67545	complestatin	complestatin-like	7	this study
60	Y	JOJL01000117.1_84-1589	JOJL01000014.1:187021-118363	unknown	GPA		this study
61	Y	FNTE01000001.1:6502056-6503569	FNTE01000002.1:570535-476296	complestatin	complestatin	7	this study

62	Y	N/A	AXVU01000019.1:91966-129952 AXVU01000134.1:1-2401 AXVU01000075.1:11484-1606 AXVU01000095.1:1-5175 AXVU01000075.1:2729-1 AXVU01000035.1:1-32560	unknown	WAC06738-like	9	this study
63	Y	N/A	KL370766.1:95096-1	unknown	WAC06738-like	9	this study
64	Y	N/A	AUKP01000017.1:114786-156847 AUKP01000055.1:12-30486 AUKP01000017.1:155109-179896	unknown	WAC06738-like	9	this study
65	Y	N/A	AZWZ01000018.1:114109-167525 AZWZ01000073.1:7346-1 AZWZ01000051.1:33278-45	unknown	WAC06738-like	9	this study
66	Y	N/A	KB897732.1:86781-1	unknown	WAC06738-like	9	this study
67	Y	LT629775.1_7898131-7899642	LT629775.1:2963242-2843423	unknown	enduracidin, ramoplanin-like	?	this study
68	Y	BBXF01000018.1_123-1630	BBXF01000001.1:264604-137000	unknown	enduracidin, ramoplanin-like	?	this study
69	Y	NR_044892.1	FMIB01000002.1:35607-142931	unknown	enduracidin, ramoplanin-like	?	this study
70	Y	OAOJ01000001.1_1486024-1487537	OAOJ01000001.1:4307563-4193402	unknown	enduracidin, ramoplanin-like	?	this study
71	Y	CP017717.1_11588372-11589881	CP017717.1:11969581-12040280	kistamicin	complestatin-like	7	(Nazari et al., 2017)

Supplementary Table 3-2: Time tree node details. All node heights provided as Millions of Years (Ma). Tree names as in main text.

	Battistuzzi *et al.* (2004)		McDonald and Currie (2017)		stree04 WAG + strict + BD				stree05 WAG + strict + Coal			
Taxa at Node	Median	CI	Median	CI	Node	Posterior	Median	95% HPD	Node	Posterior	Median	95% HPD
root	3977	3434-4464	N/A	3500-3800	116	1.0	3999	3994-4000	116	1.0	3999	3994-4000
EC+SALM	102	57-176	87	45-130	115	1.0	117	110-124	115	1.0	117	111-124
FIRM+ACTINO	3051	2738-3434	3452	1806-5121	112	1.0	3434	3433-3434	112	1.0	3433	3433-3434
ACTINOBACTERIA	2743	2512-3076	2578	1345-3811	110	1.0	2512	2512-2513	110	1.0	2512	2512-2513
STREPTOMYCES	1375	1032-1727	382	250-514	95	1.0	993	973-1014	95	1.0	993	973-1014
Mc+Ac+Kb+Amy					48	1.0	2007	1976-2037	48	1.0	2006	1976-2036
Mc+Ac					47	1.0	825	806-844	47	1.0	824	806-844
ACTINOPLANES					45	1.0	554	539-568	45	1.0	554	539-568
Ac					43	1.0	406	394-419	43	1.0	406	393-419
Amy+Kb					40	1.0	900	879-921	40	1.0	900	878-921
KIBDELLOSPORANGIUM					39	1.0	1.62	0.87-2.5	39	1.0	1.63	0.89-2.49
AMYCOLATOPSIS					36	1.0	376	364-388	36	1.0	376	364-388
Amy					35	1.0	1.8	1.01-2.67	35	1.0	1.78	1.05-2.71
Amy					32	1.0	125	120-130	32	1.0	125	120-130
Amy					30	1.0	103	99-107	30	1.0	103	99-107
					--	--	--	--	--	--	--	--
					--	--	--	--	--	--	--	--
Amy					29	1.0	26	23-29	29	1.0	26	23-30
Amy					27	1.0	9	7.0-11	27	1.0	8.78	7.03-10.7
					--	--	--	--	--	--	--	--
Amy					24	1.0	94	90-98	24	1.0	94	90-98
Amy					22	1.0	87	83-90	22	1.0	86	83-90
Amy					21	1.0	61	58-65	21	1.0	23	20.6-26.2
					--	--	--	--	--	--	--	--
Amy					19	1.0	23	21-26	19	1.0	23	21-26
Amy					16	1.0	63	60-67	16	1.0	63	60-67
Amy					14	1.0	53	49-56	14	1.0	53	49-56
Amy					12	1.0	32	29-35	12	1.0	32	29-35
Amy					11	1.0	13	11.2-15.2	11	1.0	13.2	11.3-15.2
Amy					9	1.0	9	7.0-10.3	9	1.0	8.66	7.02-10.3
Amy					6	1.0	26	23-28	6	1.0	26	23-28
Amy					4	1.0	11	9.6-13	4	1.0	11.4	9.6-13.3
Amy					2	1.0	0.22	0.03-0.53	2	1.0	0.23	0.03-0.54
No+St					109	1.0	1860	1831-1888	109	1.0	1860	1831-1888
Acm+Hb+No					108	1.0	1194	1169-1219	108	1.0	1194	1169-1219
ACTINOMADURA					107	1.0	4.8	3.5-6.2	107	1.0	4.8	3.5-6.2

Hb+No			104	1.0	654	636-671	104	1.0	654	636-672
NONOMURAEA			102	1.0	241	231-250	102	1.0	240	232-250
NONOMURAEA			100	1.0	194	186-202	100	1.0	194	186-202
No			98	1.0	83	78-89	98	1.0	83	77-89
St			93	1.0	683	669-697	93	1.0	683	670-698
St			92	1.0	119	114-125	92	1.0	119	114-125
St			90	1.0	100	95-105	90	1.0	110	95-105
St			89	1.0	19	16-22	89	1.0	18.9	16.4-21.7
St			86	1.0	17	16-20	86	1.0	17.8	15.8-20.0
St			84	1.0	12.5	10.8-14.2	84	1.0	12.6	10.9-14.3
St			82	1.0	4.3	3.1-5.7	82	1.0	4.34	3.10-5.70
St			79	1.0	579	567-591	79	1.0	578	567-591
St			77	1.0	513	502-524	77	1.0	513	503-524
St			76	1.0	3.4	2.3-4.6	76	1.0	3.4	2.31-4.63
St			73	1.0	424	414-433	73	1.0	424	503-524
St			71	1.0	366	358-374	71	1.0	366	358-374
St			70	1.0	156	150-163	70	1.0	156	150-163
St			69	1.0	1.67	0.87-2.57	69	1.0	1.67	0.88-2.57
St			66	1.0	114	108-120	66	1.0	114	109-120
St			63	1.0	292	285-299	63	1.0	292	285-300
St			61	1.0	241	235-248	61	1.0	241	235-248
St			60	1.0	192	185-199	60	1.0	192	185-199
St			58	1.0	114	109-120	58	1.0	114	109-120
St			55	1.0	175	169-183	55	1.0	175	168-182
St			53	1.0	38	36-43	53	1.0	39.8	36.2-43.3
St			51	1.0	12.5	10.7-14.3	51	1.0	12.5	10.7-14.3

Supplementary Table 3-2 (Continued): Time tree node details. All node heights provided as Millions of Years (Ma). Tree names as in main text.

	Battistuzzi *et al.* (2004)		McDonald and Currie (2017)		stree06 WAG + urc + BD				stree09 WAG + urc + Coal			
Taxa at Node	Median	CI	Median	CI	Node	Posterior	Median	95% HPD	Node	Posterior	Median	95% HPD
root	3977	3434-4464	N/A	3500-3800	116	1.0	3714	3500-3968	116	1.0	3741	3500-3972
EC+SALM	102	57-176	87	45-130	115	1.0	92.9	51.0-158	115	1.0	101	51.0-161
FIRM+ACTINO	3051	2738-3434	3452	1806-5121	112	1.0	3401	3301-3434	112	1.0	3400	3300-3434
ACTINOBACTERIA	2743	2512-3076	2578	1345-3811	110	1.0	2537	2512-2615	110	1.0	2538	2512-2616
STREPTOMYCES	1375	1032-1727	382	250-514	95	1.0	1002	625-1419	46	1.0	1049	766-1408
Mc+Ac+Kb+Amy					48	1.0	1988	1293-2358	109	1.0	1869	1061-2365
Mc+Ac					47	1.0	731	284-1185	108	1.0	665	346-1102
ACTINOPLANES					45	1.0	404	143-726	106	1.0	366	169-614
Ac					43	1.0	224	69.0-485	104	1.0	218	95.1-387
Amy+Kb					40	1.0	703	411-1197	101	1.0	794	377-1261
KIBDELLOSPORANGIUM					39	1.0	1.48	0.32-3.58	100	1.0	1.58	0.36-3.93
AMYCOLATOPSIS					36	1.0	303	179-518	97	1.0	354	207-655
Amy					35	1.0	1.68	0.43-4.00	96	1.0	1.85	0.45-4.46
Amy					32	1.0	140	87.0-209	92	1.0	152	92.3-227
Amy					26	1.0	63	57.0-127	91	1.0	115	75.5-168
					30	1.0	106	87.0-209	85	1.0	94.3	63.7-139
					29	1.0	63	27.8-109	84	1.0	58.6	26.3-99.4
Amy					25	1.0	21.9	8.96-41.3	90	1.0	26.3	9.11-55.1
Amy					23	1.0	6.18	1.94-13.1	88	1.0	7.27	1.99-16.1
					--	--	--	--	81	1.0	69.6	46.4-101
Amy					--	--	--	--	--	--	--	--
Amy					20	1.0	67.9	43.6-98.0	79	1.0	59	38.8-87.2
Amy					--	--	--	--	--	--	--	--
					--	--	--	--	78	1.0	40	22.6-62.5
Amy					19	1.0	31.3	10.8-57.9	76	1.0	18.7	7.76-33.9
Amy					14	1.0	39.7	23.6-58.6	--	--	--	--
Amy					16	1.0	53.9	33.4-78.1	--	--	--	--
Amy					12	1.0	27	16.1-40.6	73	1.0	35.8	20.8-57.8
Amy					11	1.0	12	5.67-20.6	72	1.0	13.7	6.07-25.4
Amy					9	1.0	7.21	2.59-13.9	70	1.0	7.93	2.76-16.5
Amy					6	1.0	19.6	10.6-31.4	67	1.0	25.3	12.4-42.1
Amy					4	1.0	8.88	3.71-16.4	65	1.0	10.4	3.27-20.3
Amy					2	1.0	0.03	0.01-0.75	63	1.0	0.26	0.01-0.84
No+St					109	1.0	1913	1311-2407	60	1.0	1864	1264-2402
Acm+Hb+No					108	1.0	964	527-1554	59	1.0	904	438-1468
ACTINOMADURA					107	1.0	4.09	1.15-9.30	58	1.0	4.45	1.17-10.4

Hb+No			104	1.0	504	261-883	55	1.0	456	272-816
NONOMURAEA			100	1.0	140	67.1-235	53	1.0	202	123-304
NONOMURAEA			102	1.0	202	106-330	51	1.0	141	81.4-222
No			98	1.0	60	21.8-116	49	1.0	61.6	26.4-114
St			93	1.0	718	480-965	44	1.0	740	518-1013
St			92	1.0	119	60.6-206	43	1.0	126	67.9-209
St			90	1.0	83.7	41.7-150	41	1.0	87.3	48.6-142
St			89	1.0	16.9	5.38-34.9	40	1.0	18.4	5.2-40.6
St			86	1.0	18.5	9.01-32.2	37	1.0	20.1	9.87-35.9
St			84	1.0	10.5	4.18-18.7	35	1.0	11.4	4.87-21.4
St			82	1.0	3.52	0.98-7.62	33	1.0	3.8	1.05-8.67
St			79	1.0	596	480-965	30	1.0	616	435-862
St			77	1.0	503	333-678	28	1.0	519	361-706
St			76	1.0	3.08	0.78-7.13	27	1.0	3.41	0.85-8.21
St			73	1.0	393	267-526	24	1.0	406	288-570
St			71	1.0	318	211-436	22	1.0	327	220-473
St			70	1.0	135	70.3-226	21	1.0	142	36.6-153
St			69	1.0	1.3	0.29-3.16	20	1.0	1.38	0.27-3.45
St			66	1.0	80.1	31.8-151	17	1.0	81.8	36.6-153
St			63	1.0	245	161-345	14	1.0	253	167-343
St			61	1.0	190	120-273	12	1.0	195	131-268
St			60	1.0	135	76.0-200	11	1.0	141	87.6-205
St			58	1.0	72.4	37.4-125	9	1.0	75.8	36.5-121
St			55	1.0	125	66.7-199	6	1.0	132	76.1-203
St			53	1.0	28.8	13.5-46.6	4	1.0	33.1	18.0-52.6
St			51	1.0	10.8	4.02-20.3	2	1.0	11.9	4.02-23.4

Supplementary Table 3-3: Kendall-Colijn distance between species trees. Tree names as in main text.

tree	WAG + strict + BD	WAG + strict + Coal	WAG + urc + BD
WAG + strict + coal	0.00	N/A	N/A
WAG + urc + BD	10.95	10.95	N/A
WAG + urc + coal	9.00	9.00	11.62

Supplementary Table 3-4: Normalized Robinson-Foulds distance between species trees. Tree names as in main text.

tree	WAG + strict + BD	WAG + strict + Coal	WAG + urc + BD
WAG + strict + coal	0.00	N/A	N/A
WAG + urc + BD	0.09	0.09	N/A
WAG + urc + coal	0.09	0.09	0.05

Supplementary Table 3-5: Node reconciliation details. All node heights provided as Millions of Years (Ma). Tree names as in main text.

Class	Family	g_{node}	g_{node} support	s_{node}	s_{node} support	median	s_{node} 95% HPD	g_{parent}	g_{parent} support	s_{parent}	s_{parent} support	median	s_{parent} 95% HPD	Rec. event	Rec. support	combined HPD range
Precursor	DAHPS I	44	1.00	48	1.00	1988	1293-2358	122	N/A	110	1.00	2537	2512-2615	S	1.00	1293-2615
	DAHPS II	29	0.97	45	1.00	404	143-726	30	0.47	104	1.00	504	261-883	T	0.33	143-883
	PDH St	103	0.77	110	1.00	2537	2512-2615	104	0.49	112	1.00	3401	3301-3434	S	1.00	2512-3434
	PDH No+Amy	227	0.99	102	1.00	202	106-330	228	1.00	104	1.00	504	261-883	S	0.31	106-883
	CM I	45	1.00	95	1.00	1002	625-1419	46	0.35	109	1.00	1913	1311-2407	S	1.00	625-2407
	CM II	236	0.64	110	1.00	2537	2512-2615	N/A	N/A	112	1.00	3401	3301-3434	S	1.00	2512-3434
	HmaS	100	0.97	110	1.00	2537	2512-2615	N/A	N/A	112	1.00	3401	3301-3434	S	1.00	2512-3434
	Hmo	102	0.87	110	1.00	2537	2512-2615	N/A	N/A	112	1.00	3401	3301-3434	S	1.00	2512-3434
	HpgT	104	0.90	71	1.00	318	211-436	N/A	N/A	73	1.00	393	267-526	S	1.00	211-526
	DpgA	92	0.89	109	1.00	1913	1311-2407	N/A	N/A	110	1.00	2537	2512-2615	S	1.00	1311-2615
	DpgC	94	0.94	109	1.00	1913	1311-2407	N/A	N/A	110	1.00	2537	2512-2615	S	1.00	1311-2615
	DpgB	92	0.80	110	1.00	2537	2512-2615	N/A	N/A	112	1.00	3401	3301-3434	S	1.00	2512-3434
	DpgD	66	0.67	110	1.00	2537	2512-2615	N/A	N/A	112	1.00	3401	3301-3434	S	1.00	2512-3434
	OxyD	44	0.90	40	1.00	703	411-1197	N/A	N/A	48	1.00	1988	1293-2358	S	1.00	411-2358
	BpsD	44	0.96	40	1.00	703	411-1197	N/A	N/A	48	1.00	1988	1293-2358	S	1.00	411-2358
	Bhp	44	0.28	36	1.00	303	179-518	N/A	N/A	40	1.00	703	411-1197	S	1.00	179-1197
	BOH	20	1.00	110	1.00	2537	2512-2615	N/A	N/A	112	1.00	3401	3301-3434	S	1.00	2512-3434
	EvaA	40	0.76	48	1.00	1988	1293-2358	N/A	N/A	110	1.00	2537	2512-2615	S	1.00	1293-2615
	EvaB	40	0.81	48	1.00	1988	1293-2358	N/A	N/A	110	1.00	2537	2512-2615	S	1.00	1293-2615
	EvaC	20	0.18	32	1.00	140	87-209	N/A	N/A	36	1.00	303	179-518	S	1.00	87-518
	EvaD	40	0.95	45	1.00	404	143-726	N/A	N/A	47	1.00	731	284-1185	S	1.00	143-1185
	KR	32	0.96	20	1.00	67.9	43.6-98	N/A	N/A	26	1.00	63	57.0-127	S	1.00	43.6-127
NRPS	A mod1	509	0.98	109	1.00	1913	1311-2407	510	0.98	110	1.00	2537	2512-2615	S	0.98	1311-2615
	A mod2	727	1.00	45	1.00	404	143-726	728	0.98	110	1.00	2537	2512-2615	S	1.00	143-2615
	A mod3	451	0.99	108	1.00	964	527-1554	511	1.00	109	1.00	1913	1311-2407	S	0.98	527-2407
	A pek3	616	1.00	29_	1.00	0	0-0	617	0.87	39	1.00	1.48	0.32-3.58	T	0.25	0-3.58
	A van3	1002	1.00	40	1.00	703	411-1197	1003	0.96	S67->47	1.00	731	284-1185	S	1.00	411-1185
	A mod4	414	1.00	45	1.00	404	143-726	415	0.95	104	1.00	504	261-883	T	0.53	143-883
	A mod5	259	1.00	45	1.00	404	143-726	513	0.91	47	1.00	731	284-1185	S	1.00	143-1185

	A mod6	787	1.00	102	1.00	202	106-330	788	0.52	104	1.00	504	261-883	D	0.62	106-883
	A mod7	195	1.00	40	1.00	703	411-1197	515	0.37	48	1.00	1988	1293-2358	S	0.95	411-2358
	Cond1-2	698	1.00	102	1.00	202	106-330	699	0.95	S10->45	1.00	404	143-726	T	0.15	106-726
	Cond2-3	877	0.99	60	1.00	135	76-200	878	1.00	61	1.00	190	120-273	S	0.92	76-273
	van Cond2-3	20	0.97	40	1.00	703	411-1197	276	0.88	48	1.00	1988	1293-2358	S	0.60	411-2358
	Cond 3-4	398	1.00	45	1.00	404	143-726	399	0.79	47	1.00	731	284-1185	S	1.00	143-1185
	Cond 4-5	777	0.96	45	1.00	404	143-726	778	0.99	47	1.00	731	284-1185	S	1.00	143-1185
	Cond 5-6	598	0.06	102	1.00	202	106-330	599	1.00	104	1.00	504	261-883	T	0.30	106-883
	Cond 6-7	498	1.00	102	1.00	202	106-330	499	0.95	104	1.00	504	261-883	D	1.00	106-883
Tailoring	NMT	24	1.00	40	1.00	703	411-1197	N/A	N/A	45	1.00	404	143-726	S	1.00	411-726
	gtf root	164	0.98	112	1.00	3401	3301-3434	N/A	N/A	116	1.00	3714	3500-3968	S	1.00	3301-3968
	GtfB GtfE	159	0.42	43	1.00	224	69.0-485	160	0.91	45	1.00	404	143-726	T	0.93	69-726
	gtf Rha	97	0.86	19	1.00	31.3	10.8-57.9	98	0.96	20	1.00	67.9	43.6-98	T	0.44	10.8-98
	GtfC	161	0.98	45	1.00	404	143-726	162	0.90	47	1.00	731	284-1185	D	0.93	143-1185
	GtfD	161	0.98	45	1.00	404	143-726	162	0.90	47	1.00	731	284-1185	D	0.93	143-1185
	GtfA	78	0.99	14	1.00	39.7	23.6-58.6	98	0.90	20	1.00	67.9	43.6-98	T	0.42	23.6-98
	Hal	72	0.89	47	1.00	731	284-1185	N/A	N/A	48	1.00	1988	1293-2358	S	1.00	284-2358
	Hal comp	16	0.89	53	1.00	28.8	13.5-46.6	17	0.94	S67->47	1.00	731	284-1185	T	1.00	13.5-1185
	Hal GPA	71	0.89	45	1.00	404	143-726	72	0.89	47	1.00	731	284-1185	S	1.00	143-1185
	P450 root	348	0.28	110	1.00	2537	2512-2615	N/A	N/A	112	1.00	3401	3301-3434	S	1.00	2512-3434
	OxyA	59	1.00	100	1.00	140	67.1-235	95	0.72	102	1.00	202	106-330	S	0.60	67.1-330
	OxyB	283	1.00	102	1.00	202	106-330	284	1.00	104	1.00	504	261-883	S	1.00	106-883
	OxyC	345	N/A	45	1.00	404	143-726	346	1.00	47	1.00	731	284-1185	S	1.00	143-1185
	OxyE	94	1.00	28	1.00	0	0-0	95	0.72	102	1.00	202	106-330	T	0.10	0-330
	OxyF	141	0.96	95	1.00	1002	625-1419	142	0.28	109	1.00	1913	1311-2407	S	1.00	625-2407
	OxyH	218	0.78	79	1.00	596	480-965	286	0.82	93	1.00	718	480-965	S	1.00	480-965
	OxyI	223	1.00	51	1.00	10.8	4.02-20.3	285	0.99	107	1.00	4.09	1.15-9.30	S	1.00	4.02-9.30
	deAc	50	0.99	102	1.00	202	106-330	N/A	N/A	104	1.00	504	261-883	S	1.00	106-883
	AcylTfII	10	1.00	102	1.00	202	106-330	N/A	N/A	104	1.00	504	261-883	S	1.00	106-883
Other	ABC ATP root	196	0.78	109	1.00	1913	1311-2407	N/A	N/A	110	1.00	2537	2512-2615	S	1.00	1311-2615
	ABC ATP GPA	193	0.98	77	1.00	503	333-678	194	0.99	79	1.00	596	480-965	S	1.00	333-965
	StrR	92	0.72	40	1.00	703	411-1197	N/A	N/A	48	1.00	1988	1293-2358	S	0.84	411-2358
	VanH	46	0.81	45	1.00	404	143-726	N/A	N/A	47	1.00	731	284-1185	S	1.00	143-1185

VanA	42	1.00	32	1.00	140	87-209	N/A	N/A	36	1.00	303	179-518	S	1.00	87-518
VanX	42	0.89	45	1.00	404	143-726	N/A	N/A	47	1.00	731	284-1185	S	1.00	143-1185
VanY	6	1.00	01_	1.00	0	0-0	N/A	N/A	98	1.00	60	21.8-116	T	0.10	0-116
RespReg root	92	0.95	109	1.00	1913	1311-2407	N/A	N/A	110	1.00	2537	2512-2615	S	1.00	1311-2615
RespReg VanR	89	0.77	45	1.00	404	143-726	90	1.00	47	1.00	731	284-1185	S	0.95	143-1185
HisK root	90	0.88	109	1.00	1913	1311-2407	N/A	N/A	110	1.00	2537	2512-2615	D	1.00	1311-2615
Hisk VanS	81	1.00	45	1.00	404	143-726	82	0.21	47	1.00	731	284-1185	S	1.00	143-1185

CHAPTER FOUR: Phylogenetics predict diverse members of the glycopeptide biosynthetic gene cluster family

CHAPTER FOUR PREFACE

Portions of the work presented in this chapter have been published as:

Culp EJ, Waglechner N, Wang W, Fiebig-Comyn AA, Hsu Y, Koteva K, Sychantha D, Coombes BK, Van Nieuwenhze MS, Brun Y, Wright GD. 2020. Evolution-guided discovery of antibiotics that inhibit peptidoglycan remodelling. *Nature* **578**:582-587.

Author Contributions

EJC, NW and GDW conceived this study and designed experiments. NW performed phylogenetic analysis, structural predictions, initial determination of MIC values and mutant gene analysis. WW fermented strains, extracted performed structural elucidation of complestatin and corbomycin. EJC generated resistant mutants, prepared nucleic acids for sequencing, and confirmed identified mutations with conventional sequencing. All other analysis in this chapter was performed by NW. EJC and GDW wrote the published manuscript. The methods for experiments performed by EJC were written by EJC and adapted by NW from the published manuscript, see Appendix 2.

Acknowledgements

This research was funded by a Canadian Institutes of Health Research grant (FRN-148463), the Ontario Research Fund, and by a Canada Research Chair to G.D.W.; the work was also supported by National Institute of Health grants R35GM122556 to Y.V.B. and 5R01GM113172 to M.S.V. and Y.V.B., and a Canada 150 Research Chair in Bacterial Cell Biology to Y.V.B. E.J.C. was supported by a CIHR Vanier Canada Graduate Scholarship. N.W. was supported by a CIHR Canada Graduate Scholarship Doctoral Award.

ABSTRACT

Amongst a set of 71 previously identified GPA biosynthetic gene clusters were several encoding scaffolds distinct from the traditional D-Ala-D-Ala binding glycopeptides. Some heptapeptides are previously known, such as complestatin and kistamicin, and some encode longer peptide scaffolds like enduracidin and ramoplanin with different mechanisms of action. We have produced an expanded phylogenetic analysis of these BGCs and develop a new classification scheme to organize these scaffolds into four major types based on their apparent evolutionary relationships. This scheme and analysis of these BGCs allows us to make putative structural predictions for novel compound encoding by strains in our strain collection. These predictions allowed us to purify a GPA from our strain collection expected to be complestatin, and another compound with a novel scaffold we have named corbomycin. Both compounds possess antibacterial activity. Indicator strains raised to be resistant to these compounds suggest a novel mechanism of action distinct from the known or proposed mechanism of action of other compounds in the extended GPA-family.

INTRODUCTION

In Chapter 3, a phylogenetic history of glycopeptide antibiotics was presented. 71 biosynthetic gene clusters were identified from pure bacterial isolates and from environmental DNA (Waglechner, McArthur, & Wright, 2019). The significance of GPAs lies in their clinical history as drugs of last resort, once reserved for severe infections (Demain, 2009) but also in the fact that an unprecedented and growing number of BGCs have been sequenced from several genera of bacteria including *Micromonospora, Actinomadura, Actinoplanes, Nonomuraea, Amycolatopsis, Herbidospora, Kibdellosporangium,* and *Streptomyces* (Doroghazi et al., 2014; Doroghazi & Metcalf, 2013; Ju et al., 2015). The breadth of organisms to which these BGCs have spread is comparable to the chemical diversity that can be observed in the known compounds produced by these BGCs. While the evolutionary history of a subset of these compounds has been explored in detail, how these features connect to the overall sequence diversity of the entire family has not been presented.

Evolutionary analysis of these BGCs is challenging because they are expected to change at multiple levels of organization through a diverse set of mechanisms that have not been integrated in single framework. BGCs are organized in a modular fashion, and the GPA BGCs are no exception (van Wageningen et al., 1998; Yim et al., 2016). At the highest level of organization, multiple genera produce compounds in this family. Abstracting to this level would focus on questions of why only some of the members of these genera produce GPAs and not all, and what forces are responsible for the gain of these BGCs and what prevents the loss of all copies from these lineages. This may be best modeled as

the dynamics of gain and loss of binary traits on a species phylogeny (Kannan, Li, Rubinstein, & Mushegian, 2013). Abstracting to the level of the individual BGC it is possible to ask questions about the changing composition of BGCs. This again may be modeled as the dynamics of an unordered collection of binary traits representing BGC components (Cohen & Pupko, 2011), or one can use techniques such as the double-cut-join to measure the distance between two ordered collections (Fertin, Jean, & Tannier, 2017). Distances or similarity measures between BGCs (such as the Jaccard distance, see chapter 2) (Lin et al., 2006) are useful for pairwise comparisons in an analogy to BLAST searching of nucleotide and protein sequences and have many potential applications to BGC databases (B. Liu & Pop, 2009). Matrices of distances between a collection of BGCs calculated in this manner may be used as input to well-known methods for phylogenetic tree inference, but have equally well-known drawbacks that can lead to incorrect tree inference (Joseph Felsenstein, 2003; John P. Huelsenbeck, 1995). Questions have been raised about whether tree-like representations of these dynamics are even possible or desired (Kunin, Goldovsky, Darzentas, & Ouzounis, 2005; C. Liu, Wright, Allen-Vercoe, Gu, & Beiko, 2018).

At lower levels of organization, we can abstract to gene and domain sequences. Multi-domain genes are subject to the same limitations encountered when studying BGC compositions, namely that complicated series of events are potentially needed to explain how multi-domain NRPS sequence can transform into each other. This has been a barrier to studying the evolution of eukaryotic proteins, where the dynamics of duplication, loss, and transfer are known to be important mechanisms of protein evolution (Bansal, Kellis,

Kordi, & Kundu, 2018; Ravenhall, Skunca, Lassalle, & Dessimoz, 2015; Stolzer et al., 2015). At the lowest organizational level, sequences of whole or partial genes may be aligned at the protein and/or nucleotide level and subjected to probabilistic phylogenetic analysis (Bayesian or maximum likelihood) using standard techniques describing molecular sequence evolution (Drummond & Rambaut, 2007; J. Felsenstein, 1981; J. P. Huelsenbeck & Ronquist, 2001). At the sequence level various strong assumptions are made about the nature, source, and informational content of gaps or insertions (indels), and about the independence of positions in the sequence (Joseph Felsenstein, 2003; Maiolo, Zhang, Gil, & Anisimova, 2018; Nasrallah, Mathews, & Huelsenbeck, 2011). These features of molecular evolution are often ignored in favour of more tractable mathematics.

Lacking an integrated and rigorous framework that accounts for evolution at all organizational levels, a major reason the D-Ala-D-Ala binding GPAs are attractive targets for phylogenetic reconciliation is that the NRPS domains which code for and are predictive of the basic structure of these compounds in their BGC sequences, have single robustly supported common ancestors (Waglechner et al., 2019) (Chapter 3). This made it possible to reconcile and confidently date their evolutionary emergence in terms of taxonomic location and age with respect to a dated species phylogeny. The remaining BGCs in this extended family are primarily found in the genus *Streptomyces*, which is known to be especially diverse among the other bacterial genera producing GPAs (Chevrette et al., 2019; McDonald & Currie, 2017). This genetic and phenotypic diversity was observed as increased branch lengths in the *Streptomyces* clade of the previously

produced species tree, leading attempts to reconcile gene families and produce concise reconciliation dates to be more difficult than for the traditional GPA BGC sequences.

The composition of the D-Ala-D-Ala binding BGCs distributed among several genera are also more complex. Precursor biosynthesis for non-proteinogenic amino acids 4-hydroxyphenylglycine (Hpg) and 3,5-dihydroxyphenylglycine (Dpg) is more uniform in *Streptomyces*, while β-hydroxytyrosine (Bht) and aminosugar biosynthesis is absent in these uncharacterized *Streptomyces*-associated GPA-like BGCs, leaving none, few, or unsuitable genes as phylogenetic markers. There are no obvious resistance enzymes located in these GPA-like BGCs which puts nearly all the burden for distinguishing these clusters on the composition of their NRPS sequences and the complement of tailoring enzymes, if present.

We wished develop a meaningful classification scheme to categorize these BGCs based on the phylogenetic analysis of their BGC components. The goal of such a scheme should be to harmonize the sequence information with the chemical structure and biological activity of each BGC product in a way that best explains the evolution of these compounds. This synthesis extends the previous use of marker sequences to map out the extended GPA BGC family and prioritizes which BGCs and compounds are more likely to be novel (Thaker et al., 2013). Such a system should ideally assist in compound purification by predicting the structure of the putative products of these clusters where it is not known if a specific molecule has biological activity.

METHODS

GPA-like BGC sequences

The 71 GPA and GPA-like BGC sequences identified in Chapter 3 were used for this study (Waglechner et al., 2019). Briefly, a GPA or GPA-like BGC is one that has been previously published as a GPA BGC, or possesses a sequence with similarity to one of the GPA fingerprint genes (*oxyB, halI, oxyC, dpgC,* and *oxyE*) via BLAST (at least 75% length, e-value < $1.0x10^{-5}$) (Thaker et al., 2013). Putative BGCs are then analysed with antiSMASH and manually inspected for the following characteristics: NRPS scaffold biosynthesis, a scaffold that contains one or both of the amino acids Hpg and Dpg along with their biosynthetic enzymes, glycopeptide-type condensation domains ('Cglyc' in antiSMASH (Medema et al., 2011), Supplementary Table III), one or more tailoring and/or crosslinking enzymes such as halogenases, glycosyltransferases, methyltransferases, cytochrome P450 monooxygenases, acyltransferases. BGCs sequences were identified from the literature, from public sequence databases, and from genome sequences produced from strains in our in-house strain collection.

GPA BGC analysis

The set of GPA and GPA-like BGCs were subjected to analysis by the python module evoc (Waglechner et al., 2019) (Chapter 3). ORF sequences from these BGCs were divided into single- (gene) and multi-domain sequences (domains), as previously described. These sequences were subjected to a clustering step using USEARCH (at least 60% identity) (Edgar, 2010). For clarity, we use the term cluster to refer to BGCs, and family to refer to the clustered gene/domain families. Each family can be described by its

centroid sequence, a centrally representative sequence of the entire family. The centroid sequences were used to annotate each family, and these annotations were used as labels to describe each family. These family trees together with sequence and chemical structure comparison with known GPA BGCS and compounds were used to prepare initial predictions of the molecular structures of products of WAC01325 and WAC01529.

Phylogenetic trees

Gene/domain families having at least three members were aligned using MUSCLE, each alignment was manually inspected for the presence of excessive gaps (Edgar, 2004). 92 of these aligned families were subjected to phylogenetic analysis using fasttree 2 using the WAG substitution model, and the default number of categories for the CAT rate approximation (Price et al., 2010; Whelan & Goldman, 2001).

Growth of organisms

Streptomyces sp. WAC1325 and *Streptomyces* sp. WAC01529 spores stored at -80°C were streaked out on Bennett's agar (1% potato starch, 0.2% casamino acids, 0.18% yeast extract, 0.02% KCl, 0.02% $MgSO_4 \cdot 7H_2O$, 0.024% $NaNO_3$, $4x10^{-4}$ % $FeSO_4 \cdot 7H_2O$) and incubated for 7 days at 30°C. Seed cultures were prepared from colonies by sub-culture into 50mL of Trypic Soy Broth (BD biosciences) and incubated for 5 days under 250rpm shaking at 30°C.

WAC1529 mycelium from 50 mL TSB seed culture was inoculated into each of eighteen 2.8 L flasks containing 600 mL Bennett's media. After 4 days, fermentations were fed with 0.2 mM each cysteine, histidine, glutamine and tyrosine. Amino acid supplements

were prepared as 100x stocks in water (Cys, His, Glu) or 10 mM HCl (Tyr) and neutralized with two equivalents of $NaHCO_3$ after addition to fermentations.

Spent media was extracted with 8% (W/V) HP-20 (Diaion) resin. Cell pellets were extracted twice with 500 mL methanol (MeOH) and was concentrated under vacuum with 100 g HP-20 (Diaion) resin. These resins were combined and eluted with H_2O (2 L), 20% MeOH (2 L), 40% MeOH (2 L), and 100% MeOH (4 L). Analysis of fractions by HPLC and liquid chromatography-mass spectrometry (LC-MS) identified a peak with molecular weight (~1300-1600 Da) and UV-profile maxima (220 nm, 280 nm) consistent with the predicted structure. This fraction (100% MeOH) was extracted with ethyl acetate, MeOH/H_2O (1:4) and DMSO. The DMSO subfraction was found to contain the predicted glycopeptide and was applied to reverse-phase CombiFlash ISCO (RediSep Rf C18, Teledyne) and eluted with a linear gradient system (5-100% water/acetonitrile, 0.1% formic acid) to give 136 fractions. Fractions containing the predicted glycopeptide were combined and subject to Sephadex LH-20 column (400 mL), eluting with MeOH/Acetonitrile/H_2O (1:2:1), to yield 36 subfractions. Identified subfractions were combined and further purified with Agilent Eclipse XDB-C8 column (5 μm, 9.4 × 250 mm), to yield 23.6 mg of corbomycin.

Fermentation and purification of complestatin

WAC01325 was fermented using conditions identical to WAC01529 except for amino acid feeding. After 3 days growth at 30°C, 250 rpm, fermentations were fed with 0.2 mM each 4-hydroxyphenylglycine, tryptophan and tyrosine. Amino acids were prepared as a 100x stock solution in 10 mM HCl and neutralized with two equivalents of $NaHCO_3$ after

addition to fermentations. Growth was allowed to continue for a total of 8 days and complestatin was purified from spent media and cell pellet in a similar manner to corbomycin. High resolution-electrospray ionisation-mass spectrometry (HR-ESI-MS) and ^{1}H- nuclear magnetic resonance (NMR) confirmed the compound as complestatin.

Initial MIC determination

MICs for triclosan (Sigma), complestatin and corbomycin were determined following the broth microdilution method in Mueller Hinton broth at 37°C, except for *Enterococcus* where Brain Heart Infusion media (BD Biosciences) was used.

Raising resistance mutants through serial passage in the presence of antibiotic

Complestatin and corbomycin resistant mutants were raised beginning with the laboratory strain *B. subtilis* 168 or *S. aureus* ATCC 29213. To begin, a single colony was inoculated into 1 mL MHB in a sterile test tube with 0.25xMIC, 0.5xMIC, 1xMIC and 2xMIC where MIC = 1 µg/mL for both complestatin and corbomycin, and 0.0625 µg/mL for rifampicin. After 24 hr growth with shaking, the lowest concentration with no growth was taken as the new MIC, and cells were subcultured into fresh tubes 1 in 100 from the highest concentration that supported growth. This process was continued for 25 days, and glycerol stocks were taken whenever there was a shift in MIC. At the end of 25 days, glycerol stocks were streaked on non-selective media (Mueller Hinton agar) and single colonies were isolated for two generations. The MIC of purified strains was measured by microbroth dilution. Serial passaging was performed in biological duplicate using two independent lines.

For one line of the serially passaged cells, whole genome sequencing was performed on the strain on the final day (COM25 and COR25), or at the earliest time point that the highest MIC was reached. For corbomycin (WAC01529), this strain arose at day 14 (*B. subtilis* COR14), and for complestatin, day 20 (*B. subtilis* COM20). Whole genome sequencing on resistant mutants, as well as our laboratory *B. subtilis* 168, was performed with Illumina MiSeq (300 bp, paired end reads) by the Farncombe Genomics Facility (McMaster University). To identify mutations unique to our evolved mutants versus wildtype, each of the three sequenced strains were compared to the published *B. subtilis* 168 reference genome (accession number AL009126.3) using breseq (version 0.33.1) (Deatherage & Barrick, 2014) to generate a list of differences. Changes in protein coding regions that were unique to resistant mutants and not present in our laboratory *B. subtilis* 168 strain were identified for follow up. Sequencing two individual colonies isolated from day 25 gave identical genotypes.

RESULTS

Several distinct scaffold types are encoded by GPA-like BGCs in Actinobacteria

The work of Waglechner *et al* (2019) (Chapter 3) focused on the analysis of 71 BGCs related to GPAs, particularly the evolution of the D-Ala-D-Ala-binding GPAs. Unlike these GPAs, the potential evolutionary relationships between the extended GPA-like BGCs appears to be more complex. There are no universal patterns that can point to clear ancestral relationships between the various scaffold components. Analysis of NRPS components can help to determine the basic scaffold encoded by a BGC. Consideration of the overall domain architecture of the NRPS sequences in these BGCs allow a preliminary evolutionary classification of these scaffolds. Specifically, we propose a system of four classes (I, II, III and IV) that augments the existing GPA type nomenclature (Nicolaou et al., 1999) based on the compositional and genetic diversity as well as the apparent evolutionary relationships of the expanded set of GPA BGC sequences. An overview of these working criteria is provided in Table 4-1, while details and examples for each BGC class are subsequently discussed. BGCs may first be divided into simple or complex NRPSs classes based whether antiSMASH analysis identifies more than the common NRPS domains: adenylation (A), condensation (C), epimerization (E), peptidyl-carrier protein (or alternately thiolation) (PCP, or T respectively) and thioesterase (TE) domains.

Table 4-1: GPA BGC classification criteria

Class	Subclass	Positive criteria	Negative criteria
GPA		NRPS peptide scaffold incorporating Hpg, Dpg, components more closely related to the exclusion of other NRPS sequences	
I		Linear peptide, variable sequence Variable length, >7 amino acids Highly repetitive NRPS modules, putatively duplicated (shared MRCA)	Cyclization (peptide or depsipeptide)
II		> 10 amino acids Cyclized depsipeptide, variable sequence Repetitive NRPS components/modules, putatively duplicated (share MRCA)	Linear peptide
III		Crosslinked scaffold, presence of NRPS X domain NRPS X-domain shares MRCA with known GPA BGCs ≥ 1 cytochrome P450 monooxygenase sequences in BGC Heptapeptide scaffold ≥ 5 NRPS modules share MRCA with known GPAs Follows Nicoloau (1999) classification Type I-IV	Crosslinked Trp in scaffold
IV		Crosslinked scaffold, presence of NRPS X domain NRPS X-domain shares MRCA with known GPA BGCs ≥ 1 cytochrome P450 monooxygenase in BGC Scaffold Trp participates in crosslinks Presence of putative inactive A-domain after X-domain in final module Variable length peptide, ≥ 7 amino acids	
	a	Heptapeptide scaffold, complestatin-type peptide (Hpg-Trp-Hpg-Hpg-Hpg-Tyr-Hpg) NRPS HPG modules are duplicated Loss of DPG in scaffold and DPG biosynthesis in BGC NRPS components share MRCA	
	b	Variable length peptide, ≥ 7 amino acids Variable peptide sequence at N- and C-terms, shared core (crosslinked) scaffold sequence with corbomycin (X-Dpg-Trp-Hpg-Hpg-Tyr-Tyr-Dpg-X) partly repetitive scaffold, > 1 NRPS modules share MRCA	

	NRPS components share MRCA with corbomycin NRPS
	>2 intramolecular oxidative crosslinks
	acylated scaffold N-term
c	nonapeptide
	shares scaffold sequence with GP6738 (Dpg-Dpg-Val-Trp-Dpg-Hpg-Dpg-Tyr-Dpg)
	partly repetitive scaffold, >1 NRPS modules share MRCA
	NRPS components shares MRCA with GP6738 NRPS
	2 intramolecular oxidative crosslinks

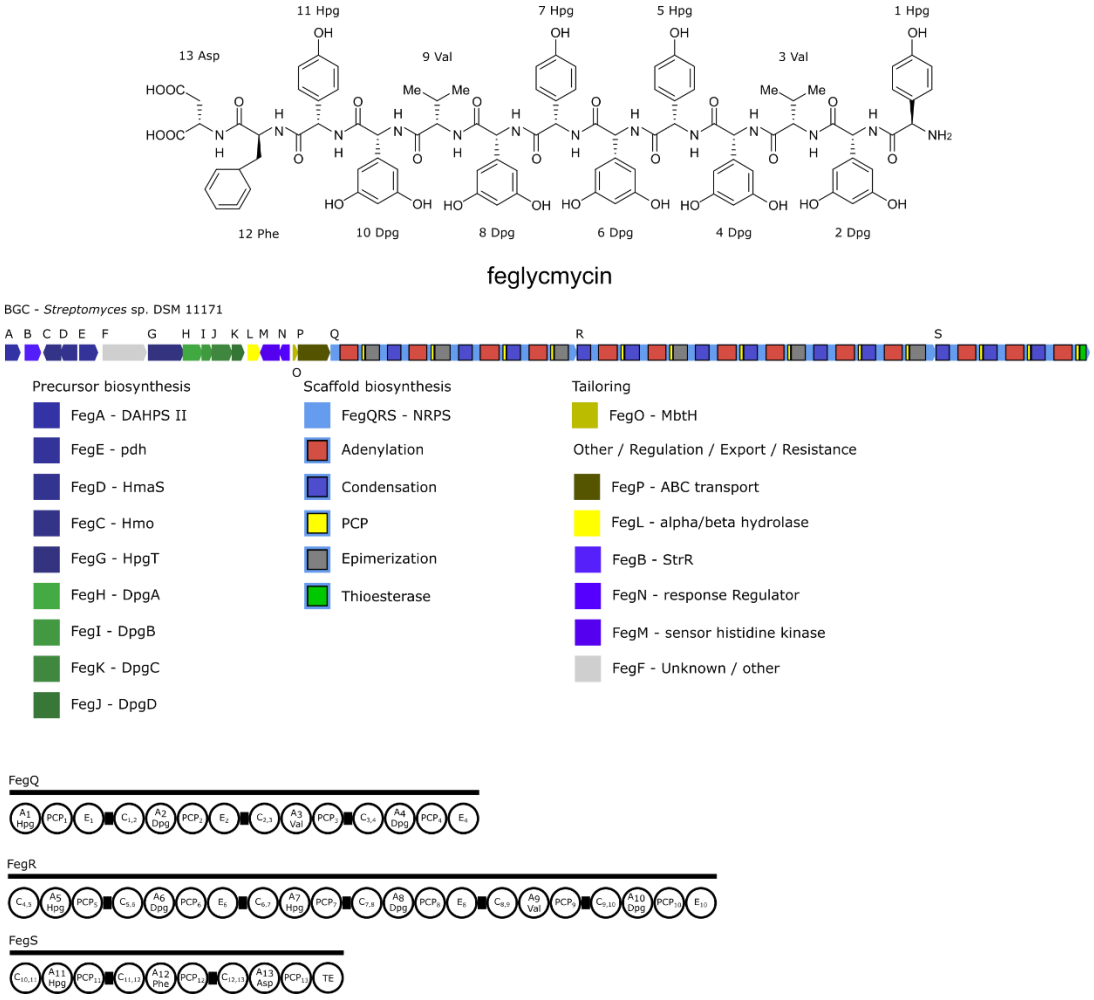


Figure 4-1: Feglymycin structure, BGC, and NRPS configuration. Feglymcin is a 13 amino acid linear peptide with repeated Dpg residues. The BGC includes the necessary precursor biosynthesis genes for Dpg and Hpg, but lacks tailoring and resistance genes.

The simple category is known to produce both linear (class I, example feglymycin (Gonsior et al., 2015), Figure 4-1) and cyclic depsipeptides (class II, examples ramoplanin (Hoertz et al., 2012) and enduracidin (Yin & Zabriskie, 2006), Figure 4-2) scaffolds. Depsipeptide cyclization in class II occurs via hydroxyl groups on the β-

carbons of amino acids, and both of these BGCs have the unusual feature of an incomplete NRPS module in the middle of a polypeptide being supplied in trans.

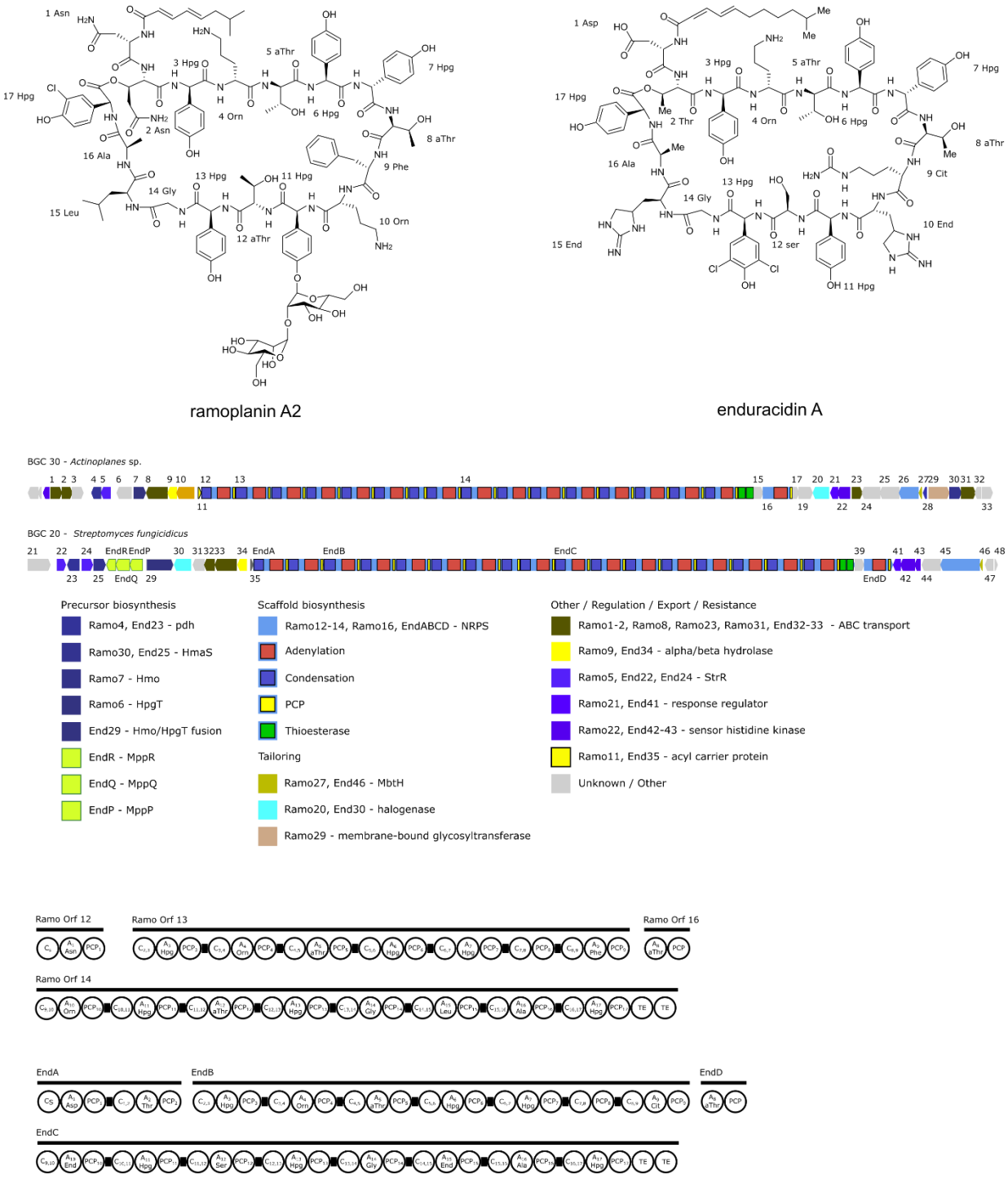


Figure 4-2: Ramoplanin and enduracidin structures, BGCs, and NRPS configurations. These cyclic depsipeptides lack Dpg residues, but do encode the non-proteinogenic amino acids ornithine, citrulline and enduracidin. Both are acylated on the N-terminus but differ in tailoring. Ramo Orf 13 and EndB share structural homology

having an A-domain encoded in trans on Ramo Orf 16 and EndD, respectively, supply allo-Thr during scaffold biosynthesis, however this is not a defining feature of all class II scaffolds.

Complex NRPS sequences include additional domains like the X domain, and inactive A-domains. In the nomenclature of Nicoloau *et al.*, the type I-IV (D-Ala-D-Ala binding GPAs) and type V (Trp-containing) scaffolds are products of the complex category by virtue of possessing an X domain (Nicolaou et al., 1999). All BGCs with an X-domain containing NRPS appear to include one or more cytochrome P450 monooxygenase sequences, suggesting they have one or more intramolecular oxidative crosslinks. The D-Ala-D-Ala binding GPAs are crosslinked heptapeptides (Class III, example vancomycin (Xu et al., 2014) Figure 4-3). We have previously used the term ‘true glycopeptides (Chapter 3) (Demain, 2014) to refer to this subset of the GPAs that bind D-Ala-D-Ala, and for this purpose we might resurrect the name dalbaheptides *sensu* (Parenti & Cavalleri, 1989). Other complex scaffolds are known or predicted to incorporate Trp and, with the exception of kistamicin, they all have an inactive A-domain in their NRPS sequences downstream of the X-domain. This group of scaffolds includes known the heptapeptides complestatin and kistamicin (kistamicin Figure 4-4) (Nazari et al., 2017) and two groups of compounds with longer scaffolds with a distinct distribution in the genera *Streptomyces* (collectively class IV) that will be treated in some detail below.

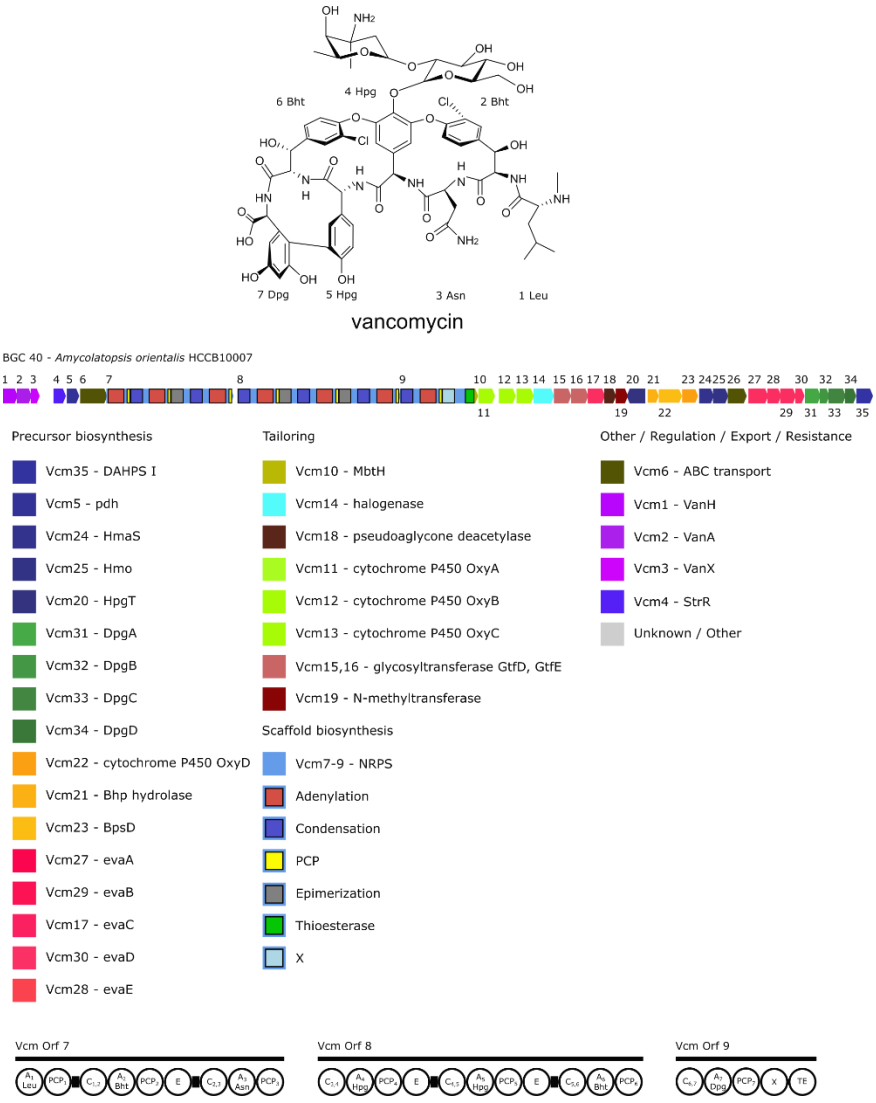


Figure 4-3: Vancomycin structure, BGC, and NRPS configuration. Despite being the archetypal clinical GPA, the BGC was not published until 2011 after many features of GPA biosynthesis were investigated through other clusters. Significantly, the seven NRPS modules have a 3-3-1 configuration which is a derived character from the ancestral 2-1-3-1 configuration observed in GPA BGCs. Unlike Class I and II scaffolds, the NRPS BGCs have an additional X domain in the terminal module which recruits several tailoring cytochrome P450 monooxygenases to install intramolecular bi-aryl and bi-aryl ether crosslinks in these scaffolds.

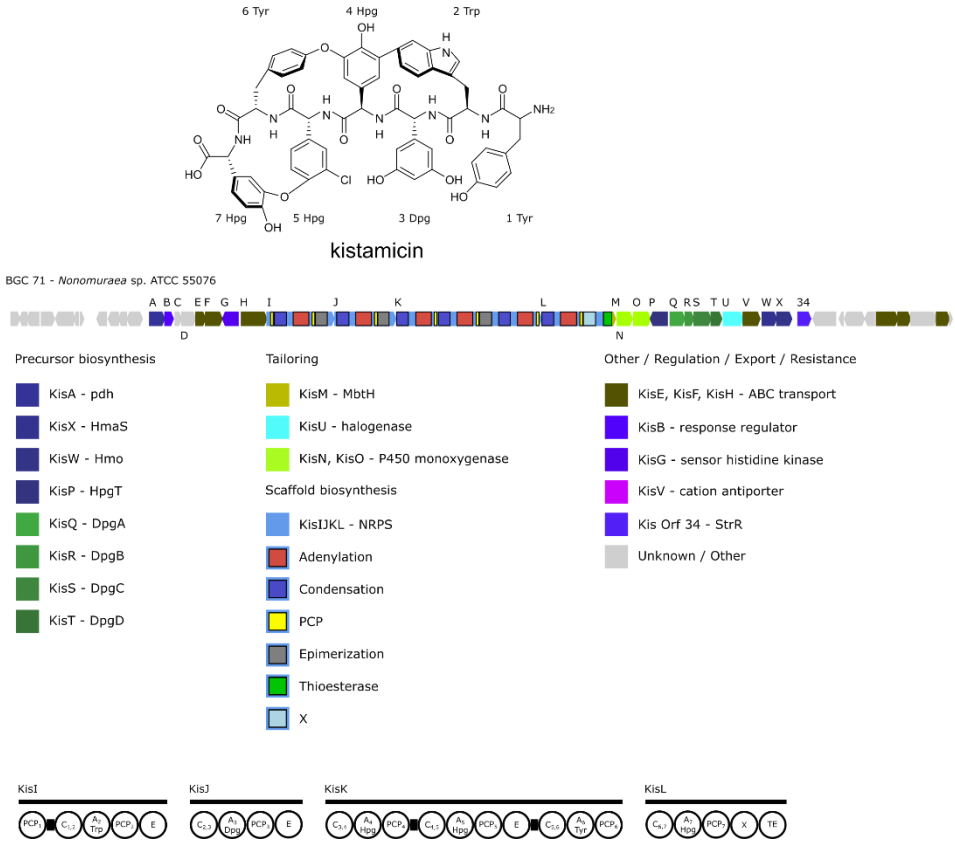


Figure 4-4: Kistamicin structure, BGC, and NRPS configuration. Kistamicin shares a heptapeptide scaffold structure with Class III GPAs with Trp in position 2 leading to its classification by Nicolau *et al* as a Type V GPA. Significantly, the NRPS KisI module 1 is incomplete, leading to the hypothesis that Tyr_1 is supplied to the scaffold in trans via another mechanism. The NRPS KisL possesses an X-domain, however there are only two cytochrome P450 monooxygenases, KisN and KisO, available to install the 3 crosslinks. The specificity of the halogenase, KisU, also differs compared to the Class III scaffolds.

A-domain sequence analysis in conjunction with the collinearity rule have been successfully used to predict the scaffold peptide sequence, and is a key element in the bioinformatic analysis of NRPS type BGCs (Eppelmann, Stachelhaus, & Marahiel, 2002; Medema et al., 2011; Stachelhaus, Mootz, & Marahiel, 1999). As previously reported, the topology of the tree suggests that positions corresponding to each position in the true GPA scaffold are monophyletic (Waglechner et al., 2019) (Chapter 3). A phylogenetic

tree using the antiSMASH identified A-domains is depicted illustrating the scaffold classifications in Figure 4-5. The positions for the other scaffold classes do not follow the same pattern. Rather than each position evolving once, the other BGCs are characterized by extensive repetitiveness, putatively the result of duplication. Scaffold peptide sequence predictions and putative classifications based on antiSMASH analysis of A-domains using the SandPUMA ensemble classification tool (Chevrette, Aicheler, Kohlbacher, Currie, & Medema, 2017) for all BGCs are provided in Table 4-2.

4-5

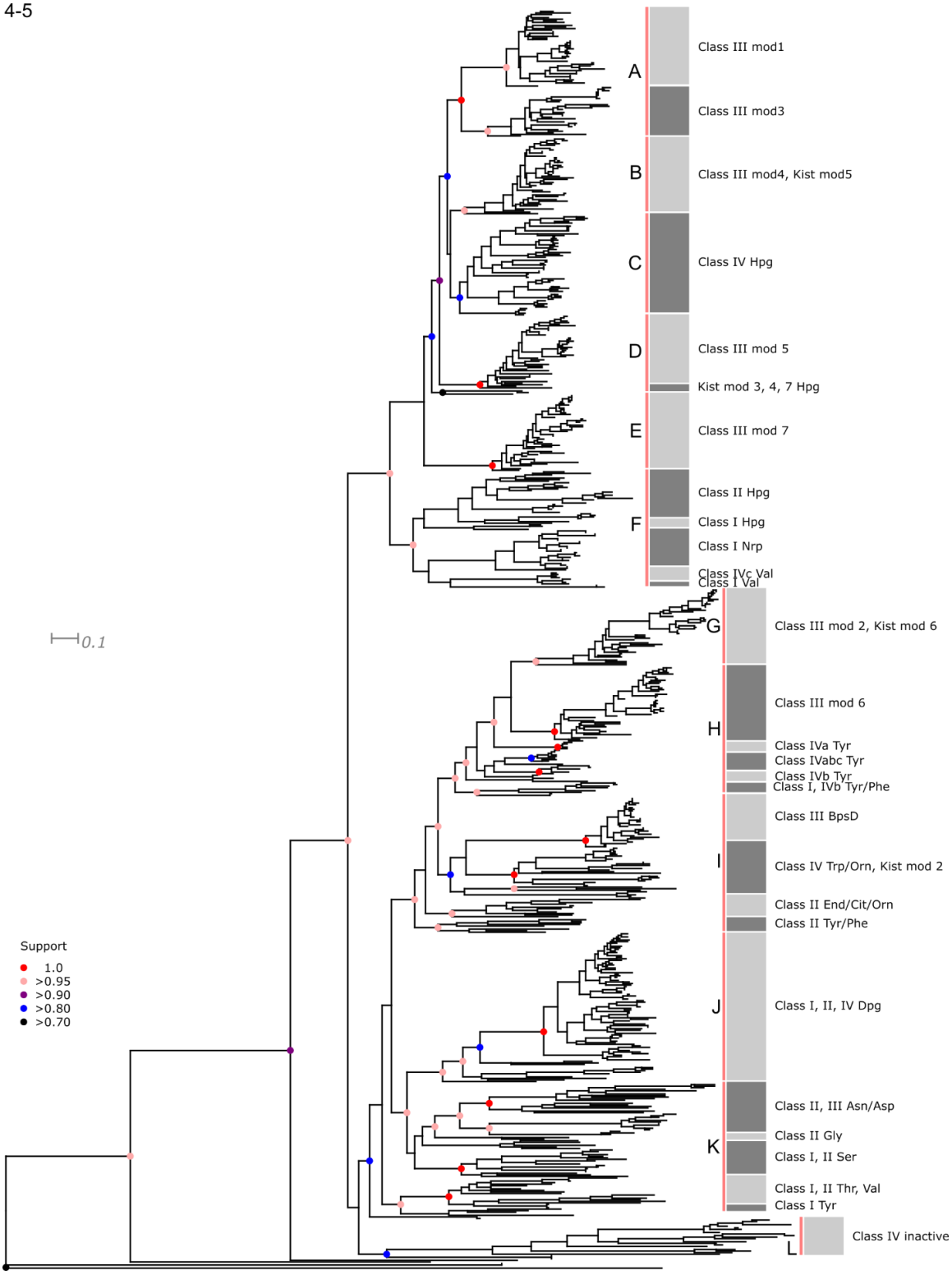

A
Class III mod1
Class III mod3
B
Class III mod4, Kist mod5
C
Class IV Hpg
D
Class III mod 5
Kist mod 3, 4, 7 Hpg
E
Class III mod 7
F
Class II Hpg
Class I Hpg
Class I Nrp
Class IVc Val
Class I Val
0.1
G
Class III mod 2, Kist mod 6
H
Class III mod 6
Class IVa Tyr
Class IVabc Tyr
Class IVb Tyr
Class I, IVb Tyr/Phe
I
Class III BpsD
Class IV Trp/Orn, Kist mod 2
Class II End/Cit/Orn
Class II Tyr/Phe
J
Class I, II, IV Dpg
Class II, III Asn/Asp
K
Class II Gly
Class I, II Ser
Class I, II Thr, Val
Class I Tyr
L
Class IV inactive
Support
1.0
>0.95
>0.90
>0.80
>0.70

Figure 4-5: Approximate maximum likelihood phylogeny of GPA BGC adenylation domains. GPA BGC A-domains amino acid sequences identified by antiSMASH HMMs, aligned and subjected to phylogenetic analysis under the WAG substitution model with the default CAT approximate rate categories using fasttree2 (see methods). The overall structure of the phylogeny is shown using a midpoint rooting. SH-like support values for basal branches are summarized. Insets for major clades are indicated by lowercase letters and include numeric support values. All scale bars indicate expected number of substitutions per site. Subset A - Class III modules 1 and 3. Subset B - Class III module 4 and kistamicin module 5. Subset C - Class IV A_{Hpg} domains. Subset D - Class III module 5 and kistamicin modules 3, 4, and 7. Subset E - Class III module 7. Subset F - Class II A_{Hpg}, Class I A_{nrp} (domains not predicted by antiSMASH), class IVc A_{Val}, Class I A_{Val}. Subset G - Class III module 2 and kistamicin module 6. Subset H - Class III module 6, Class IV A_{Tyr}. Subset I - Class III BpsD A_{Tyr}, Class IV A_{Trp}/A_{Orn}, kistamicin module 2, Class II $A_{End}/A_{Orn}/A_{Cit}$, Class II A_{Tyr}/A_{Phe}. Subset J - Class I, Class II, and Class IV A_{Dpg} modules not predicted by antiSMASH. Subset K - Class II, Class II A_{Asn}/A_{Asp}, Class II A_{Gly}, Class I, Class II A_{Ser}, Class I, Class II A_{Thr}, A_{Val}, Class I A_{Tyr}. Subset L - Class IV inactive A-domains and outgroup A-domains not part of GPA BGCs.

Table 4-2: Scaffold peptides and classes

Class	Cluster	Prefix	scaffold **gene_id** (domain_id) amino acid
I	34	feglymycin	**1732**: (4) nrp; (8) nrp; (12) Val; (15) nrp **1733**: (19) Hpg; (22) nrp; (26) Hpg; (29) nrp; (33) Val; (36) nrp **1734**: (41) Hpg; (44) Phe; (47) Asp
	12	Arif	**657**: (1) nrp; (4) Thr; (7) nrp **658**: (12) nrp; (15) nrp; (18) nrp **659**: (1) Ser; (4) nrp; (7) nrp; (11) nrp; (15) Tyr **661**: (3) nrp; (6) nrp; (9) nrp; (12) nrp
	56	WAC06369	**2950**: (5) Ser; (8) nrp; (11) nrp; (15) nrp; (19) Tyr **2951**: (25) nrp **2952**: (31) nrp; (34) Thr; (37) nrp **2953**: (42) nrp; (45) nrp; (48) nrp; (51) Ser; (54) nrp; (57) nrp; (60) nrp; (63) nrp; (66) nrp
II	20	End	**1061**: (4) Asp; (7) aThr **1062**: (10) Hpg; (13) Orn; (16) aThr; (19) Hpg; (22) Hpg **1065**: (55) aThr; (27) Cit **1063**: (30) End; (33) Hpg; (36) Ser; (39) Dpg; (42) Gly; (45) End; (48) Ala; (51) Hpg
	23	esnapd22	**1164**: (1) Thr; (4) Tyr; (7) nrp; (10) Tyr
	30	Ram	**1566**: (2) Asn **1567**: (5) Hpg; (8) Orn; (11) aThr; (14) Hpg; (17) Hpg **1571**: (50) aThr **1567**: (22) Phe **1568**: (25) Orn; (28) Hpg; (31) aThr; (34) Hpg; (37) Gly; (40) Leu; (43) Asn; (46) Hpg
	49	WAC01438	**2520**: (5) Hpg; (8) Ser; (11) nrp; (14) Asp; (17) nrp; (20) Thr; (23) Hpg; (26) Hpg; (29) Asp; (32) Hpg; (35) Ser **2521**: (38) Hpg; (41) nrp; (44) nrp; (47) Asp; (50) Ser **2522**: (53) Asn; (56) Ser; (59) Hpg; (62) nrp; (65) Phe **2522**: (68) nrp; (71) Ser
	67	STLI053	**3496**: (8) Asp; (11) Asp **3497**: (17) Glu; (20) nrp; (23) nrp; (26) nrp; (31) Phe **3498**: (34) nrp; (37) Hpg; (40) Thr; (43) nrp; (46) Gly; (49) Val; (52) nrp; (55) Leu **3532**: (2) nrp; (5) Asp; (8) Asp
	68	Hdaliensis	**3547**: (11) nrp; (14) Asp; (17) Ser; (20) nrp; (23) nrp; (26) nrp; (29) nrp; (32) nrp; (35) nrp; (38) Tyr; (41) Asp **3548**: (44) Asn; (47) Tyr; (50) nrp; (53) Tyr; (56) Asp; (59) Asp; (62) nrp; (65) Aso
	69	Mchersina	**3595**: (4) Asn; (7) Asn **3596**: (10) Hpg; (13) nrp **3596**: (16) Thr; (19) Hpg; (22) Hpg; (27) Phe **3597**: (30) nrp; (33) Hpg; (36) Thr; (39) nrp; (42) Gly; (45) Val; (48) Ala; (51) Hpg; (55) Thr
III	1	A40926	**13**: (13) Hpg; (10) Tyr **12**: (6) Dpg **23**: (16) Hpg; (20) Tyr; (24) Bht **24**: (28) Dpg
	2	A47934	**63**: (4) Hpg; (7) Tyr **64**: (11) Dpg **65**: (17) Hpg; (21) Hpg; (25) Bht **66**: (29) Dpg
	4	Aazuraea	**157**: (1) Hpg; (4) Bht **158**: (8) Dpg **159**: (12) Hpg; (16) Hpg; (20) Bht **160**: (24) Dpg
	6	Ajap-rist	**284**: (1) Hpg; (4) Bht **285**: (9) Dpg **286**: (13) Hpg; (17) Hpg; (21) Bht **288**: (25) Dpg
	7	AkerNogbecin	**338**: (1) Hpg; (4) Hpg **339**: (9) Phe **340**: (13) Hpg; (17) Hpg; (21) Bht **31**: (25) Dpg
	8	AlurRist	**431**: (32) Hpg; (29) Bht **430**: (25) Dpg **431**: (21) Hpg; (17) Hpg; (13) Bht **432**: (9) Dpg
	10	Arect	**519**: (1) Hpg; (4) Tyr **520**: - **521**: (9) Hpg; (13) Hpg; (17) Bht **522**: (21) Dpg
	16	CA878	**833** **835** **836** **837**

		(1) Hpg	(4) Bht	(9) Dpg	(13) Hpg	(17) Hpg	(21) Bht	(25) Ala
17	CA915	**901**	**902**	**903**	**904**			**905**
		(1) Hpg	(4) Tyr	(8) Dpg	(12) Hpg	(16) Hpg	(20) Bht	(24) Dpg
21	esnapd15	**1080**		**1081**	**1082**			**1083**
		(1) Hpg	(4) Tyr	(10) Ala	(14) Hpg	(18) Hpg	(22) Bht	(26) Dpg
22	esnapd16	**1153**		**1155**	**1156**		**1157**	**1158**
		(4) Hpg	(7) Tyr	(1) Dpg		(4) Hpg	(11) Bht	(15) Dpg
25	JPLW	**1281**		**1282**	**1283**			**1284**
		(1) Hpg	(4) Bht	(9) Dpg	(13) Hpg	(17) Hpg	(20) Bht	(25) Dpg
27	KF88Rist	**1414**		**1415**	**1416**			**1417**
		(1) Hpg	(4) Bht	(9) Dpg	(13) Hpg	(17) Hpg	(21) Bht	(25) Dpg
28	Ncox	**1493**		**1494**	**1495**			**1496**
		(4) Hpg	(7) Tyr	(12) Dpg	(16) Hpg	(20) Hpg	(24) Tyr	(28) Dpg
29	Pek	**1526**		**1527**	**1529**			**1530**
		(1) nrp	(4) Tyr	(8) Leu	(12) Hpg	(16) Hpg	(20) Bht	(24) Dpg
36	TEG	**1784**		**1785**	**1787**			**1788**
		(1) Hpg	(4) Bht	(8) Ala	(13) Hpg	(17) Hpg	(21) Bht	(25) Dpg
37	Teico	**1819**		**1820**	**1821**			**1822**
		(1) Hpg	(4) Tyr	(8) Dpg	(12) Hpg	(16) Hpg	(20) Bht	(24) Dpg
38	Teico2	**1870**		**1871**	**1872**			**1873**
		(1) Hpg	(4) Tyr	(8) Dpg	(12) Hpg	(16) Hpg	(20) Bht	(24) Dpg
39	UK68597	**1909**		**1910**	**1912**			**1913**
		(1) Hpg	(4) Tyr	(8) Dpg	(14) Hpg	(18) Hpg	(21) Bht	(26) Dpg
41	VEG	**1989**		**1990**	**1994**			**1992**
		(1) Hpg	(4) Bht	(9) Hpg	(12) Hpg	(16) Hpg	(20) Bht	(24) Ala
44	WAC01375	**2127**		**2128**	**2129**			**2130**
		(1) Hpg	(4) Bht	(8) Dpg	(12) Hpg	(16) Hpg	(20) Bht	(24) Dpg
45	WAC01376	**2263**		**2262**	**2261**			**2259**
		(31) Hpg	(28) Bht	(24) Dpg	(20) Hpg	(16) Hpg	(12) Bht	(9) Dpg
46	WAC01416	**2304**		**2305**	**2306**			**2307**
		(1) Hpg	(4) Bht	(8) Dpg	(12) Hpg	(16) Hpg	(20) Bht	(23) Dpg
48	WAC01424	**2485**		**2484**	**2483**			**2482**
		(30) Hpg	(27) Tyr	(24) Dpg	(18) Hpg	(14) Hpg	(10) Tyr	(7) Dpg
52	WAC04182	**2743**		**2742**	**2741**			**2740**
		(33) Hpg	(30) Bht	(25) Dpg	(21) Hpg	(17) Hpg	(17) Bht	(9) Dpg
53	WAC04197	**2787**		**2788**	**2789**			**2790**
		(1) Hpg	(4) Bht	(8) Dpg	(12) Hpg	(16) Hpg	(20) Bht	(24) Dpg
55	WAC05977	**2902**		**2903**	**2904**			**2905**
		(1) Hpg	(4) Bht	(8) Dpg	(12) Hpg	(16) Hpg	(20) Bht	(24) Dpg
60	Asub	**3098**		**3099**	**3100**			**3101**
		(1) Hpg	(4) Tyr	(9) Dpg	(14) Hpg	(18) Hpg	(22) Bht	(26) Dpg
15	CA37	**762**		**763**	**764**			**765**
		(1) Hpg	(4) 3OH Gln	(8) Dpg	(14) Hpg	(18) Hpg	(22) Bht	(26) Dpg
3	Aalba	**95**			**96**			**97**
		(1) nrp	(4) Bht	(8) Asn	(13) Hpg	(17) Hpg	(21) Bht	(25) Dpg
5	ABalhFH1894	**230**			**231**			**232**
		(1) Leu	(4) Bht	(8) Asn	(12) Hpg	(16) Hpg	(20) Bht	(24) Dpg
9	AOHO	**461**			**462**			**463**
		(1) Leu	(4) Bht	(8) Asn	(12) Hpg	(16) Hpg	(20) Bht	(24) Dpg
11	Areg	**601**			**602**			**603**
		(1) nrp	(4) Bht	(8) Asn	(13) Hpg	(17) Hpg	(21) Bht	(25) Dpg
14	Balh	**708**			**709**			**710**
		(1) Leu	(4) Bht	(8) Asn	(12) Hpg	(16) Hpg	(20) Bht	(24) Dpg
18	Chloroeremomycin	**960**			**961**			**962**
		(1) Leu	(4) Bht	(8) Asn	(12) Hpg	(16) Hpg	(20) Bht	(24) Dpg
26	Karid	**1354**			**1355**			**1356**
		(1) nrp	(4) Bht	(8) Asn	(12) Hpg	(16) Hpg	(20) Bht	(24) Dpg
40	Vanco	**1955**			**1956**			**1957**
		(1) Leu	(4) Bht	(8) Asn	(12) Hpg	(16) Hpg	(20) Bht	(24) Dpg
47	WAC01420	**2415**			**2418**			**2419**
		(1) nrp	(4) Tyr	(8) Leu	(12) Tyr	(16) Tyr	(20) Bht	(24) Dpg
51	WAC04169	**2640**			**2638**			**2637**
		(32) nrp	(29) Bht	(25) Asn	(21) Hpg	(17) Hpg	(13) Bht	(9) Dpg

71	Kistamicin	**3710**		**37411**		**3712**	**3713**
		(3) nrp	(7) nrp	(12) nrp	(15) nrp	(19) Tyr	(23) nrp

IVa

19	complestatin	**1005**		**1006**	**1007**			**1008**	
		(1) Hpg	(4) Trp	(9) Hpg	(15) Hpg	(19) Hpg	(23) Tyr	(28) Hpg	(31) inactive
43	WAC01325	**2078**		**2079**	**2080**			**2081**	
		(1) Hpg	(4) Trp	(9) Hpg	(15) Hpg	(19) Hpg	(23) Tyr	(28) Hpg	(31) inactive
54	WAC05379	**2847**		**2849**	**2851**			**2852**	
		(1) Hpg	(4) Trp	(1) Hpg	(4) Hpg	(8) Hpg	(12) Tyr	(17) Ala	(20) Inactive
61	FNTE	**3135**		**3136**	**3137**			**3138**	
		(1) Hpg	(4) Trp	(9) Hpg	(15) Hpg	(19) Hpg	(23) Tyr	(28) Hpg	(31) inactive
42	WAC00631	**2040**		**2041**	**2042**			**2043**	
		(2) Hpg	(5) Trp	(11) Hpg	(17) Hpg	(21) Hpg	(25) Tyr	(31) Hpg	(34) inactive

57	WAC06725	**2976**		**2977**	**2978**			**2979**	
		(2) Hpg	(6) Trp	(11) Hpg	(17) Hpg	(21) Hpg	(25) Tyr	(31) Hpg	(34) inactive
59	WAC06783	**3057**		**3058**	**3059**			**3060**	
		(2) Hpg	(6) Trp	(11) Hpg	(17) Hpg	(21) Hpg	(25) Tyr	(31) Hpg	(34) inactive

IVb			**686**	**687**		**689**		**690**			**691**	
13	AspOK074		(4) Hpg	(7) nrp	(10) Trp	(14) Hpg	(19) Hpg	(23) Tyr	(27) Tyr	(32) nrp	(36) nrp	(39) inactive
		1608				**1609**	**1610**				**1611**	
31	Slut	(5) nrp	(8) nrp	(11) nrp	(14) Trp	(18) Hpg	(24) Hpg	(28) Tyr	(32) Tyr	(37) nrp	(41) nrp	(44) inactive
		1651				**1653**	**1654**		**1655**		**1656**	
32	Ssil	(6) nrp	(8) nrp	(11) nrp	(14) Trp	(1) Hpg	(7) Hpg	(11) Tyr	(1) Tyr	(6) nrp	910) nrp	(13) inactive
			1756		**1759**			**1760**				
35	SspNRRLS1521		(4) Hpg	-	(1) Trp	-	-	(3) Tyr	(7) Tyr	(12) Tyr	(16) nrp	(19) inactive
			2565		**2566**	**2567**					**2568**	
50	WAC01529		(4) Hpg	(7) nrp	(10 Trp	(14) Hpg	(20) Hpg	(24) Tyr	(28) Tyr	(33) nrp	(37) nrp	(40) inactive
		3658		**3659**								
70	TLI55	(2) nrp	(6) nrp	(12) Tyr	(15) nrp	(18) nrp	(21) Tyr	(24) nrp	(27) nrp	(30) nrp	(33) nrp	(36) inactive

IVc		**1691**			**1692**	**1693**	**1694**			**1695**	
33	SspCNQ509	(3) Tyr	(6) nrp	(9) val	(15) Trp	(19) nrp	(25) Hpg	(29) nrp	(33) Tyr	(39) nrp	(42) inactive
		3020			**3021**	**3022**	**3023**			**3024**	
58	WAC06738	(1) Tyr	(4) nrp	(7) Val	(12) Trp	(16) nrp	(22) Hpg	(26) nrp	(30) Tyr	(36) nrp	(39) inactive
		3213			**3214**	**3215**	**3216**			**3217**	
62	SspCNQ329	(3) nrp	(6) nrp	(9) nrp	(14) Orn	(18) nrp	(25) Hpg	(27) nrp	(31) Tyr	(37) nrp	(40) inactive
		3269			**3270**	**3271**	**3272**			**3273**	
63	SspCNQ525	(3) nrp	(6) nrp	(9) nrp	(14) Orn	(19) nrp	(25) Hpg	(29) nrp	(33) Tyr	(39) nrp	(42) inactive
		3325			**3226**	**3227**	**3328**			**3329**	
64	SspCNQ865	(3) nrp	(6) nrp	(9) nrp	(14) Orn	(19) nrp	(25) Hpg	(29) nrp	(33) Tyr	(39) nrp	(42) inactive
		3381			**3382**	**3283**	**3324**			**3385**	
65	SspCNT371	(3) nrp	(6) nrp	(9) nrp	(15) Trp	(19) nrp	(25) Hpg	(29) nrp	(33) Tyr	(39) nrp	(42) inactive
		3435			**3436**	**3437**	**3438**			**3439**	
66	SspCNY243	(3) nrp	(6) nrp	(9) nrp	(15) Trp	(19) nrp	(25) Hpg	(29) nrp	(33) Tyr	(39) nrp	(42) inactive

The prediction of the structure of some of the amino acids in several BGCs was facilitated by the A-domain phylogeny and the structure of feglymycin. Incorporation of 3,5-dihydroxy phenylglycine was not unambiguously predicted for the A-domains in the feglymycin BGC (Gonsior et al., 2015). Clades of similar A-domains were identified in the phylogeny consistent with the presence of a set of biosynthetic genes for Dpg in these BGCs. The Dpg-activating A-domains are found in distinct regions in the phylogeny, differentiated by being located in either class I (Supplementary Figure 4-6,

Supplementary Figure 4-10), II (Supplementary Figure 4-6), IV (Supplementary Figure 4-10), or class III (Dpg found in the D-Ala-D-Ala GPA scaffolds module 3 (Supplementary Figure 4-1) and 7 (Supplementary Figure 4-5)).

In addition to the large number of similar Dpg A-domains, the feglymycin BGC encodes A-domains with similarity to other positions that will be noted as those scaffolds are analyzed. Unlike the A-domains, the C-domains (Figure 4-6) synthesizing the feglymycin scaffold come in two varieties found in different regions of the tree (Supplementary Figures 4-20 and 4-23, Group 1 and 2 respectively). Neither group is monophyletic, but group 1 is responsible for forming peptide bonds with the residue following a Dpg. A similar profile is observed with the feglymcin PCP sequences (Figure 4-7) being divided into two related groups, however the Dpg-associated PCP groups are not monophyletic, being ancestral to PCP sequences found in the WAC06738-type scaffold NRPS (Supplementary Figure 4-32), while the non-Dpg related PCP domains are monophyletic (Supplementary Figure 4-34).

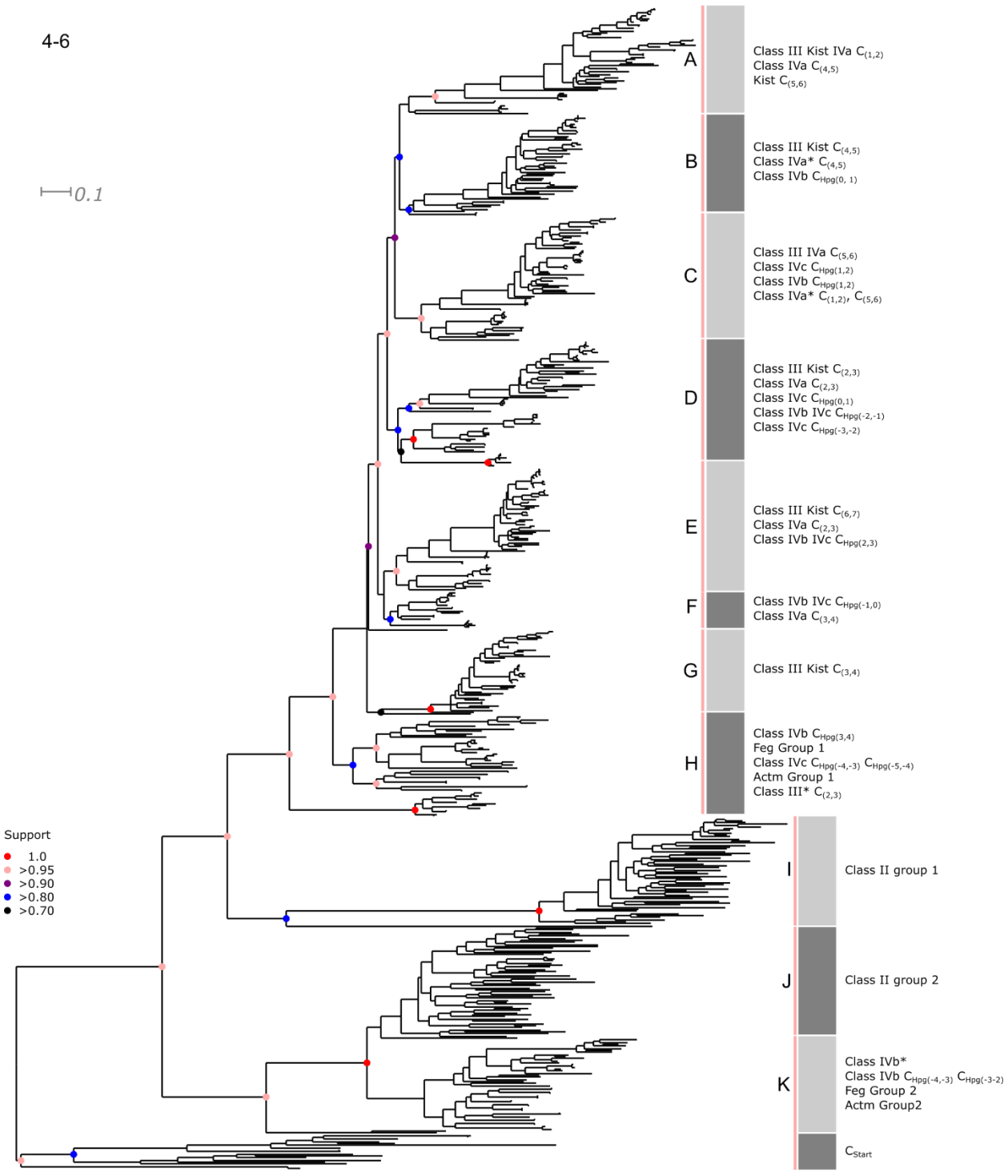


Figure 4-6: Approximate maximum likelihood phylogeny of GPA BGC condensation domains. The amino acid sequences of condensation domains identified by the antiSMASH HMM in the GPA BGCs were aligned and subjected to the WAG substitution model with the default CAT approximation rate categories as implemented

by fasttree2. For scaffold classes with variable lengths, C_{Hpg} refers to the central Hpg participating in intramolecular crosslinks labeled with coordinate 0. All other coordinates are absolute. The overall structure of the C-domain phylogeny as a mid-point rooted tree. SH-like support values are summarized. Insets for major clades are indicated with lowercase letters and include numeric node support values. All scale bars indicate expected number of substitutions per site. Subset A - Class III and kistamicin $C_{(1,2)}$, Class IVa $C_{(4,5)}$. Subset B - Class III and kistamicin $C_{(4,5)}$, IVa* $C_{(4,5)}$, IVb $C_{Hpg(0,1)}$. Class IVa* refers to the complestatin-variable scaffolds (see main text). Subset C - Class III and IVa $C_{(5,6)}$, Class IVc and IVb $C_{Hpg(1,2)}$, Class IVa $C_{(1,2)}$ and $C_{(5,6)}$. Subset D - Class III, kistamicin and IVa $C_{(2,3)}$, Class IVc $C_{Hpg(0,1)}$, Class IVb and IVc $C_{Hpg(-2,-3)}$, IVc $C_{Hpg(-3,-2)}$. Subset E - Class III and kistamicin $C_{(6,7)}$, Class IVa $C_{(2,3)}$, Class IVb and IVc $C_{Hpg(2,3)}$. Subset F - Class IVb and IVc $C_{Hpg(-1,0)}$, Class IVa $C_{(3,4)}$. Subset G - Class III and kistamicin $C_{(3,4)}$. Subset H - Class IVb $C_{Hpg(3,4)}$, feglymycin group 1, Class IVc $C_{Hpg(-4,-3)}$ and $C_{Hpg(-5,-4)}$, *Actinomadura* group 1, Class III* $C_{(2,3)}$. Class III* refers to the GPA scaffolds from *Amycolatopsis* that have replaced an aromatic with aliphatic amino acid at position 3. Subset I - Class II group 1. Subset J - Class II group 2. Subset K - Class IVb*, Class IVb $C_{Hpg(-4,-3)}$ and $C_{Hpg(-3,-2)}$, feglymycin group 2, *Actinomadura* group 2, starter condensation domains.

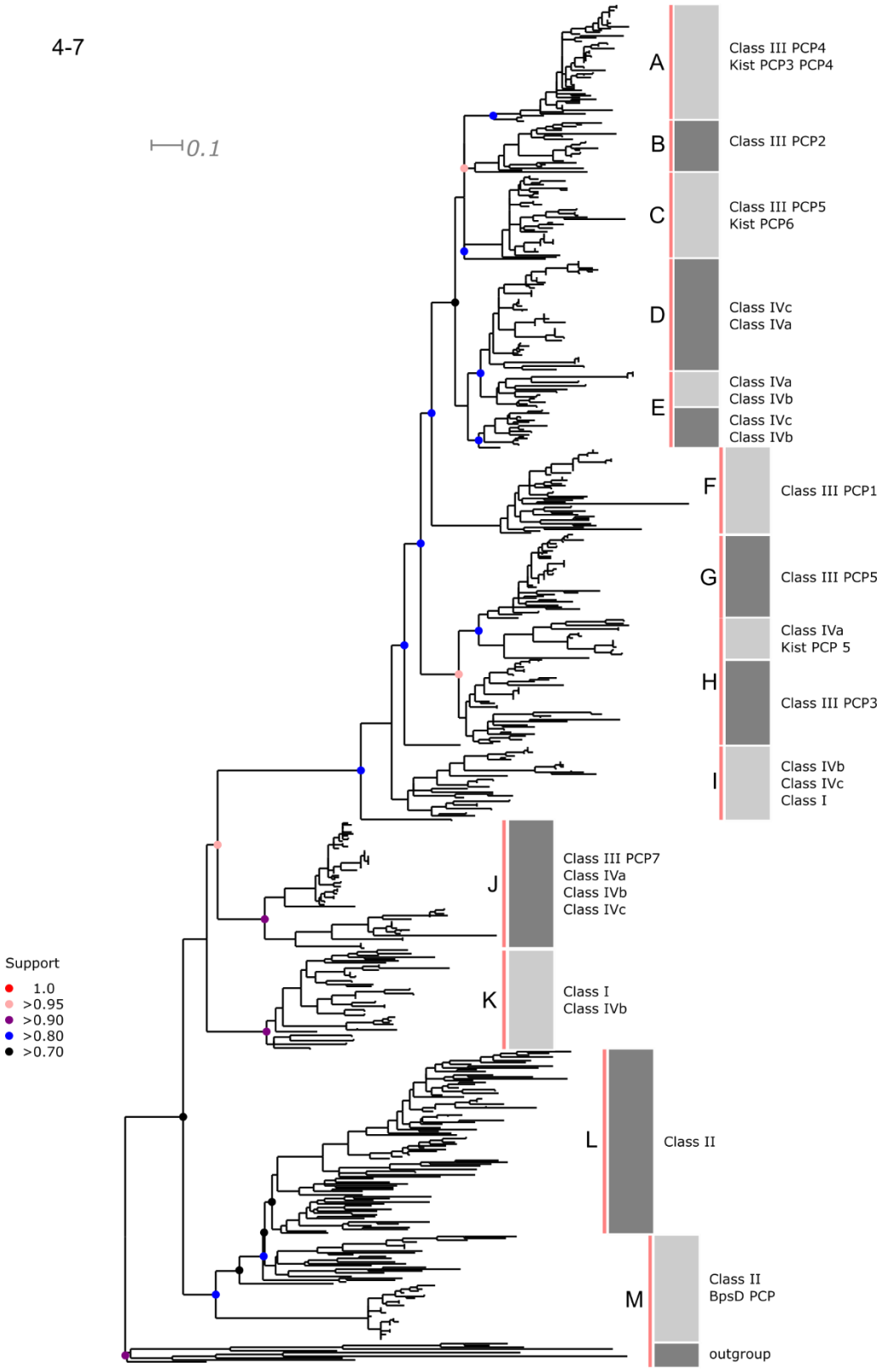

4-7
0.1
A
Class III PCP4
Kist PCP3 PCP4
B
Class III PCP2
C
Class III PCP5
Kist PCP6
D
Class IVc
Class IVa
E
Class IVa
Class IVb
Class IVc
Class IVb
F
Class III PCP1
G
Class III PCP5
H
Class IVa
Kist PCP 5
Class III PCP3
I
Class IVb
Class IVc
Class I
J
Class III PCP7
Class IVa
Class IVb
Class IVc
K
Class I
Class IVb
L
Class II
M
Class II
BpsD PCP
outgroup
Support
1.0
>0.95
>0.90
>0.80
>0.70

Figure 4-7: Approximate maximum likelihood phylogeny of GPA BGC peptidyl-carrier-protein domains. The amino acid sequences of PCP domains identified by the antiSMASH HMM in the GPA BGCs were aligned and subjected to the WAG substitution model, the default CAT rate approximation as implemented by fasttree2. The overall phylogeny of this domain, tree as a mid-point rooted tree. SH-like support values for basal branches are summarized. Insets for major clades are indicated with lowercase letters and include numeric support values. All scale bars indicate expected number of substitutions per site. Subset A - Class III PCP4, kistamicin PCP3 and PCP4. Subset B - Class III PCP2. Subset C - Class III PCP5, kistamicin PCP6. Subset D - Multiple Class IVc, Class IVa PCP domains. Subset E - Multiple Class IVa and IVb, Class IVc and IVb PCP domains. Subset F - Class III PCP1. Subset G - Class III PCP5. Clade H - Multiple Class IVa PCP domains and kistamicin PCP 5, Class III PCP3. Subset I - Multiple Class IVb, IVc and Class I PCP domains. Subset J - Class III PCP7, multiple Class IVa, IVb and IVc PCP domains. Subset K - Multiple Class I and Class IVb PCP domains. Subset L - Multiple Class II PCP domains. Subset M - Multiple Class II PCP domains, Class III BpsD PCP domains from *Amycolatopsis* spp. and outgroup PCPs not involved in GPA biosynthesis.

The kistamicin scaffold is a hybrid Class III and IV

The kistamicin heptapeptide is challenging to classify. While the kistamicin NRPS sequences possesses the hallmarks of the complestatin-type scaffolds (Trp_2-crosslinked to Hpg_4, X-domain followed by inactive A-domain), the first residue, Tyr_1, is incorporated as a starting unit to the NRPS (Nazari et al., 2017). The phylogenetic profile of the A-domain for kistamicin Trp_2 is indeed related to the other Trp-activating domains in the Nicolau type V scaffolds (Supplementary Figure 4-9). Hpg5 and Tyr6 are found as outgroups to the GPA module 4 (Hpg) and GPA module 2 (Tyr/Bht) clades, respectively (Supplementary Figure 4-2 and 4-9). A-domains encoding the incorporation of Hpg3, Hpg4, and Hpg7 form a distinct clade with moderate support (0.775, Supplementary Figure 4-4).

The phylogenetic profile of the $C_{(1,2)}$, $C_{(2,3)}$, $C_{(3,4)}$, $C_{(4,5)}$ and $C_{(6,7)}$ domains are basal to the clades encoding the equivalent C-domains in the GPA scaffolds (Figure 4-6). The exception to this is the kistamicin $C_{(5,6)}$ domain, which is found to be most closely related to complestatin-type scaffold with virtually no phylogenetic support (Supplementary Figure 4-13). The PCP sequences for the kistamicin NRPS (Figure 4-7) show similar profiles are most closely related to the GPA PCP sequences (Supplementary Figure 4-24 - 4-26, 4-31) with the exception of PCP5 (Supplementary Figure 4-31). Kistamicin PCP3 appears to be a duplication of PCP4, most closely related GPA PCP4. Kistamicin PCP5 is most closely related to a clade of complestatin-type PCP sequences with good support (0.842, Supplementary Figure 4-31). The X-domain in the kistamicin BGC is a basal branch of the monophyletic clade that includes the domains from the true GPAs with high support (Figure 4-8).

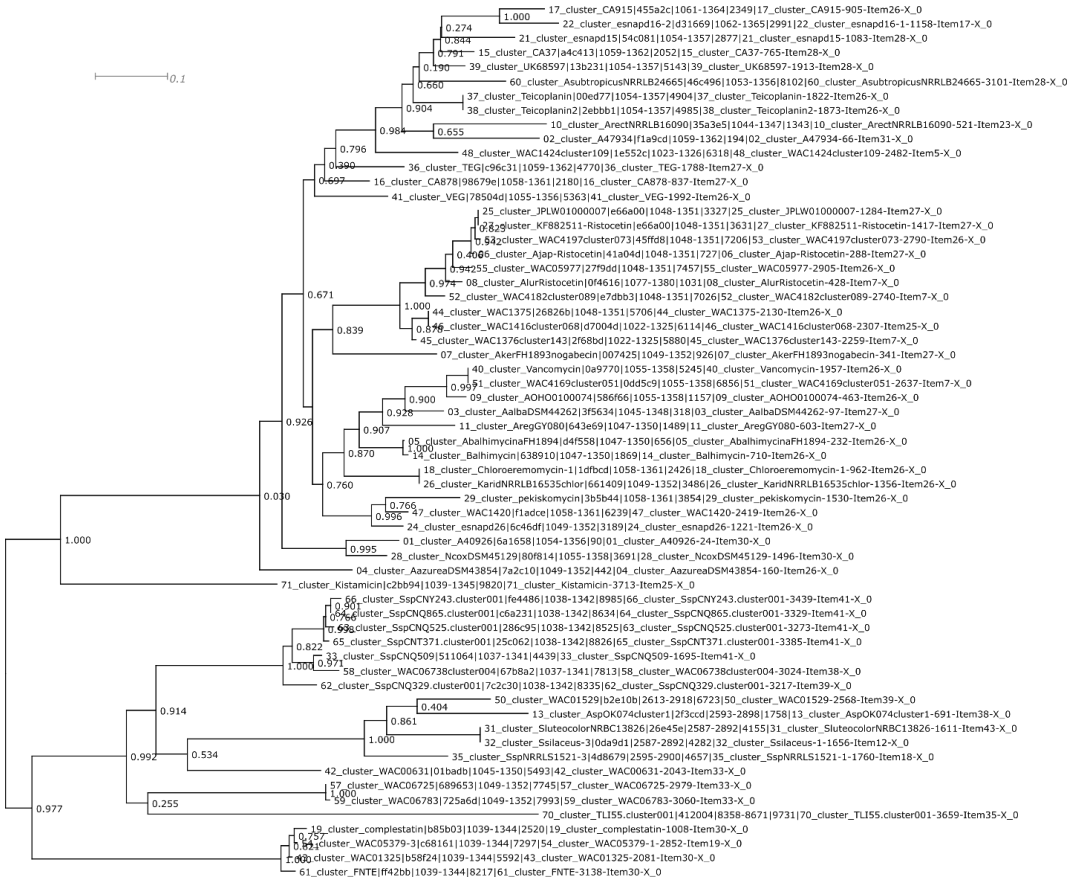


Figure 4-8: Approximately maximum likelihood phylogeny of GPA BGC X domains. The amino acid sequences for GPA X domains identified by the antiSMASH HMM were aligned and analyzed using the WAG substitution model and default number of CAT approximation rate categories using fasttree2 (see methods). The overall phylogeny of this family as a mid-point rooted tree. SH-like support values are indicated. Scale bare indicates expected number of substitutions per site.

Analysis of class II scaffolds

The scaffolds of the class II products enduracidin and ramoplanin code for the non-proteinogenic amino acids citrulline (Cit), ornithine (Orn) and the cyclic amino acid enduracididine (End) in addition to Hpg and Dpg (Figure 4-2). The class II scaffold encoded by *Streptomyces* sp. WAC01438, in addition to the putative Dpg activating A-

domains with similarity to those in the feglymycin BGC, is encoded by several A-domains activating amino acids that are not predicted by the ensemble of tools in antiSMASH. The WAC01438 NRPS sequences have A-domains that share an MRCA with the enduracidin and ramoplanin A-domains but lacks the genes *endP*, *endQ* and *endR* for enduracididine biosynthesis found in the enduracidin BGC and other NRPS BGCs coding for incorporation of this amino acid (Magarvey, Haltli, He, Greenstein, & Hucul, 2006). The A-domains activating these unusual amino acids in enduracidin, A_{Orn}, A_{Cit}, two A_{End}, and two A_{Orn} in ramoplanin are monophyletic with two domains in *Micromonospora chersina* and STLI053 scaffolds and a putative A_{Phe} from *Streptomyces.* sp. WAC01438 (Supplementary Figure 4-9). The domains from each BGC are more similar to each other than between these BGCs.

Asn or Asp at the N-terminus is a common feature of these scaffolds. These positions are N-acylated in both enduracidin and ramoplanin, in agreement with the presence of a C-domain at the N-terminus of the first module that presumably catalyzes the addition of a lipid from a loaded acyl-carrier protein (ACP) supplied in trans by another sequence in the BGC (Hoertz et al., 2012) (Figure 4-2, Supplementary Figure 4-23). A large clade of A-domains known or predicted to add Asp or Asn, including two monophyletic subclades of duplicated A_{Asp} found in the *Herbidospora daliensis* and *Streptomyces* sp. WAC01438 scaffolds (Supplementary Figure 4-11). We previously identified a monophyletic clade of A_{Asn} domains in vancomycin-like GPA BGCs in several *Amycolatopsis* sp. genomes that arrived via HGT during the evolution of the class III scaffolds. In the wider phylogenetic context, that clade appears to have originated from within a larger Asp/Asn clade

showing that these domains can flow between distant members of this BGC family (Waglechner et al., 2019) (Supplementary Figure 4-11).

A single Gly is found in the peptide scaffolds for enduracidin, ramoplanin, STLI053 and the GPA from *Micromonospora chersina*. The A-domain encoding its incorporation is not monophyletic in this phylogeny however as the A_{Ala} domain in enduracidin is included in this clade with low support (Supplementary Figure 4-11). Other A_{Ala} domains from ramoplanin and *Micromonospora* are found as an outgroup to the A_{Asp} and A_{Asn} clade with low support, suggesting they are not good marker sequences for this family (Supplementary Figure 4-11). Similarly, A_{Gly}, a group of A_{Leu} domains and A_{Val} are found in a clade in *Micromonospora*, ramoplanin, *H. daliensis* and STLI053 BGCs, sister to a large clade of A_{Thr} domains found in every member of this scaffold class (Supplementary Figure 4-11). A single A_{Thr} from this clade is found in the *Actinomadura* sp. scaffolds. The last unusual feature is a clade of duplicated A_{Ser} domains in the *Streptomyces* sp. WAC01438 scaffold that is sister to a pair of A_{Ser} domains found in the class I *Actinomadura* sp. scaffolds (Supplementary Figure 4-11). Finally, we propose a structural prediction for GP1438 (Figure 4-9).

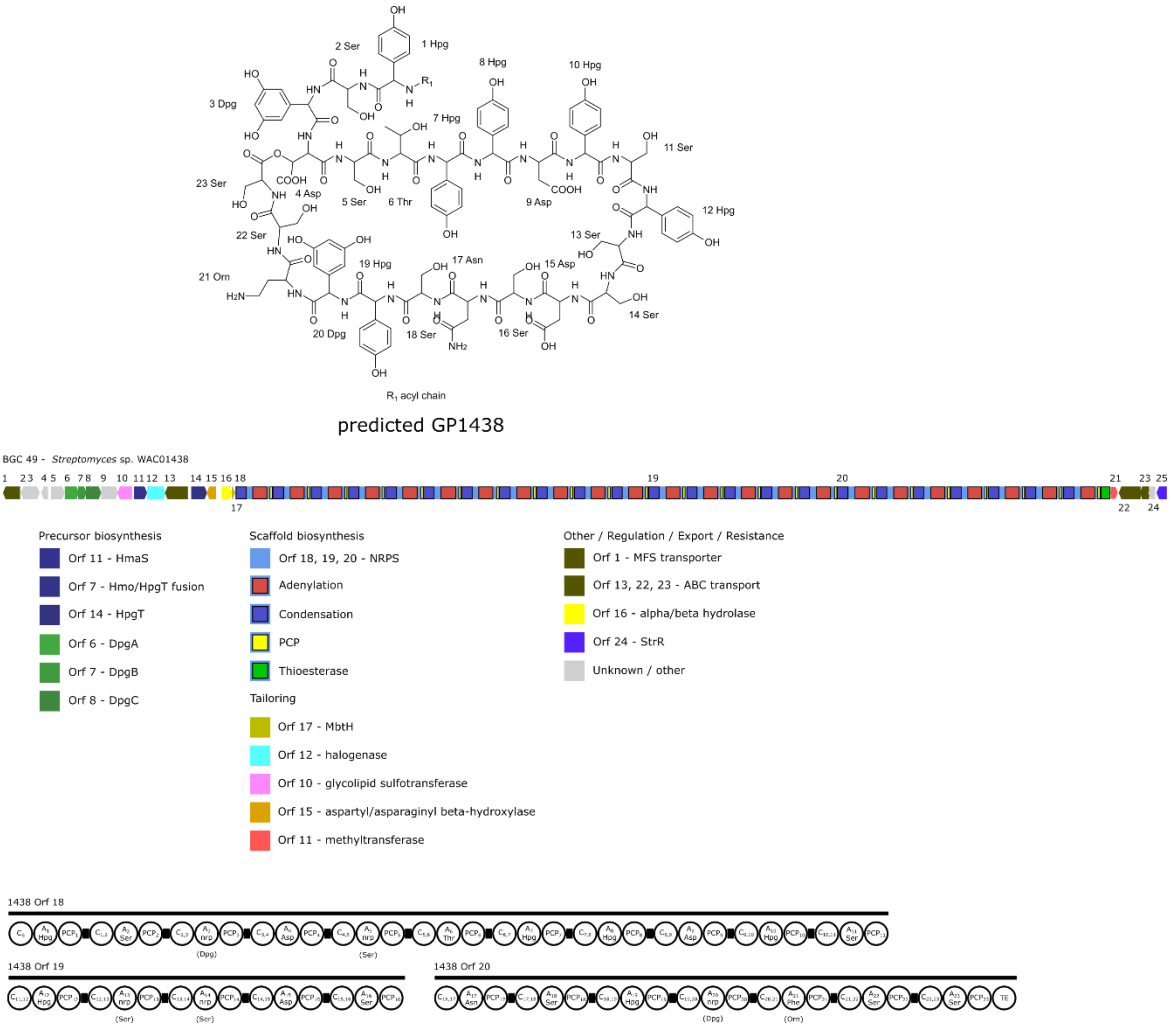


Figure 4-9: GP1438 predicted structure, BGC, and NRPS configuration. *Streptomyces* sp. WAC01438 encodes a GPA BGC with the following features. Minimal Hpg and Dpg biosynthesis genes are included, both residues predicted in the 23 amino acid scaffold. Predicted tailoring reactions include N-acylation of unknown length (by virtue of a condensation starter domain) along with sulfation, halogenation, and methylation though the exact position of these potential modifications is unpredictable. The supplied β-hydroxylase (Orf 15) suggests modification of one or more Asp or Asn residues, depicted here as β-OH Asp_4 participating in depsipeptide cyclization in analogy with ramoplanin A2. Based on nearest-neighbour phylogenetic analysis the cryptic amino acid predictions are indicated, notably Dpg_3, Ser_5, Ser_{13}, Ser_{14}, Dpg_{20} and Orn_{21}.

The predicted peptide sequence for the scaffold encoded by the two *Actinomadura rifamycini* DSM 43936 and *Actinomadura* sp. WAC06369 (BGCs 56 and 12 respectively) are challenging to compare as the DSM 43936 sequence from the public database is incomplete. From the WAC06369 sequence there is a lack of clear evidence for either putative peptide or depsipeptide cyclization, there is no β-hydroxylase as in the ramoplanin and WAC01438 BGCs. While there are putative Ser or Thr -OH group available in the scaffold, the TE domain which may be responsible for cyclization is more similar to those found in the class IV scaffolds leading us to tentatively classify this BGC as a class I scaffold but distinct from feglymycin (Figure 4-10) (Kohli, Trauger, Schwarzer, Marahiel, & Walsh, 2001). The A-, C- and PCP domains for these sequences are largely distinct from the other scaffold classes in the phylogeny suggesting that this scaffold represents a novel type in the GPA-like family (A-domains: Figure 4-5, Supplementary Figures 4-6, 4-8, 4-11) (C-domains: Figure 4-6, Supplementary Figures 4-20, 4-23) (PCP domains – Supplementary Figures 4-32, 4-34). The two BGCs in this class possess starter C-domains, suggesting these scaffolds are acylated at their N-terminus (Supplementary Figure 4-23). Based on this information, we propose a structure for GP6369 in Figure 4-11.

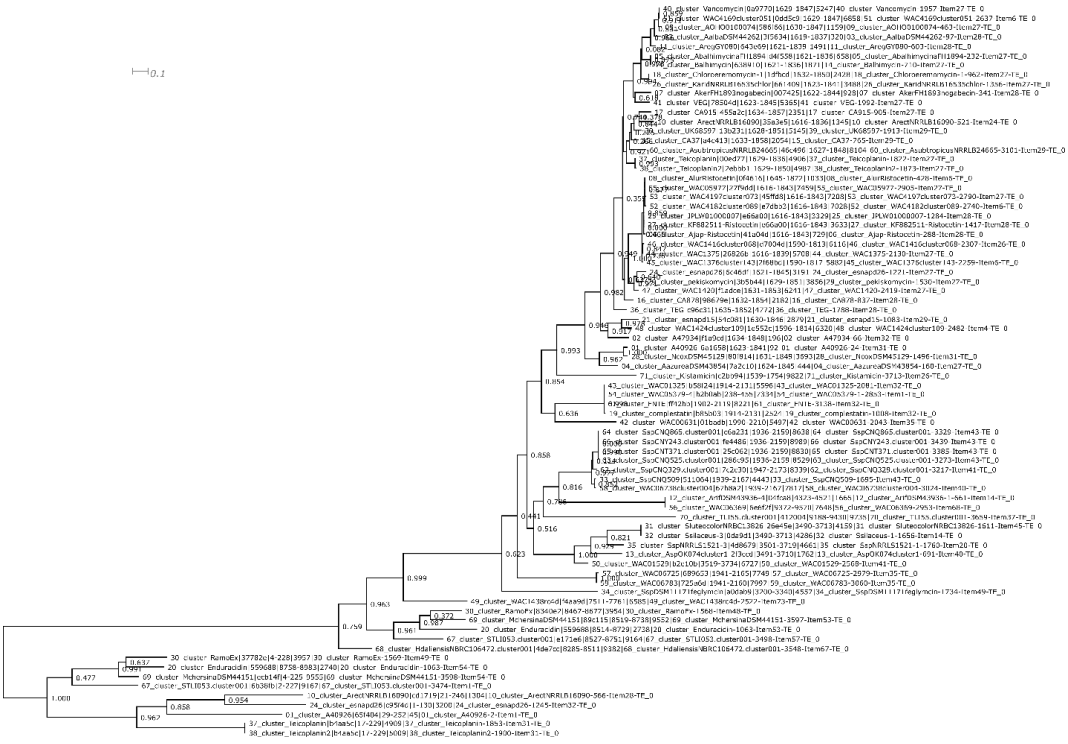


Figure 4-10: Approximate maximum likelihood phylogeny of GPA BGC thioesterase domains. Amino acid sequences for thioesterase domains predicted by the integrated antiSMASH HMM were aligned and analyzed with the WAG substitution model using the default number of CAT approximate rate categories with fasttree2 (see methods). The tree is mid-point rooted. SH-like support values are indicated. The scale bar represents the expected number of substitutions per site.

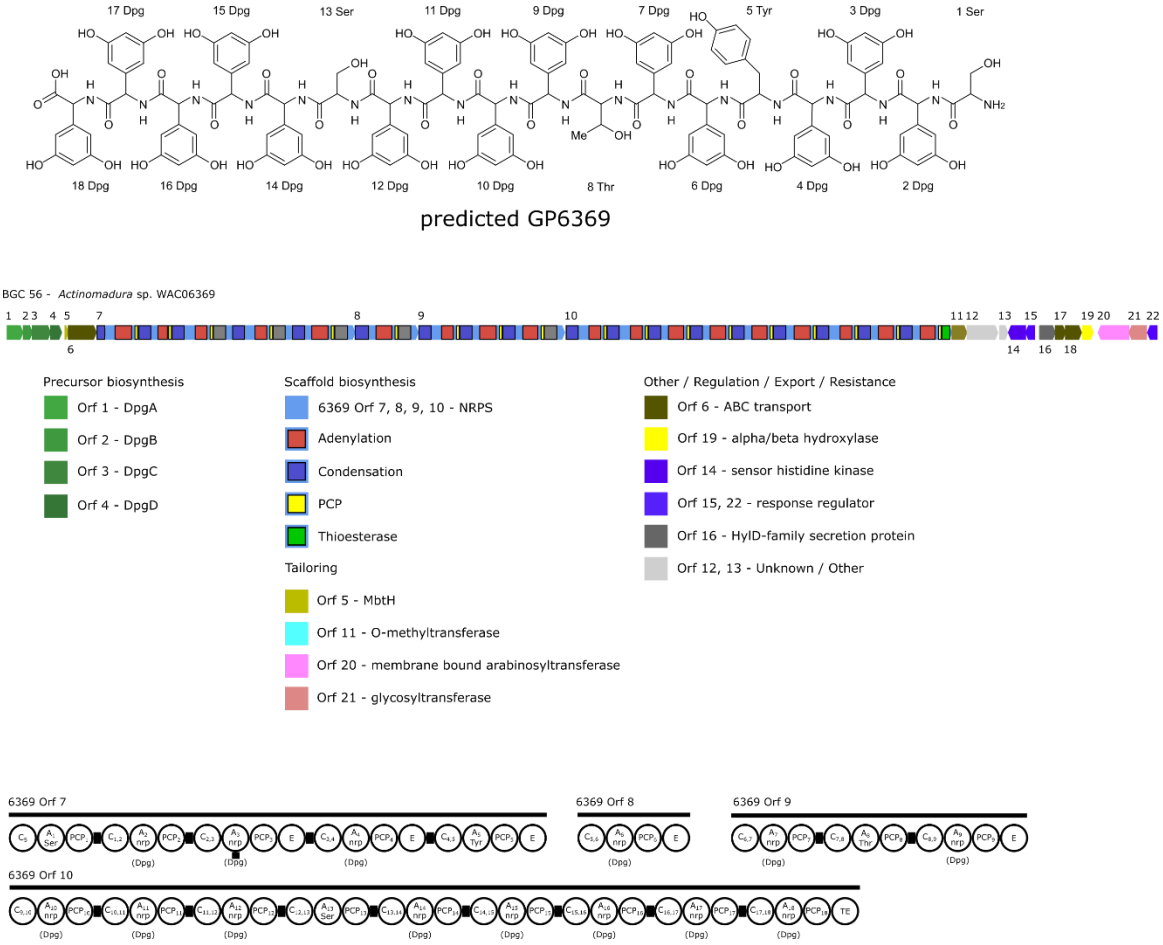


Figure 4-11: GP6369 predicted structure, BGC and NRPS configuration.
Actinomadura sp. WAC06369 encodes a GPA-like BGC with the following structure. This scaffold is not predicted to include Hpg. The scaffold NRPS is highly repetitive. Phylogenetic analysis was used to predict the identity of the amino acids adenylated by the cryptic A-domains as Dpg at positions 2, 3, 4, 6, 7, 9, 10, 11, 12, 14, 15, 16, 17 and 18, though the nearest neighbor sequences in the tree are aliphatic amino acids (Val in class IV and class I feglymycin) and Hpg in class I feglymycin and class II depsipeptide scaffolds. The scaffold is predicted to be O-methylated at an unknown position. If Orf 20 and 21 are considered part of this BGC, the scaffold is predicted to be glycosylated at two sites.

Analysis of class IV scaffolds

According to the NRPS domain phylogenies, the class IV scaffolds may be divided into three groups which do not include the kistamicin scaffold despite the shared features of

oxidative crosslinking and a singly putatively inactive A-domain following the X domain in the terminal module of the NRPS. Class IVa includes scaffolds related to complestatin, while class IVb is includes scaffolds encoded by the BGC found in *Streptomyces* sp. WAC01529. Class IVc consists of scaffolds related to *Streptomyces* sp. WAC06738 from our strain collection and a group of related organisms identified by sequence analysis from public sequence databases (Waglechner et al., 2019) (See Chapter 3).

The class IVc scaffolds lack an N-terminal condensation starter domain, suggesting the N-terminus is a primary amine. All scaffolds in this class are predicted to have 9 active and 1 terminal inactive A-domain. The phylogenetic profile for NRPS domains is not uniform, suggesting that some positions have an MRCA like the dalbaheptides, while others have differentiated in more complex ways. Each of these peptide scaffolds have multiple ambiguous A-domains predicted by phylogenetic analysis to adenylate Dpg (Supplementary Figure 4-10). The presence of an X domain suggests these scaffolds are oxidatively crosslinked, while the phylogenetic pattern observed for the X domain is monophyletic for these BGCs (Figure 4-9).

Four of the seven BGC sequences available for this class are predicted to incorporate Trp at position 4, while all are predicted to encode Hpg at position 6 and Tyr at position 8. This arrangement is analogous to the trio of crosslinked residues observed at positions 2, 4 and 6 in complestatin and kistamicin, suggesting that these residues are oxidatively crosslinked. In support of this idea, there are two candidate cytochrome P450 monooxygenases observed downstream of the NRPS in each of these BGCs which are

both more closely related to the homologous sequences of the same scaffold class than outside of this class (Figure 4-12, Supplementary Figure 4-37 and 4-40).

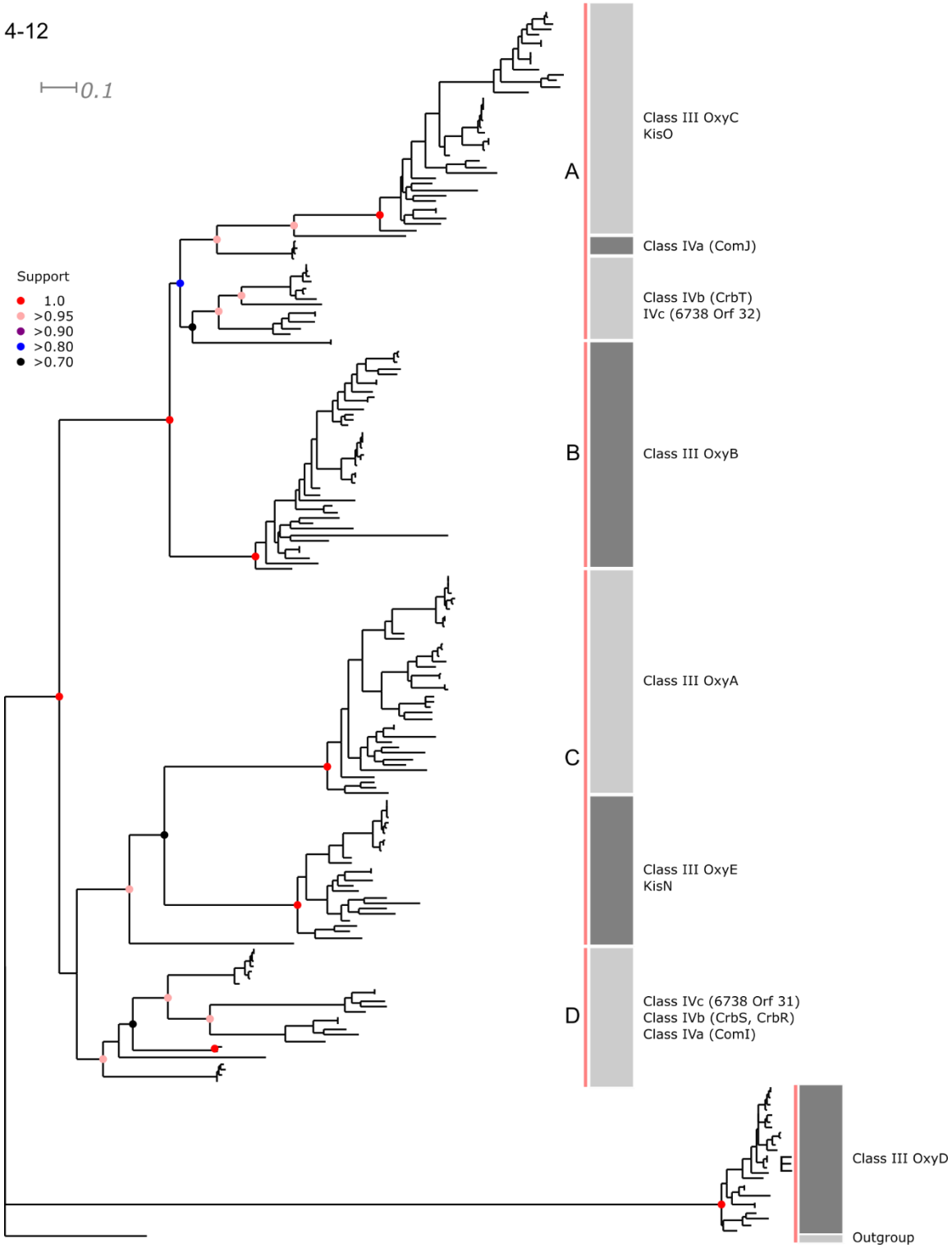


Figure 4-12: Approximate maximum likelihood phylogeny of GPA BGC cytochrome P450 monooxygenases. The amino acid sequences of GPA BGC cytochrome P450

monoxygenases were aligned and analyzed using the WAG substitution model with the default number of CAT approximation rate categories as implemented by fasttree2 (see methods). **a** The overall structure of the phylogeny is shown as a mid-point rooted tree. SH-like support values are summarized. All scale bars represent the expected number of substitutions per site. Insets for major clades are indicated using lowercase letters and include numerical support values. Subset A - Class III OxyC and KisO. Class IVa (including ComJ), Class IVb (including CrbT) and Class IVc including 6738 Orf 32. Subset B - Class III OxyB. Subset C - Class III OxyA and OxyE, KisN. Subset D - Class IVc including 6738 Orf 31, Class IVb including both CrbS and CrbR which appear to be a duplication as both are more similar to each other than each is to the Class IVc enzymes, and Class IVa which includes ComI. Subset E - Class III OxyD which is involved in β-OH tyrosine biosynthesis in *Amycolatopsis* and outgroup sequence not part of GPA biosynthesis.

The three BGCs that are not predicted to activate Trp at position 4 instead are predicted to activate Orn, however all A_{Trp} and A_{Orn} sequences from class IVc for this position form a monophyletic clade, and are sister to the A_{Trp} domains from the other class IV scaffolds as well as kistamicin, suggesting this domain is an ancestral feature regardless of the amino acid (Orn or Trp) being adenylated (Supplementary Figure 4-9). The class IVc A_{Hpg6} and A_{Tyr8} sequences are similarly monophyletic, highlighting their conservation and use as marker sequences for this class (Supplementary Figure 4-3, and 4-8). The A-domain encoding Val_3 also appears conserved in these scaffolds as these sequences form a monophyletic group sister to the repeated A_{Val} from the class I scaffold of feglymycin (Supplementary Figure 4-6). The phyletic profile of the remaining domains, primarily repeated A_{Dpg}, consist of a monophyletic clade with conflicting branching making it difficult to determine if there is an MRCA for these positions for the entire class or only within each BGC, or a mixture of both. Either way these domains are distinct but related to the A_{Dpg} domains from other scaffolds supporting the identity of this scaffold class

(Supplementary Figure 4-10). Based on this analysis, a structural prediction for product from *Streptomyces* sp. WAC6738 is provided in Figure 4-13.

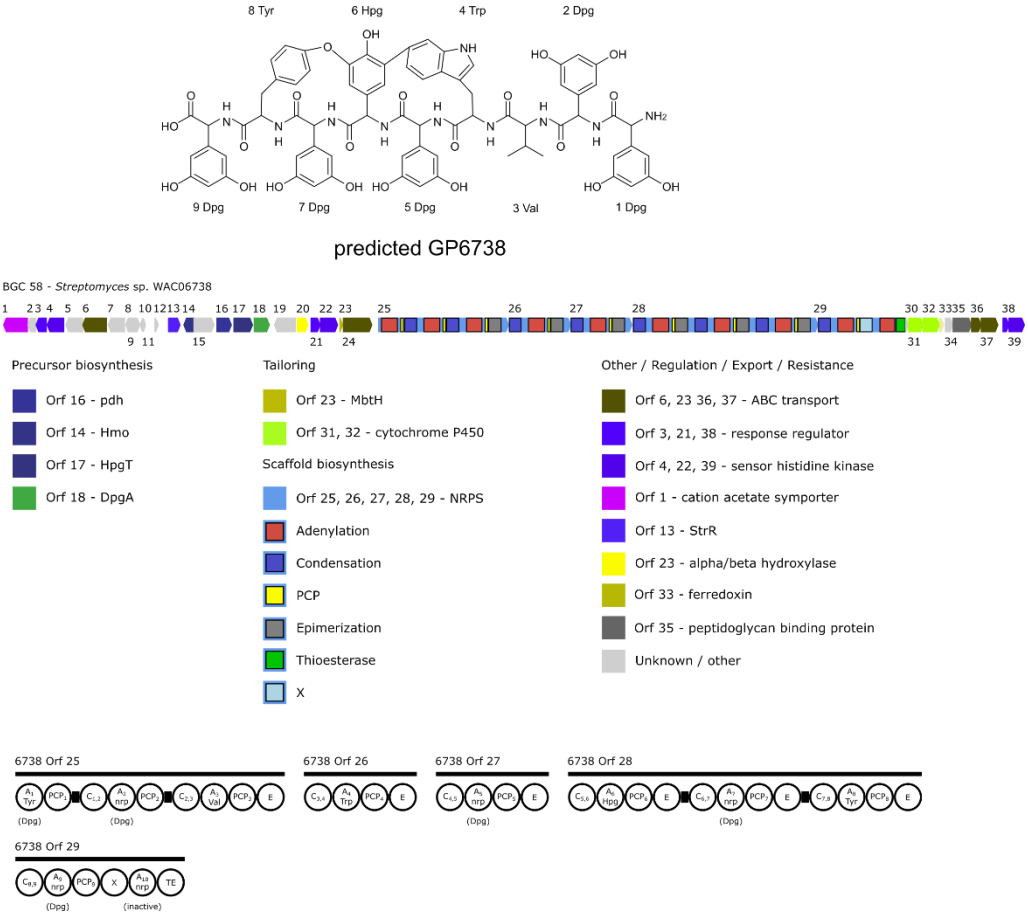


Figure 4-13: GP6738 predicted structure, BGC, and NRPS configuration. The BGC encoded by *Streptomyces* sp. WAC06738 encodes a class IVc scaffold. All scaffolds predicted in this class are highly similar. Note that Orfs 15 and 19 are erroneously identified by prodigal (as part of the antiSMASH pipeline) resulting in homologues of HmaS, DpgB and DpgC being missing from this BGC despite Hpg and Dpg being incorporated in this scaffold. The predicted structure is not likely to be modified due to the lack of obvious tailoring reactions. Two crosslinks are predicted to be installed in this scaffold, both linking to the central Hpg in analogy to the GPAs with known structures, particularly kistamicin and complestatin. No obvious resistance genes are identified in this cluster. These BGCs include the hallmark inactive A-domain in the terminal NRPS module.

Streptomyces sp. WAC01325 produces the class IVa scaffold complestatin

The BGC encoding the production of the Nicolau type V GPA complestatin has been previously reported in *Streptomyces lavendulae* (Chiu et al., 2001). Sequence analysis and antiSMASH analysis revealed complestatin BGCs in the genomes of *Streptomyces* sp. WAC01325, *Streptomyces* sp. WAC05379, and *Streptomyces* sp. KS 5 (not shown) (Figure 4-14). Phylogenetic analysis of the A, C, PCP, X and TE domains of the NRPS sequences predicted in these clusters reveal monophyletic clades consisting of these four BGCs for every component to the exclusion of all other BGCs (A-domains: Supplementary Figures 4-3, and 4-8 to 4-10) (C-domains: Figure 4-6, Supplementary 4-13, and 4-15 to 4-18) (PCP domains: Figure 4-7, Supplementary Figure 4-27, 4-28, 4-31, and 4-33) (Figure 4-8). The TE phylogeny is presented in Figure 4-10. Complestatin was successfully extracted from the spent media *Streptomyces* sp. WAC01325 (Culp et al., 2020) (Appendix 2).

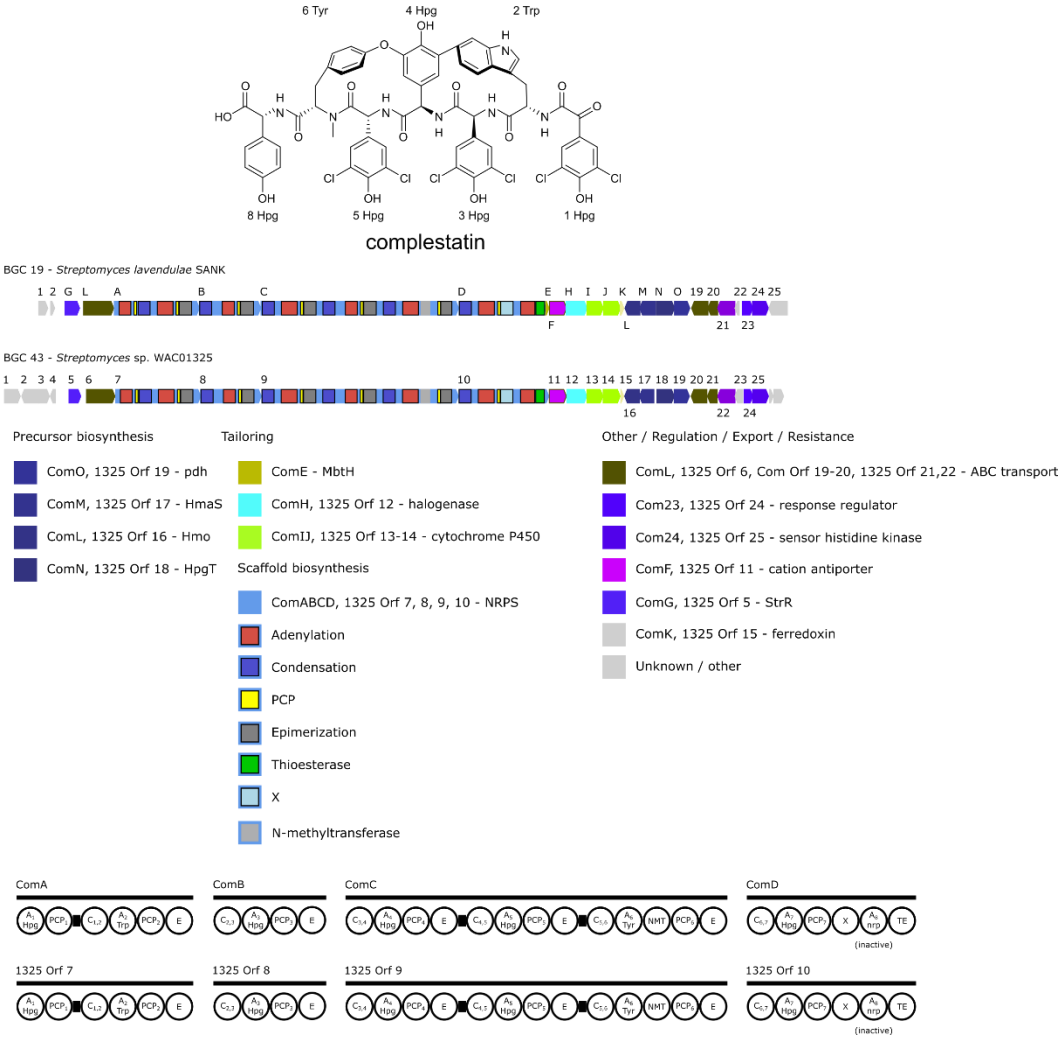


Figure 4-14: Complestatin structure, BGC, and NRPS configuration. *Streptoymces* sp. WAC01325 encodes a BGC with similarity to the BGC reported in *Streptomyces lavendulae* known to produce the type IVa scaffold complestatin. These scaffolds lack Dpg, but include repeated Hpg amino acids. The N-terminus of this scaffold is deaminated via an unknown process. This NRPS includes an N-methyltransferase (N-MT) as part of module 6, and is halogenated in multiple positions but does not include other tailoring enzymes. Another distinguishing characteristic of Class IV scaffolds is the presence of a putatively inactive A-domain in the terminal NRPS module.

Other BGCs in the GPA family are predicted to produce similar scaffolds to complestatin. *Streptomyces* sp. strains WAC06725, WAC00631, and WAC06783, while

predicted to produce a similar scaffold, the repeated A_{Hpg} domains in these three BGCs are more closely related to A_{Hpg} domains from the class IVb and IVc BGCs than IVa (complestatin) (Supplementary Figure 4-8). Interestingly, the NRPS in WAC00631 appears to have lost the N-methyltransferase domain that methylates the α nitrogen in the amide bond between the 5^{th} and 6^{th} positions of complestatin and the other class IVa scaffolds (Figure 4-14, Figure 4-15). All three of these strains have a starter unit condensation domain predicted to acylate the N-terminus of the scaffold but the nature of this acyl group modification is unknown (Supplementary Figure 4-23). These starter condensation domains belong to a monophyletic clade in the C-domain tree along with those observed in class IVb and the class II scaffolds (enduracidin and ramoplanin). Similarly, the four complestatin-like BGCs and three complestatin-variant BGCs are all predicted to have a halogenase enzyme. These halogenases form a monophyletic group in the halogenase phylogeny while recapitulating the division between the four complestatin and the three other complestatin-variant BGCs (Figure 4-16). In a similar fashion, the putative crosslinking cytochrome P450 monoxygenases in these three variant BGCs are more closely related to the class IVb and IVc scaffolds than the class IVa scaffolds (Figure 4-12, Supplementary Figure 4-37, and 4-40).

Figure 4-15: Variant class IVa BGC comparison. *Streptomyces* sp. WAC06725, WAC06783 and WAC00631 encode variant class IVa scaffolds with similarity to complestatin but having a different phylogenetic profile (see main text). The position of the Hpg biosynthesis genes downstream of the NRPS orfs differs from the other Class IVa BGCs. WAC00631 additionally has transposon inserted downstream of the NRPS orfs and appears to have lost both the starter condensation domain and the N-methyltransferase domain relative to the other BGCs in this group. As with the other Class IV scaffolds, the terminal modules of these NRPS sequences include a putatively inactive A-domain.

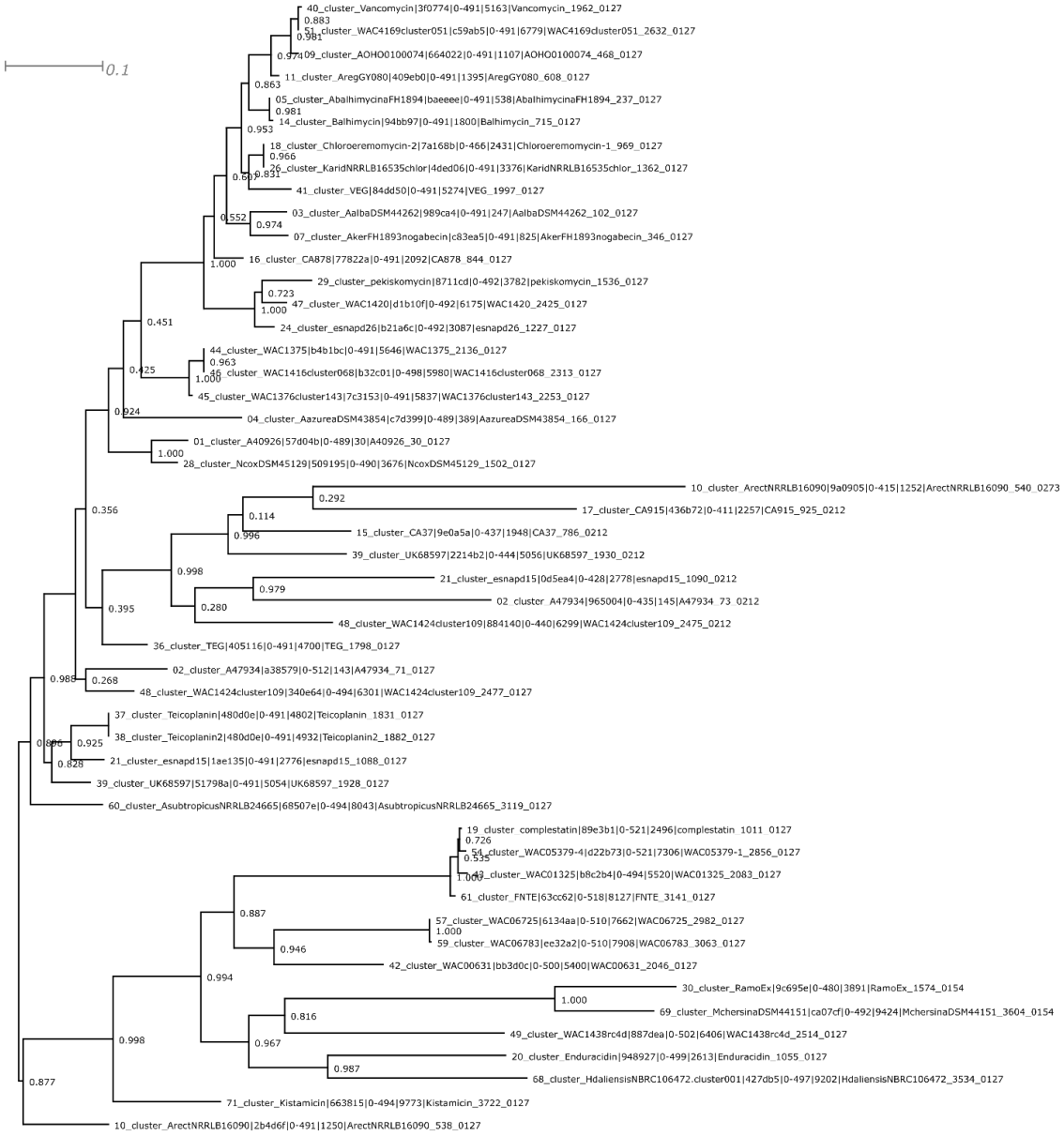


Figure 4-16: Approximate maximum likelihood phylogeny of GPA BGC halogenases. The amino acid sequences identified by the integrated antiSMASH HMM in the GPA BGC sequences were aligned and analyzed using the WAG substitution model with the default number of CAT approximate rate categories implemented in fasttree2. The phylogeny is represented as a mid-point rooted tree. SH-like support values are indicated. The scale bar represents the expected number of substitutions per site. Aside from one sequence from *Actinoplanes rectilineatus* NRRL B-6090 the sequences from Classes IVa, Class I and kistamicin are more similar to each other than the sequences from the Class III scaffolds.

Streptomyces sp. WAC01529 produces corbomycin

Based on the overall phylogenetic and sequence analysis of the BGC clusters, specifically the class IVb scaffolds, the putative structure of GP1529 (the product of the WAC1529 BGC) is depicted in Figure 14-17. The product extracted from this strain based on this prediction was named corbomycin in reference to *Nid de corbeau* after Crowsnest Pass, Alberta, the location of the soil sample from which *S.* sp. WAC1529 was isolated. The genes in the BGC labeled with the symbol *crb* (Culp et al., 2020) (Appendix 2).

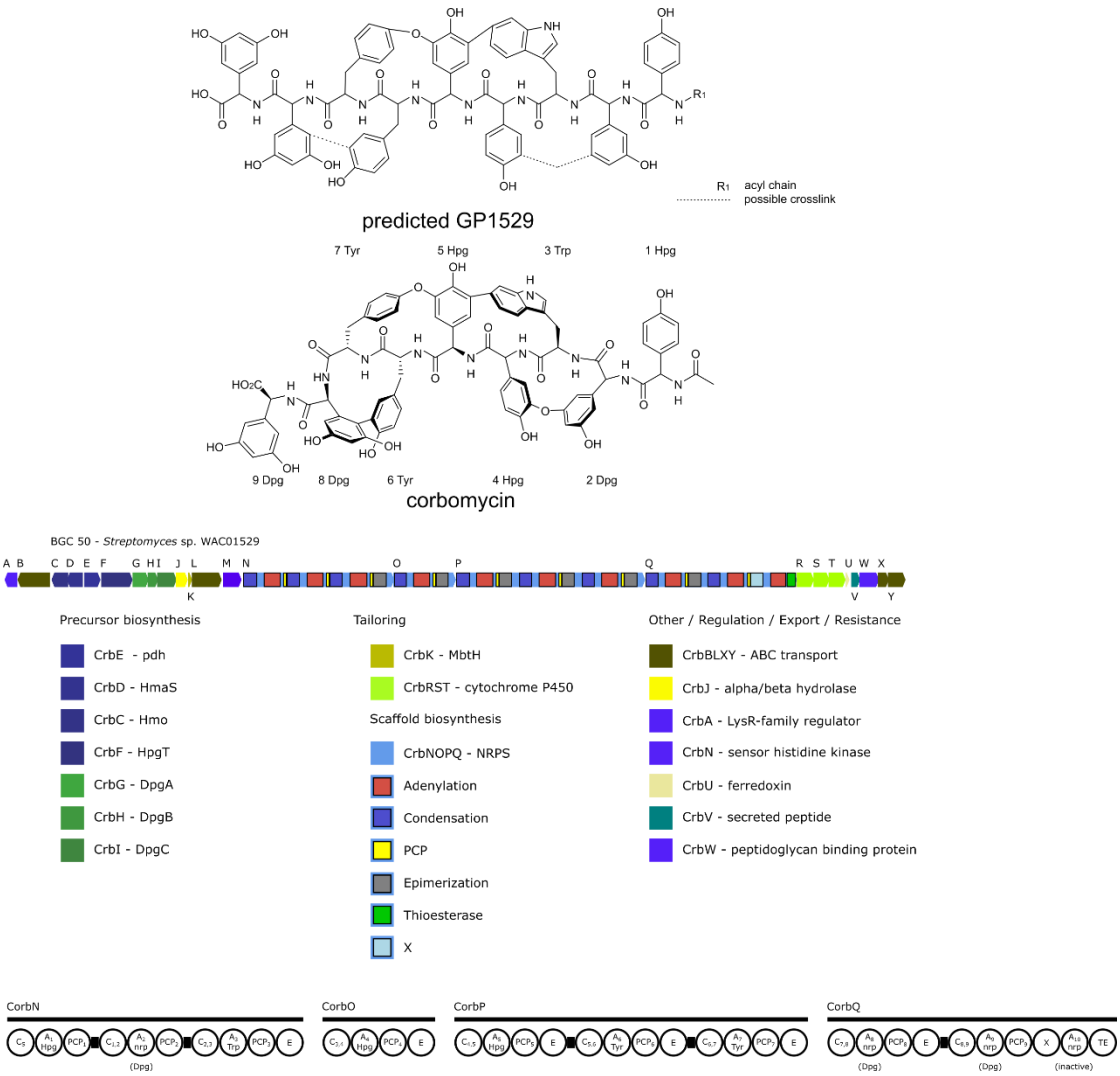


Figure 14-17: GP1529 prediction and structure of corbomycin, the corbomycin BGC and NRPS configuration. *Streptomyces* sp. WAC01529 a nonapeptide NRPS BGC. Based on nearest-neighbour phylogenetic analysis the cryptic A-domains were identified as A_{Dpg}. The NRPS includes the hallmark inactive A-domain in the terminal module of the NRPS. We predict the N-terminus is acylated, and that the scaffold has the hallmark Class IV Trp-Hpg and Hpg-Tyr crosslinks. The presence of an additional cytochrome P450 monooxygenase suggested possible crosslinks between Dpg-Hpg or Tyr-Dpg. This BGC does not include apparent resistance genes. The structure of corbomycin agrees well with the prediction. The acyl group on the N-terminus was observed to be acetate, and both predicted crosslinks were observed in the structure of the main product extracted from *Streptomyces* sp. WAC01529.

The acyl-group on the N-terminus, predicted based on the presence of a starter condensation domain was determined to be an acetyl group in the major product (Supplementary Figure 4-23). There are three different cytochrome P450 monooxygenase sequences identified in the corbomycin BGC, *crbR*, *crbS* and *crbT* (Figure 4-12, Supplementary Figures 4-37, and 4-40). The prediction of possible crosslinking was based on the geometry of (Hpg_7, Tyr_9) and (Hpg_3, Dpg_5) pairs in addition to the well-conserved (Tyr_8-Hpg_6) and (Hpg_6, Trp_4) pairs also observed in complestatin and kistamicin (Figure 4-4, Figure 4-14). Work on *in vitro* reconstructed minimal NRPS systems containing the X domain and PCP-linked scaffolds *in vitro* suggests the identity of the P450 responsible for the ring in complestatin and kistamicin (Greule et al., 2019; Mollo et al., 2017). Corbomycin was observed to have intramolecular ring systems suggesting a bi-functional cytochrome P450 responsible for two crosslinks as in the kistamicin BGC (Nazari et al., 2017) though the P450 phylogeny does not clearly suggest which enzyme is responsible for which crosslink (Figure 4-12, Supplementary 4-37, and 4-40).

Notably, no sequences with similarity to the known cell-wall modifying glycopeptide resistance genes were observed in these BGCs (Marcone et al., 2010; Yim, Thaker, et al., 2014). There are sensor histidine kinase and response regulators sequences present which are not included in the monophyletic clade of VanS and VanR sequences (Figure 4-19, 4-20) (Arthur, Molinas, & Courvalin, 1992; Hong, Hutchings, Buttner, Biotechnology, & Biological Sciences Research Council, 2008; Kilian et al., 2016). The class IVb and IVc

clusters are observed to have multiple pairs but it is unclear which, if any, downstream sequences are under the control of these regulators.

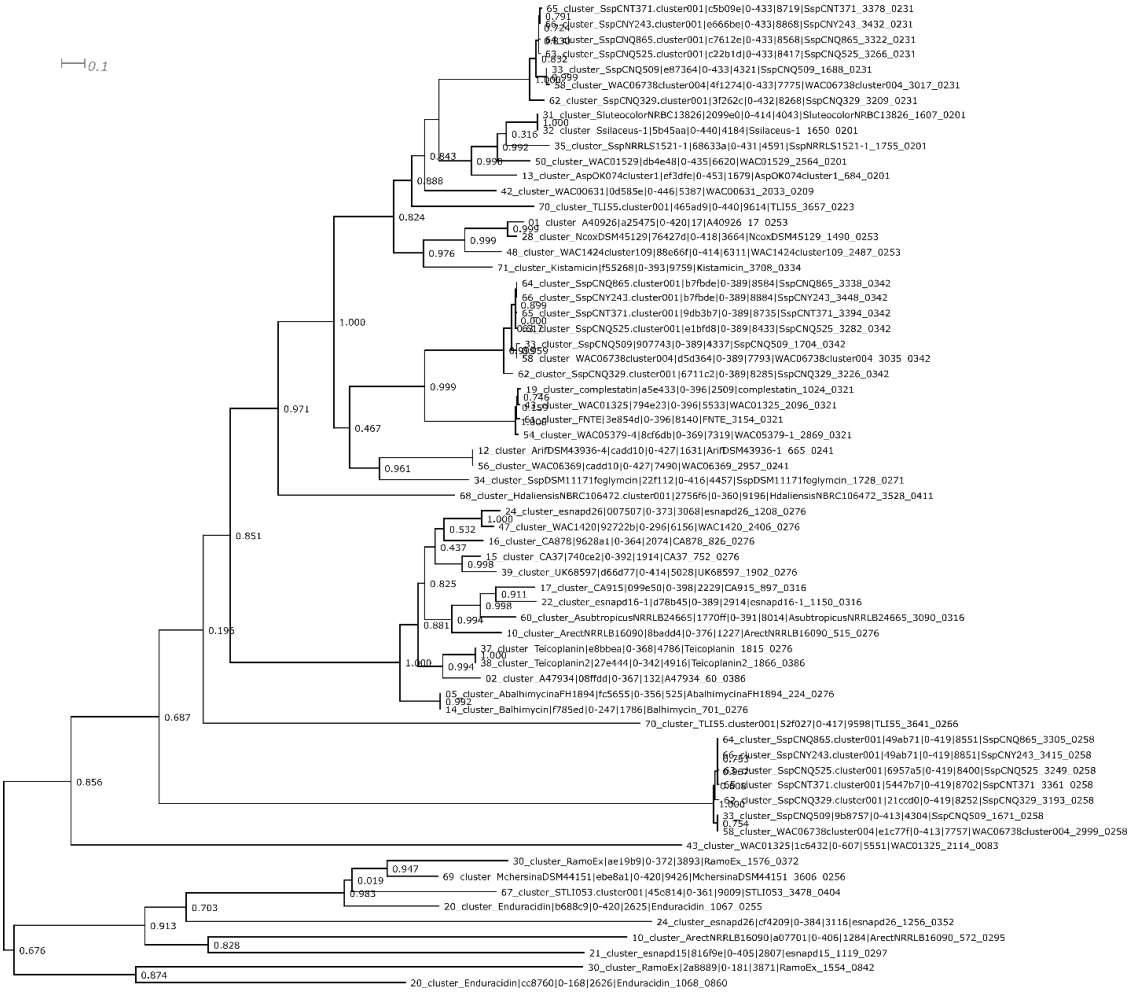


Figure 14-18: Approximate maximum likelihood phylogeny of GPA BGC sensor histidine kinases. The amino acid sequences of sensor histidine kinases were extracted from the set of GPA BGCs and subjected to hierarchical clustering using USEARCH with a 60% identity cutoff (see methods). The largest and most widely distributed family of sequences was aligned and analyzed using the WAG substitution model with the default number or CAT approximate rate categories as implemented by fasttree2. The phylogeny is shown as a mid-point rooted tree. The scale bar represents the expected number of substitutions per site.

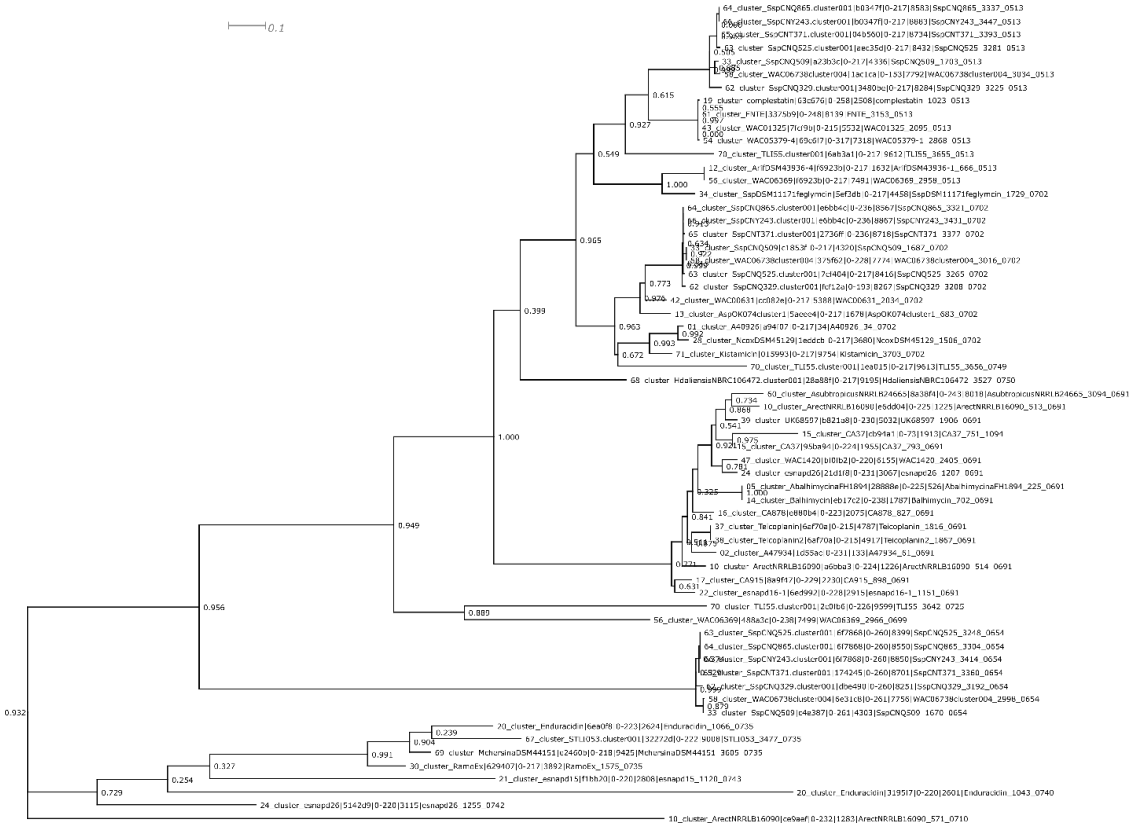


Figure 14-19: Approximate maximum likelihood phylogeny of GPA BGC response regulators. The amino acid sequences of DNA-binding response regulators were extracted from the GPA BGCs and subjected to a hierarchical clustering step using USEARCH with a 60% identity cutoff (see methods). Sequences from the largest and most widely distributed family were aligned and analyzed using the WAG substitution model and the default number of CAT approximate rate categories as implemented by fasttree2 (see methods). The phylogeny is depicted as a mid-point rooted tree. The scale bar indicates the expected number of substitutions per site.

Corbomycin and complestatin have antibiotic activity

Investigators have previously attributed the antimicrobial activity of complestatin to inhibition of bacterial fatty acid biosynthesis through FabI, even reporting slight cross resistance with the compound triclosan (Kwon, Kim, & Kim, 2015). MICs for

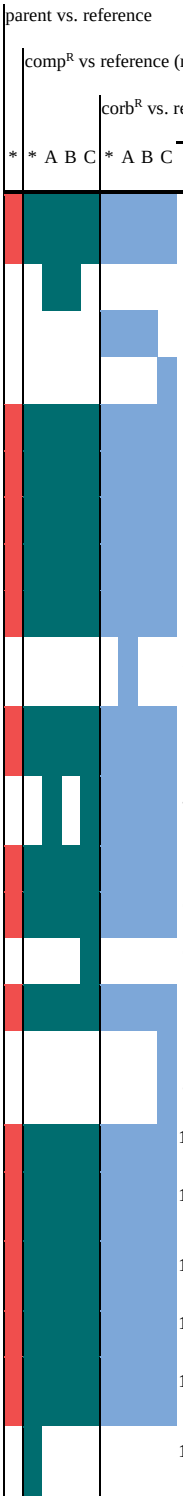
corbomycin, complestatin and triclosan were determined against several pathogen strains, shown in Table 4-3.

Table 4-3: Initial broth dilution MIC values (mg/ml). All MICs determined in cation-adjusted Mueller-Hinton broth (CAMHB), except *E. faecium*, which was repeated in brain-heart infusion broth (BHI).

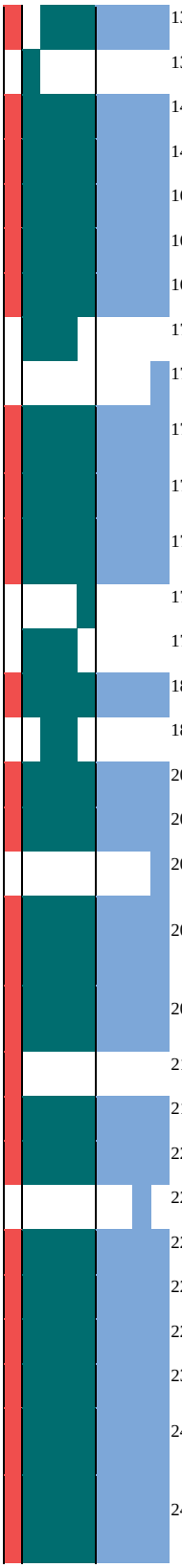
strain	complestatin	triclosan (ng/ml)	vancomycin	corbomycin
Pseudomonas aeruginosa PA01	>128 *	>128000	>128	>128
Escherichia coli ATCC 25922	>128	125	>128	>128
Acinetobacter baumanii ATCC 17987	64	250	>128	64
Enterobacter aerogenes ATCC 13048	>128	500	>128	>128
Enterococcus faecium ATCC 19434 (CAMHB)	2	8000	0.5	NT
Enterococcus faecium ATCC 19434 (BHI)	>128	4000	2	8
Escherichia coli BW25113 Δ*bamB* Δ*tolC*	4	3.9	8	>128
Staphylococcus aureus ATCC 29213	2	16	1	1
Klebsiella pneumoniae ATCC 33495	>128	500	>128	>128
Bacillus subtilis 168	1	250	0.25	1

These MIC values were later recapitulated and expanded to include vancomycin-intermediate and resistant strains of *Staphylococcus* and *Enterococcus* (Culp et al., 2020) (Appendix 2). The lack of activity against these strains suggested a different mechanism of action than both typical glycopeptide antibiotics like vancomycin and triclosan. Resistant mutants of *B. subtilis* 168 and *S. aureus* ATCC 29213 were raised in liquid media over the course of three weeks demonstrating a stable increase of 4x MIC over the parental strains (Culp et al., 2020) (Appendix 2). Sequencing of the parental and replicates of resistant mutant strains followed by comparison against wildtype reference sequences revealed a constellation of variants that suggest a novel activity against the bacterial cell-wall (*B. subtilis* Table 4-4, *S. aureus* Table 4-5).

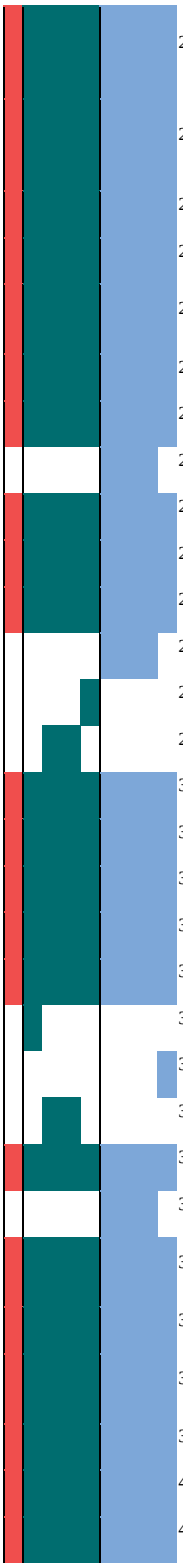
Table 4-4: Genotypes of *B. subtilis* 168 strains raised to be resistant to complestatin and corbomycin compared to parental and NCBI reference genome (accession NC_000964.1).



bp	detected genotype	variant	context	affected locus
52646	C→T	intergenic (+94/-117)	*yabG* → / → *veg*	sporulation-specific protease YabG/protein Veg
77756	T→C	F258S (TTC→TCC)	*ftsH* →	ATP-dependent zinc metalloprotease FtsH
77788	C→T	R269C (CGT→TGT)	*ftsH* →	ATP-dependent zinc metalloprotease FtsH
117650	(TTATCTTTTTTG)2→1	coding (119-130/180 nt)	*secE* →	preprotein translocase subunit SecE
165749	Δ1 bp	intergenic (+42/-5)	*rrnI-5S* → / → *trnI-Asn*	5S ribosomal RNA/tRNA-Asn
165751	Δ1 bp	intergenic (+44/-3)	*rrnI-5S* → / → *trnI-Asn*	5S ribosomal RNA/tRNA-Asn
165826	+C	intergenic (+1/-4)	*trnI-Asn* → / → *trnI-Thr*	tRNA-Asn/tRNA-Thr
166037	+T	intergenic (+4/-27)	*trnI-Gly* → / → *trnI-Arg*	tRNA-Gly/tRNA-Arg
166345	Δ1 bp	intergenic (+17/-155)	*trnI-Ala* → / → *rrnH-16S*	tRNA-Ala/16S ribosomal RNA
217697	(A)9→8	coding (1/1344 nt)	*skfF* →	bacteriocin-SkfA transport system permease SkfF
376025	(G)8→7	intergenic (-227/-7)	*hxlA* ← / → *hxlR*	3-hexulose-6-phosphate synthase/HxlR family transcriptional regulator
490597	(T)17→16	intergenic (+43/+180)	*ydaP* → / ← *ydzK*	thiamine pyrophosphate-containing protein YdaP/hypothetical protein
557870	(T)5→6	intergenic (+421/-3)	*yddT* → / → *ydzN*	hypothetical protein/hypothetical protein
608219	(A)5→6	intergenic (-144/+27)	*ydgF* ← / ← *dinB*	transporter/protein DinB
679521	(A)6→5	coding (132/372 nt)	*ydjM* →	hypothetical protein
774704	(A)7→6	coding (567/654 nt)	*yesY* →	rhamnogalacturonan acetylesterase YesY
796377	Δ5 bp	coding (64-68/1920 nt)	*ltaSA* →	lipoteichoic acid synthase
942138	A→G	F31S (TTC→TCC)	*ygaE* ←	hypothetical protein
1073117	C→A	V201L (GTA→TTA)	*hpr* ←	MarR family transcriptional regulator
1073879	(T)7→6	intergenic (-162/+16)	*hpr* ← / ← *yhaH*	MarR family transcriptional regulator/hypothetical protein
1224524	T→G	V357G (GTT→GGT)	*oppD* →	oligopeptide transport ATP-binding protein OppD
1264284	T→A	K216N (AAA→AAT)	*yjcM* ←	hypothetical protein
1317152	+GT	intergenic (+23/+5)	*phrA* → / ← *yjpA*	phosphatase RapA inhibitor/hypothetical protein
1317154	+CTT	intergenic (+25/+3)	*phrA* → / ← *yjpA*	phosphatase RapA inhibitor/hypothetical protein



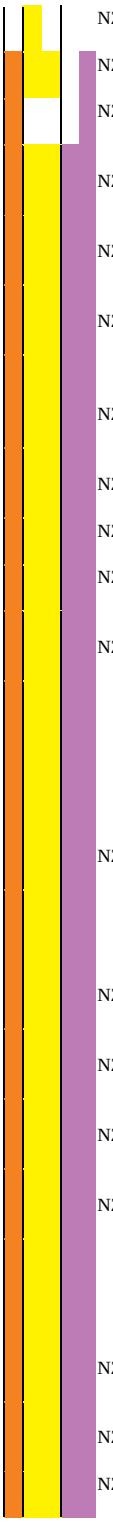
1317158	(T)5→6	coding (257/258 nt)	*yjpA* ←	hypothetical protein
1350817	C→T	A95T (GCC→ACC)	*ykaA* ←	hypothetical protein
1412484	T→G	L198R (CTG→CGG)	*sigI* →	RNA polymerase sigma factor SigI
1424639	T→G	intergenic (-199/+128)	*mtnK* ← / ← *mtnU*	methylthioribose kinase/hydrolase
1610896	T→C	M11T (ATG→ACG)	*sepF* →	cell division machinery factor
1618068	C→T	P287S (CCG→TCG)	*rluD* →	RNA pseudouridine synthase YlyB
1675849	C→T	H227Y (CAC→TAC)	*trmD* →	tRNA (guanine-N(1)-)-methyltransferase
1734164	(A)6→7	coding (156/2151 nt)	*infB* →	translation initiation factor IF-2
1735977	C→T	R657C (CGT→TGT)	*infB* →	translation initiation factor IF-2
1741610	(G)7→6	intergenic (+110/-8)	*pnpA* → / → *ylxY*	polyribonucleotide nucleotidyltransferase/hypothetical protein
1756603	G→A	A319T (GCT→ACT)	*ymfD* →	bacillibactin exporter
1764558	C→T	intergenic (+86/-87)	*cinA* → / → *recA*	competence-damage inducible protein/recombinase RecA
1769091	(A)6→7	coding (151/795 nt)	*ymdB* →	hypothetical protein
1769203	C→A	P88Q (CCG→CAG)	*ymdB* →	hypothetical protein
1841174	(A)6→7	coding (6766/16467 nt)	*pksN* →	polyketide synthase PksN
1890172	A→G	K467E (AAG→GAG)	*xynB* →	beta-xylosidase
2011091	G→A	A1181V (GCA→GTA)	*gltA* ←	glutamate synthase [NADPH] large chain
2041099	A→G	D43D (GAT→GAC)	*yozT* ←	hypothetical protein
2047140	(T)7→6	coding (272/432 nt)	*yoaW* ←	hypothetical protein
2096253	(A)8→7	intergenic (-468/+97)	*yocI* ← / ← *yocJ*	ATP-dependent DNA helicase RecQ/FMN-dependent NADH-azoreductase 1
2097087	(A)7→8	intergenic (-111/+39)	*yocJ* ← / ← *yocK*	FMN-dependent NADH-azoreductase 1/general stress protein 16O
2153580	(C)6→7	coding (73/213 nt)	*yotJ* ←	hypothetical protein
2174758	(CT)4→3	intergenic (-24/+93)	*yorN* ← / ← *yorM*	hypothetical protein/hypothetical protein
2201408	A→G	L23P (CTG→CCG)	*yoqA* ←	hypothetical protein
2249090	T→C	V225A (GTC→GCC)	*yomJ* →	hypothetical protein
2271424	T→C	R78R (AGA→AGG)	*uvrX* ←	UV-damage repair protein UvrX
2271505	C→T	V51V (GTG→GTA)	*uvrX* ←	UV-damage repair protein UvrX
2271523	A→C	D45E (GAT→GAG)	*uvrX* ←	UV-damage repair protein UvrX
2366016	T→C	H87R (CAT→CGT)	*ypiB* ←	hypothetical protein
2421606	T→C	Q189Q (CAA→CAG)	*rluB* ←	ribosomal large subunit pseudouridine synthase B
2480646	2 bp→AT	intergenic (+21/+103)	*zwf* → / ← *yqjI*	glucose-6-phosphate 1-dehydrogenase/6-phosphogluconate dehydrogenase



2480654	Δ1 bp	intergenic (+29/+96)	*zwf* → / ← *yqjI*	glucose-6-phosphate 1-dehydrogenase/6-phosphogluconate dehydrogenase
2480667	Δ1 bp	intergenic (+42/+83)	*zwf* → / ← *yqjI*	glucose-6-phosphate 1-dehydrogenase/6-phosphogluconate dehydrogenase
2546160	(C)7→6	coding (717/1467 nt)	*gcvPB* ←	glycine dehydrogenase subunit 2
2560902	C→T	E181K (GAA→AAA)	*yqxL* ←	magnesium transport protein CorA
2581732	(T)6→7	intergenic (-115/+39)	*pstS* ← / ← *pbpA*	phosphate-binding protein PstS/hypothetical protein
2619105	C→T	A225A (GCG→GCA)	*yqeZ* ←	hypothetical protein
2814468	C→T	intergenic (-61/+275)	*yrvM* ← / ← *aspS*	hypothetical protein/aspartate--tRNA ligase
2837088	C→A	V112L (GTG→TTG)	*bofC* ←	general stress protein BofC
2893906	G→A	A268A (GCC→GCT)	*ilvC* ←	ketol-acid reductoisomerase
2982417	T→A	intergenic (-148/+186)	*citZ* ← / ← *ytwI*	citrate synthase 2/hypothetical protein
2982437	T→C	intergenic (-168/+166)	*citZ* ← / ← *ytwI*	citrate synthase 2/hypothetical protein
2985711	G→A	Q279* (CAA→TAA)	*pyk* ←	pyruvate kinase
2985815	C→T	R244H (CGC→CAC)	*pyk* ←	pyruvate kinase
2986083	(T)6→7	coding (463/1758 nt)	*pyk* ←	pyruvate kinase
3051461	G→A	P375S (CCA→TCA)	*ytpS* ←	DNA translocase SftA
3253956	T→C	E628G (GAG→GGG)	*comP* ←	sensor histidine kinase
3319154	C→T	A37T (GCA→ACA)	*yutE* ←	hypothetical protein
3391676	A→G	T299A (ACA→GCA)	*gerAA* →	spore germination protein A1
3527377	G→A	A276V (GCG→GTG)	*epsC* ←	polysaccharide biosynthesis protein EpsC
3574563	G→A	Q408* (CAA→TAA)	*cwlO* ←	peptidoglycan DL-endopeptidase CwlO
3575475	(T)6→5	coding (310/1422 nt)	*cwlO* ←	peptidoglycan DL-endopeptidase CwlO
3624531	G→A	Q100* (CAG→TAG)	*ftsX* ←	cell division protein FtsX
3696869	T→C	L210L (CTA→CTG)	*pgdS* ←	gamma-dl-glutamyl hydrolase
3748096	G→A	R54C (CGT→TGT)	*mbl* ←	MreB-like protein
3770066	(A)8→9	intergenic (-958/+38)	*ureA* ← / ← *csbD*	urease subunit gamma/stress response protein CsbD
3902306	C→A	L448F (TTG→TTT)	*sacA* ←	sucrose-6-phosphate hydrolase
3935825	Δ1 bp	coding (1190/1191 nt)	*ywbD* ←	ribosomal RNA large subunit methyltransferase YwbD
3993539	G→T	G126G (GGG→GGT)	*yxjM* →	sensor histidine kinase
4095811	C→T	V9I (GTA→ATA)	*yxbD* ←	hypothetical protein
4155395	(A)5→6	intergenic (-186/+38)	*trnY-Lys* ← / ← *purA*	tRNA-Lys/adenylosuccinate synthase

Table 4-5: Genotypes of *S. aureus* ATCC 29213 strains raised to be resistant to complestatin and corbomycin compared to parental and reference NCBI genome (accession NZ_MOPB0100000000).

parent vs reference	comp^R vs ref (replicates)		corb^R vs ref (replicates)							
*	A	B	A	B	contig accession	bp	detected genotype	variant	context	affected locus
					NZ_MOPB01000004	128537	G→A	W275* (TGG→TGA)	*BJI72_RS00850* →	glycosyltransferase family 2 protein
					NZ_MOPB01000004	128612	+A	coding (900/1722 nt)	*BJI72_RS00850* →	glycosyltransferase family 2 protein
					NZ_MOPB01000004	128895	A→T	K395* (AAG→TAG)	*BJI72_RS00850* →	glycosyltransferase family 2 protein
					NZ_MOPB01000008	14257	G→A	E167K (GAA→AAA)	*BJI72_RS02450* →	pur operon repressor
					NZ_MOPB01000010	44011	N→C	?807D (GAN→GAC)	*BJI72_RS02875* →	hydrolase
					NZ_MOPB01000011	91300	C→T	Q86* (CAA→TAA)	*BJI72_RS03380* →	DNA-binding response regulator
					NZ_MOPB01000011	92676	C→A	T322K (ACA→AAA)	*BJI72_RS03385* →	sensor histidine kinase
					NZ_MOPB01000011	94903	(A)6→5	coding (1257/1890 nt)	*BJI72_RS03395* →	bacitracin ABC transporter permease
					NZ_MOPB01000013	74205	N→C	?663S (TCN→TCC)	*BJI72_RS04140* →	clumping factor A
					NZ_MOPB01000013	74328	N→C	?704D (GAN→GAC)	*BJI72_RS04140* →	clumping factor A
					NZ_MOPB01000013	74598	N→C	?794D (GAN→GAC)	*BJI72_RS04140* →	clumping factor A
					NZ_MOPB01000013	74661	N→C	?815S (TCN→TCC)	*BJI72_RS04140* →	clumping factor A
					NZ_MOPB01000014	184069	C→A	Q26K (CAA→AAA)	*BJI72_RS05215* →	phosphoribosylformylglycinamidine synthase subunit PurL
					NZ_MOPB01000016	31132	A→G	V110A (GTA→GCA)	*BJI72_RS05550* ←	heme uptake protein IsdB
					NZ_MOPB01000028	157210	C→A	R119L (CGT→CTT)	*BJI72_RS08705* ←	autolysin
					NZ_MOPB01000031	1166	G→A	S137S (AGC→AGT)	*BJI72_RS09005* ←	serine protease SplF
					NZ_MOPB01000033	27139	C→T	intergenic (-14/+60)	*BJI72_RS09765* ← / ← *BJI72_RS15725*	methionine aminopeptidase/hypothetical protein
					NZ_MOPB01000036	47206	(TTA)2→3	coding (152/450 nt)	*BJI72_RS10715* ←	hypothetical protein
					NZ_MOPB01000042	2313	Δ30 bp	coding (40-69/1260 nt)	*BJI72_RS11965* →	lysostaphin resistance protein A
					NZ_MOPB01000042	2558	+T	coding (285/1260 nt)	*BJI72_RS11965* →	lysostaphin resistance protein A
					NZ_MOPB01000046	8793	N→A	?917T (ACN→ACT)	*BJI72_RS13170* ←	fibronectin-binding protein A
					NZ_MOPB01000046	148616	N→A	?707D (GAN→GAT)	*BJI72_RS13920* ←	clumping factor B
					NZ_MOPB01000046	148640	N→A	?699D (GAN→GAT)	*BJI72_RS13920* ←	clumping factor B



NZ_MOPB01000046	148679	N→A	?686S (AGN→AGT)	*BJI72_RS13920* ←	clumping factor B
NZ_MOPB01000046	148712	N→C	?675E (GAN→GAG)	*BJI72_RS13920* ←	clumping factor B
NZ_MOPB01000051	338	A→G	pseudogene (338/1574 nt)	*BJI72_RS14735* →	hypothetical protein
NZ_MOPB01000060	1	Δ1,543 bp		*BJI72_RS15055–BJI72_RS15060*	*BJI72_RS15055, BJI72_RS15060*
NZ_MOPB01000063	1	Δ1,839 bp		*[BJI72_RS15105]–[BJI72_RS15110]*	*[BJI72_RS15105], [BJI72_RS15110]*
NZ_MOPB01000064	1	Δ1,311 bp		*BJI72_RS15115–BJI72_RS15125*	*BJI72_RS15115, BJI72_RS15120, BJI72_RS15125*
NZ_MOPB01000066	1	Δ2,251 bp		*BJI72_RS15145–[BJI72_RS15160]*	*BJI72_RS15145, BJI72_RS15805, BJI72_RS15155, [BJI72_RS15160]*
NZ_MOPB01000068	1	Δ1,109 bp		*BJI72_RS15175–[BJI72_RS15180]*	*BJI72_RS15175, [BJI72_RS15180]*
NZ_MOPB01000071	1	Δ1,889 bp		*[BJI72_RS15215]*	*[BJI72_RS15215]*
NZ_MOPB01000072	1	Δ1,042 bp		*BJI72_RS15220*	BJI72_RS15220
NZ_MOPB01000073	1	Δ1,737 bp		*[BJI72_RS15225]–BJI72_RS15235*	*[BJI72_RS15225], BJI72_RS15230, BJI72_RS15235*
NZ_MOPB01000074	1	Δ1,452 bp		*BJI72_RS15240–BJI72_RS15305*	*BJI72_RS15240, BJI72_RS15245, BJI72_RS15250, BJI72_RS15255, BJI72_RS15260, BJI72_RS15265, BJI72_RS15270, BJI72_RS15275, BJI72_RS15280, BJI72_RS15285, BJI72_RS15290, BJI72_RS15295, BJI72_RS15300, BJI72_RS15305*
NZ_MOPB01000076	1	Δ1,258 bp		*[BJI72_RS15320]–[BJI72_RS15360]*	*[BJI72_RS15320], rrf, BJI72_RS15330, BJI72_RS15335, BJI72_RS15340, BJI72_RS15345, BJI72_RS15350, BJI72_RS15355, [BJI72_RS15360]*
NZ_MOPB01000077	1	Δ1,241 bp		*[BJI72_RS15365]–[BJI72_RS15375]*	*[BJI72_RS15365], rrf, [BJI72_RS15375]*
NZ_MOPB01000078	1	Δ1,125 bp		*[BJI72_RS15380]–BJI72_RS15385*	*[BJI72_RS15380], BJI72_RS15385*
NZ_MOPB01000079	1	Δ1,448 bp		*[BJI72_RS15390]–[BJI72_RS15400]*	*[BJI72_RS15390], BJI72_RS15395, [BJI72_RS15400]*
NZ_MOPB01000083	1	Δ1,575 bp		*[BJI72_RS15440]–rrf*	*[BJI72_RS15440], BJI72_RS15445, BJI72_RS15450, BJI72_RS15455, BJI72_RS15460, BJI72_RS15465, BJI72_RS15470, BJI72_RS15475, BJI72_RS15480, rrf*
NZ_MOPB01000084	1	Δ1,055 bp		*BJI72_RS15490–BJI72_RS15495*	*BJI72_RS15490, BJI72_RS15495*
NZ_MOPB01000085	1	Δ1,298 bp		*BJI72_RS15500*	BJI72_RS15500

NZ_MOPB01000086	1	Δ1,181 bp	*BJI72_RS15505–[BJI72_RS15560]*	*BJI72_RS15505, BJI72_RS15510, BJI72_RS15515, BJI72_RS15520, BJI72_RS15525, BJI72_RS15530, BJI72_RS15535, BJI72_RS15540, BJI72_RS15545, BJI72_RS15550, rrf, [BJI72_RS15560]*	

DISCUSSION

Having previously shown a plausible natural history of the GPAs in Chapter 3(Waglechner et al., 2019), major unresolved questions remain about other members of the extended GPA-like BGC family. This family of BGCs encodes products with at least four different mechanisms of action: the well-known D-Ala-D-Ala binding of Nicolau class I, II, III and IV glycopeptides like vancomycin and teicoplanin (Leclercq et al., 1988; Stegmann et al., 2015), lipid II binding of enduracidin and ramoplanin (Fang et al., 2006; Lo et al., 2000; Sugimoto, Maeda, Itto, & Arimoto, 2017), the reported FabI binding of complestatin (Kwon et al., 2015) and feglymycin has been reported to inhibit binding of the HIV envelope protein gp120 to CD4, to inhibit both MurA and MurC in peptidoglycan biosynthesis (Ferir et al., 2012; Hanchen et al., 2013; Rausch et al., 2011). Clearly, the various sequence components of these BGCs represent a common pool that have been reconfigured in many ways among several different Actinobacterial genera to achieve these activities. There are trends observed in these BGCs and accordingly we have constructed a putative classification of the predicted scaffolds of these BGCs based on known structures and a careful phylogenetic analysis of their BGC components. Our scheme supports 4 major scaffold classes and 8 distinct subclasses, several of which encode products with no confirmed structures or activities.

Predictions made using this analysis have led to isolation of complestatin from *Streptomyces* sp. WAC01325, and the novel product corbomycin from *Streptomyces* sp. WAC01529, named after the location of sampled soil from which this strain was isolated (Crowsnest Pass, AB or *Pas Nid-de-Corbeau*). Initial tests of both compounds

demonstrated antibiotic activity against primarily Gram-positive organisms. Lacking obvious resistance genes in their respective BGCs, mutant strains selected for resistance to these compounds were raised and analyzed suggesting they both target the bacterial cell-wall. Subsequent characterization revealed that these compounds retained activity against vancomycin-resistant *Enterococcus* (A- and B-type), and vancomycin-intermediate *S. aureus*, suggesting a distinct mechanism of action from GPAs (Culp et al., 2020) (Appendix 2). Fluorescence microscopy, cell-wall precursor accumulation assays, peptidoglycan binding assays, and in vitro peptidoglycan clearance assays all suggest that complestatin and corbomycin bind peptidoglycan in a manner that prevents autolysins from carrying out the normal digestion and recycling of bacterial cell-wall during growth and cell division (Culp et al., 2020). For complestatin, cell-wall targeting is more consistent in comparison to other compounds in the GPA family, and is more plausible than contradictory reports requiring the compound to penetrate cells in order to inhibit FabI-mediated fatty-acid biosynthesis (Kwon et al., 2015).

The significance of observing multiple distinct modes of action of compounds in the extended GPA family directly relate to the observed diversity in their corresponding BGCs. We had previously used marker gene sequences to identify the divergent GPA pekiskomycin in *Streptomyces* sp. WAC04229 and WAC01420, but because these marker genes are only conserved among the Nicolau type I-IV GPAs our view was simultaneously confined to compounds which all happen to share a major scaffold type and mechanism of action (Thaker et al., 2013). The phylogenetic analysis provided here and previously demonstrates rationalizes a strong link between the phylogenetic

emergence of a major scaffold type and a mechanism of action (Waglechner et al., 2019) (Chapter 3).

Compared to the typical GPAs, we have observed a reduced number of tailoring enzymes. Among the class IVa complestatin scaffolds, methylation is achieved through an *N*-methyltransferase domain encoded in the NRPS rather than an enzyme supplied in trans, while chlorination is predicted by a divergent halogenase with homology to those found in typical GPAs. Glycosylation is known or predicted only in the class II scaffolds, similar to membrane-bound mannosylation observed in GPAs (Margherita Sosio et al., 2003; Wu et al., 2015). Acylation is predicted to be achieved by the presence of starter condensation domains found in class II, IVa, and IVb scaffolds. None of these features are conserved in BGCs for all scaffold types. These observations clearly delineate the sources of chemical diversity between major scaffold classes. It is tempting to speculate that tightly conserving a scaffold contributes to selection for greater diversity of tailoring enzymes as in the GPAs. In the non-GPA scaffold classes, expansion, deletion, and repetition of NRPS modules contributes to a greater amount of diversity than tailoring reactions with more of these steps like acylation and N-methylation being performed on the NRPS itself than by enzyme supplied in trans. The sequences of many of the NRPS domains in Class I, II and IV scaffolds are more similar to others within the same scaffold than they are to sequences in other scaffolds, which has the unfortunate side-effect of making these BGCs difficult to reconstruct using short DNA sequencing technology.

The relationships between these major and minor scaffold types are difficult to disentangle. The reconciliation approach taken in Chapter 3 work less well because it is harder to define singular MRCAs for the various scaffold classes (Waglechner et al., 2019). An attempt to diagram these relationships, while not strictly phylogenetic, is supplied in figure 4-20.

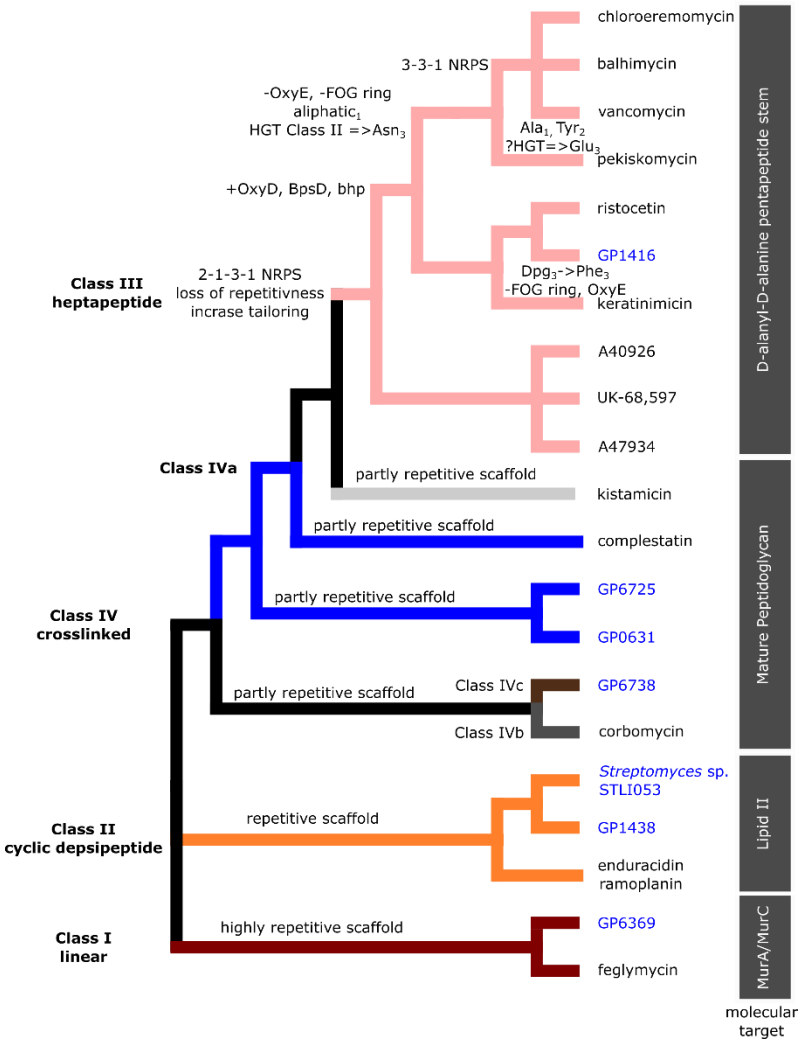


Figure 4-20: GPA scaffold class relationships. This cladogram is not phylogenetic, though these relationships are derived from the relationships between the scaffold sequences from the other trees. Three major classes emerge (I - dark red, II - orange and IV (IVa – blue, IVb – dark grey and IVc – brown) all with observed bioactivity targeting various aspects of bacterial cell-wall biology. Compounds with blue labels are only hypothetical or predicted based on BGC. Kistamicin shares characteristics with Class IV and Class III (pink) and is depicted as an ancestor of Class III, though the actual relationship cannot be known (ancestral to or hybridization/recombination with Class III). Major transitions occurring in Class IV include the development of a heptapeptide structure, loss of repetitive NRPS modules leading to defined MRCAs for the each NRPS module in Class III. Major developments in Class III include the expansion of tailoring enzymes, the reconfiguration of the NRPS from 2-1-3-1 to 3-3-1 organization, the loss of aromatic amino acids at positions 1 and 3 and loss of the fourth ring system along with the OxyE cytochrome P450 which appear to be coupled.

Two clear paths of investigation emerge from this work. On the practical side, it is not obvious if the extended GPA BGC family is characteristic of BGC evolution in general. How many other BGC families can be expected to mirror the evolution of the GPAs, and can this be used to identify other larger families of compounds that evolved from a different set of gene families? Can the tailoring enzyme diversity found in GPAs be ported into the GPA-like compounds, particularly the class IV scaffolds, using the tools of synthetic biology (Yim et al., 2016), or are the enzymes too adapted to the class III scaffold to efficiently modify class IV scaffolds? An obvious place to start is glycosylation of the hydroxyl group on the centrally crosslinked Hpg which is conserved in nearly all of the crosslinked scaffolds. Another avenue would be to attempt incorporation of β-OH Tyr, either by modification of A_{Tyr} into an A_{Bht} domain using the GPA domains as a template and supplying the exogenous amino acid or supplying the β-hydroxylase enzyme from a *Nonomuraea* BGC in trans.

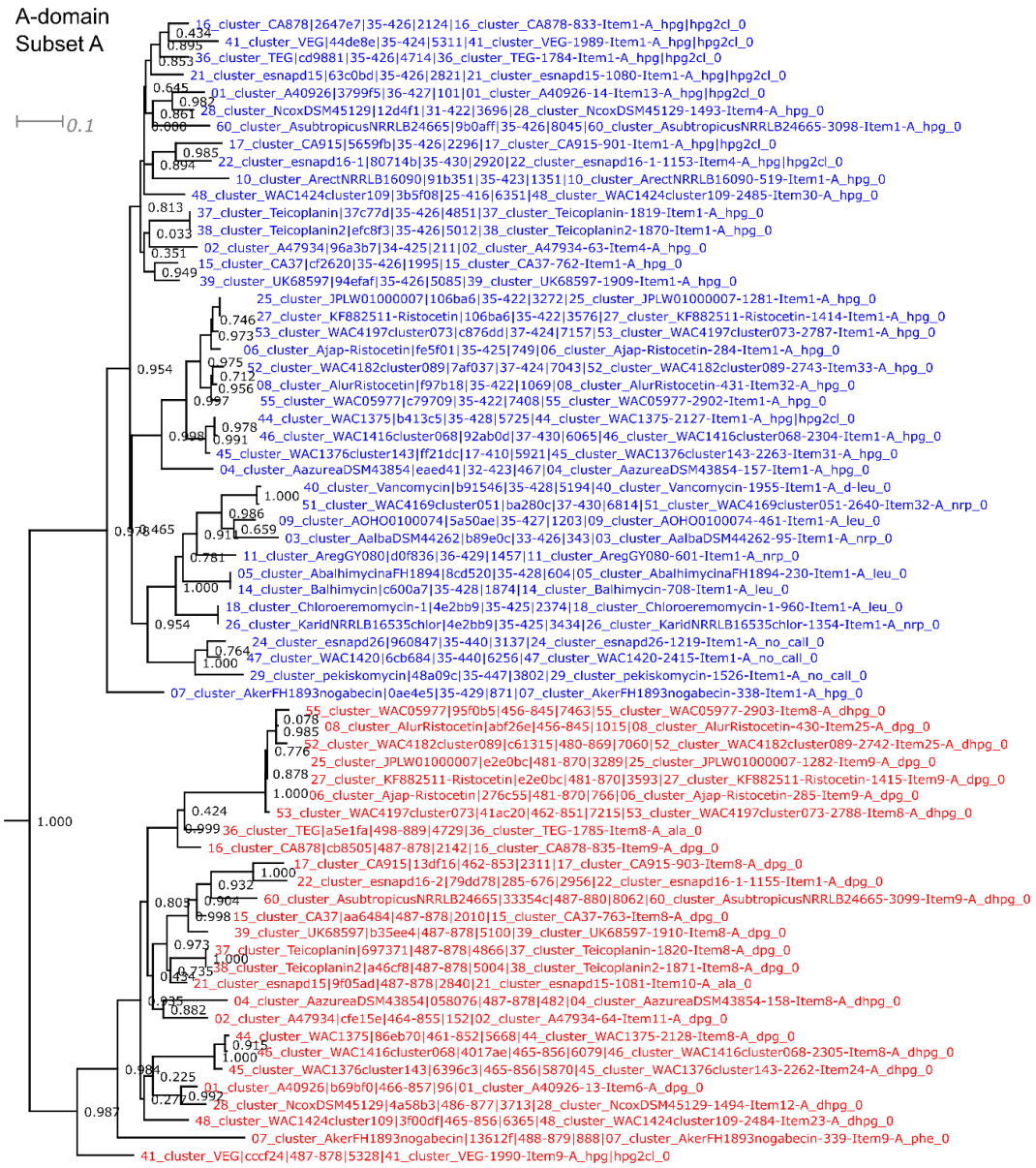
On a theoretical level it is not obvious that a fully dated reconciliation is possible or interpretable for every BGC component in the entire family. If it is, it becomes possible to measure the rates of divergence of these BGCs if there is a corresponding rate at which new activities emerge. The larger picture of BGC evolution, particularly antibiotic BGC evolution in the context of intra-bacterial competition, gives the impression that potential chemical diversity outweighs the number of potential targets. Good targets of broad-spectrum antibiotics are those that are widely conserved and therefore likely to be ancient in bacteria. All of the presumptive targets of GPAs fit this description. During the course of extended GPA scaffold evolution, we cannot tell if the mechanism of activity switches

smoothly from scaffold to scaffold, or jumps rapidly when a new scaffold configuration emerges. The only data that can ever be expected to test these ideas are known or inferred structures produced by BGCs at the tips of species trees representing a snapshot in time and not the ancestral reconstruction of precursor BGCs which might be impossible to produce. This idea has the potential to be tightly coupled with the concept of intrinsic resistance (Cox & Wright, 2013) which as a general concept begins as an inherited difference between two organisms (physiological, behavioural, but ultimately genetic) that a hypothetical natural product could exploit. These pre-existing differences are the stuff that drive the selection of antibiotic biosynthesis, and only become recognized an intrinsic resistance once those molecules exist.

The expansion of this BGC family is an important first step towards more fully exploring the complex relationship between the evolution of antibiotic production and resistance and has the practical effect of resulting in the identification of a new mechanism of antibiotic activity along with several avenues for the discovery of novel compounds. The number and placement of each BGC family for unknown compounds with respect to observed mechanisms of action and known compounds might prove to be incorrect. As new BGCs are discovered, the still-growing family of GPA BGCs may yet be a continuing embarrassment of riches.

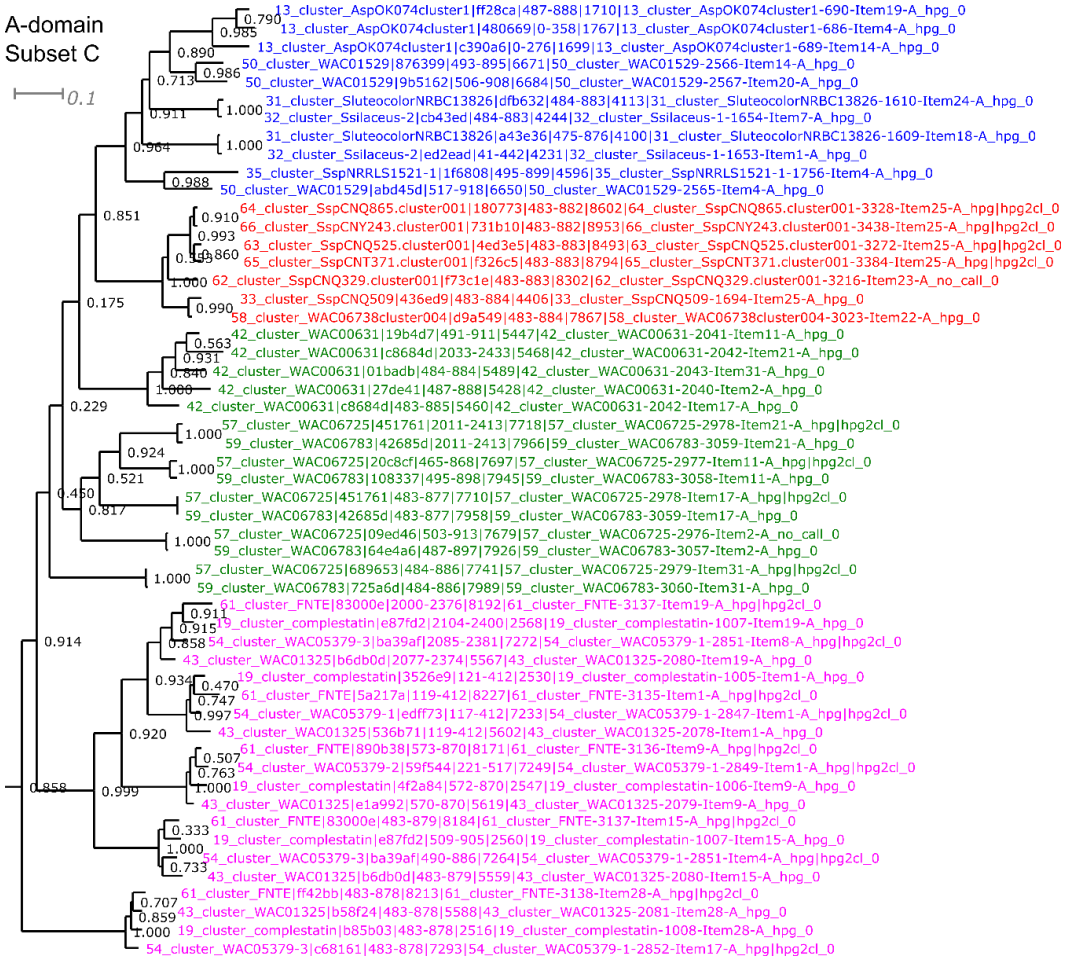
SUPPLEMENTARY MATERIAL

Supplementary Figures

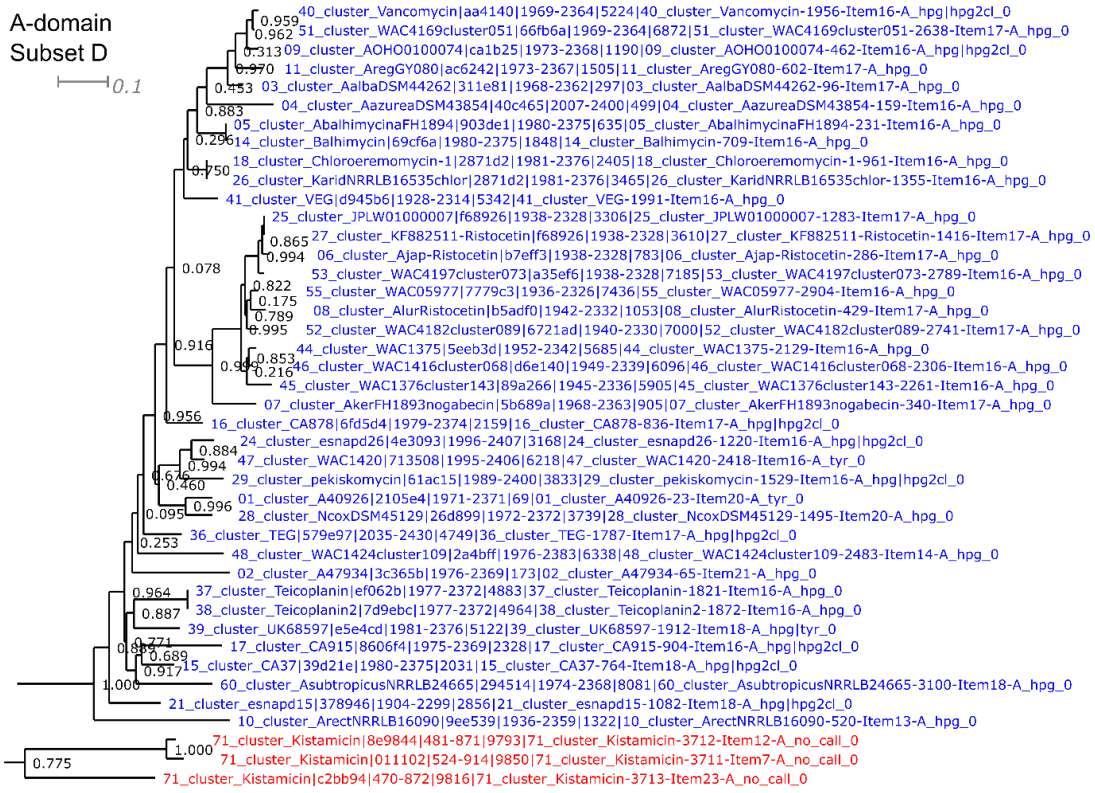


Supplementary Figure 4-1: A-domain phylogeny subset A. This subset of leaves is derived from Figure 4-5. Blue – class III GPAs, dalbaheptides, module 1. Red – class III GPAs, dalbaheptides, module 3.

Supplementary Figure 4-2: A-domain phylogeny subset B. This subset of leaves is derived from Figure 4-5. Blue – class III GPAs, dalbaheptides, module 4. Red – kistamicin module 5.

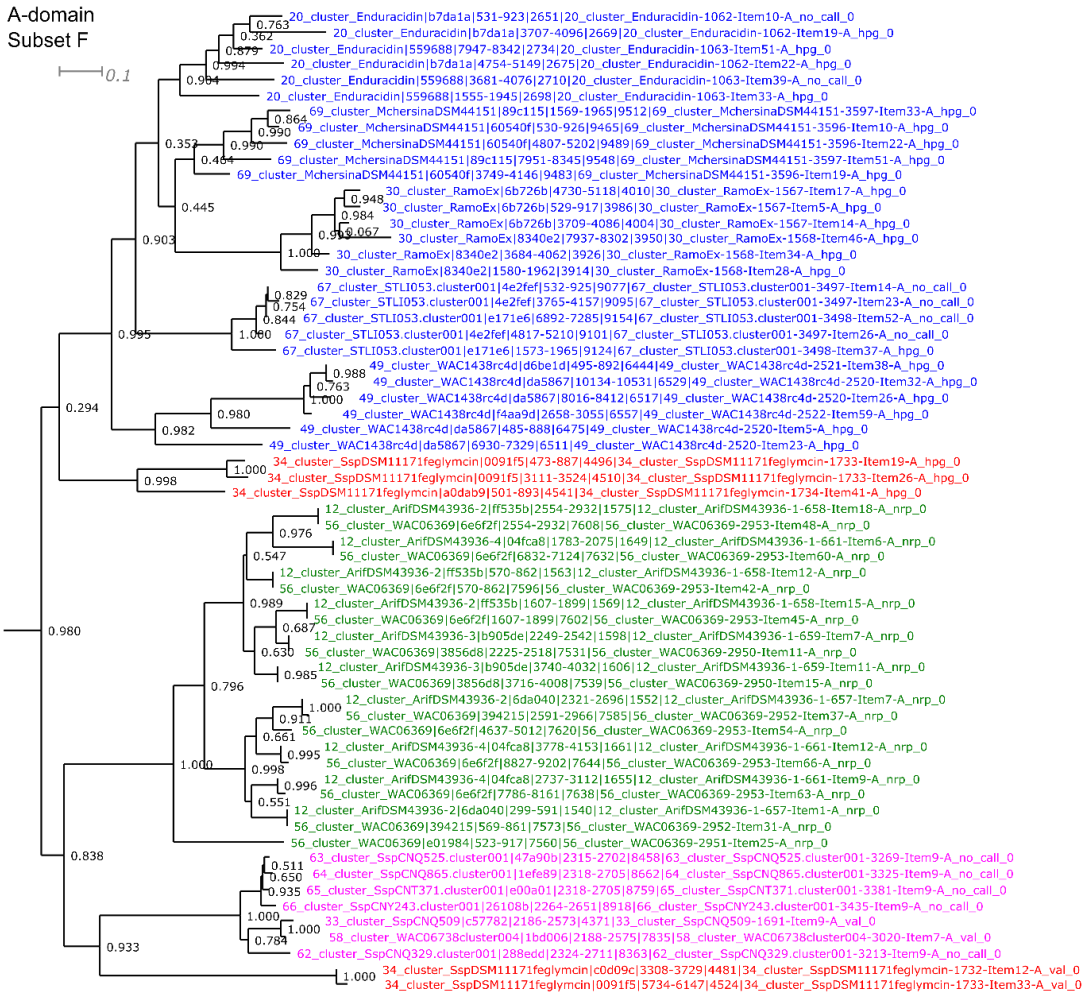


Supplementary Figure 4-3: A-domain phylogeny subset C. This subset of leaves is derived from Figure 4-5. Blue – class IVb GPAs. Red – class IVc GPAs. Green – class IVa complestatin-variant GPAs (see text). Note these sequences are not monophyletic. Magenta – class IVa complestatin.



Supplementary Figure 4-4: A-domain phylogeny subset D. This subset of leaves is derived from Figure 4-5. Blue – class III GPAs, dalbaheptides, module 5. Red – kistamicin, modules 3, 4, and 7. These three kistamicin modules are an outgroup to all of the leaves in Figures 4-1, 4-2, and 4-3.

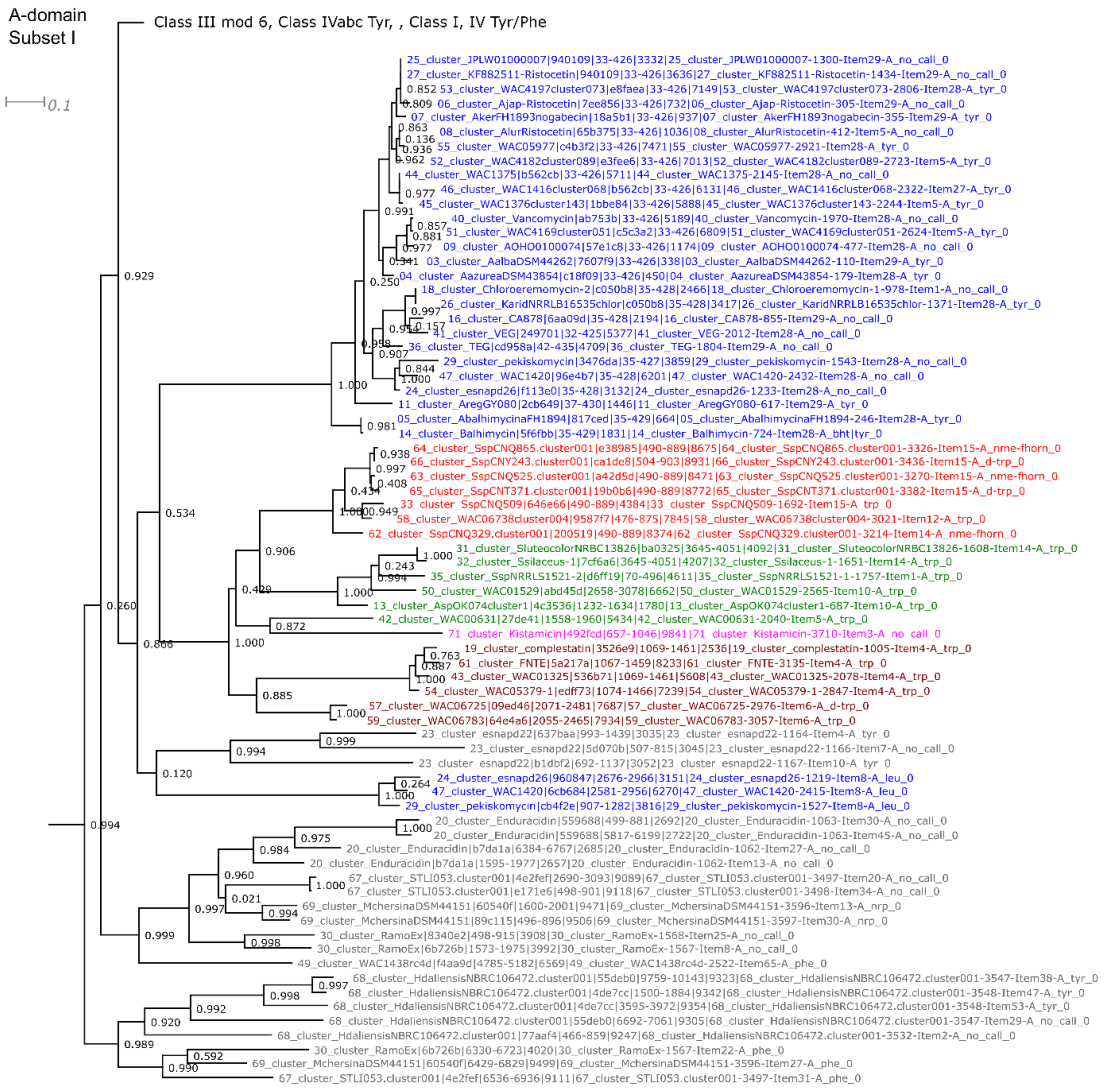
Supplementary Figure 4-5: A-domain phylogeny subset E. The subset of leaves was derived from Figure 4-5. Blue – class III GPAs, dalbaheptides, module 7.



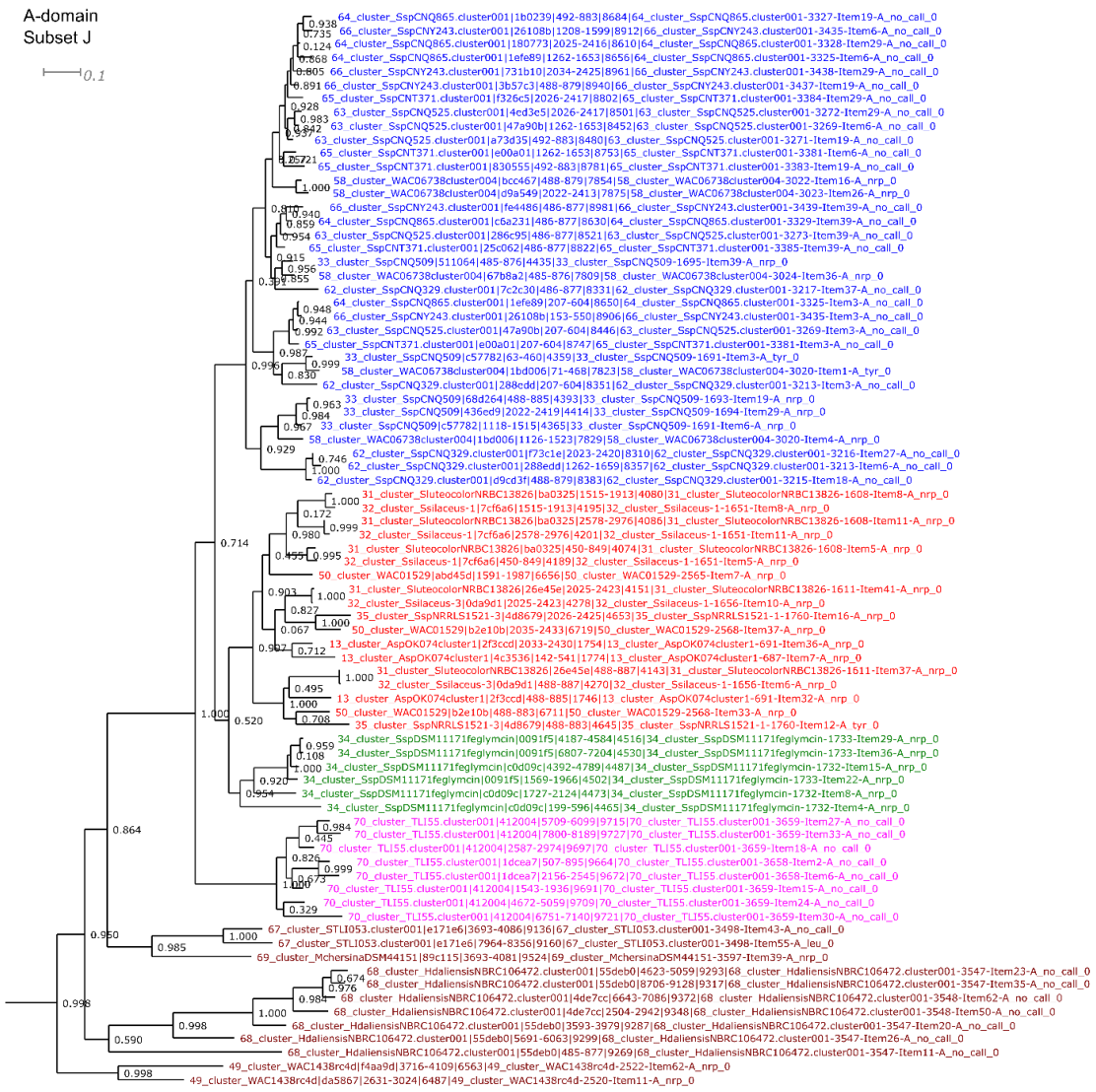
Supplementary Figure 4-6: A-domain phylogeny subset F. This subset of leaves was derived from Figure 4-5. Blue – class II scaffolds, repeated Hpg modules. Red – class I, feglymycin, repeated Hpg modules and repeated Val modules. Green – class I, scaffolds from *Actinomadura* sp., proposed to be Dpg. Magenta – class IVc GPAs, predicted to be Val.

Supplementary Figure 4-7: A-domain phylogeny subset G. This subset of leaves is an inset derived from Figure 4-5. Blue – class III GPAs, dalbaheptides, module 2. Red – kistamicin, module 2.

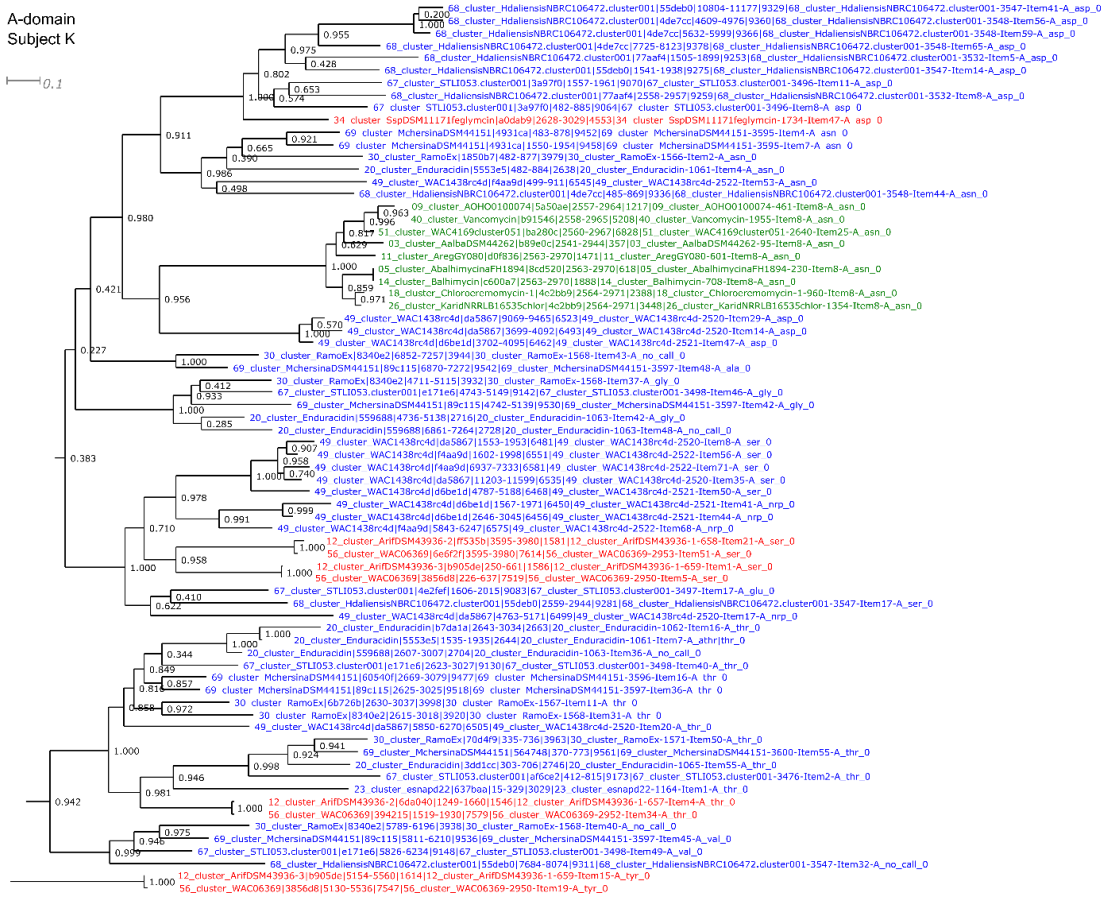
Supplementary Figure 4-8: A-domain phylogeny subset H. This subset of leaves is derived from Figure 4-5. Blue – class III GPAs, dalbaheptides, module 6. Red – class IVa complestatin repeated Tyr modules. Note the class complestatin-variant scaffolds, WAC00631, WAC06275 and WAC06783 do not form a monophyletic clade with the complestatin A_{Tyr} sequnces. Green – class IVc GPAs. Magenta – class IVb GPAs. Dark red – class I scaffolds A_{Phe} from feglymycin, proposed in GP6369. Note the class I scaffolds are also not monophyletic.



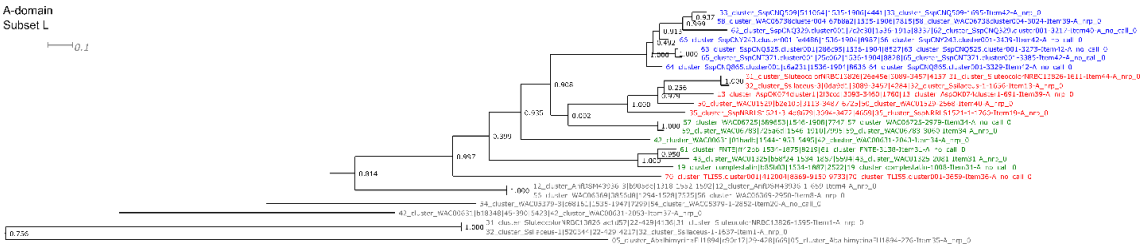
Supplementary Figure 4-9: A-domain phylogeny subset I. This set of leaves is a subset derived from Figure 4-5. Blue – class III GPAs, dalbahetides, A_{Tyr} domains from external NRPS BpsD that synthesizes β-OH Tyr, and a small clade of A_{Leu} found in the pekiskomycin BGCs. Red – class IVc GPAs A_{Trp}/A_{Orn}. Green – class IVb GPAs A_{Trp}. Magenta – kistamicin A_{Trp}. Dark red – class IVa A_{Trp}. Grey – class I depsipeptides, polyphyletic clades including A_{Orn} and A_{End} in ramoplanin and enduracidin (see main text).



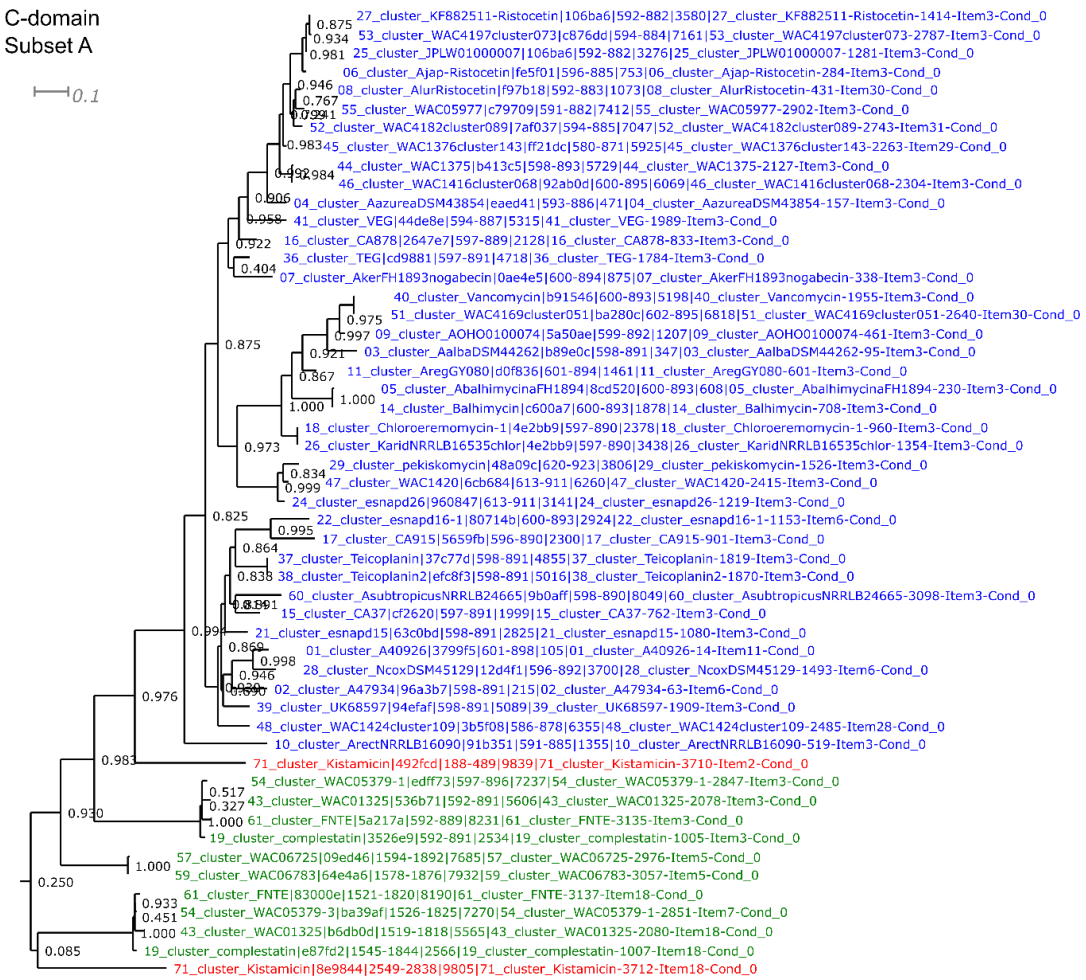
Supplementary Figure 4-10: A-domain phylogeny subset J. Blue – class IVb GPAs, repeated proposed A_{Dpg}. Red – class IVc GPAs repeated proposed A_{Dpg}. Green – class I feglymcin repeated A_{Dpg}. Magenta – class IVb GPAs repeated unclassified A-domain proposed to be A_{Dpg}. Dark red – class II depsipeptides, polyphyletic clades of proposed A_{Dpg}, repetitive in some BGCs.



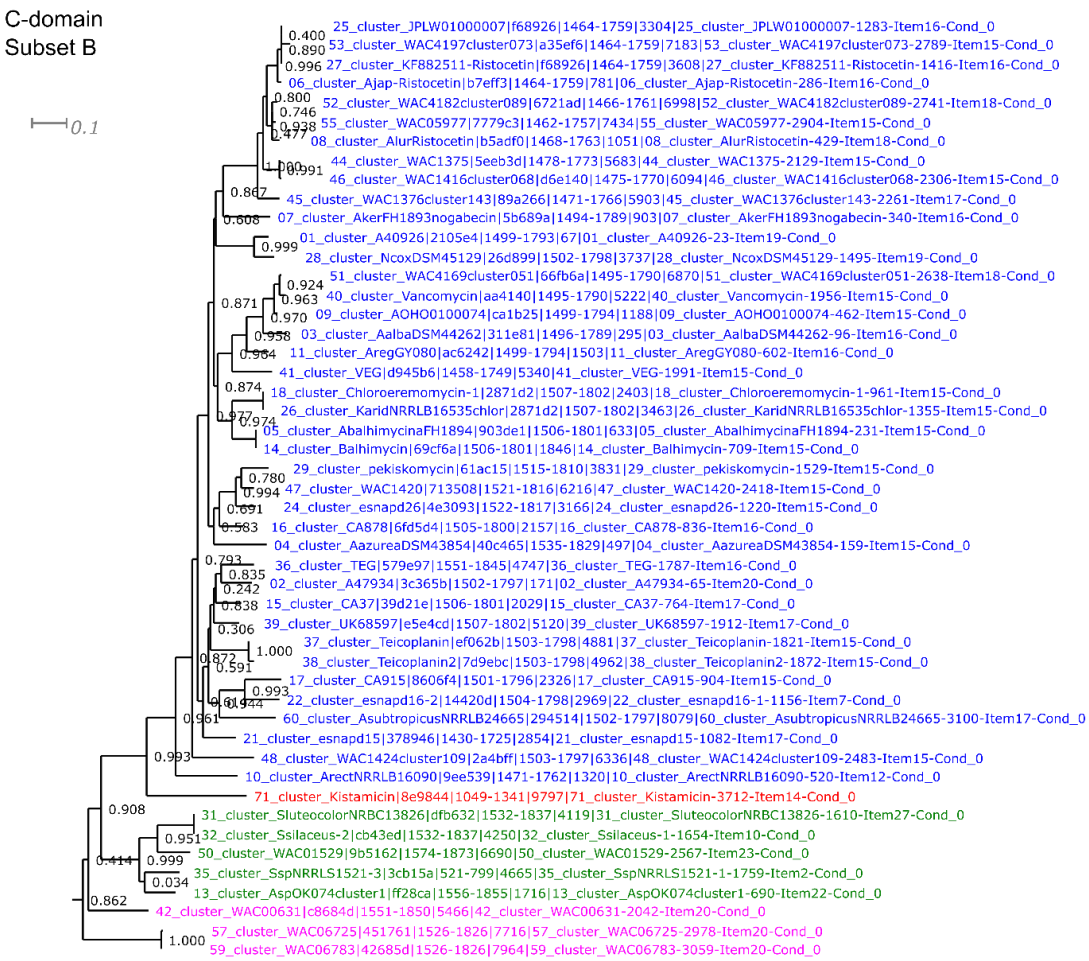
Supplementary Figure 4-11: A-domain phylogeny subset K. This subset of leaves was derived from Figure 4-5. Blue – class II depsipeptide scaffolds, polyphyletic clades of sequences known or predicted to activate a variety of amino acids including Asp, Asn, Gly, Ser and Thr. Green – class III GPAs, dalbaheptides module 3, proposed to be gained by HGT from class II scaffolds. Red – class I scaffold domains, polyphyletic. Note feglymcin A_{Asp} being most similar to sequences from class II scaffolds.



Supplementary Figure 4-12: A-domain phylogeny subset L. This subset of leaves is derived from Figure 4-5. Grey – A-domains associated with GPA BGC loci but not actually part of GPA BGCs. The remaining sequences are labeled as putatively inactive A-domains in class IV scaffold BGCs. Blue – class IVc BGCs. Red – class IVb scaffolds. The sequence from BGC 70 is not monophyletic with the other class IVc sequences. Green – class IVa complestatin scaffolds. Note the class IVa sequences are not monophyletic, the sequences in the complestatin-variant BGCs appear to be most closely related to the class IVc sequences.

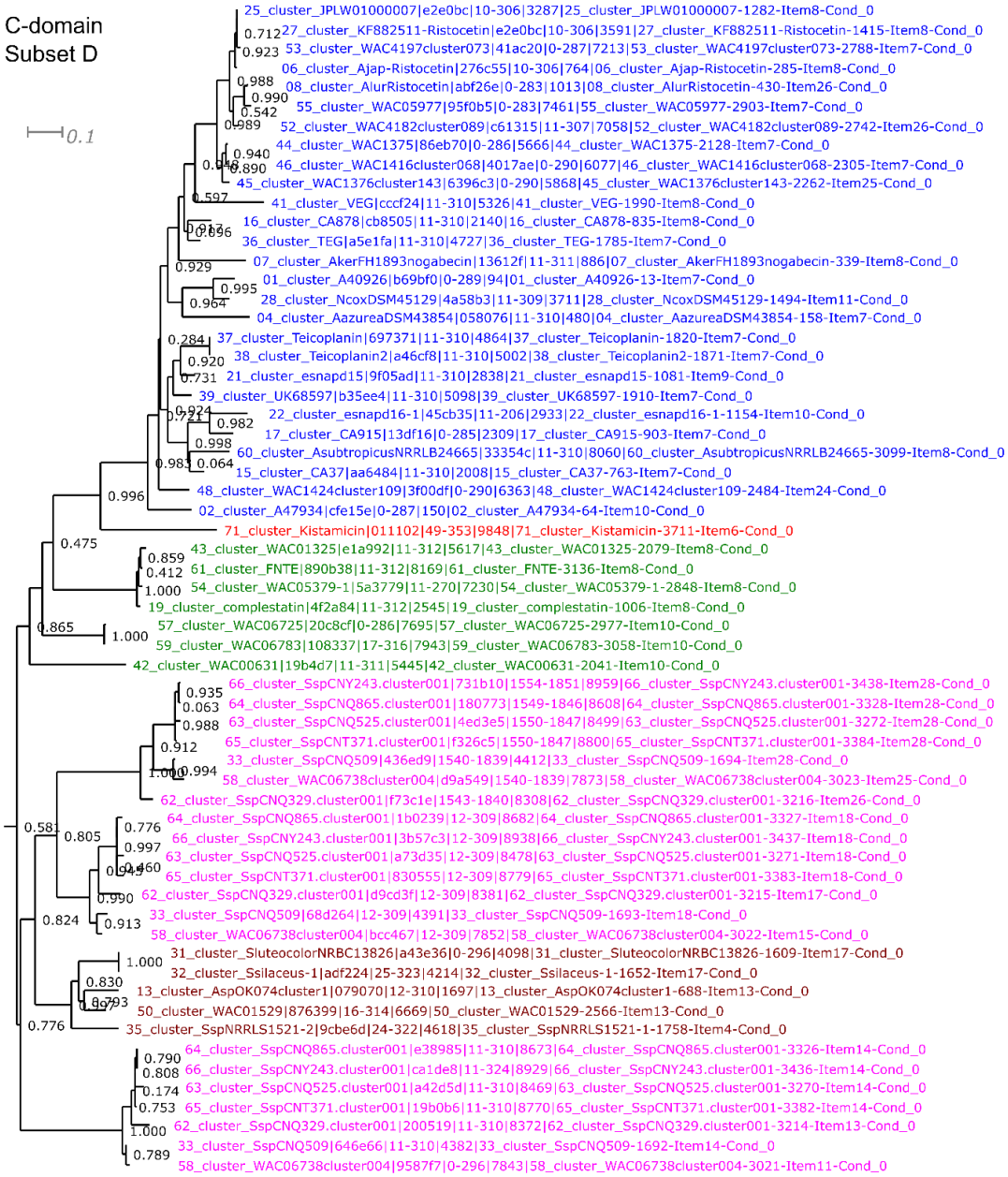


Supplementary Figure 4-13: Condensation domain phylogeny subset A. This subset of leaves is derived from Figure 4-6. Blue – class III GPA BGCs, dalbaheptides $C_{(1,2)}$. Red - kistamicin $C_{(1,2)}$ and $C_{(5,6)}$. Green - class IVa complestatin and complestatin-variant $C_{(4,5)}$ domains. Notes these domains are not monophyletic.



Supplementary Figure 4-14: Condensation domain phylogeny subset B. This subset of leaves was derived from Figure 4-6. Blue – class III GPA BGCs, dalbaheptides, $C_{(4,5)}$. Red – kistamicin $C_{(4,5)}$. Green – class IVb GPAs $C_{Hpg(0,1)}$. Coordinates from central Hpg. See main text. Magenta – class IVa complestatin-variant GPAs. Note these sequences are not monophyletic.

Supplementary Figure 4-15: Condensation domain phylogeny subset C. This subset of leaves was derived from Figure 4-6. Blue – class III GPA BGCs, dalbaheptides C(5,6). Red – class IVa complestatin and complestatin-variant GPAs C(1,2) and C(5,6). Note the sequences mostly closely related to the class III sequences are not monophyletic. Green – class IVc Hpg(1,2). Magenta – class IVb Hpg(1,2). Coordinates are relative to central Hpg. See main text.

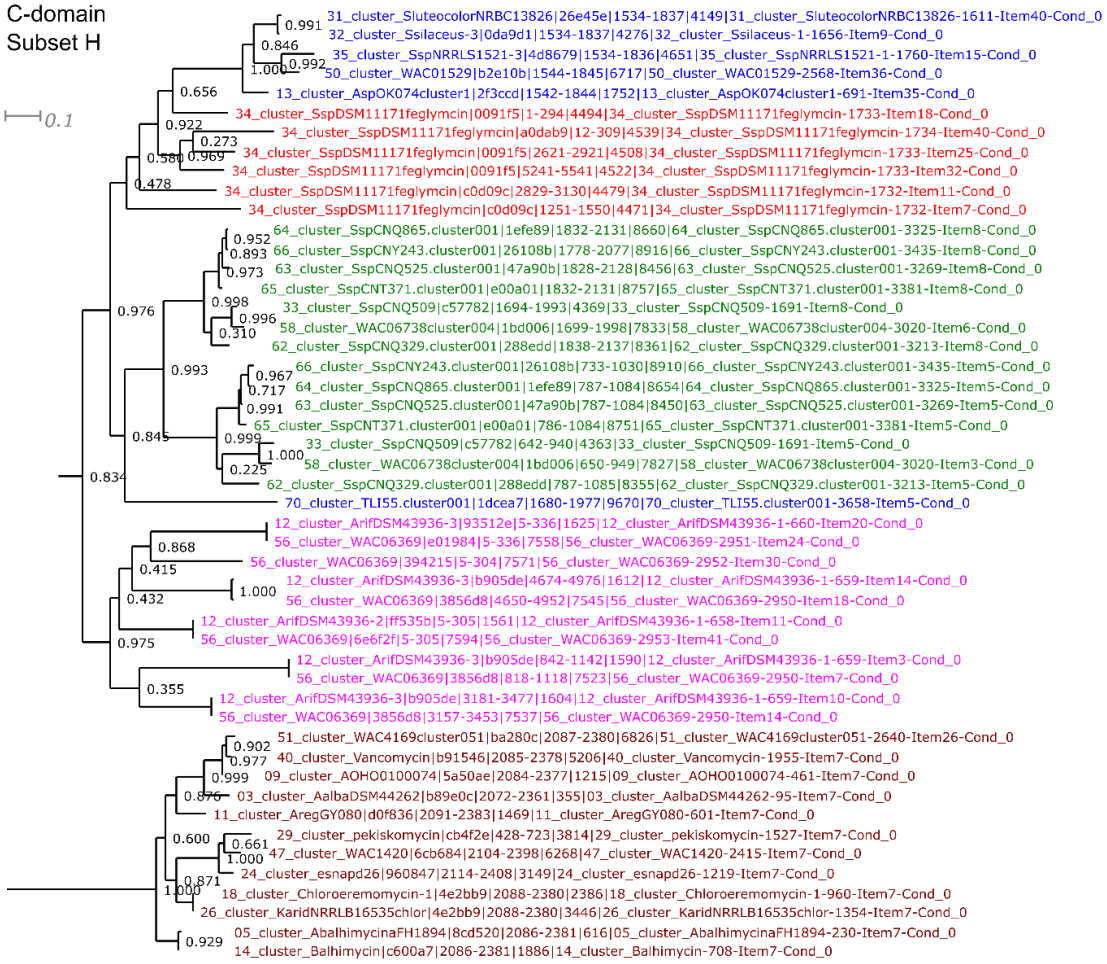


Supplementary Figure 4-16: Condensation domain phylogeny subset D. These leaves were derived from Figure 4-6. Blue – class III GPA BGCs, dalbaheptides $C_{(2,3)}$. Red – kistamicin $C_{(2,3)}$. Green – class IVa complestatin and complestatin-variant BGCs $C_{(4,5)}$. Magenta – class IVc GPA BGCs $C_{Hpg(0,1)}$, $C_{Hpg(-2,-1)}$, $C_{Hpg(-3,-2)}$. These coordinates are relative to the central Hpg. See text. Dark red – class IVb GPAs $C_{Hpg(1,2)}$.

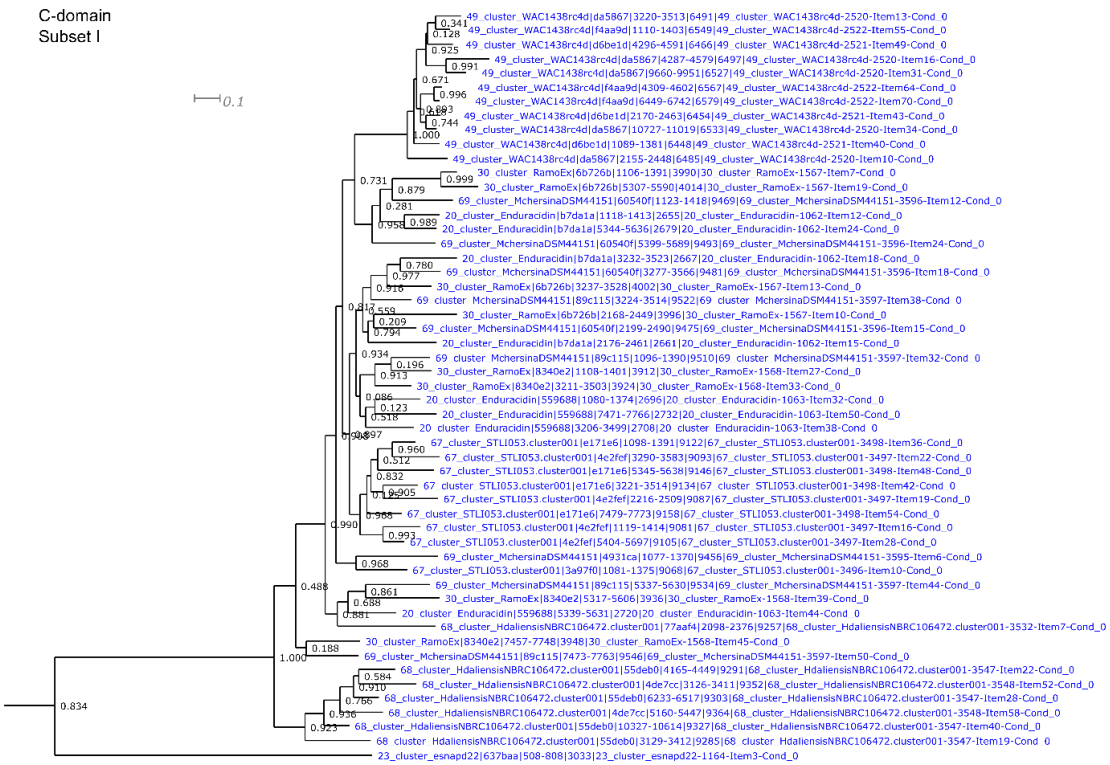
Supplementary Figure 4-17: Condensation domain phylogeny subset E. These leaves were derived from Figure 4-6. Blue – class III GPA BGCs, dalbaheptides, $C_{(6,7)}$. Red – kistamicin $C_{(2,3)}$. Green – class IVa complestatin and complestatin-variant $C_{(2,3)}$. Magenta – class IVc GPA $C_{Hpg(2,3)}$. Coordinates are relative to central Hpg. See main text. Dark red – class IVb GPA $C_{Hpg(2,3)}$. Coordinates are relative to central Hpg.

Supplementary Figure 4-18: Condensation domain phylogeny subset F. These leaves are derived from Figure 4-6. Blue – class IVc GPAs $C_{Hpg(-1,0)}$. Coordinates relative to central Hpg. See main text. Red – class IVb $C_{Hpg(-1,0)}$. Coordinates relative to central Hpg. Sequence from BGC 70 are unrelated, likely a sequence or alignment artifact. Green – class IVa complestatin and complestatin-variant $C_{(3,4)}$.

Supplementary Figure 4-19: Condensation domain phylogeny subset G. These leaves were derived from Figure 4-6. Blue – class III GPA BGCs, dalbaheptides $C_{(3,4)}$. Red – kistamicin $C_{(3,4)}$.

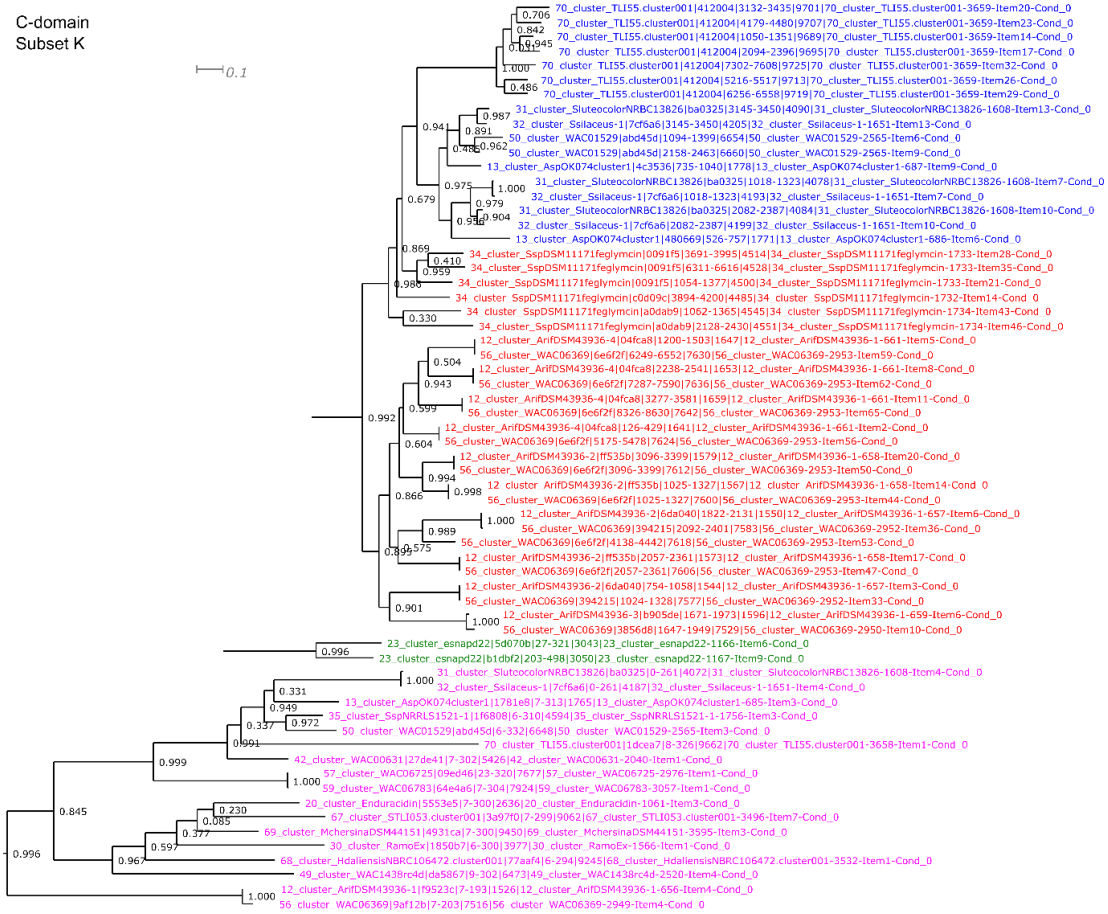


Supplementary Figure 4-20: Condensation domain phylogeny subset H. These leaves were derived from Figure 4-6. Blue – class IVb GPAs $C_{Hpg(3,4)}$. Sequence from BGC 70 appears to be distantly related to the main group of sequences. Red – class I feglymycin group 1 condensation domains. See text. Green – class IVc GPAs $C_{Hpg(-4,-3)}$, $C_{Hpg(-5,-4)}$. Magenta – class I *Actinomadura* sp. sequences group 1. See text. Dark red – class III GPA BGCs, dalbaheptides from *Amycolatopsis* $C_{(2,3)}$. These sequences are from BGCs which have altered NRPS module organization from 2-1-3-1 to 3-3-1.



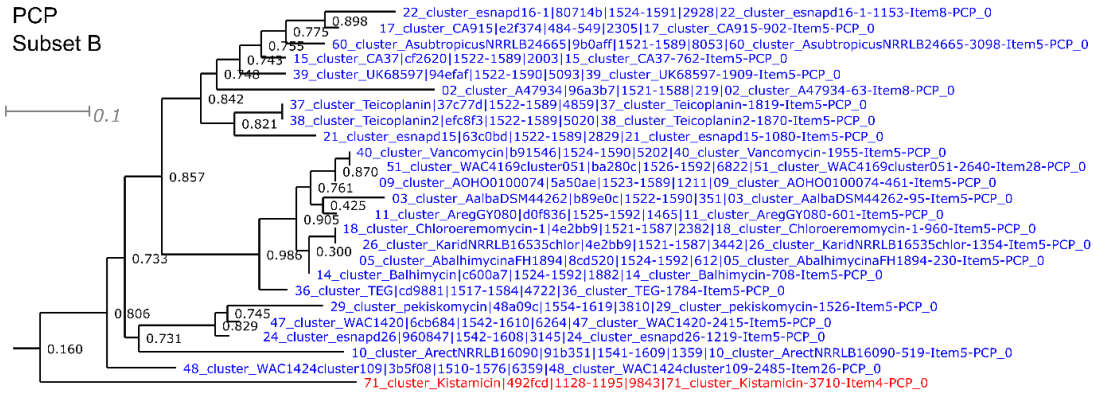
Supplementary Figure 4-21: Condensation domain phylogeny subset I. These domains were derived from Figure 4-6. Blue – class II depsipeptides, various condensation domains. Some of these sequences are repeated in some scaffolds, such as BGC 67.

Supplementary Figure 4-22: Condensation domain phylogeny subset J. These domains were derived from Figure 4-6. Blue – class II depsipeptides group 2, various condensation domains. Some of these condensation domains are repeated, such as in BGC 68, 49, and 67.

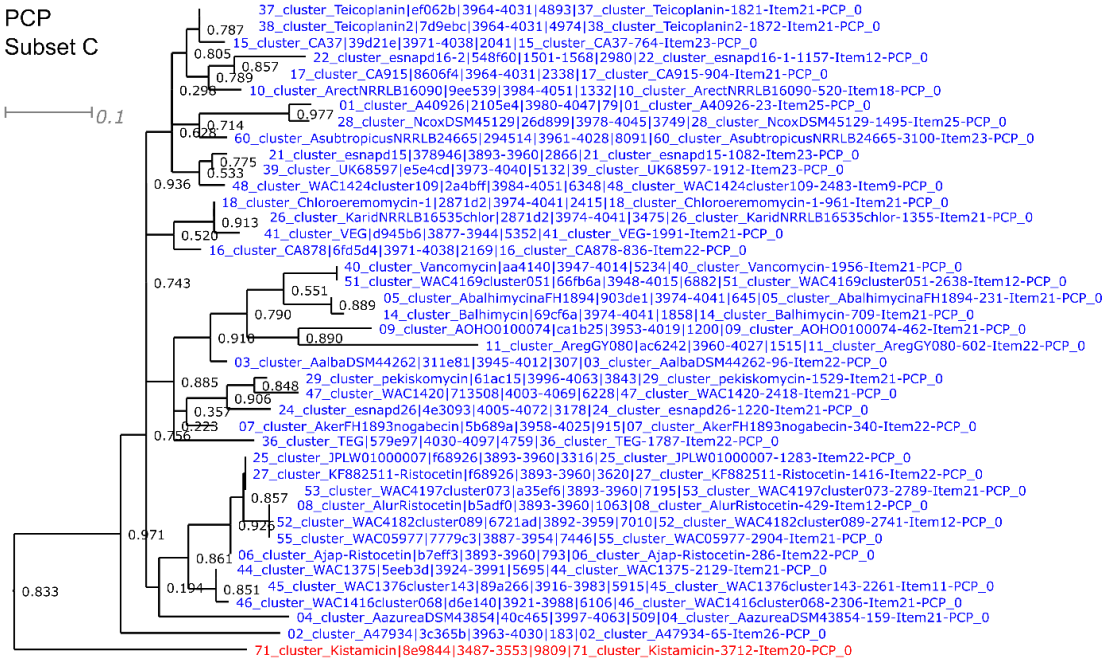


Supplementary Figure 4-23: Condensation domain phylogeny subset K. These leaves were derived from Figure 4-6. Blue – class IVb GPAs, a repeated family of condensation domains in GBC 70, $C_{Hpg(-4,-3)}$, and $C_{Hpg(-3,-2)}$. Coordinates relative to centrally linked Hpg. See text. Red – class I, repeated condensation domains in feglymcin and *Actinomadura* sp. group 2. Green – repeated condensation domains from BGC 23. These sequences may be an artifact of sequencing or alignment from an incomplete BGC. Magenta – Condensation starter domains from class I, II, IVa, IVb BGCs. These domains are predicted to produce a modified N-terminal amine group by addition of an acyl-group.

Supplementary Figure 4-24: PCP domain phylogeny subset A. These leaves are derived from Figure 4-7. Blue – class III GPA BGC, dalbaheptides, PCP4. Red – kistamicin PCP3 and PCP4.

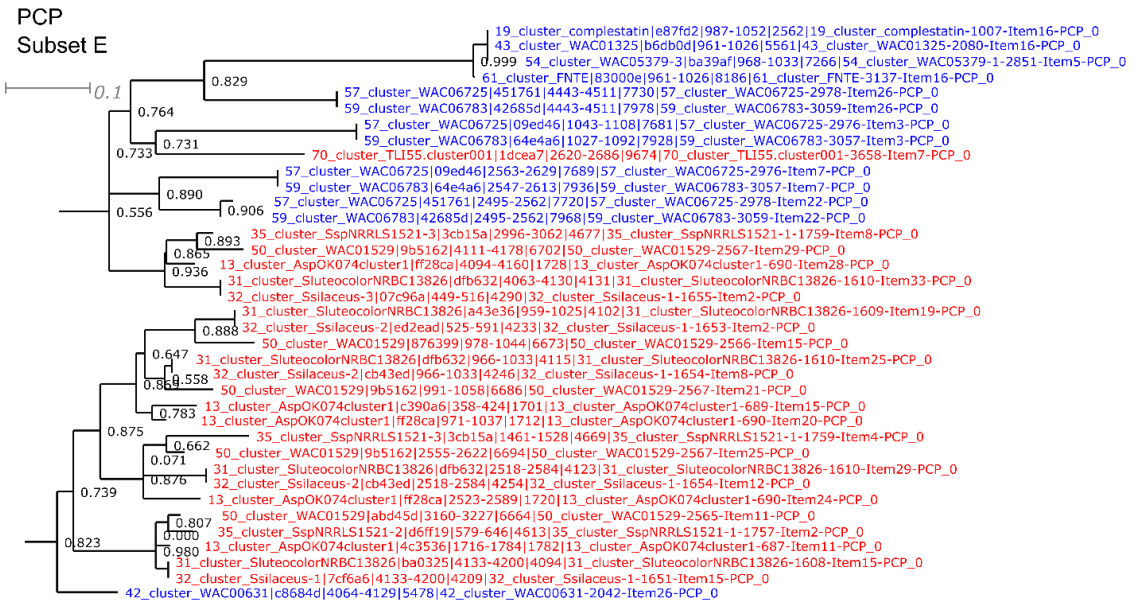


Supplementary Figure 4-25: PCP domain phylogeny subset B. These leaves are derived from Figure 4-7. Blue – class III GPA BGCs, dalbaheptides, PCP2. Red – kistamicin PCP2.

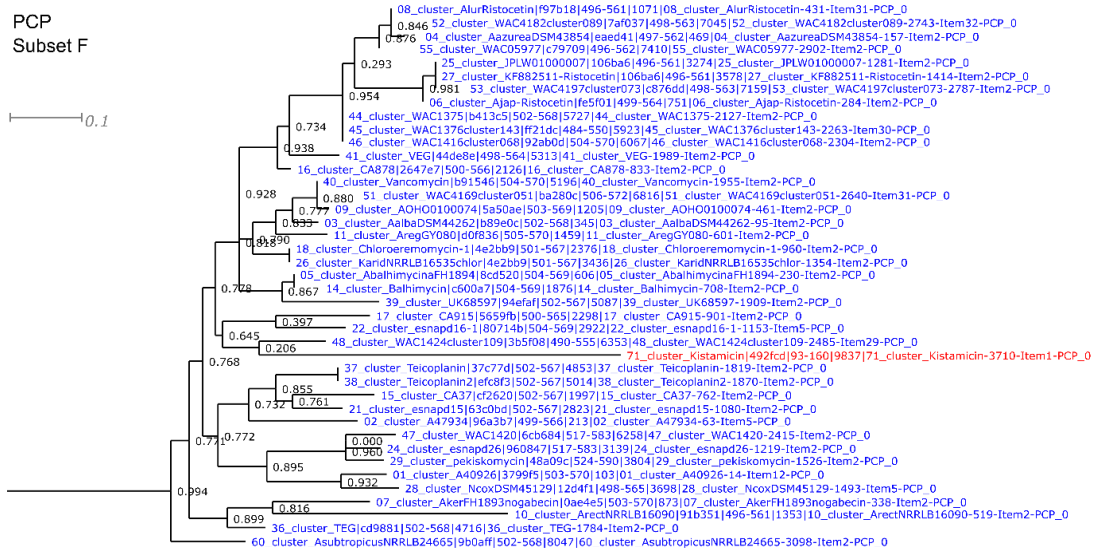


Supplementary Figure 4-26: PCP domain phylogeny subset C. These leaves are derived from Figure 4-7. Blue – class III GPA BGCs, dalbaheptides PCP5. Red – kistamicin PCP6.

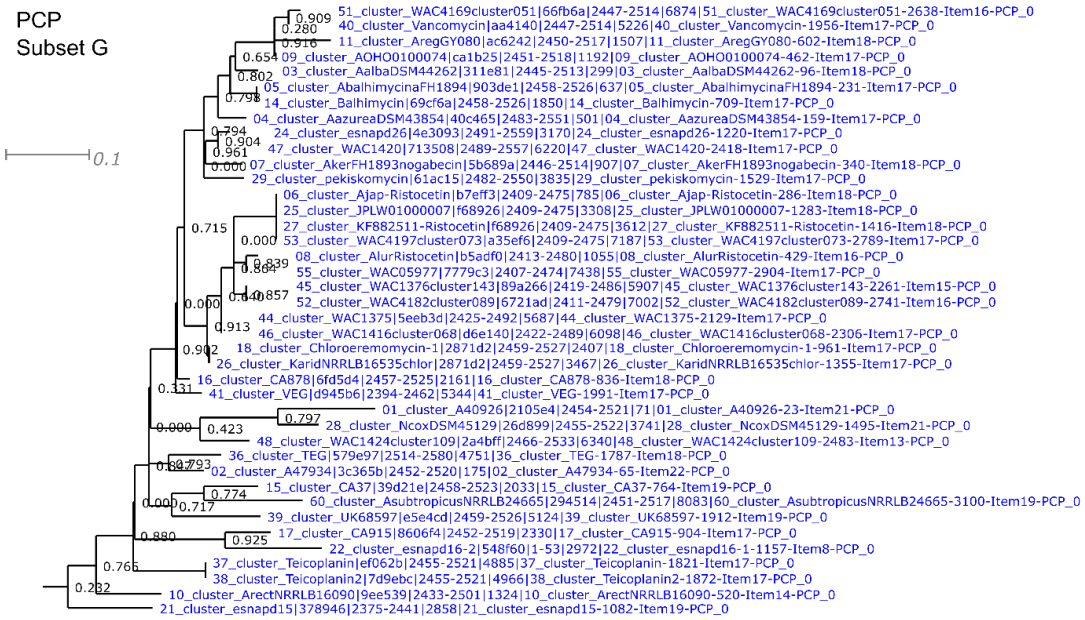
Supplementary Figure 4-27: PCP domain phylogeny subset D. These leaves are derived from Figure 4-7. Blue – class IVc, six families of PCP domains. Red – class IVa complestatin and complestatin-variant domains, two PCP families.



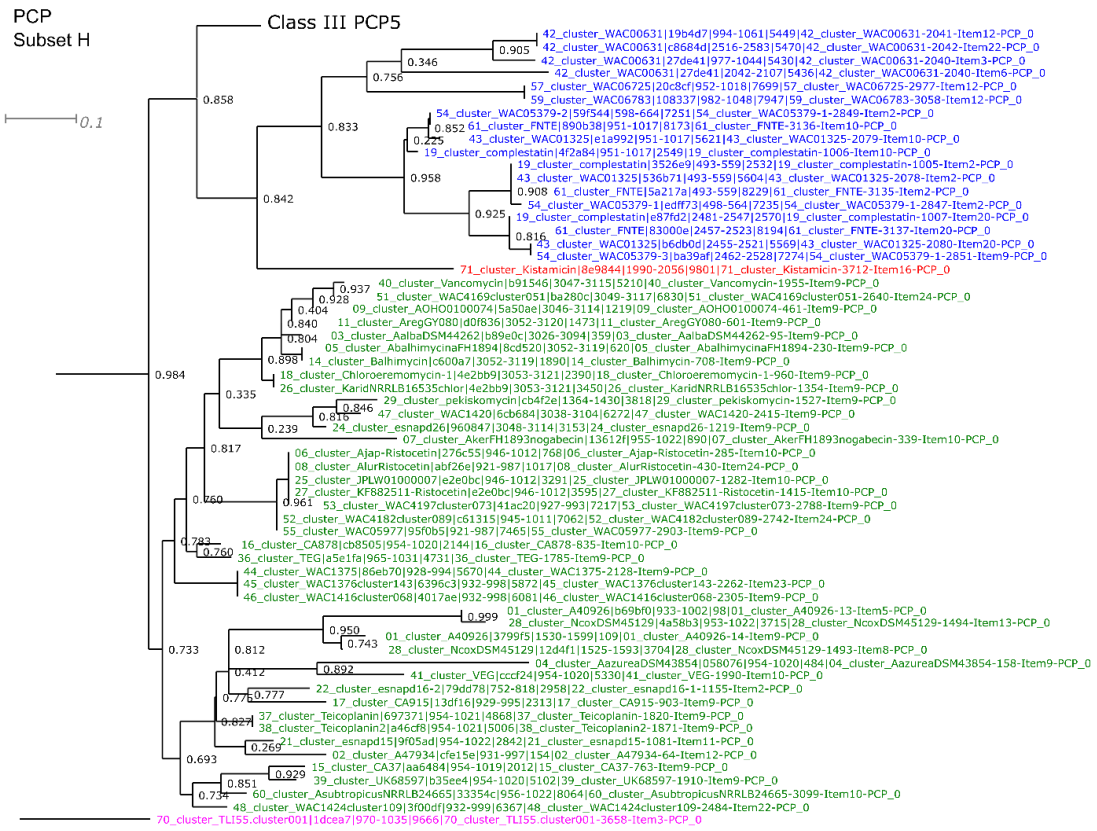
Supplementary Figure 4-28: PCP domain phylogeny subset E. These leaves were derived from Figure 4-7. Blue – class IVa complestatin and complestatin-variant. Note that complestatin-variant BGC 57 and 59 have four families of PCP vs the single family most closely related to the complestatin BGC 19, and a sequence from BGC 42 is not placed with the rest of these sequences. PCP sequences are short, and this may be an artifact of sequencing and/or alignment. Red – class IVb, four PCP domain families. A sequence from BGC 70 is separated from the other class IVb sequences.



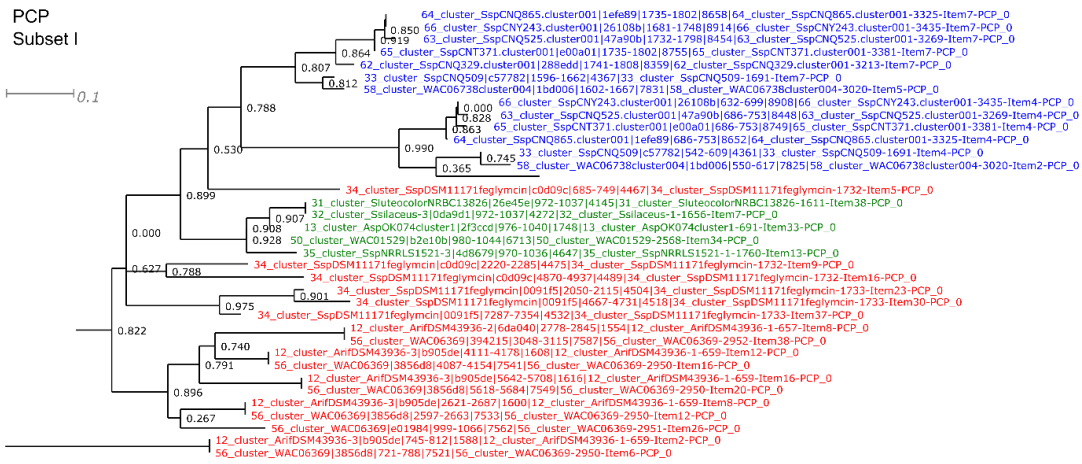
Supplementary Figure 4-29: PCP domain phylogeny subset F. These domains are derived from Figure 4-7. Blue – class III GPA BGCs, dalbaheptides PCP1. Red – kistamicin PCP 1.



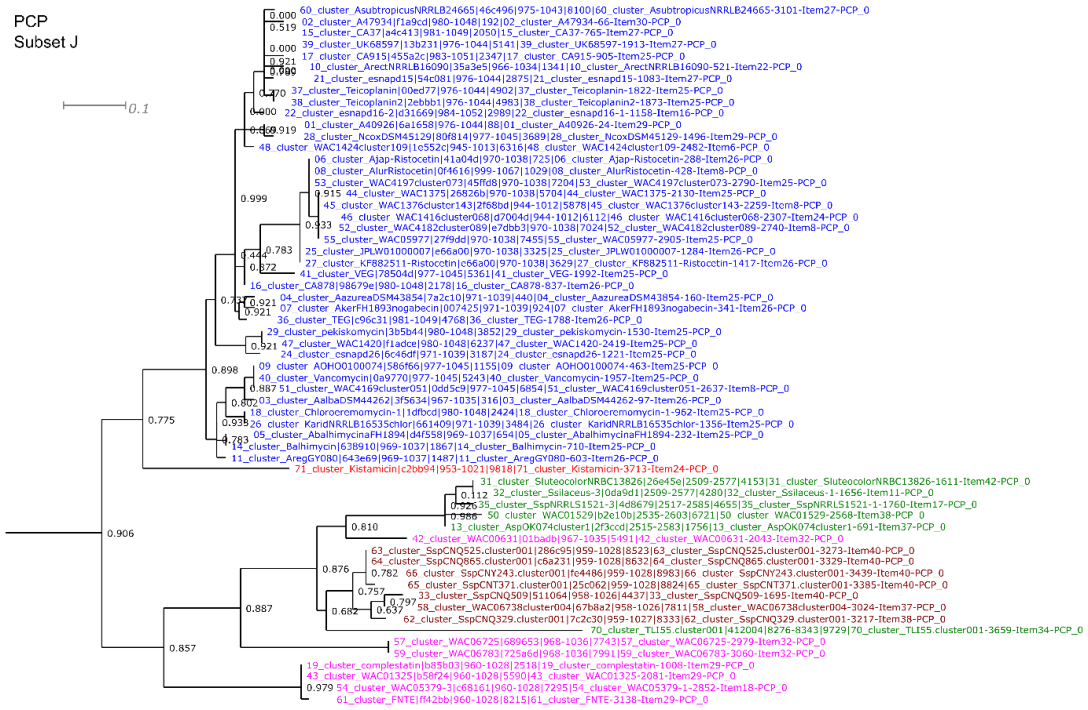
Supplementary Figure 4-30: PCP domain phylogeny subset G. This subset of leaves were derived from Figure 4-7. Blue – class III GPA BGCs, dalbaheptides PCP5.



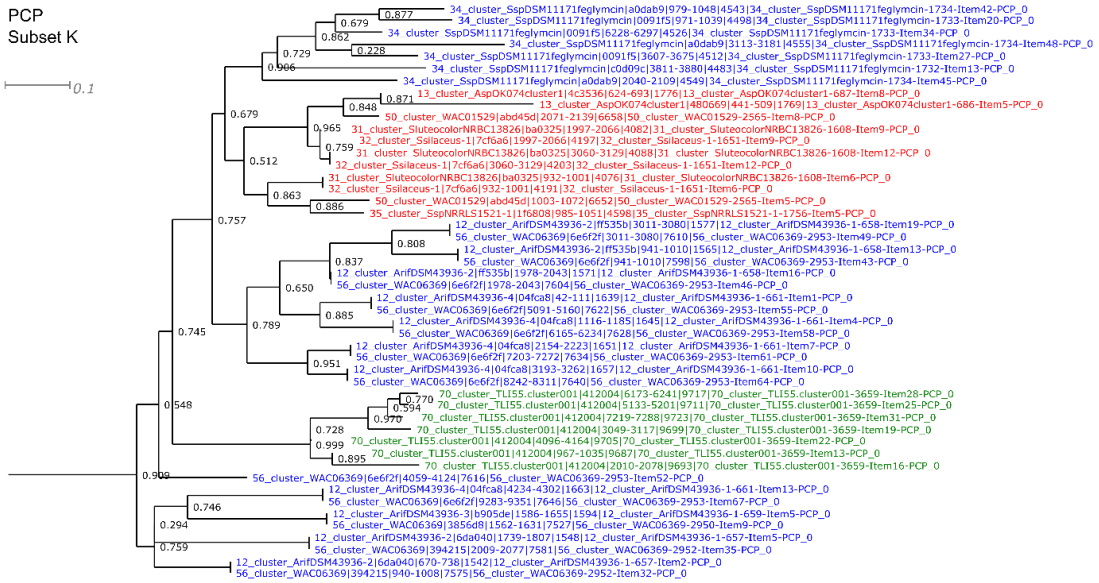
Supplementary Figure 4-31: PCP domain phylogeny subset H. This subset of leaves was derived from Figure 4-7. Blue – class IVa complestatin and complestatin-variant PCP domains. Red – kistamicin PCP5. Green – class III GPA BGCs, dalbaheptides PCP3. Magenta – class IVb BGC 70.



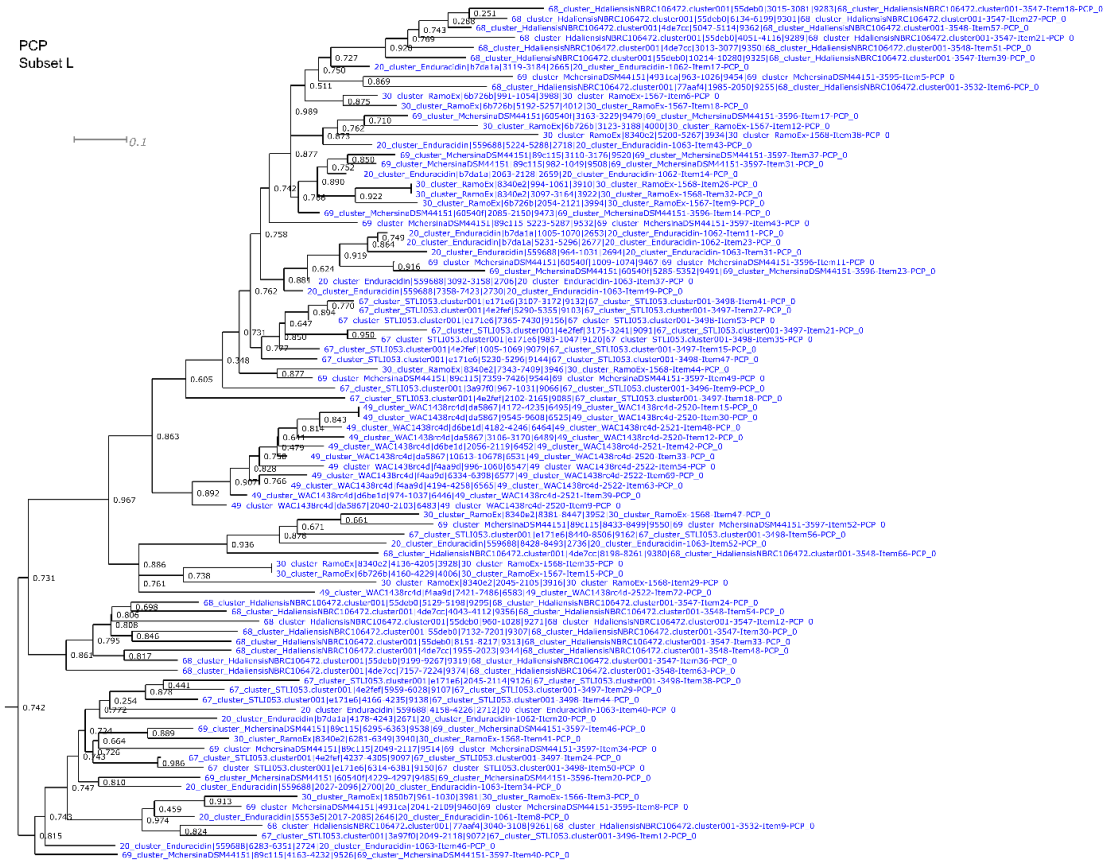
Supplementary Figure 4-32: PCP domain phylogeny subset I. This subset of leaves was derived from Figure 4-7. Blue – class IVc GPAs, two PCP families. Red – class I feglymycin and *Actinomadura* sp. repeated PCP domains. A single domain from BGC 34 is placed closer to the class IVc domains which may be a sequence or alignment artifact. Green – class IVb GPAs.



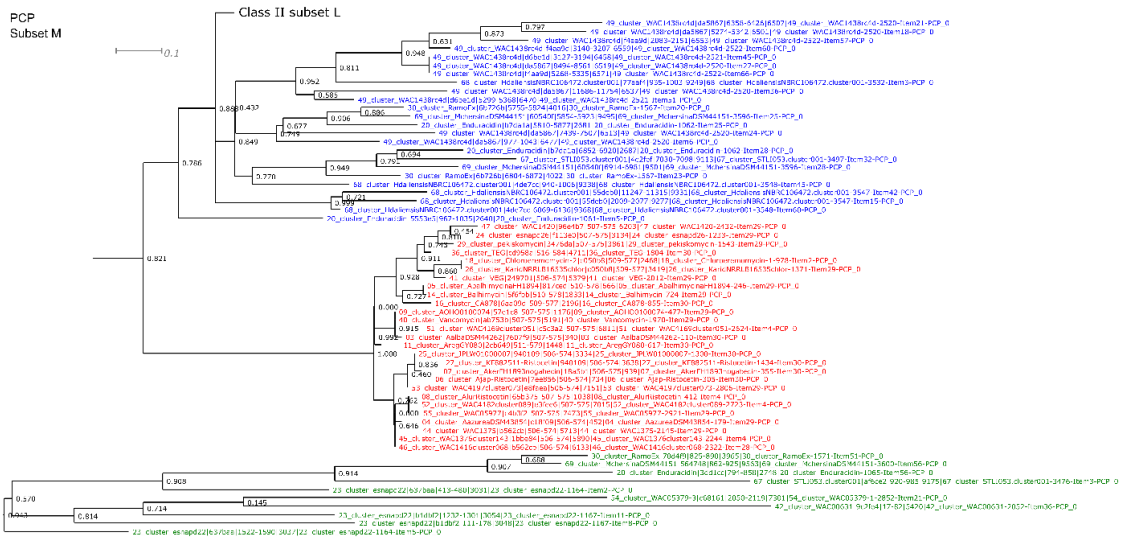
Supplementary Figure 4-33: PCP domain phylogeny subset J. This subset of sequences is derived from Figure 4-7. Blue – class III GPA BGCs, dalbaheptides PCP7. Red – kistamicin PCP7. Green – class IVb. Magenta – class IVa complestatin and complestatin-variant BGCs. Note a sequence from BGC 42 is grouped with the IVb sequences and this group of PCP sequences is not monophyletic.



Supplementary Figure 4-34: PCP domain phylogeny subset K. This subset of leaves was derived from Figure 4-7. Blue – class I feglymycin and *Acintomadura* sp. repeated PCP sequences. Red – class IVb GPA BGCs, two PCP sequences. Green – class IVb repetitive PCP sequence from BGC 70.

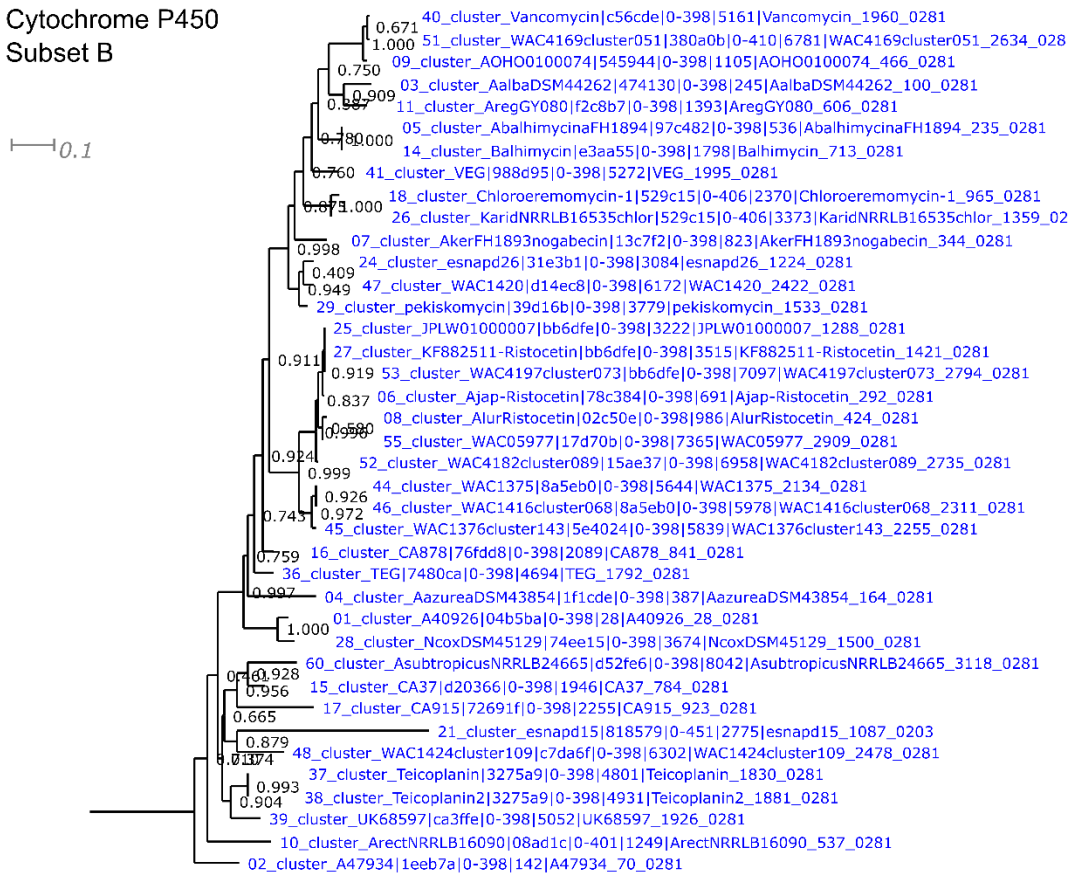


Supplementary Figure 4-35: PCP domain phylogeny subset L. This subset of leaves was derived from Figure 4-7. Blue – class II depsipeptides, multiple PCP sequences. BGCs 68, 30, 67, and 49 in particular contain several sequences with related copies in their respective BGCs.

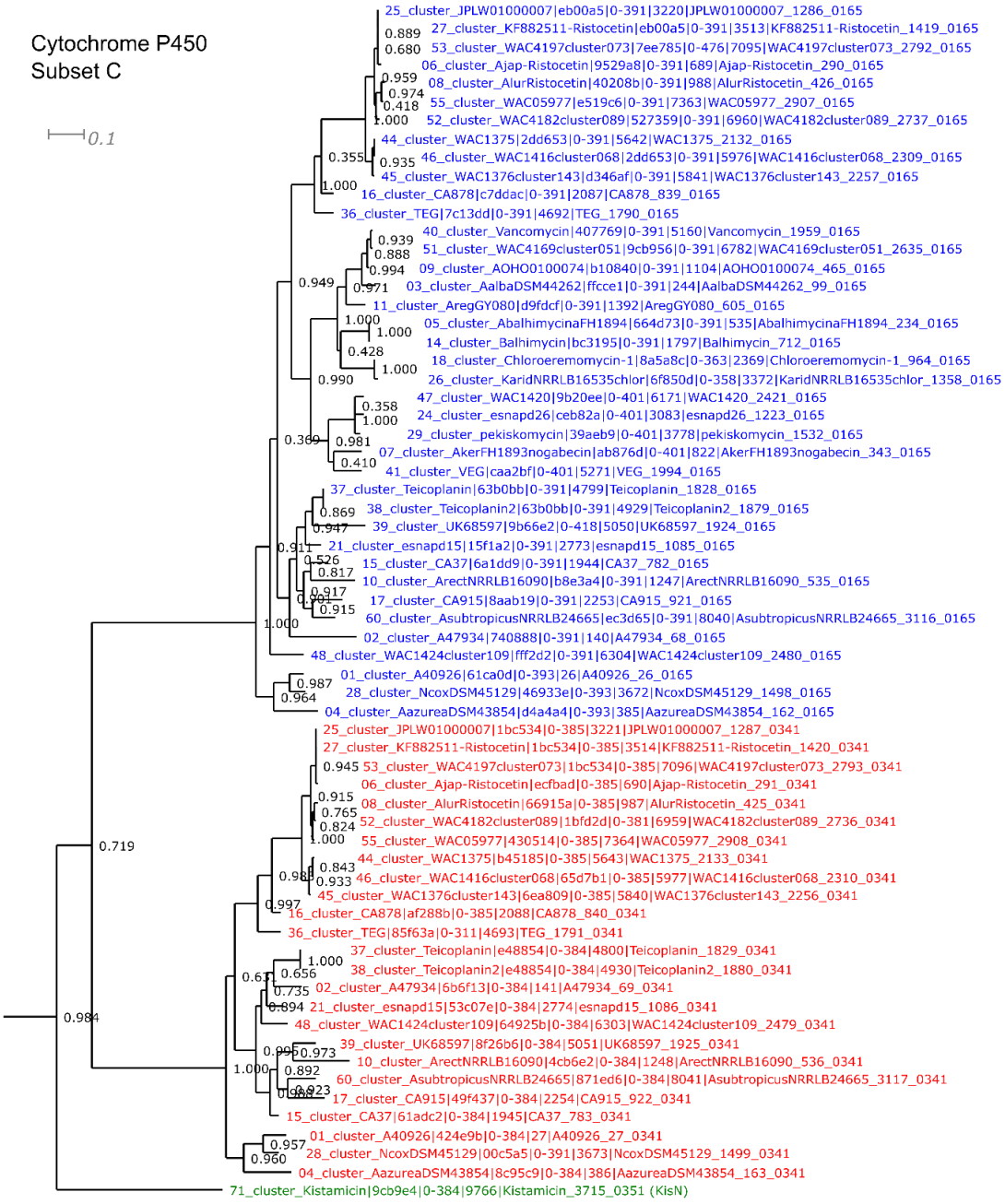


Supplementary Figure 4-36: PCP domain phylogeny subset M. This subset of sequences was derived from Figure 4-7. Blue – class II depsipeptide PCP sequences. Note the repeated sequences found in BGC 49. Red – PCP sequences found in the minimal NRPS BpsD involved in β-OH tyrosine biosynthesis in a subset of class III dalbaheptide BGCs. Green – outgroup PCP sequences found near GPA BGCs.

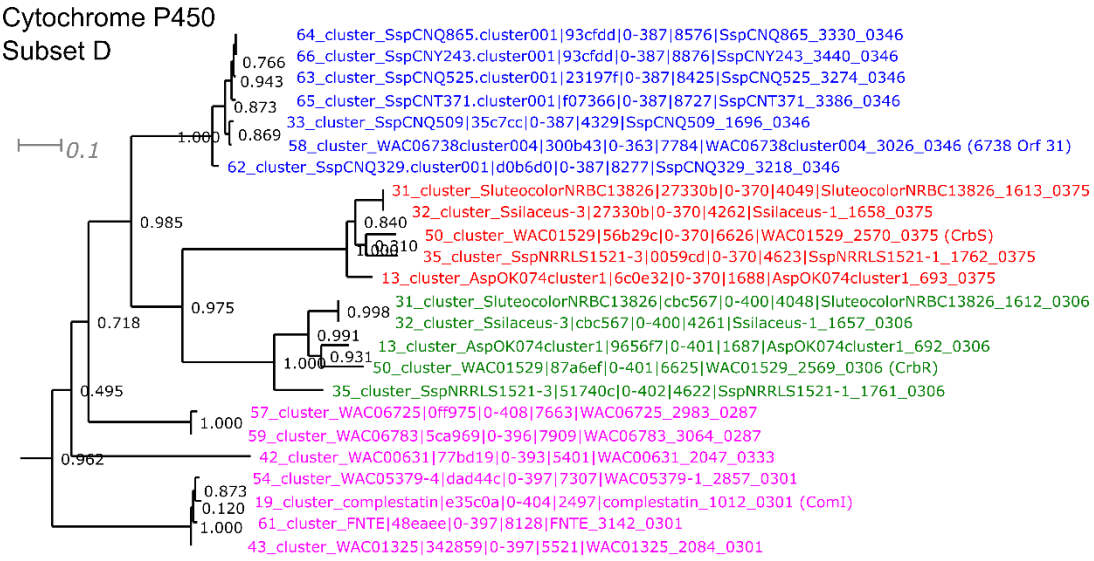
Supplementary Figure 4-37: Cytochrome P450 monooxygenase phylogeny subset A. This set of leaves was derived from Figure 4-12. Blue – class III GPA BGCs, dalbaheptides OxyC. Red – kistamicin KisO. Green – class IVa complestatin and complestatin-variant sequences. Sequences from BGC 42, 57 and 59 are not monophyletic with ComJ from BGC 19. Magenta – class IVc GPA BGC sequences including WAC06738 orf 32. Dark red – class IVb GPA BGCs including CrbT.



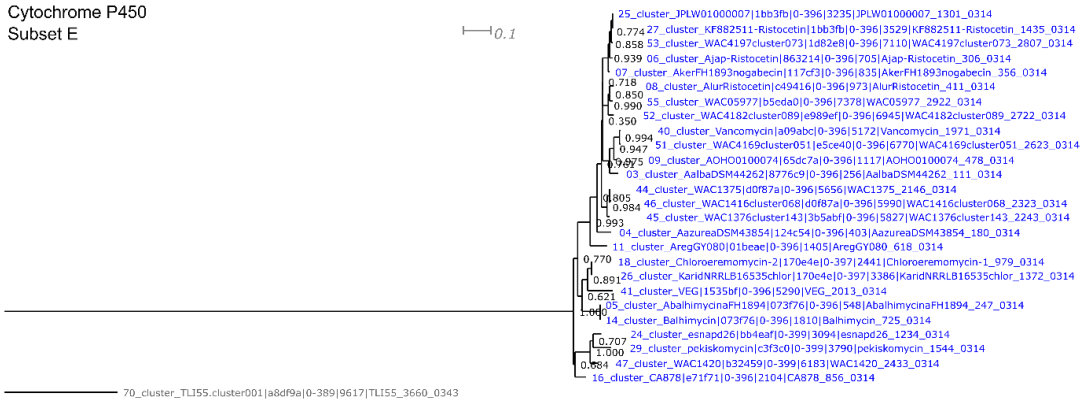
Supplementary Figure 4-38: Cytochrome P450 monooxygenase phylogeny subset B. This set of leaves was derived from Figure 4-12. Blue – class III GPA BGCs, dalbaheptides OxyB.



Supplementary Figure 4-39: Cytochrome P450 monooxygenase phylogeny subset C. This set of leaves was derived from Figure 4-12. Blue – class III GPA BGCs, dalbaheptides OxyA. Red – class III GPA BGCs, dalbaheptides OxyE. Green – kistamicin KisN.



Supplementary Figure 4-40: Cytochrome P450 monooxygenase phylogeny subset D. This set of leaves was derived from Figure 4-12. Blue – class IVc GPA BGCs including WAC06738 orf 31. Red – class IVb GPA BGCs including CrbS. Green – class IVc GPA BGCs including CrbR. Magenta – class IVa complestatin and complestatin-variant BGCs including ComI. Note these sequences are not monophyletic.



Supplementary Figure 4-41: Cytochrome P450 monooxygenase phylogeny subset E. This set of leaves was derived from Figure 4-12. Blue – class III GPA BGCs, dalbaheptides, OxyD associated with β-OH tyrosine biosynthesis. Grey – outgroup cytochrome P450 monooxygenase from BGC 70, it is the only P450 in the BGC and is completely different from all other sequences in class IVb.

CHAPTER FIVE: Discussion and future directions

Discussion

Serendipity has characterized the discovery of natural products, and antibiotics specifically. Early researchers benefitted from easy access to abundant 'low-hanging fruit', while attempting to understand why such bioactive molecules should exist at all. The hypothesized obvious consequence of the existence of antibiotics—resistance (Fleming, 1945)—became a reality shortly after the large-scale introduction of every known antibiotic. Bacteria have managed to exploit and solve the problem of differential susceptibility to a wide variety of chemicals long before humans were aware of the conflict. Serendipity is also a crutch as long as new frontiers are within easy reach. New culturing techniques to identify new organisms from new materials producing new compounds can only continue until there is nothing new. The wildest frontier for antimicrobial discovery is now in the sequence databases. It has been clear since the initial genome sequences were produced for the model organisms serving as the workhorses for natural product discovery that there is a mismatch between the biosynthetic potential observed in these sequences and what is observed in compounds extracted from these strains under laboratory conditions (Bentley et al., 2002). It is good news for the antibiotic resistance crisis that this potential exists in other bacteria as well (Doroghazi et al., 2014).

Following in the footsteps of early pioneers, we need to explain why the history of antibiotic discovery proceeded as it did. Screening will remain an essential tool because it is the only way to sift through thousands of compounds and identify activity, but it will

not be the most efficient way to identify novel compounds since the well-known problem of dereplication can now be rationalized by the genome sequences of producing organisms. So many antibiotics are rediscovered again and again because biosynthetic potential is distributed non-randomly in bacterial genomes. The frequency that strains are observed to produce various compounds provided by Baltz – for example, 10^{-1}–10^{-3} for actinomycin, tetracycline, streptothricin and streptomycin, 10^{-4} for chloramphenicol, 10^{-5} for glycopeptides and 10^{-6} for daptomycin – have been recapitulated in part by others (Baltz, 2006; Cox et al., 2017), but not yet via genome mining.

BGCs are more complex entities than individual genes or domains because natural products are the result of the combinatorial interactions of BGC components. In a short time, the number of bacterial genome sequences available will cross from 10^5 to 10^6 with a concomitant increase in the number of BGCs encoded in these genomes. Tools to manipulate and analyze these BGC sequences enable their exploitation for basic research, biotechnological and medical applications (Culp et al., 2019). Identifying BGCs similar to known BGCs is trivial with current methods like ClusterBlast (Medema et al., 2011; Medema et al., 2013). These techniques focus on similarities between BGCs based on shared conservation of their components, however measures that are more focused on the accumulation of differences might be better guides. The ultimate goal is to decouple the identification of novel chemical matter from the laborious steps of isolating strains and sequencing genomes. Once sequences are in digital form, the work towards this goal changes, data may be combined from many sources and becomes more powerful in the hands of many workers beyond those that work at the bench (Stevens, 2013)(Stevens,

2013)(Stevens, 2013)(Stevens, 2013)(Stevens, 2013) (Stevens, 2013) (Stevens, 2013) (Stevens, 2013).

One of the more profound questions that can be asked of this data is about the origin of these compounds. The history of the natural products synthesized by these BGCs is written in the path each of these components took from their origin to the configuration observed in the temporal snapshot represented by the genome sequence of each extant organism. The observation that GPA resistance shares an evolutionary trajectory with the GPA BGCs prompts broader questions about the relationship between antibiotic resistance, the targets of antibiotics and antibiotic biosynthesis (Waglechner et al., 2019)(Waglechner et al., 2019)(Waglechner et al., 2019)(Waglechner et al., 2019)(Waglechner et al., 2019) (Waglechner et al., 2019) (Waglechner et al., 2019) (Waglechner et al., 2019). Unlike other classes of natural products, biosynthesis of compounds with inhibitory activity may only evolve in a producer that is resistant to such compounds. Maintaining the capability to produce these compounds presumably depends on the advantage conferred upon producers in the context of competition with susceptible organisms. Canonical GPA resistance conferred by the *vanHAX* operon or *vanY* differs from other resistance mechanisms that act by modifying or metabolizing an antibiotic into an inactive form in the sense that a GPA molecule is not directly affected by the presence of a resistant organism. It remains to be seen if similar evolutionary relationships between the biosynthesis of, and resistance to, other antibiotic molecules exist beyond the GPAs.

The continued development of a systems approach coupled with biotechnological engineering of both strains and BGCs is the way forward for the identification of novel natural products. Heterologous expression in engineered hosts is a logical step to more fully explore the large numbers of BGCs being discovered in new genomes. In the same way that detailed examination of the genetic diversity has helped to prioritize which BGCs are interesting and worth characterizing, it will guide the systematic exploration cryptic BGCs that are not linked to any known production conditions.

These developments do not address the largest outstanding questions regarding natural products in general, or antibiotics in particular. The uncontroversial view is that natural products exist because they confer an advantage to the producing organism, an idea well within the bounds of the Modern Synthesis. The evolution of antibiotics is conventionally cited as an example of bacterial chemical warfare, though more nuanced explanations have been put forth (Davies, 1990; Kallifidas, Jiang, Ding, & Luesch, 2018; Yim, Wang, & Davies). Central to considering whether antibiotics solely exist to kill other bacteria is the surprising lack of observations of these molecules at inhibitory concentrations in complex natural environments such as the soil. The arms-race metaphor also suggests that stockpiling the capability to produce different classes of antibiotic could be a successful strategy, especially since it is well-known that genera like *Streptomyces* can maintain up to dozens of BGCs in their genomes and selectively express these genes. But this stockpiling is not observed in the bacterial genomes available so far, at least in the sense that strains do not generally harbour multiple BGCs capable of making clinically useful natural products. The compounds that humans have found most useful are not the

same as the compounds the producers have found most useful. The challenge of cryptic antibiotic production by these strains suggests that bacteria may not couple the production of a particular compound in order to inhibit a specific competing neighbor population under standard laboratory conditions. It remains to be discovered if this is also true in natural environments.

Another problem is accounting for the asymmetry between the prevalence of production and resistance. Antibiotics are often the products of a complex set of biosynthetic genes working in concert, that in many cases are rendered ineffective by single mutations or the acquisition of single genes. There may be a finite number of ways to inhibit a microbe in any environment, a finite number of biosynthetic genes, combinations of which produce a finite number of antibiotics. There are a finite number of ways a susceptible organism can become resistant to one of these antibiotics. Understanding the dynamics of each of these elements is necessary to explain the observations of BGCs and resistance determinants in bacterial genomes.

Future Directions

I) Refine and apply the evoc approach to more families of natural products. Good candidates are BGC families with large clusters and high chemical diversity deriving from a shared genetic pool having at least one well-understood mechanism of action. This will better inform how often and where new activities can arise.

II) A dedicated theoretical framework needs to be constructed to model the evolution of multi-domain gene families and multi-gene clusters as they

evolve inside bacterial genomes. The reconciliation approach used in chapter three is an *ad hoc* application of existing tools to a complex problem which calls for new models to better study these complex biological phenomena specifically, and more generally the dynamics of bacterial genome content.

III) Undertake a study of the paired evolution of antibiotic biosynthesis and antibiotic resistance at the scale of at least 10^5 bacterial genomes to identify critical phylogenetic locations (temporal and taxonomic contexts). This is the frame on which to hang a future ecological survey of antibiotic resistance and to estimate the total selective pressure of the anthropogenic use of antibiotics on the bacterial pan-genome.

Concluding Remarks

The remarkable pace of sequencing technology allows unprecedented data generation in the biological sciences. It is possible to leverage this technology against the problem of antibiotic resistance, one of the most pressing issues of this generation through the development of new tools with an eye towards evolution. We show that glycopeptide antibiotic biosynthesis and canonical glycopeptide resistance are linked over millions of years. Starting with the ability to place a family of known antibiotics and their resistance mechanism in an evolutionary genetic context, we have used this information as a steppingstone to discover new antibiotics with a novel mechanism of action.

REFERENCES

Adamek, M., Alanjary, M., Sales-Ortells, H., Goodfellow, M., Bull, A. T., Winkler, A., Wibberg, D., Kalinowski, J., & Ziemert, N. (2018). Comparative genomics reveals phylogenetic distribution patterns of secondary metabolites in Amycolatopsis species. *BMC Genomics, 19*(1), 426. doi:10.1186/s12864-018-4809-4

Adamek, M., Alanjary, M., & Ziemert, N. (2019). Applied evolution: phylogeny-based approaches in natural products research. *Nat Prod Rep, 36*(9), 1295-1312. doi:10.1039/c9np00027e

Alanjary, M., Kronmiller, B., Adamek, M., Blin, K., Weber, T., Huson, D., Philmus, B., & Ziemert, N. (2017). The Antibiotic Resistant Target Seeker (ARTS), an exploration engine for antibiotic cluster prioritization and novel drug target discovery. *Nucleic Acids Res, 45*(W1), W42-W48. doi:10.1093/nar/gkx360

Alcock, B. P., Raphenya, A. R., Lau, T. T. Y., Tsang, K. K., Bouchard, M., Edalatmand, A., Huynh, W., Nguyen, A. V., Cheng, A. A., Liu, S., Min, S. Y., Miroshnichenko, A., Tran, H. K., Werfalli, R. E., Nasir, J. A., Oloni, M., Speicher, D. J., Florescu, A., Singh, B., Faltyn, M., Hernandez-Koutoucheva, A., Sharma, A. N., Bordeleau, E., Pawlowski, A. C., Zubyk, H. L., Dooley, D., Griffiths, E., Maguire, F., Winsor, G. L., Beiko, R. G., Brinkman, F. S. L., Hsiao, W. W. L., Domselaar, G. V., & McArthur, A. G. (2020). CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. *Nucleic Acids Res, 48*(D1), D517-D525. doi:10.1093/nar/gkz935

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res, 25*(17), 3389-3402.

Antibiotic Resistance Threats in the United States, 2013. (2013). Atlanta, GA: U.S. Department of Health and Human Services, CDC Retrieved from https://www.cdc.gov/drugresistance/threat-report-2013/pdf/ar-threats-2013-508.pdf

Arango-Argoty, G., Garner, E., Pruden, A., Heath, L. S., Vikesland, P., & Zhang, L. (2018). DeepARG: a deep learning approach for predicting antibiotic resistance genes from metagenomic data. *Microbiome, 6*(1), 23. doi:10.1186/s40168-018-0401-z

Arthur, M., Molinas, C., & Courvalin, P. (1992). The VanS-VanR two-component regulatory system controls synthesis of depsipeptide peptidoglycan precursors in Enterococcus faecium BM4147. *J Bacteriol, 174*(8), 2582-2591. doi:10.1128/jb.174.8.2582-2591.1992

Ayres, D. L., Darling, A., Zwickl, D. J., Beerli, P., Holder, M. T., Lewis, P. O., Huelsenbeck, J. P., Ronquist, F., Swofford, D. L., Cummings, M. P., Rambaut, A., & Suchard, M. A. (2012). BEAGLE: an application programming interface and high-performance computing library for statistical phylogenetics. *Syst Biol, 61*(1), 170-173. doi:10.1093/sysbio/syr100

Baltz, R. H. (2005). Antibiotic discovery from Actinomycetes: Will a renaissance follow the decline and fall. *SIM News, 55*, 189-196.

Baltz, R. H. (2006). Marcel Faber Roundtable: is our antibiotic pipeline unproductive because of starvation, constipation or lack of inspiration? *J Ind Microbiol Biotechnol, 33*(7), 507-513. doi:10.1007/s10295-005-0077-9

Baltz, R. H. (2018). Natural product drug discovery in the genomic era: realities, conjectures, misconceptions, and opportunities. *J Ind Microbiol Biotechnol*. doi:10.1007/s10295-018-2115-4

Banik, J. J., & Brady, S. F. (2008). Cloning and characterization of new glycopeptide gene clusters found in an environmental DNA megalibrary. *Proc Natl Acad Sci U S A, 105*(45), 17273-17277. doi:10.1073/pnas.0807564105

Banik, J. J., Craig, J. W., Calle, P. Y., & Brady, S. F. (2010). Tailoring enzyme-rich environmental DNA clones: a source of enzymes for generating libraries of unnatural natural products. *J Am Chem Soc, 132*(44), 15661-15670. doi:10.1021/ja105825a

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko, S. I., Pham, S., Prjibelski, A. D., Pyshkin, A. V., Sirotkin, A. V., Vyahhi, N., Tesler, G., Alekseyev, M. A., & Pevzner, P. A. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol, 19*(5), 455-477. doi:10.1089/cmb.2012.0021

Bansal, M. S., Kellis, M., Kordi, M., & Kundu, S. (2018). RANGER-DTL 2.0: rigorous reconstruction of gene-family evolution by duplication, transfer and loss. *Bioinformatics, 34*(18), 3214-3216. doi:10.1093/bioinformatics/bty314

Battistuzzi, F. U., Feijao, A., & Hedges, S. B. (2004). A genomic timescale of prokaryote evolution: insights into the origin of methanogenesis, phototrophy, and the colonization of land. *BMC Evol Biol, 4*, 44. doi:10.1186/1471-2148-4-44

Bentley, S. D., Chater, K. F., Cerdeno-Tarraga, A. M., Challis, G. L., Thomson, N. R., James, K. D., Harris, D. E., Quail, M. A., Kieser, H., Harper, D., Bateman, A., Brown, S., Chandra, G., Chen, C. W., Collins, M., Cronin, A., Fraser, A., Goble, A., Hidalgo, J., Hornsby, T., Howarth, S., Huang, C. H., Kieser, T., Larke, L., Murphy, L., Oliver, K., O'Neil, S., Rabbinowitsch, E., Rajandream, M. A., Rutherford, K., Rutter, S., Seeger, K., Saunders, D., Sharp, S., Squares, R., Squares, S., Taylor, K., Warren, T., Wietzorrek, A., Woodward, J., Barrell, B. G., Parkhill, J., & Hopwood, D. A. (2002). Complete genome sequence of the model actinomycete Streptomyces coelicolor A3(2). *Nature, 417*(6885), 141-147. doi:10.1038/417141a

Bhullar, K., Waglechner, N., Pawlowski, A., Koteva, K., Banks, E. D., Johnston, M. D., Barton, H. A., & Wright, G. D. (2012). Antibiotic resistance is prevalent in an isolated cave microbiome. *PLoS One, 7*(4), e34953. doi:10.1371/journal.pone.0034953

Blin, K., Medema, M. H., Kottmann, R., Lee, S. Y., & Weber, T. (2017). The antiSMASH database, a comprehensive database of microbial secondary metabolite biosynthetic gene clusters. *Nucleic Acids Res, 45*(D1), D555-D559. doi:10.1093/nar/gkw960

Blin, K., Shaw, S., Steinke, K., Villebro, R., Ziemert, N., Lee, S. Y., Medema, M. H., & Weber, T. (2019). antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline. *Nucleic Acids Res, 47*(W1), W81-W87. doi:10.1093/nar/gkz310

Blin, K., Wolf, T., Chevrette, M. G., Lu, X., Schwalen, C. J., Kautsar, S. A., Suarez Duran, H. G., de Los Santos, E. L. C., Kim, H. U., Nave, M., Dickschat, J. S., Mitchell, D. A., Shelest, E., Breitling, R., Takano, E., Lee, S. Y., Weber, T., & Medema, M. H. (2017). antiSMASH 4.0-improvements in chemistry prediction and gene cluster boundary identification. *Nucleic Acids Res, 45*(W1), W36-W41. doi:10.1093/nar/gkx319

Boolchandani, M., D'Souza, A. W., & Dantas, G. (2019). Sequencing-based methods and resources to study antimicrobial resistance. *Nat Rev Genet, 20*(6), 356-370. doi:10.1038/s41576-019-0108-4

Bozhuyuk, K. A. J., Fleischhacker, F., Linck, A., Wesche, F., Tietze, A., Niesert, C. P., & Bode, H. B. (2018). De novo design and engineering of non-ribosomal peptide synthetases. *Nat Chem, 10*(3), 275-281. doi:10.1038/nchem.2890

Bozhuyuk, K. A. J., Linck, A., Tietze, A., Kranz, J., Wesche, F., Nowak, S., Fleischhacker, F., Shi, Y. N., Grun, P., & Bode, H. B. (2019). Modification and de novo design of non-ribosomal peptide synthetases using specific assembly points within condensation domains. *Nat Chem, 11*(7), 653-661. doi:10.1038/s41557-019-0276-z

Brock, W. H. (2000). *The Chemical Tree: A History of Chemistry*. New York: W. W. Norton & Company.

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

Bud, R. (2007). Antibiotics: the epitome of a wonder drug. *BMJ, 334 Suppl 1*, s6. doi:10.1136/bmj.39021.640255.94

Challis, G. L. (2014). Exploitation of the Streptomyces coelicolor A3(2) genome sequence for discovery of new natural products and biosynthetic pathways. *J Ind Microbiol Biotechnol, 41*(2), 219-232. doi:10.1007/s10295-013-1383-2

Challis, G. L., & Hopwood, D. A. (2003). Synergy and contingency as driving forces for the evolution of multiple secondary metabolite production by Streptomyces species. *Proceedings of the National Academy of Sciences, 100*(suppl 2), 14555-14561.

Charlop-Powers, Z., Owen, J. G., Reddy, B. V., Ternei, M. A., & Brady, S. F. (2014). Chemical-biogeographic survey of secondary metabolism in soil. *Proc Natl Acad Sci U S A, 111*(10), 3757-3762. doi:10.1073/pnas.1318021111

Charlop-Powers, Z., Owen, J. G., Reddy, B. V., Ternei, M. A., Guimaraes, D. O., de Frias, U. A., Pupo, M. T., Seepe, P., Feng, Z., & Brady, S. F. (2015). Global biogeographic sampling of bacterial secondary metabolism. *Elife, 4*, e05048. doi:10.7554/eLife.05048

Charlop-Powers, Z., Pregitzer, C. C., Lemetre, C., Ternei, M. A., Maniko, J., Hover, B. M., Calle, P. Y., McGuire, K. L., Garbarino, J., Forgione, H. M., Charlop-Powers, S., & Brady, S. F. (2016). Urban park soil microbiomes are a rich reservoir of natural product biosynthetic diversity. *Proc Natl Acad Sci U S A, 113*(51), 14811-14816. doi:10.1073/pnas.1615581113

Chen, H., Thomas, M. G., Hubbard, B. K., Losey, H. C., Walsh, C. T., & Burkart, M. D. (2000). Deoxysugars in glycopeptide antibiotics: enzymatic synthesis of TDP-L-epivancosamine in chloroeremomycin biosynthesis. *Proc Natl Acad Sci U S A, 97*(22), 11942-11947. doi:10.1073/pnas.210395097

Chen, H., Tseng, C. C., Hubbard, B. K., & Walsh, C. T. (2001). Glycopeptide antibiotic biosynthesis: enzymatic assembly of the dedicated amino acid monomer (S)-3,5-dihydroxyphenylglycine. *Proc Natl Acad Sci U S A, 98*(26), 14901-14906. doi:10.1073/pnas.221582098

Chevrette, M. G., Aicheler, F., Kohlbacher, O., Currie, C. R., & Medema, M. H. (2017). SANDPUMA: ensemble predictions of nonribosomal peptide chemistry reveal biosynthetic diversity across Actinobacteria. *Bioinformatics, 33*(20), 3202-3210. doi:10.1093/bioinformatics/btx400

Chevrette, M. G., Carlos-Shanley, C., Louie, K. B., Bowen, B. P., Northen, T. R., & Currie, C. R. (2019). Taxonomic and Metabolic Incongruence in the Ancient Genus Streptomyces. *Front Microbiol, 10*, 2170. doi:10.3389/fmicb.2019.02170

Chevrette, M. G., & Currie, C. R. (2018). Emerging evolutionary paradigms in antibiotic discovery. *J Ind Microbiol Biotechnol*. doi:10.1007/s10295-018-2085-6

Chin, C. S., Alexander, D. H., Marks, P., Klammer, A. A., Drake, J., Heiner, C., Clum, A., Copeland, A., Huddleston, J., Eichler, E. E., Turner, S. W., & Korlach, J. (2013). Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat Methods, 10*(6), 563-569. doi:10.1038/nmeth.2474

Chiu, H. T., Hubbard, B. K., Shah, A. N., Eide, J., Fredenburg, R. A., Walsh, C. T., & Khosla, C. (2001). Molecular cloning and sequence analysis of the complestatin biosynthetic gene cluster. *Proc Natl Acad Sci U S A, 98*(15), 8548-8553. doi:10.1073/pnas.151246498

Cimermancic, P., Medema, M. H., Claesen, J., Kurita, K., Wieland Brown, L. C., Mavrommatis, K., Pati, A., Godfrey, P. A., Koehrsen, M., Clardy, J., Birren, B. W., Takano, E., Sali, A.,

Linington, R. G., & Fischbach, M. A. (2014). Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters. *Cell, 158*(2), 412-421. doi:10.1016/j.cell.2014.06.034

Cock, P. J., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., & de Hoon, M. J. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics, 25*(11), 1422-1423. doi:10.1093/bioinformatics/btp163

Cohen, O., & Pupko, T. (2011). Inference of gain and loss events from phyletic patterns using stochastic mapping and maximum parsimony--a simulation study. *Genome Biol Evol, 3*, 1265-1275. doi:10.1093/gbe/evr101

Cole, J. R., Wang, Q., Fish, J. A., Chai, B., McGarrell, D. M., Sun, Y., Brown, C. T., Porras-Alfaro, A., Kuske, C. R., & Tiedje, J. M. (2014). Ribosomal Database Project: data and tools for high throughput rRNA analysis. *Nucleic Acids Res, 42*(Database issue), D633-642. doi:10.1093/nar/gkt1244

Cornforth, D. M., & Foster, K. R. (2015). Antibiotics and the art of bacterial war. *Proc Natl Acad Sci U S A, 112*(35), 10827-10828. doi:10.1073/pnas.1513608112

Cox, G., Sieron, A., King, A. M., De Pascale, G., Pawlowski, A. C., Koteva, K., & Wright, G. D. (2017). A Common Platform for Antibiotic Dereplication and Adjuvant Discovery. *Cell Chem Biol, 24*(1), 98-109. doi:10.1016/j.chembiol.2016.11.011

Cox, G., & Wright, G. D. (2013). Intrinsic antibiotic resistance: mechanisms, origins, challenges and solutions. *Int J Med Microbiol, 303*(6-7), 287-292. doi:10.1016/j.ijmm.2013.02.009

Crits-Christoph, A., Diamond, S., Butterfield, C. N., Thomas, B. C., & Banfield, J. F. (2018). Novel soil bacteria possess diverse genes for secondary metabolite biosynthesis. *Nature, 558*(7710), 440-444. doi:10.1038/s41586-018-0207-y

Cruz-Morales, P., Kopp, J. F., Martinez-Guerrero, C., Yanez-Guerra, L. A., Selem-Mojica, N., Ramos-Aboites, H., Feldmann, J., & Barona-Gomez, F. (2016). Phylogenomic Analysis of Natural Products Biosynthetic Gene Clusters Allows Discovery of Arseno-Organic Metabolites in Model Streptomycetes. *Genome Biol Evol, 8*(6), 1906-1916. doi:10.1093/gbe/evw125

Culp, E. J., Waglechner, N., Wang, W., Fiebig-Comyn, A. A., Hsu, Y. P., Koteva, K., Sychantha, D., Coombes, B. K., Van Nieuwenhze, M. S., Brun, Y. V., & Wright, G. D. (2020). Evolution-guided discovery of antibiotics that inhibit peptidoglycan remodelling. *Nature, 578*(7796), 582-587. doi:10.1038/s41586-020-1990-9

Culp, E. J., Yim, G., Waglechner, N., Wang, W., Pawlowski, A. C., & Wright, G. D. (2019). Hidden antibiotics in actinomycetes can be identified by inactivation of gene clusters for common antibiotics. *Nat Biotechnol, 37*(10), 1149-1154. doi:10.1038/s41587-019-0241-9

D'Costa, V. M., King, C. E., Kalan, L., Morar, M., Sung, W. W., Schwarz, C., Froese, D., Zazula, G., Calmels, F., Debruyne, R., Golding, G. B., Poinar, H. N., & Wright, G. D. (2011). Antibiotic resistance is ancient. *Nature, 477*(7365), 457-461. doi:10.1038/nature10388

Davies, J. (1990). What are antibiotics? Archaic functions for modern activities. *Mol Microbiol, 4*(8), 1227-1232.

Davis, J. J., Boisvert, S., Brettin, T., Kenyon, R. W., Mao, C., Olson, R., Overbeek, R., Santerre, J., Shukla, M., Wattam, A. R., Will, R., Xia, F., & Stevens, R. (2016). Antimicrobial Resistance Prediction in PATRIC and RAST. *Sci Rep, 6*, 27930. doi:10.1038/srep27930

Deatherage, D. E., & Barrick, J. E. (2014). Identification of mutations in laboratory-evolved microbes from next-generation sequencing data using breseq. *Methods Mol Biol, 1151*, 165-188. doi:10.1007/978-1-4939-0554-6_12

Demain, A. L. (1974). How do antibiotic-producing microorganisms avoid suicide? *Ann N Y Acad Sci, 235*(0), 601-612. doi:10.1111/j.1749-6632.1974.tb43294.x

Demain, A. L. (2009). Antibiotics: natural products essential to human health. *Med Res Rev, 29*(6), 821-842. doi:10.1002/med.20154

Demain, A. L. (2014). Importance of microbial natural products and the need to revitalize their discovery. *J Ind Microbiol Biotechnol, 41*(2), 185-201. doi:10.1007/s10295-013-1325-z

Doroghazi, J. R., Albright, J. C., Goering, A. W., Ju, K. S., Haines, R. R., Tchalukov, K. A., Labeda, D. P., Kelleher, N. L., & Metcalf, W. W. (2014). A roadmap for natural product discovery based on large-scale genomics and metabolomics. *Nat Chem Biol, 10*(11), 963-968. doi:10.1038/nchembio.1659

Doroghazi, J. R., & Metcalf, W. W. (2013). Comparative genomics of actinomycetes with a focus on natural product biosynthetic genes. *BMC Genomics, 14*, 611. doi:10.1186/1471-2164-14-611

Doster, E., Lakin, S. M., Dean, C. J., Wolfe, C., Young, J. G., Boucher, C., Belk, K. E., Noyes, N. R., & Morley, P. S. (2020). MEGARes 2.0: a database for classification of antimicrobial drug, biocide and metal resistance determinants in metagenomic sequence data. *Nucleic Acids Res, 48*(D1), D561-D569. doi:10.1093/nar/gkz1010

Drummond, A. J., & Rambaut, A. (2007). BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol Biol, 7*, 214. doi:10.1186/1471-2148-7-214

Eddy, S. R. (2009). A new generation of homology search tools based on probabilistic inference. *Genome Inform, 23*(1), 205-211.

Eddy, S. R. (2011). Accelerated Profile HMM Searches. *PLoS Comput Biol, 7*(10), e1002195. doi:10.1371/journal.pcbi.1002195

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res, 32*(5), 1792-1797. doi:10.1093/nar/gkh340

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics, 26*(19), 2460-2461. doi:10.1093/bioinformatics/btq461

Eppelmann, K., Stachelhaus, T., & Marahiel, M. A. (2002). Exploitation of the selectivity-conferring code of nonribosomal peptide synthetases for the rational design of novel peptide antibiotics. *Biochemistry, 41*(30), 9718-9726.

Fang, X., Tiyanont, K., Zhang, Y., Wanner, J., Boger, D., & Walker, S. (2006). The mechanism of action of ramoplanin and enduracidin. *Mol Biosyst, 2*(1), 69-76. doi:10.1039/b515328j

Farnet, C. M. S., A.; Zazopoulos, E. . (2002). WO2002031155A2. WIPO.

Federal Action Plan on Antimicrobial Resistance and Use in Canada. (140538). (2015). Ottawa, ON, Canada: Public Health Agency of Canada

Feldgarden, M., Brover, V., Haft, D. H., Prasad, A. B., Slotta, D. J., Tolstoy, I., Tyson, G. H., Zhao, S., Hsu, C. H., McDermott, P. F., Tadesse, D. A., Morales, C., Simmons, M., Tillman, G., Wasilenko, J., Folster, J. P., & Klimke, W. (2019). Validating the AMRFinder Tool and Resistance Gene Database by Using Antimicrobial Resistance Genotype-Phenotype Correlations in a Collection of Isolates. *Antimicrob Agents Chemother, 63*(11). doi:10.1128/AAC.00483-19

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol, 17*(6), 368-376. doi:10.1007/bf01734359

Felsenstein, J. (2003). *Inferring Phylogenies*: Sinauer.

Ferir, G., Hanchen, A., Francois, K. O., Hoorelbeke, B., Huskens, D., Dettner, F., Sussmuth, R. D., & Schols, D. (2012). Feglymycin, a unique natural bacterial antibiotic peptide, inhibits HIV entry by targeting the viral envelope protein gp120. *Virology, 433*(2), 308-319. doi:10.1016/j.virol.2012.08.007

Fertin, G., Jean, G., & Tannier, E. (2017). Algorithms for computing the double cut and join distance on both gene order and intergenic sizes. *Algorithms Mol Biol, 12*, 16. doi:10.1186/s13015-017-0107-y

Finn, R. D., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Mistry, J., Mitchell, A. L., Potter, S. C., Punta, M., Qureshi, M., Sangrador-Vegas, A., Salazar, G. A., Tate, J., & Bateman, A. (2016). The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res, 44*(D1), D279-285. doi:10.1093/nar/gkv1344

Fischbach, M. A., & Walsh, C. T. (2006). Assembly-line enzymology for polyketide and nonribosomal Peptide antibiotics: logic, machinery, and mechanisms. *Chem Rev, 106*(8), 3468-3496. doi:10.1021/cr0503097

Fleming, A. (1945). *Penicillin. Nobel Lecture. 1945*. Paper presented at the Nobel Lecture, Stockholm, Sweden.

Forsberg, K. J., Reyes, A., Wang, B., Selleck, E. M., Sommer, M. O., & Dantas, G. (2012). The shared antibiotic resistome of soil bacteria and human pathogens. *Science, 337*(6098), 1107-1111. doi:10.1126/science.1220761

Gensini, G. F., Conti, A. A., & Lippi, D. (2007). The contributions of Paul Ehrlich to infectious disease. *J Infect, 54*(3), 221-224. doi:10.1016/j.jinf.2004.05.022

Gibson, M. K., Forsberg, K. J., & Dantas, G. (2015). Improved annotation of antibiotic resistance determinants reveals microbial resistomes cluster by ecology. *ISME J, 9*(1), 207-216. doi:10.1038/ismej.2014.106

Gibson, M. K., Wang, B., Ahmadi, S., Burnham, C. A., Tarr, P. I., Warner, B. B., & Dantas, G. (2016). Developmental dynamics of the preterm infant gut microbiota and antibiotic resistome. *Nat Microbiol, 1*, 16024. doi:10.1038/nmicrobiol.2016.24

Gonsior, M., Muhlenweg, A., Tietzmann, M., Rausch, S., Poch, A., & Sussmuth, R. D. (2015). Biosynthesis of the Peptide Antibiotic Feglymycin by a Linear Nonribosomal Peptide Synthetase Mechanism. *Chembiochem, 16*(18), 2610-2614. doi:10.1002/cbic.201500432

Goodman, M. C., J., Moore, G. W., Romero-Herrera, A. E., & Matsuda, G. (1979). Fitting the Gene Lineage into its Species Lineage, a Parsimony Strategy Illustrated by Cladograms Constructed from Globin Sequences. *Systematic Zoology, 28*(2), 132-163.

Gottlieb, D. (1976). The production and role of antibiotics in soil. *The Journal of antibiotics, 29*(10), 987-1000.

Greule, A., Izore, T., Iftime, D., Tailhades, J., Schoppet, M., Zhao, Y., Peschke, M., Ahmed, I., Kulik, A., Adamek, M., Goode, R. J. A., Schittenhelm, R. B., Kaczmarski, J. A., Jackson, C. J., Ziemert, N., Krenske, E. H., De Voss, J. J., Stegmann, E., & Cryle, M. J. (2019). Kistamicin biosynthesis reveals the biosynthetic requirements for production of highly crosslinked glycopeptide antibiotics. *Nat Commun, 10*(1), 2613. doi:10.1038/s41467-019-10384-w

Gupta, S. K., Padmanabhan, B. R., Diene, S. M., Lopez-Rojas, R., Kempf, M., Landraud, L., & Rolain, J. M. (2014). ARG-ANNOT, a new bioinformatic tool to discover antibiotic resistance genes in bacterial genomes. *Antimicrob Agents Chemother, 58*(1), 212-220. doi:10.1128/AAC.01310-13

Hadjithomas, M., Chen, I. M., Chu, K., Ratner, A., Palaniappan, K., Szeto, E., Huang, J., Reddy, T. B., Cimermancic, P., Fischbach, M. A., Ivanova, N. N., Markowitz, V. M., Kyrpides, N. C., & Pati, A. (2015). IMG-ABC: A Knowledge Base To Fuel Discovery of Biosynthetic Gene Clusters and Novel Secondary Metabolites. *MBio, 6*(4), e00932. doi:10.1128/mBio.00932-15

Hall, B. G., & Barlow, M. (2005). Revised Ambler classification of {beta}-lactamases. *J Antimicrob Chemother, 55*(6), 1050-1051. doi:10.1093/jac/dki130

Hanchen, A., Rausch, S., Landmann, B., Toti, L., Nusser, A., & Sussmuth, R. D. (2013). Alanine scan of the peptide antibiotic feglymycin: assessment of amino acid side chains contributing to antimicrobial activity. *Chembiochem, 14*(5), 625-632. doi:10.1002/cbic.201300032

Hao, Y., Pierce, E., Roe, D., Morita, M., McIntosh, J. A., Agarwal, V., Cheatham, T. E., 3rd, Schmidt, E. W., & Nair, S. K. (2016). Molecular basis for the broad substrate selectivity of a peptide prenyltransferase. *Proc Natl Acad Sci U S A, 113*(49), 14037-14042. doi:10.1073/pnas.1609869113

Harms, A., Maisonneuve, E., & Gerdes, K. (2016). Mechanisms of bacterial persistence during stress and antibiotic exposure. *Science, 354*(6318). doi:10.1126/science.aaf4268

Haydock, S. F., Aparicio, J. F., Molnar, I., Schwecke, T., Khaw, L. E., Konig, A., Marsden, A. F., Galloway, I. S., Staunton, J., & Leadlay, P. F. (1995). Divergent sequence motifs correlated with the substrate specificity of (methyl)malonyl-CoA:acyl carrier protein transacylase domains in modular polyketide synthases. *FEBS Lett, 374*(2), 246-248. doi:10.1016/0014-5793(95)01119-y

Hoertz, A. J., Hamburger, J. B., Gooden, D. M., Bednar, M. M., & McCafferty, D. G. (2012). Studies on the biosynthesis of the lipodepsipeptide antibiotic Ramoplanin A2. *Bioorg Med Chem, 20*(2), 859-865. doi:10.1016/j.bmc.2011.11.062

Hong, H. J., Hutchings, M. I., Buttner, M. J., Biotechnology, & Biological Sciences Research Council, U. K. (2008). Vancomycin resistance VanS/VanR two-component systems. *Adv Exp Med Biol, 631*, 200-213. doi:10.1007/978-0-387-78885-2_14

Huelsenbeck, J. P. (1995). Performance of Phylogenetic Methods in Simulation. *Systematic Biology, 44*(1), 17-48. doi:10.2307/2413481

Huelsenbeck, J. P., & Ronquist, F. (2001). MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics, 17*(8), 754-755. doi:10.1093/bioinformatics/17.8.754

Huerta-Cepas, J., Serra, F., & Bork, P. (2016). ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Mol Biol Evol, 33*(6), 1635-1638. doi:10.1093/molbev/msw046

Hyatt, D., Chen, G. L., Locascio, P. F., Land, M. L., Larimer, F. W., & Hauser, L. J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics, 11*, 119. doi:10.1186/1471-2105-11-119

Ikeda, H., Ishikawa, J., Hanamoto, A., Shinose, M., Kikuchi, H., Shiba, T., Sakaki, Y., Hattori, M., & Omura, S. (2003). Complete genome sequence and comparative analysis of the industrial microorganism Streptomyces avermitilis. *Nat Biotechnol, 21*(5), 526-531. doi:10.1038/nbt820

Jacox, E., Chauve, C., Szöllősi, G. J., Ponty, Y., & Scornavacca, C. (2016). ecceTERA: comprehensive gene tree-species tree reconciliation using parsimony. *Bioinformatics, 32*(13), 2056-2058.

Jia, B., Raphenya, A. R., Alcock, B., Waglechner, N., Guo, P., Tsang, K. K., Lago, B. A., Dave, B. M., Pereira, S., Sharma, A. N., Doshi, S., Courtot, M., Lo, R., Williams, L. E., Frye, J. G., Elsayegh, T., Sardar, D., Westman, E. L., Pawlowski, A. C., Johnson, T. A., Brinkman, F. S., Wright, G. D., & McArthur, A. G. (2017). CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database. *Nucleic Acids Res, 45*(D1), D566-D573. doi:10.1093/nar/gkw1004

Jiang, H., Lei, R., Ding, S. W., & Zhu, S. (2014). Skewer: a fast and accurate adapter trimmer for next-generation sequencing paired-end reads. *BMC Bioinformatics, 15*, 182. doi:10.1186/1471-2105-15-182

Jombart, T., Kendall, M., Almagro-Garcia, J., & Colijn, C. (2017). treespace: Statistical exploration of landscapes of phylogenetic trees. *Mol Ecol Resour, 17*(6), 1385-1392. doi:10.1111/1755-0998.12676

Joynt, R., & Seipke, R. F. (2018). A phylogenetic and evolutionary analysis of antimycin biosynthesis. *Microbiology, 164*(1), 28-39. doi:10.1099/mic.0.000572

Ju, K. S., Gao, J., Doroghazi, J. R., Wang, K. K., Thibodeaux, C. J., Li, S., Metzger, E., Fudala, J., Su, J., Zhang, J. K., Lee, J., Cioni, J. P., Evans, B. S., Hirota, R., Labeda, D. P., van der Donk, W. A., & Metcalf, W. W. (2015). Discovery of phosphonic acid natural products by mining the genomes of 10,000 actinomycetes. *Proc Natl Acad Sci U S A, 112*(39), 12175-12180. doi:10.1073/pnas.1500873112

Kallifidas, D., Jiang, G., Ding, Y., & Luesch, H. (2018). Rational engineering of Streptomyces albus J1074 for the overexpression of secondary metabolite gene clusters. *Microb Cell Fact, 17*(1), 25. doi:10.1186/s12934-018-0874-2

Kannan, L., Li, H., Rubinstein, B., & Mushegian, A. (2013). Models of gene gain and gene loss for probabilistic reconstruction of gene content in the last universal common ancestor of life. *Biol Direct, 8*, 32. doi:10.1186/1745-6150-8-32

Kaur, N., Kumar, S., Bala, M., Raghava, G. P., & Mayilraj, S. (2013). Draft Genome Sequence of Amycolatopsis decaplanina Strain DSM 44594T. *Genome Announc, 1*(2), e0013813. doi:10.1128/genomeA.00138-13

Kautsar, S. A., Suarez Duran, H. G., Blin, K., Osbourn, A., & Medema, M. H. (2017). plantiSMASH: automated identification, annotation and expression analysis of plant biosynthetic gene clusters. *Nucleic Acids Res, 45*(W1), W55-W63. doi:10.1093/nar/gkx305

Khosla, C., Gokhale, R. S., Jacobsen, J. R., & Cane, D. E. (1999). Tolerance and specificity of polyketide synthases. *Annu Rev Biochem, 68*, 219-253. doi:10.1146/annurev.biochem.68.1.219

Kieser, T., Bibb, M. J., Buttner, M. J., Chater, K. F., & Hopwood, D. A. (2000). *Practical streptomyces genetics* (Vol. 291): John Innes Foundation Norwich.

Kilian, R., Frasch, H. J., Kulik, A., Wohlleben, W., & Stegmann, E. (2016). The VanRS Homologous Two-Component System VnlRSAb of the Glycopeptide Producer Amycolatopsis balhimycina Activates Transcription of the vanHAXSc Genes in Streptomyces coelicolor, but not in A. balhimycina. *Microb Drug Resist, 22*(6), 499-509. doi:10.1089/mdr.2016.0128

Kirst, H. A., Thompson, D. G., & Nicas, T. I. (1998). Historical yearly usage of vancomycin. *Antimicrob Agents Chemother, 42*(5), 1303-1304.

Kohli, R. M., Trauger, J. W., Schwarzer, D., Marahiel, M. A., & Walsh, C. T. (2001). Generality of peptide cyclization catalyzed by isolated thioesterase domains of nonribosomal peptide synthetases. *Biochemistry, 40*(24), 7099-7108. doi:10.1021/bi010036j

Komaki, H., Ichikawa, N., Oguchi, A., Hamada, M., Tamura, T., & Fujita, N. (2015). Genome-based analysis of non-ribosomal peptide synthetase and type-I polyketide synthase gene clusters in all type strains of the genus Herbidospora. *BMC Res Notes, 8*, 548. doi:10.1186/s13104-015-1526-9

Krogh, A., Brown, M., Mian, I. S., Sjolander, K., & Haussler, D. (1994). Hidden Markov models in computational biology. Applications to protein modeling. *J Mol Biol, 235*(5), 1501-1531. doi:10.1006/jmbi.1994.1104

Kunin, V., Goldovsky, L., Darzentas, N., & Ouzounis, C. A. (2005). The net of life: reconstructing the microbial phylogenetic network. *Genome Res, 15*(7), 954-959. doi:10.1101/gr.3666505

Kwon, Y. J., Kim, H. J., & Kim, W. G. (2015). Complestatin exerts antibacterial activity by the inhibition of fatty acid synthesis. *Biol Pharm Bull, 38*(5), 715-721. doi:10.1248/bpb.b14-00824

Kwun, M. J., Cheng, J., Yang, S. H., Lee, D. R., Suh, J. W., & Hong, H. J. (2014). Draft Genome Sequence of Ristocetin-Producing Strain Amycolatopsis sp. Strain MJM2582 Isolated in South Korea. *Genome Announc, 2*(5). doi:10.1128/genomeA.01091-14

Kwun, M. J., & Hong, H. J. (2014a). Draft Genome Sequence of Amycolatopsis lurida NRRL 2430, Producer of the Glycopeptide Family Antibiotic Ristocetin. *Genome Announc, 2*(5). doi:10.1128/genomeA.01050-14

Kwun, M. J., & Hong, H. J. (2014b). Genome Sequence of Streptomyces toyocaensis NRRL 15009, Producer of the Glycopeptide Antibiotic A47934. *Genome Announc, 2*(4). doi:10.1128/genomeA.00749-14

Lagesen, K., Hallin, P., Rodland, E. A., Staerfeldt, H. H., Rognes, T., & Ussery, D. W. (2007). RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res, 35*(9), 3100-3108. doi:10.1093/nar/gkm160

Leclercq, R., Derlot, E., Duval, J., & Courvalin, P. (1988). Plasmid-mediated resistance to vancomycin and teicoplanin in Enterococcus faecium. *N Engl J Med, 319*(3), 157-161. doi:10.1056/NEJM198807213190307

Lewis, K. (2012). Antibiotics: Recover the lost art of drug discovery. *Nature, 485*(7399), 439-440. doi:10.1038/485439a

Li, T. L., Huang, F., Haydock, S. F., Mironenko, T., Leadlay, P. F., & Spencer, J. B. (2004). Biosynthetic gene cluster of the glycopeptide antibiotic teicoplanin: characterization of two glycosyltransferases and the key acyltransferase. *Chem Biol, 11*(1), 107-119. doi:10.1016/j.chembiol.2004.01.001

Libeskind-Hadas, R., Wu, Y. C., Bansal, M. S., & Kellis, M. (2014). Pareto-optimal phylogenetic tree reconciliation. *Bioinformatics, 30*(12), i87-95. doi:10.1093/bioinformatics/btu289

Lin, K., Zhu, L., & Zhang, D. Y. (2006). An initial strategy for comparing proteins at the domain architecture level. *Bioinformatics, 22*(17), 2081-2086. doi:10.1093/bioinformatics/btl366

Liu, B., & Pop, M. (2009). ARDB--Antibiotic Resistance Genes Database. *Nucleic Acids Res, 37*(Database issue), D443-447. doi:10.1093/nar/gkn656

Liu, C., Wright, B., Allen-Vercoe, E., Gu, H., & Beiko, R. (2018). Phylogenetic Clustering of Genes Reveals Shared Evolutionary Trajectories and Putative Gene Functions. *Genome Biol Evol, 10*(9), 2255-2265. doi:10.1093/gbe/evy178

Lo Grasso, L., Maffioli, S., Sosio, M., Bibb, M., Puglia, A. M., & Alduina, R. (2015). Two Master Switch Regulators Trigger A40926 Biosynthesis in Nonomuraea sp. Strain ATCC 39727. *J Bacteriol, 197*(15), 2536-2544. doi:10.1128/JB.00262-15

Lo, M.-C., Men, H., Branstrom, A., Helm, J., Yao, N., Goldman, R., & Walker, S. (2000). A New Mechanism of Action Proposed for Ramoplanin. *Journal of the American Chemical Society, 122*(14), 3540-3541. doi:10.1021/ja000182x

Magarvey, N. A., Haltli, B., He, M., Greenstein, M., & Hucul, J. A. (2006). Biosynthetic pathway for mannopeptimycins, lipoglycopeptide antibiotics active against drug-resistant gram-positive pathogens. *Antimicrob Agents Chemother, 50*(6), 2167-2177. doi:10.1128/AAC.01545-05

Magoc, T., & Salzberg, S. L. (2011). FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics, 27*(21), 2957-2963. doi:10.1093/bioinformatics/btr507

Maiolo, M., Zhang, X., Gil, M., & Anisimova, M. (2018). Progressive multiple sequence alignment with indel evolution. *BMC Bioinformatics, 19*(1), 331. doi:10.1186/s12859-018-2357-1

Marcone, G. L., Beltrametti, F., Binda, E., Carrano, L., Foulston, L., Hesketh, A., Bibb, M., & Marinelli, F. (2010). Novel mechanism of glycopeptide resistance in the A40926 producer Nonomuraea sp. ATCC 39727. *Antimicrob Agents Chemother, 54*(6), 2465-2472. doi:10.1128/AAC.00106-10

Marshall, C. G., Broadhead, G., Leskiw, B. K., & Wright, G. D. (1997). D-Ala-D-Ala ligases from glycopeptide antibiotic-producing organisms are highly homologous to the enterococcal vancomycin-resistance ligases VanA and VanB. *Proc Natl Acad Sci U S A, 94*(12), 6480-6483.

Martin, J. F. (1992). Clusters of genes for the biosynthesis of antibiotics: regulatory genes and overproduction of pharmaceuticals. *J Ind Microbiol, 9*(2), 73-90.

McArthur, A. G., Waglechner, N., Nizam, F., Yan, A., Azad, M. A., Baylay, A. J., Bhullar, K., Canova, M. J., De Pascale, G., Ejim, L., Kalan, L., King, A. M., Koteva, K., Morar, M., Mulvey, M. R., O'Brien, J. S., Pawlowski, A. C., Piddock, L. J., Spanogiannopoulos, P., Sutherland, A. D., Tang, I., Taylor, P. L., Thaker, M., Wang, W., Yan, M., Yu, T., & Wright, G. D. (2013). The comprehensive antibiotic resistance database. *Antimicrob Agents Chemother, 57*(7), 3348-3357. doi:10.1128/AAC.00419-13

McDonald, B. R., & Currie, C. R. (2017). Lateral Gene Transfer Dynamics in the Ancient Bacterial Genus Streptomyces. *MBio, 8*(3). doi:10.1128/mBio.00644-17

Medema, M. H., Blin, K., Cimermancic, P., de Jager, V., Zakrzewski, P., Fischbach, M. A., Weber, T., Takano, E., & Breitling, R. (2011). antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. *Nucleic Acids Res, 39*(Web Server issue), W339-346. doi:10.1093/nar/gkr466

Medema, M. H., Cimermancic, P., Sali, A., Takano, E., & Fischbach, M. A. (2014). A systematic computational analysis of biosynthetic gene cluster evolution: lessons for engineering biosynthesis. *PLoS Comput Biol, 10*(12), e1004016. doi:10.1371/journal.pcbi.1004016

Medema, M. H., Kottmann, R., Yilmaz, P., Cummings, M., Biggins, J. B., Blin, K., de Bruijn, I., Chooi, Y. H., Claesen, J., Coates, R. C., Cruz-Morales, P., Duddela, S., Dusterhus, S., Edwards, D. J., Fewer, D. P., Garg, N., Geiger, C., Gomez-Escribano, J. P., Greule, A., Hadjithomas, M., Haines, A. S., Helfrich, E. J., Hillwig, M. L., Ishida, K., Jones, A. C., Jones, C. S., Jungmann, K., Kegler, C., Kim, H. U., Kotter, P., Krug, D., Masschelein, J., Melnik, A. V., Mantovani, S. M., Monroe, E. A., Moore, M., Moss, N., Nutzmann, H. W., Pan, G., Pati, A., Petras, D., Reen, F. J., Rosconi, F., Rui, Z., Tian, Z., Tobias, N. J., Tsunematsu, Y., Wiemann, P., Wyckoff, E., Yan, X., Yim, G., Yu, F., Xie, Y., Aigle, B., Apel, A. K., Balibar, C. J., Balskus, E. P., Barona-Gomez, F., Bechthold, A., Bode, H. B., Borriss, R., Brady, S. F., Brakhage, A. A., Caffrey, P., Cheng, Y. Q., Clardy, J., Cox, R. J., De Mot, R., Donadio, S., Donia, M. S., van der Donk, W. A., Dorrestein, P. C., Doyle, S., Driessen, A. J., Ehling-Schulz, M., Entian, K. D., Fischbach, M. A., Gerwick, L., Gerwick, W. H., Gross, H., Gust, B., Hertweck, C., Hofte, M., Jensen, S. E., Ju, J., Katz, L., Kaysser, L., Klassen, J. L., Keller, N. P., Kormanec, J., Kuipers, O. P., Kuzuyama, T., Kyrpides, N. C., Kwon, H. J., Lautru, S., Lavigne, R., Lee, C. Y., Linquan, B., Liu, X., Liu, W., Luzhetskyy, A., Mahmud, T., Mast, Y., Mendez, C., Metsa-Ketela, M., Micklefield, J., Mitchell, D. A., Moore, B. S., Moreira, L. M., Muller, R., Neilan, B. A., Nett, M., Nielsen, J., O'Gara, F., Oikawa, H., Osbourn, A., Osburne, M. S., Ostash, B., Payne, S. M., Pernodet, J. L., Petricek, M., Piel, J., Ploux, O., Raaijmakers, J. M., Salas, J. A., Schmitt, E. K., Scott, B., Seipke, R. F., Shen, B., Sherman, D. H., Sivonen, K., Smanski, M. J., Sosio, M., Stegmann, E., Sussmuth, R. D., Tahlan, K., Thomas, C. M., Tang, Y., Truman, A. W., Viaud, M., Walton, J. D., Walsh, C. T., Weber, T., van Wezel, G. P., Wilkinson, B., Willey, J. M., Wohlleben, W., Wright, G. D., Ziemert, N., Zhang, C., Zotchev, S. B.,

Breitling, R., Takano, E., & Glockner, F. O. (2015). Minimum Information about a Biosynthetic Gene cluster. *Nat Chem Biol, 11*(9), 625-631. doi:10.1038/nchembio.1890

Medema, M. H., Takano, E., & Breitling, R. (2013). Detecting sequence homology at the gene cluster level with MultiGeneBlast. *Mol Biol Evol, 30*(5), 1218-1223. doi:10.1093/molbev/mst025

Meleshko, D., Mohimani, H., Tracanna, V., Hajirasouliha, I., Medema, M. H., Korobeynikov, A., & Pevzner, P. A. (2019). BiosyntheticSPAdes: reconstructing biosynthetic gene clusters from assembly graphs. *Genome Res, 29*(8), 1352-1362. doi:10.1101/gr.243477.118

Mollo, A., von Krusenstiern, A. N., Bulos, J. A., Ulrich, V., Åkerfeldt, K. S., Cryle, M. J., & Charkoudian, L. K. (2017). P450 monooxygenase ComJ catalyses side chain phenolic cross-coupling during complestatin biosynthesis. *RSC Adv., 7*(56), 35376-35384. doi:10.1039/c7ra06518c

Mungall, C. J., Emmert, D. B., & FlyBase, C. (2007). A Chado case study: an ontology-based modular schema for representing genome-associated biological information. *Bioinformatics, 23*(13), i337-346. doi:10.1093/bioinformatics/btm189

Naas, T., Oueslati, S., Bonnin, R. A., Dabos, M. L., Zavala, A., Dortet, L., Retailleau, P., & Iorga, B. I. (2017). Beta-lactamase database (BLDB) - structure and function. *J Enzyme Inhib Med Chem, 32*(1), 917-919. doi:10.1080/14756366.2017.1344235

Nasrallah, C. A., Mathews, D. H., & Huelsenbeck, J. P. (2011). Quantifying the impact of dependent evolution among sites in phylogenetic inference. *Syst Biol, 60*(1), 60-73. doi:10.1093/sysbio/syq074

Navarro-Muñoz, J., Selem-Mojica, N., Mullowney, M., Kautsar, S., Tryon, J., Parkinson, E., De Los Santos, E., Yeong, M., Cruz-Morales, P., Abubucker, S., Roeters, A., Lokhorst, W., Fernandez-Guerra, A., Teresa Dias Cappelini, L., Thomson, R., Metcalf, W., Kelleher, N., Barona-Gomez, F., & Medema, M. H. (2018). A computational framework for systematic exploration of biosynthetic diversity from large-scale genomic data. *bioRxiv*, 445270v445271. doi:10.1101/445270

Nazari, B., Forneris, C. C., Gibson, M. I., Moon, K., Schramma, K. R., & Seyedsayamdost, M. R. (2017). Nonomuraea sp. ATCC 55076 harbours the largest actinomycete chromosome to date and the kistamicin biosynthetic gene cluster. *Medchemcomm, 8*(4), 780-788. doi:10.1039/c6md00637j

Nicolaou, K. C., Boddy, C. N., Brase, S., & Winssinger, N. (1999). Chemistry, Biology, and Medicine of the Glycopeptide Antibiotics. *Angew Chem Int Ed Engl, 38*(15), 2096-2152. doi:10.1002/(sici)1521-3773(19990802)38:15<2096::aid-anie2096>3.0.co;2-f

No Time To Wait: Securing The Future From Drug-Resistant Infections. (2019). Geneva, Switzerland: World Health Organization Retrieved from https://www.who.int/docs/default-source/documents/no-time-to-wait-securing-the-future-from-drug-resistant-infections-en.pdf

Ohnishi, Y., Ishikawa, J., Hara, H., Suzuki, H., Ikenoya, M., Ikeda, H., Yamashita, A., Hattori, M., & Horinouchi, S. (2008). Genome sequence of the streptomycin-producing microorganism Streptomyces griseus IFO 13350. *J Bacteriol, 190*(11), 4050-4060. doi:10.1128/JB.00204-08

Omura, S., Tanaka, H., Tanaka, Y., Spiri-Nakagawa, P., Oiwa, R., Takahashi, Y., Matsuyama, K., & Iwai, Y. (1979). Studies on bacterial cell wall inhibitors. VII. Azureomycins A and B, new antibiotics produced by Pseudonocardia azurea nov. sp. Taxonomy of the producing organism, isolation, characterization and biological properties. *J Antibiot (Tokyo), 32*(10), 985-994.

Organization, W. H. (2015). *Global Action Plan on Antimicrobial Resistance*. Retrieved from Geneva, Switzerland:

Osterlund, T., Nookaew, I., Bordel, S., & Nielsen, J. (2013). Mapping condition-dependent regulation of metabolism in yeast through genome-scale modeling. *BMC Syst Biol, 7*, 36. doi:10.1186/1752-0509-7-36

Owen, J. G., Reddy, B. V., Ternei, M. A., Charlop-Powers, Z., Calle, P. Y., Kim, J. H., & Brady, S. F. (2013). Mapping gene clusters within arrayed metagenomic libraries to expand the structural diversity of biomedically relevant natural products. *Proc Natl Acad Sci U S A, 110*(29), 11797-11802. doi:10.1073/pnas.1222159110

Page, R. D., & Charleston, M. A. (1997). From gene to organismal phylogeny: reconciled trees and the gene tree/species tree problem. *Molecular phylogenetics and evolution, 7*(2), 231-240.

Parenti, F., & Cavalleri, B. (1989). Proposal to name the vancomycin-ristocetin like glycopeptides as dalbaheptides. *J Antibiot (Tokyo), 42*(12), 1882-1883. doi:10.7164/antibiotics.42.1882

Partridge, S. R., & Tsafnat, G. (2018). Automated annotation of mobile antibiotic resistance in Gram-negative bacteria: the Multiple Antibiotic Resistance Annotator (MARA) and database. *J Antimicrob Chemother, 73*(4), 883-890. doi:10.1093/jac/dkx513

Pelzer, S., Sussmuth, R., Heckmann, D., Recktenwald, J., Huber, P., Jung, G., & Wohlleben, W. (1999). Identification and analysis of the balhimycin biosynthetic gene cluster and its use for manipulating glycopeptide biosynthesis in Amycolatopsis mediterranei DSM5908. *Antimicrob Agents Chemother, 43*(7), 1565-1573.

Pootoolal, J., Thomas, M. G., Marshall, C. G., Neu, J. M., Hubbard, B. K., Walsh, C. T., & Wright, G. D. (2002). Assembling the glycopeptide antibiotic scaffold: The biosynthesis of A47934 from Streptomyces toyocaensis NRRL15009. *Proc Natl Acad Sci U S A, 99*(13), 8962-8967. doi:10.1073/pnas.102285099

Price, M. N., Dehal, P. S., & Arkin, A. P. (2010). FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS One, 5*(3), e9490. doi:10.1371/journal.pone.0009490

Radivojac, P., Clark, W. T., Oron, T. R., Schnoes, A. M., Wittkop, T., Sokolov, A., Graim, K., Funk, C., Verspoor, K., Ben-Hur, A., Pandey, G., Yunes, J. M., Talwalkar, A. S., Repo, S., Souza, M. L., Piovesan, D., Casadio, R., Wang, Z., Cheng, J., Fang, H., Gough, J., Koskinen, P., Toronen, P., Nokso-Koivisto, J., Holm, L., Cozzetto, D., Buchan, D. W., Bryson, K., Jones, D. T., Limaye, B., Inamdar, H., Datta, A., Manjari, S. K., Joshi, R., Chitale, M., Kihara, D., Lisewski, A. M., Erdin, S., Venner, E., Lichtarge, O., Rentzsch, R., Yang, H., Romero, A. E., Bhat, P., Paccanaro, A., Hamp, T., Kassner, R., Seemayer, S., Vicedo, E., Schaefer, C., Achten, D., Auer, F., Boehm, A., Braun, T., Hecht, M., Heron, M., Honigschmid, P., Hopf, T. A., Kaufmann, S., Kiening, M., Krompass, D., Landerer, C., Mahlich, Y., Roos, M., Bjorne, J., Salakoski, T., Wong, A., Shatkay, H., Gatzmann, F., Sommer, I., Wass, M. N., Sternberg, M. J., Skunca, N., Supek, F., Bosnjak, M., Panov, P., Dzeroski, S., Smuc, T., Kourmpetis, Y. A., van Dijk, A. D., ter Braak, C. J., Zhou, Y., Gong, Q., Dong, X., Tian, W., Falda, M., Fontana, P., Lavezzo, E., Di Camillo, B., Toppo, S., Lan, L., Djuric, N., Guo, Y., Vucetic, S., Bairoch, A., Linial, M., Babbitt, P. C., Brenner, S. E., Orengo, C., Rost, B., Mooney, S. D., & Friedberg, I. (2013). A large-scale evaluation of computational protein function prediction. *Nat Methods, 10*(3), 221-227. doi:10.1038/nmeth.2340

Rahman, S. F., Olm, M. R., Morowitz, M. J., & Banfield, J. F. (2018). Machine Learning Leveraging Genomes from Metagenomes Identifies Influential Antibiotic Resistance Genes in the Infant Gut Microbiome. *mSystems, 3*(1). doi:10.1128/mSystems.00123-17

Ramirez, M. S., & Tolmasky, M. E. (2010). Aminoglycoside modifying enzymes. *Drug Resist Updat, 13*(6), 151-171. doi:10.1016/j.drup.2010.08.003

Rausch, S., Hanchen, A., Denisiuk, A., Lohken, M., Schneider, T., & Sussmuth, R. D. (2011). Feglymycin is an inhibitor of the enzymes MurA and MurC of the peptidoglycan biosynthesis pathway. *Chembiochem, 12*(8), 1171-1173. doi:10.1002/cbic.201100120

Ravenhall, M., Skunca, N., Lassalle, F., & Dessimoz, C. (2015). Inferring horizontal gene transfer. *PLoS Comput Biol, 11*(5), e1004095. doi:10.1371/journal.pcbi.1004095

Richardson, E. J., & Watson, M. (2013). The automatic annotation of bacterial genomes. *Brief Bioinform, 14*(1), 1-12. doi:10.1093/bib/bbs007

Selengut, J. D., Haft, D. H., Davidsen, T., Ganapathy, A., Gwinn-Giglio, M., Nelson, W. C., Richter, A. R., & White, O. (2007). TIGRFAMs and Genome Properties: tools for the assignment of molecular function and biological process in prokaryotic genomes. *Nucleic Acids Res, 35*(Database issue), D260-264. doi:10.1093/nar/gkl1043

Skinnider, M. A., Merwin, N. J., Johnston, C. W., & Magarvey, N. A. (2017). PRISM 3: expanded prediction of natural product chemical structures from microbial genomes. *Nucleic Acids Res, 45*(W1), W49-W54. doi:10.1093/nar/gkx320

Smith, B., Ceusters, W., Klagges, B., Kohler, J., Kumar, A., Lomax, J., Mungall, C., Neuhaus, F., Rector, A. L., & Rosse, C. (2005). Relations in biomedical ontologies. *Genome Biol, 6*(5), R46. doi:10.1186/gb-2005-6-5-r46

Sneader, W. (2005). *Drug Discovery*: John Wiley & Sons Ltd.

Sosio, M., Kloosterman, H., Bianchi, A., de Vreugd, P., Dijkhuizen, L., & Donadio, S. (2004). Organization of the teicoplanin gene cluster in Actinoplanes teichomyceticus. *Microbiology, 150*(Pt 1), 95-102. doi:10.1099/mic.0.26507-0

Sosio, M., Stinchi, S., Beltrametti, F., Lazzarini, A., & Donadio, S. (2003). The Gene Cluster for the Biosynthesis of the Glycopeptide Antibiotic A40926 by Nonomuraea Species. *Chemistry & Biology, 10*(6), 541-549. doi:10.1016/s1074-5521(03)00120-0

Stachelhaus, T., Mootz, H. D., & Marahiel, M. A. (1999). The specificity-conferring code of adenylation domains in nonribosomal peptide synthetases. *Chem Biol, 6*(8), 493-505. doi:10.1016/S1074-5521(99)80082-9

Stegmann, E., Albersmeier, A., Spohn, M., Gert, H., Weber, T., Wohlleben, W., Kalinowski, J., & Ruckert, C. (2014). Complete genome sequence of the actinobacterium Amycolatopsis japonica MG417-CF17(T) (=DSM 44213T) producing (S,S)-N,N'-ethylenediaminedisuccinic acid. *J Biotechnol, 189*, 46-47. doi:10.1016/j.jbiotec.2014.08.034

Stegmann, E., Frasch, H. J., Kilian, R., & Pozzi, R. (2015). Self-resistance mechanisms of actinomycetes producing lipid II-targeting antibiotics. *Int J Med Microbiol, 305*(2), 190-195. doi:10.1016/j.ijmm.2014.12.015

Stevens, H. (2013). *Life Out of Sequence: A Data-Driven History of Bioinformatics*. Chicago/London: University of Chicago Press, .

Stinchi, S., Carrano, L., Lazzarini, A., Feroggio, M., Grigoletto, A., Sosio, M., & Donadio, S. (2006). A derivative of the glycopeptide A40926 produced by inactivation of the beta-hydroxylase gene in Nonomuraea sp. ATCC39727. *FEMS Microbiol Lett, 256*(2), 229-235. doi:10.1111/j.1574-6968.2006.00120.x

Stogios, P. J., Cox, G., Spanogiannopoulos, P., Pillon, M. C., Waglechner, N., Skarina, T., Koteva, K., Guarne, A., Savchenko, A., & Wright, G. D. (2016). Rifampin phosphotransferase is an unusual antibiotic resistance kinase. *Nat Commun, 7*, 11343. doi:10.1038/ncomms11343

Stolzer, M., Siewert, K., Lai, H., Xu, M., & Durand, D. (2015). Event inference in multidomain families with phylogenetic reconciliation. *BMC Bioinformatics, 16 Suppl 14*, S8. doi:10.1186/1471-2105-16-S14-S8

Sugimoto, A., Maeda, A., Itto, K., & Arimoto, H. (2017). Deciphering the mode of action of cell wall-inhibiting antibiotics using metabolic labeling of growing peptidoglycan in Streptococcus pyogenes. *Sci Rep, 7*(1), 1129. doi:10.1038/s41598-017-01267-5

Thai, Q. K., Bos, F., & Pleiss, J. (2009). The Lactamase Engineering Database: a critical survey of TEM sequences in public databases. *BMC Genomics, 10*, 390. doi:10.1186/1471-2164-10-390

Thaker, M. N., Wang, W., Spanogiannopoulos, P., Waglechner, N., King, A. M., Medina, R., & Wright, G. D. (2013). Identifying producers of antibacterial compounds by screening for antibiotic resistance. *Nat Biotechnol, 31*(10), 922-927. doi:10.1038/nbt.2685

Thibodeaux, C. J., Melancon, C. E., 3rd, & Liu, H. W. (2008). Natural-product sugar biosynthesis and enzymatic glycodiversification. *Angew Chem Int Ed Engl, 47*(51), 9814-9859. doi:10.1002/anie.200801204

To, T. H., Jacox, E., Ranwez, V., & Scornavacca, C. (2015). A fast method for calculating reliable event supports in tree reconciliations via Pareto optimality. *BMC Bioinformatics, 16*, 384. doi:10.1186/s12859-015-0803-x

Truman, A. W., Dias, M. V., Wu, S., Blundell, T. L., Huang, F., & Spencer, J. B. (2009). Chimeric glycosyltransferases for the generation of hybrid glycopeptides. *Chem Biol, 16*(6), 676-685. doi:10.1016/j.chembiol.2009.04.013

Truman, A. W., Kwun, M. J., Cheng, J., Yang, S. H., Suh, J. W., & Hong, H. J. (2014). Antibiotic resistance mechanisms inform discovery: identification and characterization of a novel amycolatopsis strain producing ristocetin. *Antimicrob Agents Chemother, 58*(10), 5687-5695. doi:10.1128/AAC.03349-14

Truman, A. W., Robinson, L., & Spencer, J. B. (2006). Identification of a deacetylase involved in the maturation of teicoplanin. *Chembiochem, 7*(11), 1670-1675. doi:10.1002/cbic.200600308

van Wageningen, A. M., Kirkpatrick, P. N., Williams, D. H., Harris, B. R., Kershaw, J. K., Lennard, N. J., Jones, M., Jones, S. J., & Solenberg, P. J. (1998). Sequencing and analysis of genes involved in the biosynthesis of a vancomycin group antibiotic. *Chem Biol, 5*(3), 155-162.

Waglechner, N., McArthur, A. G., & Wright, G. D. (2019). Phylogenetic reconciliation reveals the natural history of glycopeptide antibiotic biosynthesis and resistance. *Nat Microbiol*. doi:10.1038/s41564-019-0531-5

Waksman, S. A. (1947). What Is an Antibiotic or an Antibiotic Substance? *Mycologia, 39*(5), 565-569.

Weber, T., Blin, K., Duddela, S., Krug, D., Kim, H. U., Bruccoleri, R., Lee, S. Y., Fischbach, M. A., Muller, R., Wohlleben, W., Breitling, R., Takano, E., & Medema, M. H. (2015). antiSMASH 3.0-a comprehensive resource for the genome mining of biosynthetic gene clusters. *Nucleic Acids Res, 43*(W1), W237-243. doi:10.1093/nar/gkv437

Weber, T., & Kim, H. U. (2016). The secondary metabolite bioinformatics portal: Computational tools to facilitate synthetic biology of secondary metabolite production. *Synth Syst Biotechnol, 1*(2), 69-79. doi:10.1016/j.synbio.2015.12.002

Weber, T., Rausch, C., Lopez, P., Hoof, I., Gaykova, V., Huson, D. H., & Wohlleben, W. (2009). CLUSEAN: a computer-based framework for the automated analysis of bacterial secondary metabolite biosynthetic gene clusters. *J Biotechnol, 140*(1-2), 13-17. doi:10.1016/j.jbiotec.2009.01.007

Whelan, S., & Goldman, N. (2001). A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol Biol Evol, 18*(5), 691-699. doi:10.1093/oxfordjournals.molbev.a003851

Wright, G. D. (2010). The antibiotic resistome. *Expert Opin Drug Discov, 5*(8), 779-788. doi:10.1517/17460441.2010.497535

Wright, G. D. (2014). Something old, something new: revisiting natural products in antibiotic drug discovery. *Can J Microbiol, 60*(3), 147-154. doi:10.1139/cjm-2014-0063

Wright, L. F., & Hopwood, D. A. (1976). Actinorhodin is a chromosomally-determined antibiotic in Streptomyces coelicolar A3(2). *J Gen Microbiol, 96*(2), 289-297. doi:10.1099/00221287-96-2-289

Wu, M. C., Styles, M. Q., Law, B. J., Struck, A. W., Nunns, L., & Micklefield, J. (2015). Engineered biosynthesis of enduracidin lipoglycopeptide antibiotics using the ramoplanin mannosyltransferase Ram29. *Microbiology, 161*(7), 1338-1347. doi:10.1099/mic.0.000095

Xavier, B. B., Das, A. J., Cochrane, G., De Ganck, S., Kumar-Singh, S., Aarestrup, F. M., Goossens, H., & Malhotra-Kumar, S. (2016). Consolidating and Exploring Antibiotic Resistance Gene Data Resources. *J Clin Microbiol, 54*(4), 851-859. doi:10.1128/JCM.02717-15

Xu, L., Huang, H., Wei, W., Zhong, Y., Tang, B., Yuan, H., Zhu, L., Huang, W., Ge, M., Yang, S., Zheng, H., Jiang, W., Chen, D., Zhao, G. P., & Zhao, W. (2014). Complete genome sequence and comparative genomic analyses of the vancomycin-producing Amycolatopsis orientalis. *BMC Genomics, 15*, 363. doi:10.1186/1471-2164-15-363

Yim, G., Kalan, L., Koteva, K., Thaker, M. N., Waglechner, N., Tang, I., & Wright, G. D. (2014). Harnessing the synthetic capabilities of glycopeptide antibiotic tailoring enzymes: characterization of the UK-68,597 biosynthetic cluster. *Chembiochem, 15*(17), 2613-2623. doi:10.1002/cbic.201402179

Yim, G., Thaker, M. N., Koteva, K., & Wright, G. (2014). Glycopeptide antibiotic biosynthesis. *J Antibiot (Tokyo), 67*(1), 31-41. doi:10.1038/ja.2013.117

Yim, G., Wang, H. H., & Davies, J. (2006). The truth about antibiotics. *Int J Med Microbiol, 296*(2-3), 163-170. doi:10.1016/j.ijmm.2006.01.039

Yim, G., Wang, W., Thaker, M. N., Tan, S., & Wright, G. D. (2016). How To Make a Glycopeptide: A Synthetic Biology Approach To Expand Antibiotic Chemical Diversity. *ACS Infect Dis, 2*(9), 642-650. doi:10.1021/acsinfecdis.6b00105

Yin, X., Jiang, X. T., Chai, B., Li, L., Yang, Y., Cole, J. R., Tiedje, J. M., & Zhang, T. (2018). ARGs-OAP v2.0 with an expanded SARG database and Hidden Markov Models for enhancement characterization and quantification of antibiotic resistance genes in environmental metagenomes. *Bioinformatics, 34*(13), 2263-2270. doi:10.1093/bioinformatics/bty053

Yin, X., & Zabriskie, T. M. (2006). The enduracidin biosynthetic gene cluster from Streptomyces fungicidicus. *Microbiology, 152*(Pt 10), 2969-2983. doi:10.1099/mic.0.29043-0

Zankari, E., Hasman, H., Cosentino, S., Vestergaard, M., Rasmussen, S., Lund, O., Aarestrup, F. M., & Larsen, M. V. (2012). Identification of acquired antimicrobial resistance genes. *J Antimicrob Chemother, 67*(11), 2640-2644. doi:10.1093/jac/dks261

Ziemert, N., Lechner, A., Wietz, M., Millan-Aguinaga, N., Chavarria, K. L., & Jensen, P. R. (2014). Diversity and evolution of secondary metabolism in the marine actinomycete genus Salinispora. *Proc Natl Acad Sci U S A, 111*(12), E1130-1139. doi:10.1073/pnas.1324161111

Ziemert, N., Weber, T., & Medema, M. H. (2019). Genome Mining Approaches to Bacterial Natural Product Discovery. *Reference Module in Chemistry, Molecular Sciences and Chemical Engineering*. doi:10.1016/B978-0-12-409547-2.14627-X

APPENDICES

Appendix 1: Culp, E. and Yim, G. *et al.* (2019) Hidden antibiotics in actinomycetes can be identified by inactivation of gene clusters for common antibiotics. *Nat Biotechnol* 37(10):1149-54.

nature biotechnology

LETTERS

https://doi.org/10.1038/s41587-019-0241-9

Hidden antibiotics in actinomycetes can be identified by inactivation of gene clusters for common antibiotics

Elizabeth J. Culp[1,2], Grace Yim[1,2], Nicholas Waglechner[1], Wenliang Wang[1], Andrew C. Pawlowski[1] and Gerard D. Wright[1]*

Actinobacteria, which are one of the largest bacterial phyla and comprise between 13 and 30% of the soil microbiota, are the main source of antibiotic classes in clinical use[1]. During screens for antimicrobials, as many as 50% of actinomycete strains are discarded because they produce a known antibiotic (Supplementary Fig. 1) (ref. [2]). Despite each strain likely having the capacity to produce many compounds, strains are abandoned because the already characterized antibiotic could interfere with screening for, or purification of, newly discovered compounds[3]. We applied CRISPR-Cas9 genome engineering to knockout genes encoding two of the most frequently rediscovered antibiotics, streptothricin or streptomycin, in 11 actinomycete strains. We report that this simple approach led to production of different antibiotics that were otherwise masked. We were able to rapidly discover rare and previously unknown variants of antibiotics including thiolactomycin, amicetin, phenanthroviridin and 5-chloro-3-formylindole. This strategy could be applied to existing strain collections to realize their biosynthetic potential.

Antibiotics have been used to treat bacterial infections for the past 70 years but the emergence of multidrug-resistant pathogenic bacteria for which there are few, and in some cases no treatment options, is an urgent and pressing global problem[4]. No new classes of antibiotics have come through the clinical pipeline in the past 30 years[4] and refilling this pipeline is crucial if we are to address the challenge of multidrug-resistant pathogens.

Actinomycetes are a rich source of antibiotics with each genome containing 20–40 distinct biosynthetic gene clusters (BGCs) encoding specialized metabolites, only a minority of which have been chemically explored. Despite their potential, the canonical Waksman platform for antibiotic discovery[5], where actinobacterial extracts are screened against susceptible organisms for antimicrobial activity, has largely failed to yield novel drug scaffolds[3]. Chief among the platform's drawbacks is the rediscovery of known compounds and the need to identify these in complex mixtures, a process known as dereplication. Baltz has estimated that one antibiotic (streptothricin) is found in one in ten isolates and a handful of others at frequencies ~10^{-2} to 10^{-3} (streptomycin, tetracycline and actinomycin D)[6]. Using the Waksman platform, libraries of tens of millions of isolates must be screened to find new antibiotics. To ease the consequent discovery burden, versatile and efficient platforms have been developed that can dereplicate common antibiotics[2,7]. However, dereplication alone does not address the fact that strain collections are polluted with isolates producing common antibiotics. Triaging these strains on the basis of the production of a known molecule may result in the collateral loss of new or rare compounds that are also encoded by the same strains but whose expression is masked by these common antibiotics.

To exploit actinomycetes to their full biosynthetic potential, a toolbox of methods has been developed. These include activation of a specific BGC of interest by promoter refactoring or manipulation of pathway specific transcriptional regulators[8]. These targeted approaches require genetic manipulation and specialized constructs for each BGC of interest. Alternatively, methods for untargeted BGC activation include introduction of pleiotropic activators such as AfsQ1 (ref. [9]), chemical elicitors[10] or transcription factor decoys[11]. However, these methods still require detection of any compound of interest alongside the production of large quantities of other active metabolites.

We hypothesized that it might be possible to re-investigate discarded strains producing common antibiotics by disrupting the production of those common antibiotics, thereby providing a straightforward, untargeted strategy to re-screen the 'tailings' of historic strain collections (Fig. 1a). We reasoned that disruption of conserved biosynthetic genes of the main antibiotic produced by a strain might facilitate detection of metabolites whose activity was otherwise overlooked. It is feasible that disrupting this production could also alter BGC regulatory circuits or liberate precursors, thereby enabling the increased production of metabolites produced in low levels (or not at all) in wild-type strains, as has been observed previously[12]. We report a generalizable CRISPR-Cas9-based system capable of inactivating production of antibiotics of interest by disrupting key common biosynthetic genes.

First, we devised a pipeline to identify highly conserved single guide (sg)RNA target sequences in a given BGC, enabling application of the same CRISPR-based targeted construct in multiple strains without prior knowledge of specific BGC sequences. Targeting the most commonly produced antibiotic, streptothricin (Supplementary Fig. 1), we mined 29 streptothricin BGCs from GenBank and our in-house genome sequences (Supplementary Table 1) and identified conserved target sites using a custom Python script (Supplementary Fig. 1); sgRNA sites found in all 29 clusters were further refined (see Methods), leaving us with two targets in the streptothricin BGC, *orf15* and *orf17* (Supplementary Fig. 2 and Supplementary Table 2) (ref. [13]).

At the time of this study, two CRISPR-Cas9 systems for actinomycetes had been published: pCRISPomyces-2 (ref. [14]) and pCRISPR-Cas9 (ref. [15]) (Fig. 1a and Supplementary Fig. 1). The

[1]David Braley Center for Antibiotic Discovery, M.G. DeGroote Institute for Infectious Disease Research, Department of Biochemistry and Biomedical Sciences, McMaster University, Hamilton, Ontario, Canada. [2]These authors contributed equally: Elizabeth Culp, Grace Yim. *e-mail: wrightge@mcmaster.ca

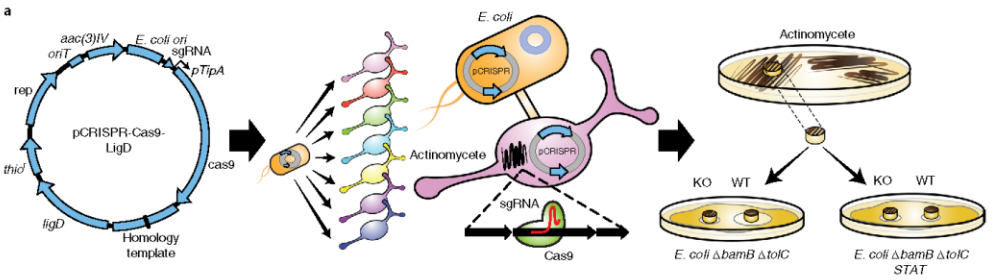


Fig. 1 | Application of CRISPR-Cas9 to inactivate streptothricin and streptomycin production. a, Summary of the CRISPR-Cas9 strategy for the inactivation of commonly found antibiotics. First, the pCRISPR-Cas9-LigD plasmid, containing both *ligD* and a homology repair template, was engineered to target a highly conserved protospacer sequence in a BGC of interest. Next, the construct was delivered, via conjugation, to a wide range of actinomycete strains allowing for efficient genome editing. Successful gene inactivations resulting in a loss of antibiotic production were identified by performing agar plug assays against susceptible and resistant *E. coli* Δ*bamB* Δ*tolC*. *oriT*, origin of transfer; *thio'*, thiostrepton resistance; rep, pSG5 replication origin; *aac(3)IV*, apramycin resistance; *pTipA*, thiostrepton inducible promoter; KO, knockout; WT, wild type. **b**, Comparative bioassays are shown for representative streptothricin producers, WAC7149 and WAC5950 and streptomycin producer, WAC4770. Two independent fermentations and bioassays gave similar results. **c**, Summary of editing efficiencies on the basis of bioassay results using the pCRISPR-Cas9 system. 'X' indicates that conjugations were unsuccessful.

systems are similar but differ in promoters and the DNA repair pathways employed. While both carry a repair template for homologous recombination (HR), we modified the pCRISPR-Cas9 plasmid to promote nonhomologous end-joining (NHEJ) as well. We tested both systems on eight streptothricin producers identified by our Antibiotic Resistance Platform[2]. Gene inactivation was confirmed by bioassay (Fig. 1b and Supplementary Fig. 3), liquid chromatography mass spectrometry (LC–MS) (Supplementary Fig. 4) and Sanger sequencing across the predicted lesion site (Supplementary Fig. 3). We disrupted streptothricin in five of the eight strains tested at efficiencies ranging from 8–75% (Supplementary Table 3 and Fig. 1c). In those strains that failed, it was largely owing to an inability to obtain exconjugants. We also targeted streptomycin, another highly abundant antibiotic, for inactivation, targeting three biosynthetic genes: *strF*, *strH* and *strI* (Supplementary Fig. 2 and Supplementary Table 2) (refs. [2,6]). Using the pCRISPR-Cas9 system, inactivation of streptomycin biosynthesis was verified in all five strains (Fig. 1, Supplementary Fig. 3, Supplementary Fig. 4 and Supplementary Table 3). From start to finish, 25–30 d were required to generate, cure and verify engineered strains, with only 6 d involving hands-on work (Supplementary Fig. 1).

As part of our verification pipeline, Sanger sequencing was performed on at least one exconjugant for every gene/strain inactivation generated (Table 1 and Supplementary Fig. 3). In roughly one-third of the strains verified (9/25 strains generated), we found the introduction of a stop codon as expected from HR with the repair template. Perhaps owing to a lower nucleotide identity in the homology arms compared to other strains, one strain had a mutation characteristic of NHEJ (Table 1). We verified the absence of off-target effects in other genomes predicted to be repaired by HR through whole-genome sequencing (WAC5374 Δ*strF*). In the remaining strains (15/25), we were unable to amplify the edited region by PCR (Supplementary Fig. 3). Whole-genome sequencing of three of these strains revealed large deletions, including the entire streptothricin or streptomycin BGC (Table 1). In WAC5374 Δ*strH*, despite repair through HR at the CRISPR-targeted locus, off-target activity led to a ~600-kb deletion elsewhere in the genome.

We observed many phenotypic differences between wild-type/engineered strain pairs including growth rate, sporulation and pigment production (Fig. 2a and Supplementary Fig. 5) suggesting a rewiring of the transcriptional and/or precursor supply pathways regulating secondary metabolism in actinomycetes[16,17]. Metabolic profiles were further investigated in five wild-type strains and their respective engineered strains by principle component analysis (PCA) of high-resolution LC–MS data to assess gross metabolomic differences (Fig. 2b and Supplementary Fig. 5). We observed effective clustering of replicates of a single strain and clear divisions

Table 1 | CRISPR-engineered mutants generated and their uncovered metabolites

WAC strain	Gene targeted	CRISPR system	Mutation repair	Uncovered metabolites (fold production increase versus WT)
Streptomycin				
4770	*strI*	pCRISPR	Undefined deletion	
	strF	pCRISPR	HR	
	strH	pCRISPR	HR	
5467	*strF*	pCRISPR	HR	
	strH	pCRISPR	HR	
8181	*strH*	pCRISPR	Undefined deletion	
5374	*strI*	pCRISPR	Deletion (~6.6 kb)	Thiolactomycin (NS)
	strF	pCRISPR	HR	Thiolactomycin (NS)
	strH	pCRISPR	HR and deletion (~600 kb)	5-Chloro-3-formylindole (NS)
8241	*strI*	pCRISPR	Deletion (~500 kb)	Phenanthroviridin aglycone (7×)
Streptothricin				
7149	*orf15*	pCRISPR	Undefined deletion	
	orf17	pCRISPR	Undefined deletion	
	orf17	pCRISPomyces	Undefined deletion	
8452	*orf15*	pCRISPR	Undefined deletion	
	orf17	pCRISPR	NHEJ, 1-bp insertion	
6273	*orf15*	pCRISPR	Deletion (~200 kb)	Amicetin (3–20×), ferrioxamines (9–68×)
	orf15	pCRISPomyces	Undefined deletion	Amicetin (4–22×), ferrioxamines (11–62×)
	orf17	pCRISPR	Undefined deletion	Amicetin (4–21×), ferrioxamines (7–50×)
6776	*orf15*	pCRISPR	HR	
	orf15	pCRISPomyces	Undefined deletion	
	orf17	pCRISPR	Undefined deletion	
	orf17	pCRISPomyces	HR	
6128	*orf17*	pCRISPomyces	Undefined deletion	
5950	*orf15*	pCRISPomyces	Deletion (~1.5 Mb)	
	orf17	pCRISPomyces	HR	

NS, not significant.

between the wild type and each engineered strain, reflecting altered metabolic profiles. Not surprisingly, greater shifts in metabolic profile were associated with large genomic deletions. Unexpectedly, targeting the same gene in the same strain also resulted in different metabolite profiles as was observed for WAC6273 Δ*orf15* strains generated by the pCRISPomyces or pCRISPR-Cas9 system (Fig. 2b). We detected large deletions in genomes of both of these engineered strains, which may have led to this distinction (Table 1).

Consistent with PCA groupings, we identified metabolites unique or upregulated in engineered strains (Supplementary Fig. 5). To characterize these upregulated molecules, we used the Global Natural Product Social Molecular Networking (GNPS) platform[18]. In all three WAC6273-inactivated strains, we identified ferrioxamines, a suite of siderophores not generally observed in *Streptomyces* under typical culture conditions[19], as a family of highly upregulated metabolites (Fig. 2c and Supplementary Fig. 6) (ref. [19]). Altered metabolism in engineered strains therefore enables compounds not detectable in wild-type backgrounds to be readily identified.

To determine the potential chemical diversity available in streptothricin producers, we performed a phylogenetic analysis to examine the distribution of BGCs present in known streptothricin producers. Whole-genome sequences of 42 streptothricin producers, pulled from in-house strains and GenBank (Supplementary Table 4), were evaluated for their biosynthetic potential (see Methods) and shown to have the capacity to produce a wide variety of specialized metabolites that rarely overlapped in identity (Supplementary Fig. 7). When compared by multilocus sequence analysis to 145 *Streptomyces* strains selected to represent diversity across the genus[20], streptothricin producers were spread across the phylogenetic tree (Supplementary Figs. 7 and 8) These findings further encouraged us to explore the bioactive compounds from our inactivated strains.

From our collection of CRISPR-inactivated streptothricin and streptomycin producers, we selected a set of eight strains that retained antimicrobial activity against our tester strain *Escherichia coli* BW25113 Δ*bamB* Δ*tolC* (data not shown). We performed fermentations under a variety of conditions and observed antimicrobial activity in all strains (Supplementary Fig. 9), suggesting that low-abundance antibiotics were masked by streptothricin and streptomycin production in wild-type strains as these wild types had shown no activity against streptothricin- and streptomycin-resistant tester strains (Fig. 1 and Supplementary Fig. 3). We chose three strains with substantial activity against our tester strain for bioactivity-guided purification: one streptothricin-inactivated producer, WAC6273, and two streptomycin-inactivated producers, WAC5374 and WAC8241.

Activity-guided purification from WAC6273 Δ*orf17* resulted in the isolation of the rare antibiotic amicetin and associated

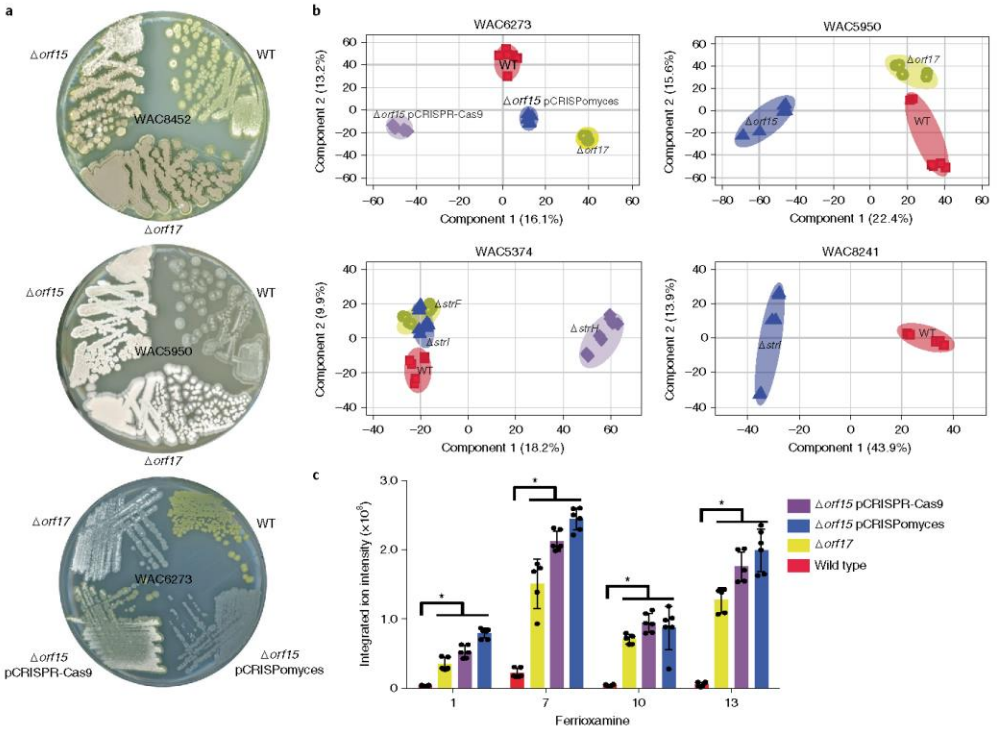


Fig. 2 | Inactivating antibiotic biosynthesis shifts the metabolic profile of producer bacteria. Metabolic flux of wild-type and biosynthetically inactivated streptomycin or streptothricin strains were assessed globally by phenotypic analysis (**a**), high-resolution LC–MS metabolomics (**b**) and on the production of select metabolites (**c**). Data are shown for representative streptothricin producers (WAC8452, WAC5950, WAC6273), and streptomycin producers (WAC5374, WAC8241). **a**, Phenotypic comparisons were made after 7-d growth on Bennett's media. Growth was repeated on three independent occasions showing similar phenotypes. **b**, LC–MS data from *n*-butanol extracts were used to construct PCA plots, revealing effective clustering of different engineered strains from their wild-type parent ($n = 6$, three independent fermentations analyzed in technical duplicate). **c**, High-resolution LC–MS analysis of WAC6273 extracts identifies a family of upregulated metabolites in engineered strains as ferrioxamines, including ferrioxamine 1, 7, 10 and 13 (Supplementary Fig. 6). Mean with error bars showing s.d. is plotted ($n = 6$, three independent fermentations analyzed in technical duplicate). Significance is tested to $P < 0.0001$ by one-way analysis of variance (ANOVA) with Tukey's post hoc analysis.

derivatives, a family of compounds that inhibit translation elongation (Fig. 3a and Supplementary Table 5) (ref. [21]). Genome analysis identified the amicetin BGC in WAC6273 (Supplementary Fig. 10). The GNPS workflow revealed many amicetin derivatives in fermentation broths, several of which were novel derivatives with modifications that could be predicted by LC–MS/MS (Supplementary Fig. 11). LC–MS quantification showed that our CRISPR-engineered strains produced higher levels of the amicetin family compounds compared to the wild-type producer WAC6273 (Fig. 3b and Table 1). PCR with reverse transcription (RT–PCR) on the amicetin BGC indicated similar transcription levels in wild-type and engineered strains, suggesting that redirection of metabolic flux may be the primary driver of improved yield (Fig. 3b).

Activity-guided purification from two different WAC5374 engineered strains, Δ*strI* and Δ*strH*, identified two different antibacterial compounds, thiolactomycin and 5-chloro-3-formylindole, respectively (Fig. 3c,d and Table 1). Thiolactomycin is a rare fatty acid synthesis inhibitor[22], while the antibiotic activity of 5-chloro-3-formylindole has not been previously described (Supplementary Table 5). The isolation of two different compounds from two mutants of the same parent strain is consistent with the distinct metabolic PCA profiles of Δ*strH* when compared to wild-type and other mutant strains (Fig. 2b). Thiolactomycin was found in crude fermentation extracts from the wild type, Δ*strF* and Δ*strI* mutants but not WAC5374 Δ*strH* (Fig. 3f), consistent with deletion of the thiolactomycin BGC in WAC5374 Δ*strH*. Genome analysis resulted in the identification of the thiolactomycin BGC in wild type WAC5374 (Supplementary Fig. 12) (ref. [23]). The predicted BGC for 5-chloro-3-formylindole was identified by searching the WAC5374 genome for a tryptophan halogenase and confirmed by deletion of this halogenase (Fig. 3g and Supplementary Fig. 13).

Activity-guided purification from WAC8241 Δ*strI* resulted in the purification of phenanthroviridin aglycone, a low-abundance derivative of the antibiotic, jadomycin (Fig. 3e) (ref. [24]) and identification of its BGC (Supplementary Fig. 14). By PCA, the metabolic profile of the Δ*strI* mutant was also distinct from the wild type (Fig. 2b), and correspondingly the mutant produced higher levels of phenanthroviridin than the wild type (Fig. 3h). Levels of

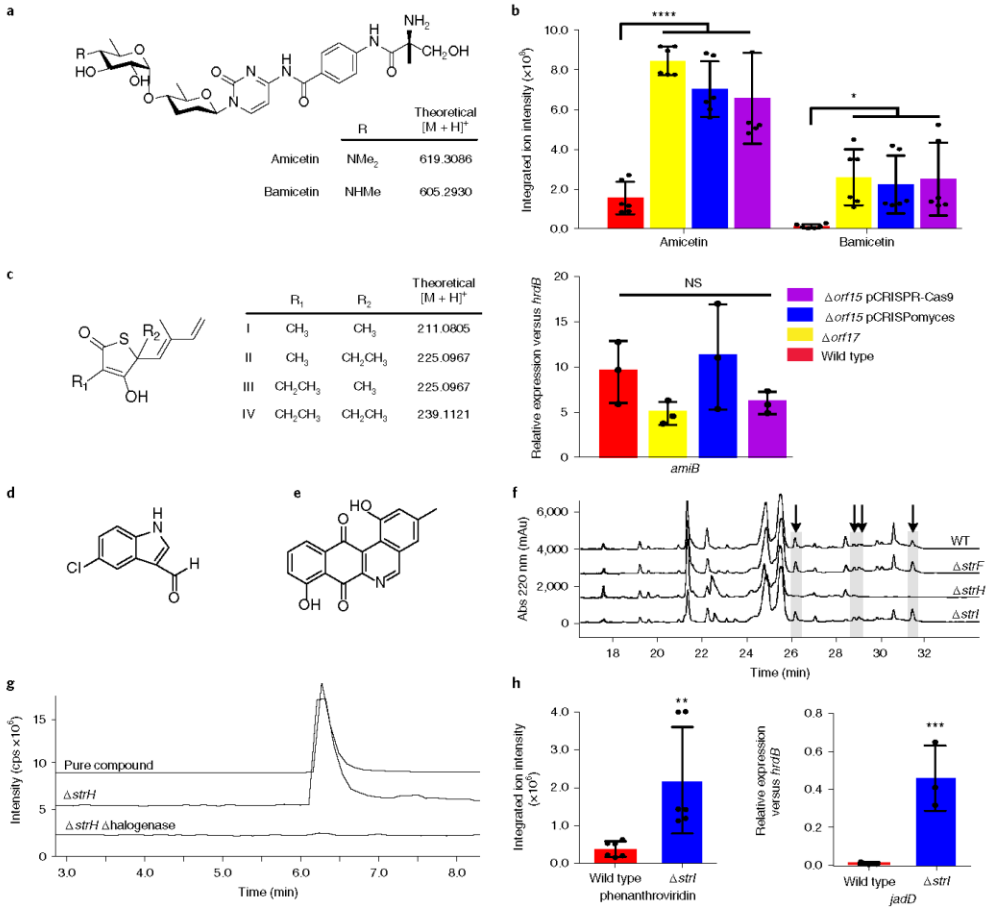


Fig. 3 | New antibiotic compounds discovered from CRISPR-inactivated strains. a, Chemical structures of amicetin and bamicetin. **b**, Relative antibiotic production or expression of a key gene in amicetin and bamicetin BGCs, *amiB*, in WAC6273 wild-type and engineered strains. WAC6273 Δ*orf15*-targeted mutants generated by the pCRISPomyces and pCRISPR-Cas9 systems were both analyzed. **c**–**e**, Chemical structure of thiolactomycins I–IV (**c**), 5-chloro-3-formylindole (**d**) and phenanthroviridin aglycone (**e**). **f**, High performance liquid chromatography analysis of WAC5374 crude extracts identified thiolactomycins I–IV, indicated by black arrows, produced in all strains except WAC5374 Δ*strH*. **g**, Relative amounts of 5-chloro-3-formylindole present in crude extracts as determined by extracted ion chromatograms (negative mode *m/z* 179.23–179.73). The WAC5374 Δ*strH* mutant was compared to derivatives with the tryptophan halogenase deleted. cps, counts per second. **h**, Relative amounts of phenanthroviridin aglycone and expression of a key gene in its BGC, *jadD*, in WAC8241 wild-type and engineered strains. **b**,**g**,**h**, For LC–MS production quantification, mean with error bars showing s.d. is plotted ($n = 6$, three independent fermentations analyzed in technical duplicate). For gene expression, mean with error bars showing s.d. is plotted ($n = 3$, three independent fermentations). Multiple comparison significance was tested to $^{****}P < 0.0001$, $^{**}P < 0.0005$ or $^{*}P < 0.05$ by one-way ANOVA with Tukey's post hoc analysis (**b**), or pairwise to $^{**}P = 0.011$ or $^{***}P = 0.0103$ by an unpaired two-sided Student's *t*-test (**f**, **g**). NS, not significant ($P = 0.18$). All experiments (**b**,**f**–**h**) were repeated on two independent occasions with similar results.

transcription of the identified BGC were similarly increased, indicating that transcriptional rewiring was at least partially responsible for improved yields in the engineered strain (Fig. 3h). Transcriptional rewiring in engineered strains may be the result of accumulation of biosynthetic intermediates that can interact with transcription factors in feed-forward mechanisms to activate expression of later-stage biosynthetic enzymes[12] or cross-talk with other biosynthetic pathways[25]. Together with the liberation of precursors, these mechanisms result in shifts in metabolic profile to allow for the identification of compounds in engineered strains not detectable in the wild type.

To assess the ability of our platform to discover rare, novel antibacterials from actinomycetes, we used BLAST to identify the number of BGCs of our discovered compounds in GenBank.

Three unique biosynthetic genes were identified for each BGC and genomes that contained all three were taken to contain the cluster. At the time of this study, only five and four amicetin and thiolactomycin producers, respectively, were identified in GenBank, much smaller than the 42 streptothricin producers found. While this proxy for compound rarity is heavily skewed by sampling bias, it does indicate that uncommon antibiotics can be found in streptothricin- and streptomycin-inactivated strains.

Our platform is widely applicable to most strains, and despite a few failures to induce mutations, we were successful in choosing conserved protospacers and inactivating streptothricin or streptomycin in 11/14 of our selected strains. While a drawback of all genetic platforms for BGC activation is the requirement to manipulate potentially intractable environmental isolates, accessing even only the easily engineered strains amounts to a substantial proportion of collections, given the number of streptothricin/streptomycin producers they contain. Furthermore, the platform could be expanded to access difficult strains using technologies to manipulate nonlaboratory or undomesticated strains[26,27].

When strains producing common compounds are triaged after initial screens, all other bioactive compounds potentially produced by these strains are simultaneously discarded. We report that CRISPR-Cas9 targeted inactivation of commonly found BGCs allows the mining of a greater proportion of strain collections for new antibiotics. The concept is not limited to antibiotic biosynthesis and could be applied to target other specialized metabolites and bioactivities, such as antibiotic adjuvants or anticancers. Our simple strategy to inactivate ubiquitous BGCs could allow researchers to tap into rare BGCs that are difficult to access by conventional screening of strain collections. Continuing to develop alternative approaches, such as these, will propel antibiotic discovery forward in the genomics age.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of code and data availability and associated accession codes are available at https://doi.org/10.1038/s41587-019-0241-9.

Received: 30 November 2018; Accepted: 25 July 2019;
Published online: 9 September 2019

References

1. Barka, E. A. et al. Taxonomy, physiology, and natural products of actinobacteria. *Microbiol. Mol. Biol. Rev.* **80**, 1–43 (2016).
2. Cox, G. et al. A common platform for antibiotic dereplication and adjuvant discovery. *Cell Chem. Biol.* **24**, 98–109 (2017).
3. Wright, G. D. Something old, something new: revisiting natural products in antibiotic drug discovery. *Can. J. Microbiol.* **60**, 147–154 (2014).
4. Wright, G. D. Solving the antibiotic crisis. *ACS Infect. Dis.* **1**, 80–84 (2015).
5. Lewis, K. Antibiotics: Recover the lost art of drug discovery. *Nature* **485**, 439–440 (2012).
6. Baltz, R. H. Marcel Faber Roundtable: Is our antibiotic pipeline unproductive because of starvation, constipation or lack of inspiration? *J. Ind. Microbiol. Biotechnol.* **33**, 507–513 (2006).
7. Allard, P.-M. et al. Integration of molecular networking and in-silico MS/MS fragmentation for natural products dereplication. *Anal. Chem.* **88**, 3317–3323 (2016).
8. Rutledge, P. J. & Challis, G. L. Discovery of microbial natural products by activation of silent biosynthetic gene clusters. *Nat. Rev. Microbiol.* **13**, 509–523 (2015).
9. Daniel-Ivad, M. et al. An engineered allele of *afsQ1* facilitates the discovery and investigation of cryptic natural products. *ACS Chem. Biol.* **12**, 628–634 (2017).
10. Mao, D., Okada, B. K., Wu, Y., Xu, F. & Seyedsayamdost, M. R. Recent advances in activating silent biosynthetic gene clusters in bacteria. *Curr. Opin. Microbiol.* **45**, 156–163 (2018).
11. Wang, B., Guo, F., Dong, S.-H. & Zhao, H. Activation of silent biosynthetic gene clusters using transcription factor decoys. *Nat. Chem. Biol.* **15**, 111–114 (2019).
12. Zhang, Y. et al. JadR*-mediated feed-forward regulation of cofactor supply in jadomycin biosynthesis. *Mol. Microbiol.* **90**, 884–897 (2013).
13. Maruyama, C. et al. A stand-alone adenylation domain forms amide bonds in streptothricin biosynthesis. *Nat. Chem. Biol.* **8**, 791–797 (2012).
14. Cobb, R. E., Wang, Y. & Zhao, H. High-efficiency multiplex genome editing of *Streptomyces* species using an engineered CRISPR/Cas system. *ACS Synth. Biol.* **4**, 723–728 (2015).
15. Tong, Y., Charusanti, P., Zhang, L., Weber, T. & Lee, S. Y. CRISPR-Cas9 based engineering of actinomycetal genomes. *ACS Synth. Biol.* **4**, 1020–1029 (2015).
16. Craney, A., Ozimok, C., Pimentel-Elardo, S. M., Capretta, A. & Nodwell, J. R. Chemical perturbation of secondary metabolism demonstrates important links to primary metabolism. *Chem. Biol.* **19**, 1020–1027 (2012).
17. Thykaer, J. et al. Increased glycopeptide production after overexpression of shikimate pathway genes being part of the balhimycin biosynthetic gene cluster. *Metab. Eng.* **12**, 455–461 (2010).
18. Wang, M. et al. Sharing and community curation of mass spectrometry data with Global natural products social molecular networking. *Nat. Biotechnol.* **34**, 828–837 (2016).
19. Sidebottom, A. M., Johnson, A. R., Karty, J. A., Trader, D. J. & Carlson, E. E. Integrated metabolomics approach facilitates discovery of an unpredicted natural product suite from *Streptomyces coelicolor* M145. *ACS Chem. Biol.* **8**, 2009–2016 (2013).
20. McDonald, B. R. & Currie, C. R. Lateral gene transfer dynamics in the ancient bacterial genus *Streptomyces*. *MBio* **8**, e00644–17 (2017).
21. Zhang, G. et al. Characterization of the amicetin biosynthesis gene cluster from *Streptomyces vinaceusdrappus* NRRL 2363 implicates two alternative strategies for amide bond formation. *Appl. Environ. Microbiol.* **78**, 2393–2401 (2012).
22. Slayden, R. A. et al. Antimycobacterial action of thiolactomycin: an inhibitor of fatty acid and mycolic acid synthesis. *Antimicrob. Agents Chemother.* **40**, 2813–2819 (1996).
23. Tang, X. et al. Identification of thiotetronic acid antibiotic biosynthetic pathways by target-directed genome mining. *ACS Chem. Biol.* **10**, 2841–2849 (2015).
24. Prendrich, G. et al. Phenanthridine derivatives, process for the preparation thereof, and compositions containing them. European patent EP0304400B1 (1990).
25. de la Fuente, A., Lorenzana, L. M., Martin, J. F. & Liras, P. Mutants of *Streptomyces clavuligerus* with disruptions in different genes for clavulanic acid biosynthesis produce large amounts of holomycin: possible cross-regulation of two unrelated secondary metabolic pathways. *J. Bacteriol.* **184**, 6559–6565 (2002).
26. Li, L. et al. CRISPR-Cpf1-assisted multiplex genome editing and transcriptional repression in *Streptomyces*. *Appl. Environ. Microbiol.* **84**, e00827–18 (2018).
27. Brophy, J. A. N. et al. Engineered integrative and conjugative elements for efficient and inducible DNA transfer to undomesticated bacteria. *Nat. Microbiol.* **3**, 1043–1053 (2018).

Acknowledgements

Computational resources for genome assembly and analysis were provided by A.G. McArthur at McMaster University. This research was funded by a Canadian Institutes of Health Research grant (no. MT-14981), the Ontario Research Fund and by a Canada Research Chair (to G.D.W.). E.C. was supported by a Canadian Institutes of Health Research Vanier Canada Graduate Scholarship. G.Y. was supported by an M.G. DeGroote Fellowship Award and a CIHR postdoctoral fellowship. N.W. was supported by a Canadian Institutes of Health Research Canada Graduate Scholarship Doctoral Award. We thank C. Groves for graphical edits to figures.

Author contributions

E.C., G.Y. and G.D.W. conceived the study and designed the experiments. W.W. performed bioactivity-guided purification. N.W. assembled whole-genome sequences and performed phylogenetic and BGC content analysis. A.C.P. designed the computer script for sgRNA identification. E.C. and G.Y. performed all other experiments. E.C., G.Y. and G.D.W wrote the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41587-019-0241-9.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to G.D.W.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

Strains and culture conditions. *E. coli* strains and plasmids are listed in Supplementary Table 6. All *Streptomyces* strains are listed in Supplementary Table 3. *E. coli* was grown at 37 °C, 250 r.p.m., in LB Broth (Bioshop, Canada). *Streptomyces* strains were grown at 30 °C, 250 r.p.m., in Tryptone–Soya–Broth (TSB, BD Biosciences) for starter cultures and genomic DNA preparation or SM–$MgCl_2$ (2% each D-mannitol, soya flour, agar, 10 mM $MgCl_2$) for sporulation and conjugation. Fermentations were performed in Bennett's media (1% potato starch, 0.2% casamino acids, 0.18% yeast extract, 0.02% KCl, 0.02% $MgSO_4{\cdot}7H_2O$, 0.024% $NaNO_3$, 4×10^{-4}% $FeSO_4{\cdot}7H_2O$) or SMP media (2.5% soluble starch, 0.146% glutamine, 0.1% K_2HPO_4, 0.1% NaCl, 0.05% $MgSO_4{\cdot}7H_2O$, 4×10^{-6}% $ZnCl_2$, 2×10^{-5}% $FeCl_3{\cdot}6H_2O$, 1×10^{-6}% each $CuCl_2{\cdot}2H_2O$, $MnCl_2{\cdot}4H_2O$, $Na_2B_4O_7{\cdot}10H_2O$, $(NH_4)_6Mo_7O_{24}{\cdot}4H_2O$), as indicated. Antibiotics were supplemented as necessary (100 μg ml^{-1} ampicillin, 50 μg/ml^{-1} apramycin, 35 μg ml^{-1} chloramphenicol, 25 μg ml^{-1} nalidixic acid).

In silico detection of conserved protospacer motifs. A Python script (Supplementary Note 1) was used to identify conserved sgRNA sites across a set of 12 or 29 exemplar streptomycin or streptothricin BGCs, respectively, pulled from GenBank and in-house genome sequences (Supplementary Table 1). First, all protospacer motifs of the form N_{20}–NGG were identified in select gene targets, and only those present in all given BGCs were kept. Off-target sites were checked by alignment of protospacer motifs to whole genomes (streptomycin, WAC4770; streptothricin, WAC5950; *Streptomyces* sp., TOR3209 accession no. NZ_AGNH00000000.1; and *Streptomyces* sp., PCS3-D2 accession no. JDUZ01000001.1) using Bowtie2 (ref. [28]), then prioritized on the basis of proximity to a gene's 5′ end (Supplementary Fig. 1). Protospacer motifs selected to target streptothricin and streptomycin BGCs are listed in Supplementary Table 2. Sequence alignments of *orf15* and *orf17* in Supplementary Fig. 1 were generated using Geneious v.8.1.

pCRISPR-Cas9-LigD and pCRISPomyces sgRNA and repair template cloning. Primers used for construction of all CRISPR constructs are listed in Supplementary Table 2. For simplicity, while Tong et al.[15] distinguish between pCRISPR-Cas9 and pCRISPR-Cas9-LigD, in this work the pCRISPR-Cas9 system refers to a plasmid containing both *ligD* (an enzyme required for NHEJ) and a homology template. Similarly, while Cobb et al.[14] distinguish between two forms of the pCRISPomyces system (pCRISPomyces-1 and 2), only the pCRISPomyces-2 system was used in this work.

pCRISPR-Cas9 was previously designed to be able to carry either *ligD* or a homology repair template, but not both. We modified the pCRISPR-Cas9 plasmid to heterologously express *ligD* and carry the homology repair template, thereby allowing either DNA repair pathway to occur. As we used the same repair template for all strains, we reasoned that if the HR template lacked sufficient identity in a particular strain's genome, LigD might allow for efficient repair via NHEJ. Therefore, we first modified the plasmid, adding a *StuI* site in front of the *ligD* expression cassette to allow a location for repair template cloning. *ligD* was PCR amplified from the original pCRISPR-Cas9-LigD construct and cloned back into *StuI*-digested pCRISPR-Cas9 by Gibson assembly (NEB), this time designing the primers to regenerate the *StuI* site upstream of *ligD*.

Repair templates for HR were designed in the same way for both pCRISPR-Cas9 and pCRISPomyces systems. Approximately 1 kb on either side of the sgRNA target site was amplified from a representative BGC (*Streptomyces* sp. WAC5950 for streptothricin, WAC4770 for streptomycin) using primers designed to remove the protospacer motif while introducing an in-frame stop codon and in some cases, a *Hind*III site. Overlap-extension PCR or Gibson assembly was used to combine the two sides of the repair template and clone them into *Xba*I-digested pCRISPomyces-2 or *StuI*-digested pCRISPR-Cas9.

Desired protospacer sequences were introduced into primers and cloned into pCRISPomyces using Golden Gate assembly[29]. Alternatively, they were cloned into pCRISPR-Cas9 between *Nco*I and *Sna*BI restriction sites using traditional ligation-based methods. Streptothricin repair templates and sgRNAs were subcloned from pCRISPomyces plasmids to pCRISPomyces. All plasmids described were sequence verified and transformed into *E. coli* ET12567/pUZ8002 (ref. [30]).

Generation and verification of CRISPR-mediated BGC inactivation strains. All constructs were moved into *Streptomyces* spp. via conjugations from *E. coli* ET12567/pUZ8002 using standard protocols[31]. Exconjugants were selected using 50 μg ml^{-1} apramycin and 25 μg ml^{-1} nalidixic acid. For the pCRISPR-Cas9 system, 12 exconjugants were subsequently re-streaked on Bennett's or nutrient broth (BD Biosciences) plates with 1 μg ml^{-1} thiostrepton to induce Cas9 expression. For both systems, single colonies were finally re-streaked onto Bennett's media, or inoculated into 3 ml of Bennett's or D media and grown for 7 d. Agar plugs or 30 μl of spent media were tested by a Kirby–Bauer assay on cation-adjusted Mueller–Hinton agar (BD Biosciences) against *E. coli* BW25113 Δ*bamB* Δ*tolC* with or without pGDP3:*aph(6)*Ia and pGDP1:*STAT*[6], for streptomycin and streptothricin detection, respectively. Exconjugants that produced no differential zone of inhibition around wild-type and resistant *E. coli* strains were predicted to be successfully engineered (Fig. 1b and Supplementary Fig. 3). Candidate engineered strains were cured of targeting plasmids, all of which carried a temperature-sensitive *ori* derived from pSG5, by growth at 37–40 °C. Curing was confirmed by PCR and apramycin susceptibility testing.

Candidate inactivated strains were verified by Sanger sequencing and LC–MS. PCR sequencing primers were designed in conserved regions roughly 500 bp on either side of the target site, allowing amplification from all strains using the same primers (Supplementary Table 2). Samples for LC–MS verification were prepared by growing strains in liquid media under production conditions, as described above, and vortexing spent media with an equal volume of chloroform then centrifuging to remove protein and cell material. Samples were analyzed in the positive mode using an Agilent 110 Series high performance liquid chromatography (HPLC) system and Applied Biosystems Q-TRAP LC–ESI/MS in the positive mode, with a Sunniest RP-Aqua 4.6×150 mm 5-μm C28 column and LC conditions as follows: 0–2 min 2% B, 2–4 min 2–15% B, 4–5 min 15–20% B, 5–7 min 20–100% B, 7–9 min 100% B, 9–9.5 min 100–2% B, 9.5–11.5 min 2% B (A, water + 0.1% formic acid; B, acetonitrile + 0.1% formic acid; 0.8 ml min^{-1}).

Preparation of crude CRISPR-mediated BGC-inactivated strain extracts. Verified streptothricin and streptomycin null strains were inoculated 1:100 from TSB seed cultures to 50 ml cultures in a 250-ml flask and grown for 7 d at 30 °C, 250 r.p.m. Bennett's media was used for metabolic profiling, and up to seven different fermentation conditions were used for bioactivity testing. Whole cultures were extracted with 15 ml of *n*-butanol, dried under vacuum, and resuspended in 100 μl of DMSO.

Metabolic profiling and PCA analysis of CRISPR-mediated BGC-inactivated strains. Extracts from three biological replicates were run in technical duplicate in the positive mode using an Agilent 1290 Infinity II Series HPLC system and 6550 iFunnel Q-TOF LC–ESI/MS, with an Agilent XDB-C8 3.5 μm 2.1×100-mm column. LC conditions were as follows: 0–1 min 5% B, 1–7 min 5–97% B, 7–7.5 min 97% B, 7.5–8 min 97–5% B, 8–10 min 5% B (A, water + 0.1% formic acid; B, acetonitrile + 0.1% formic acid; 0.4 ml min^{-1}). For compound abundance comparisons between strains, extracted ion chromatograms for the mass of interest were integrated. Standard curves were used to verify linearity over the dynamic range of interest. Full metabolomic comparisons were made using Agilent MassHunter Profinder software for feature extraction and chromatographic alignment followed by Mass Profiler Professional for PCA.

Molecular networking of CRISPR-mediated BGC-inactivated strain extracts. Extracts from three biological replicates were analyzed using the same Q-TOF LC–MS system as for metabolic profiling. LC conditions were as follows: 0–1.5 min 5% B, 1.5–3 min 5–40% B, 3–4 min 40% B, 4–10 min 40–100% B, 10–11 min 100% B, 11–11.5 min 100–5% B, 11.5–12 min 5% B (A, water + 0.1% formic acid; B, acetonitrile + 0.1% formic acid; 0.4 ml min^{-1}). Auto MS/MS acquisition settings used were as follows: positive-ion mass range 200–2,000, MS scan rate, three spectra per second, MS/MS scan rate, two spectra per second, collision energy 20 eV. The top five most abundant precursor ions were selected at every scan for fragmentation, and after three cycles these precursor masses were excluded from fragmentation for 45 s.

Molecular networks were generated using the online workflow at GNPS[18]. Data were filtered to remove MS/MS peaks within 17 Da of the precursor *m/z*, parent masses were clustered with a tolerance of 0.1 Da, and MS/MS fragments were clustered with an ion tolerance 0.1 Da. Networks were generated where edges had a cosine score >0.7 and there were more than five matched MS/MS peaks.

Genome sequencing of wild-type and CRISPR-inactivated strains. Strains for sequencing were grown in TSB at 30 °C, 250 r.p.m. to the mid-log phase. Cells were pelleted and lysed with standard lysozyme, proteinase K and SDS treatment, followed by phenol/chloroform clean-up and ethanol precipitation or column purification. Illumina MiSeq sequencing (300 bp, paired end reads) was performed by the Farncombe Genomics Facility.

Nanopore MinION sequencing was performed in-house with 1D, R9.4 chemistry. Libraries were made using standard ligation sequencing kit protocols (Nanoporetech, catalog no. SQK-LSK109), omitting the FFPE repair enzyme mix and performing clean-up with Long Fragment Buffer (LFB, Oxford Nanopore). Genomic libraries were run on the flow cell (FLO-MIN-106) and base calling was performed on the raw nanopore signals using Albacore 2.3.1 with default parameters to generate reads in fast5 and fastq format. Passing reads were adapter-trimmed using Porechop v.0.2.3 (https://github.com/rrwick/Porechop), and assembled using minimap2 v.2.11 (ref. [33]) (using the -x ava-ont switch) and miniasm 0.3-r179 (ref. [35]) with default parameters. The assembly graph in Graphic Fragment Assembly (GFA) format was manually inspected using Bandage v.0.8.0[34] to make sure there were no unusual edges in the assembly. An improved consensus sequence was generated by mapping the trimmed reads to the initial assembly with minimap2 (using the -x map-ont switch) and polishing with Racon (v.1.3.1) (ref. [35]) using default parameters. A second round of contig polishing was performed using Nanopolish 0.10.1 (ref. [36]) using the reads mapped to the improved Racon assembly as well as the passing Albacore fast5 output with the -methylation-aware = dam,dcm option.

The size of large genomic deletions in knockout strains was determined by comparing the full-genome assembly size of the wild-type and knockout strains.

Where nanopore sequencing was used to build complete genomes of knockout strains, Illumina contigs from corresponding wild-type strains were aligned onto nanopore assemblies using BLAST to determine the location and contents of deletions.

WAC5374 Δ*strH* halogenase gene inactivation. The putative tryptophan halogenase or the related nonribosomal peptide synthetase (NRPS) in WAC5374 was disrupted using pCRISPR-Cas9, without the LigD system, in a similar way to that described above. The pCRISPR-Cas9 construct was designed to introduce a 654-bp in-frame deletion using the primers listed in Supplementary Table 2 and verified by PCR and Sanger sequencing. Disruption of chlorinated indole aldehyde production was verified by high-resolution LC–MS analysis of *n*-butanolic extracts from two independent fermentations.

Assessing streptothricin producer BGC diversity. In addition to the genome sequences harboring a streptothricin BGC from our in-house strain collection, we added 145 *Streptomyces* genomes described previously[10] and 17 other actinomycete genomes that were available from public databases. The entire set of streptothricin BGC-containing genome sequences were subjected to analysis by antiSMASH v.4 using the '—smcogs' '—knownclusterblast' and '—full-hmmer' in addition to default options[37]. The BGC in each genome was extracted and examined individually by a custom Python script. Coding sequences (CDSs) in each cluster were extracted whole if they contained zero or one 'aSdomain' features. However, if they contained two or more 'aSdomain' features, the CDS was divided at the borders of each 'aSdomain' and extracted separately, including the sequences, if any, at the beginning and end of each CDS and between labeled domains. The entire set of extracted CDSs were then subjected to clustering using USEARCH v.8.1.181 (ref. [38]) using the 'cluster_fast' mode with a 60% identity cut-off. After clustering, each CDS and CDS fragment was labeled with the cluster number it was assigned, so that each BGC could be described as a sequence of labeled CDS fragments. We compared the overall biosynthetic capacity pairwise between each genome by implementing the Jaccard index component of the Lin index[39,40] using the unique elements (genes/domains) of the set of combined genome BGC CDS fragment labels.

Phylogenetic analysis. The hidden Markov models (HMMs) corresponding to every family listed under TIGRFAM genome property 0799 'bacterial core gene set, exactly 1 per genome' (https://genome-properties.jcvi.org/cgi-bin/Listing.cgi) were collected. HMMER3 was used to analyze every genome using each model's trusted cut-off[41]. The top hits for each model were retained and aligned as a group against the model HMM. If a genome lacked a hit for a model, gaps equal to the length of the missing sequence were added to the alignment for that genome. These aligned model families were subsequently concatenated into an overall alignment that was inspected manually. This alignment was used for phylogenetic analysis using fasttree2 using the WAG substitution model and otherwise default parameters[42].

RT–PCR on WAC6273, WAC5374 and WAC8241. Strains for analysis were inoculated 1:100 from a TSB seed culture into 60 ml of Bennett's (WAC6273) or SMP media (WAC8241) and pellets were taken at the early stationary phase, at 16 h (WAC8241) or at 20 h (WAC6273). Cells were lysed by bead beating mycelium with 4-mm glass beads in 5 ml of TRIzol reagent (Invitrogen), and RNA was extracted using the manufacturer's recommendations. RNA from the resulting aqueous phase was extracted a second time using acid phenol/chloroform, then combined with a half volume of anhydrous ethanol and finally purified using a PureLink RNA Mini Kit (Invitrogen). Maxima H Minus First Strand cDNA synthesis kit with dsDNase (Thermo Scientific) was used for cDNA synthesis, and PowerUp SYBR Green master mix (Applied Biosystems) was used for RT–PCR quantification on a BioRad CFX96 real time system. Primers targeting major biosynthetic operons in each BGC of interest were designed (Supplementary Table 2) and 90–100% efficiency was verified before quantification. Analysis was performed on three or four independent fermentations and quantified in technical triplicate. Technical triplicates for each biological replicate were averaged, then fold change expression for each replicate was calculated by normalizing to *hrdB* expression using the ΔCt method.

Thiolactomycin purification from WAC5374 Δ*strI*. WAC5374 Δ*strI* was inoculated 1:100 from a TSB seed culture into two 3-l flasks each with 600 ml of SMP media. After 7 d of growth at 30 °C, spent media was adjusted to pH 4 and extracted with 1 l of ethyl acetate. The ethyl acetate layer was dried under vacuum, resuspended in methanol and loaded onto a 5-g C18 prepacked cartridge. Reverse phase flash chromatography was performed as follows: 0–5 min 5% B, 5–25 min 5–75% B, 25–28 min 75% B, 28–30 min 75–95% B (A, water + 0.1% formic acid; B, acetonitrile + 0.1% formic acid; 30 ml min^{-1}). Active fractions containing pure compound were identified as thiolactomycin I–IV by analysis using Q-TOF LC–MS and LC–MS/MS under the same conditions as for metabolic profiling.

Phenanthroviridin aglycone purification from WAC8241 Δ*strI*. Phenanthroviridin aglycone was isolated by activity-guided purification from WAC8241 Δ*strI* fermented in SMP media. After 7 d of fermentation, spent media was treated with Diaion HP-20 resin (Sigma-Aldrich). The HP-20 resin was eluted with 20, 40, 80 and 100% methanol (MeOH). The 80 and 100% MeOH fractions were acidified with 0.3% acetic acid and extracted twice with hexane to remove highly hydrophobic compounds. The MeOH layer was put on a C18 vacuum liquid chromatography column and eluted with 20, 40, 60, 80 and 100% MeOH. The 60% and 80% MeOH fractions were subjected to a Sephadex LH20 (Sigma-Aldrich) column using MeOH:dichloromethane (2:1) as a running solvent. Active fractions were run on a silica gel (Sigma-Aldrich) vacuum liquid chromatography column, eluting with hexane, hexane:ethyl acetate (1:1), ethyl acetate, ethyl acetate:MeOH (9:1), ethyl acetate:MeOH (8:2), ethyl acetate:MeOH (1:1), MeOH and MeOH:water (7:3). The active fractions (ethyl acetate and ethyl acetate:MeOH (9:1)) were further purified by HPLC on a Waters Atlantis T3 prep column using a linear gradient from 95% to 5% solvent A (solvent A, water; solvent B, acetonitrile, both acidified with 0.1% trifluoroacetic acid) to yield the active compound phenanthroviridin aglycone.

Purification of amicetin and derivatives from WAC6273 Δ*orf17*. WAC6273 Δ*orf17* was fermented in Bennett's media for 5 d. Amicetin and its derivatives were isolated by activity-guided purification. Spent media was treated with Diaion HP-20 resin. The HP-20 resin was eluted with 5, 20, 50 and 100% MeOH. The active fractions were subjected to a C18 CombiFlash column using a linear gradient from 95% to 5% solvent A (solvent A, water; solvent B acetonitrile, both acidified with 0.1% formic acid). Active fractions were purified by HPLC on a Waters Atlantis T3 prep column using a linear gradient from 95% to 5% solvent A (solvent A, water; solvent B, acetonitrile, both acidified with 0.1% trifluoroacetic acid). The active fractions contained amicetin and its derivatives.

Purification of 5-chloro-3-formylindole from WAC5374 Δ*strH*. WAC5374 Δ*strH* was fermented in SMP media for 8 d. The culture was treated with Diaion HP-20 resin and the resin was eluted with 10, 20, 40, 60, 80 and 100% MeOH. The 60 and 80% MeOH fractions were run on a Sephadex LH20 column using 350 ml of MeOH, 200 ml of dichloromethane:MeOH (1:1) and 100 ml of MeOH as running solvents. The active fractions were put on a C18 vacuum column and eluted with 20, 40, 50, 60, 70, 80 and 100% MeOH. The 50 and 60% MeOH fractions were combined and subjected to a C18 CombiFlash column using a linear gradient from 95% to 5% solvent A (solvent A, water; solvent B, acetonitrile, both acidified with 0.1% formic acid). The active fractions were fractionated by HPLC on a Waters Atlantis T3 prep column using a linear gradient from 95% to 5% solvent A (solvent A, water; solvent B, acetonitrile, both acidified with 0.1% trifluoroacetic acid) to give 5-chloro-3-formylindole.

Statistical analysis. Statistical analysis of production as quantified by LC–MS and gene expression as quantified by RT–PCR was performed using GraphPad Prism v.6. Pairwise comparisons were made by an unpaired two-sided Student's *t*-test ($n = 6$ for LC–MS data and $n = 3$ for RT–PCR, as indicated) and *P* values are reported in figure legends. Multiple comparisons were made by one-way ANOVA with Tukey's post hoc analysis ($n = 6$). Parameters were as follows for LC–MS production quantification: WAC8241 ($t = 3.107$, d.f. = 10), WAC6273 ferrioxamine 1 ($F = 116.7$, d.f. = 20), ferrioxamine 7 ($F = 143$, d.f. = 20), ferrioxamine 10 ($F = 34.34$, d.f. = 20), ferrioxamine 13 ($F = 103.3$, d.f. = 20), WAC6273 amicetin ($F = 15.57$, d.f. = 20), WAC6273 bamicetin ($F = 3.67$, d.f. = 20). Parameters were as follows for gene expression quantification: WAC8241 ($t = 4.562$, d.f. = 4). All results are representative of two independent experiments.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Whole-genome sequences of actinomycete strains used in this study, including 19 wild-type streptomycin or streptothricin producers and six CRISPR-engineered derivatives (WAC5950 Δ*orf15* pCRISPomyces, WAC6273 Δ*orf15* pCRISPR-Cas9, WAC8241 Δ*strI*, WAC5374 Δ*strF*, WAC5374 Δ*strH* and WAC5374 Δ*strI*), are available in GenBank with the Bioproject accession number PRJNA504665 (Supplementary Table 1).

Code availability

A custom Python script was used to identify conserved sgRNA target sites in a BGC of interest and is provided in Supplementary Note 1 together with instructions for use.

References

28. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
29. Engler, C., Gruetzner, R., Kandzia, R. & Marillonnet, S. Golden gate shuffling: A one-pot DNA shuffling method based on type IIs restriction enzymes. *PLoS ONE* **4**, e5553 (2009).
30. Paget, M. S., Chamberlin, L., Atrih, A., Foster, S. J. & Buttner, M. J. Evidence that the extracytoplasmic function sigma factor sigmaE is required for normal cell wall structure in *Streptomyces coelicolor* A3(2). *J. Bacteriol.* **181**, 204–211 (1999).

31. Kieser, T., Bibb, M. J., Buttner, M. J., Chater, K. F. & Hopwood, D. A. *Practical Streptomyces Genetics* (John Innes Foundation, 2000).
32. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
33. Li, H. Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. *Bioinformatics* **32**, 2103–2110 (2016).
34. Wick, R. R., Schultz, M. B., Zobel, J. & Holt, K. E. Bandage: interactive visualization of de novo genome assemblies. *Bioinformatics* **31**, 3350–3352 (2015).
35. Vaser, R., Sović, I., Nagarajan, N. & Šikić, M. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res.* **27**, 737–746 (2017).
36. Simpson, J. T. et al. Detecting DNA cytosine methylation using nanopore sequencing. *Nat. Methods* **14**, 407–410 (2017).
37. Blin, K. et al. antiSMASH 4.0—improvements in chemistry prediction and gene cluster boundary identification. *Nucleic Acids Res.* **45**, W36–W41 (2017).
38. Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460–2461 (2010).
39. Lin, K., Zhu, L. & Zhang, D.-Y. An initial strategy for comparing proteins at the domain architecture level. *Bioinformatics* **22**, 2081–2086 (2006).
40. Cimermancic, P. et al. Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters. *Cell* **158**, 412–421 (2014).
41. Eddy, S. R. Accelerated profile HMM searches. *PLoS Comput. Biol.* **7**, e1002195 (2011).
42. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2—approximately maximum-likelihood trees for large alignments. *PLoS ONE* **5**, e9490 (2010).

Appendix 2: Culp, E. *et al.* (2020). Evolution-guided discovery of antibiotics that inhibit peptidoglycan remodelling. *Nature* 578:582-87.

Article

Evolution-guided discovery of antibiotics that inhibit peptidoglycan remodelling

https://doi.org/10.1038/s41586-020-1990-9

Received: 3 May 2019

Accepted: 12 December 2019

Published online: 12 February 2020

Check for updates

Elizabeth J. Culp[1], Nicholas Waglechner[1], Wenliang Wang[1], Aline A. Fiebig-Comyn[1], Yen-Pang Hsu[2], Kalinka Koteva[1], David Sychantha[1], Brian K. Coombes[1], Michael S. Van Nieuwenhze[2], Yves V. Brun[3,4] & Gerard D. Wright[1]✉

Addressing the ongoing antibiotic crisis requires the discovery of compounds with novel mechanisms of action that are capable of treating drug-resistant infections[1]. Many antibiotics are sourced from specialized metabolites produced by bacteria, particularly those of the Actinomycetes family[2]. Although actinomycete extracts have traditionally been screened using activity-based platforms, this approach has become unfavourable owing to the frequent rediscovery of known compounds. Genome sequencing of actinomycetes reveals an untapped reservoir of biosynthetic gene clusters, but prioritization is required to predict which gene clusters may yield promising new chemical matter[2]. Here we make use of the phylogeny of biosynthetic genes along with the lack of known resistance determinants to predict divergent members of the glycopeptide family of antibiotics that are likely to possess new biological activities. Using these predictions, we uncovered two members of a new functional class of glycopeptide antibiotics—the known glycopeptide antibiotic complestatin and a newly discovered compound we call corbomycin—that have a novel mode of action. We show that by binding to peptidoglycan, complestatin and corbomycin block the action of autolysins—essential peptidoglycan hydrolases that are required for remodelling of the cell wall during growth. Corbomycin and complestatin have low levels of resistance development and are effective in reducing bacterial burden in a mouse model of skin MRSA infection.

Throughout evolution, biosynthetic gene clusters (BGCs) are sculpted at least in part by selective pressure on the biological activity of the resultant metabolite. Therefore, BGCs that have evolutionarily diverged biosynthetic genes might also produce novel biological activity. Such phylogeny-guided discovery has previously been applied to identify divergent members of natural product families by generating phylogenetic trees from short sequence tags of select biosynthetic genes[3–5]. Alternatively, we have previously shown that by combining the selection of antibiotic resistance with phylogenetic analysis of concatenated biosynthetic gene segments, it is possible to identify new members of a class of antibiotics that retain the same mechanism of action[6]. These approaches are useful for prioritizing BGCs that produce undiscovered compounds with biological activity, but they do not provide any information on the mechanism of action. Antibiotic BGCs encode not only biosynthetic machinery, but also resistance genes to protect from self-intoxication. Searching for the presence of resistance genes in BGCs has previously directed genome mining for antibiotics with known or predicted mechanisms[7,8]. Interested instead in new modes of action, we suggested that by identifying phylogenetically distinct BGCs that lack known self-resistance genes, we could find antibiotics with different and possibly novel mechanisms of action.

Resistance and phylogeny guide discovery

We applied our hypothesis to the glycopeptide family of antibiotics, which was chosen for its diversity of BGCs, high degree of tailoring, and distinctive self-resistance genes acting through target modification (for example, *vanHAX*). From in-house genome sequences and public repositories, we collected 71 BGCs from glycopeptide antibiotics (GPAs) and antibiotics of the glycopeptide family, and constructed phylogenetic trees of every shared gene and gene segment found in these BGCs[9]. By integrating our knowledge about species phylogeny with the lineage of individual genes or domains, we identified trees that provided information about the conservation or divergence of components due to function (for example, non-ribosomal peptide synthase) rather than the relationship between strains on a species level (for example, precursor supply)[9]. The relationship that we observed in a tree of particular non-ribosomal peptide synthase condensation domains exemplifies divergence due to changing function, and was present in several phylogenies that we analysed (Fig. 1a, Extended Data Fig. 1a, b). Mapping the presence of self-resistance genes onto these trees, we found that BGCs that contain common resistance determinants for 'true' GPAs (for example, *vanHAX* and *vanY* for antibiotics binding D-Ala-D-Ala of lipid II) fell within a single clade (Fig. 1a). However,

[1]M. G. DeGroote Institute for Infectious Disease Research, David Braley Centre for Antibiotic Discovery, Department of Biochemistry and Biomedical Sciences, McMaster University, Hamilton, Ontario, Canada. [2]Department of Chemistry, Indiana University, Bloomington, IN, USA. [3]Department of Biology, Indiana University, Bloomington, IN, USA. [4]Département de Microbiologie, Infectiologie et Immunologie, Université de Montréal, Montreal, Quebec, Canada. ✉e-mail: wrightge@mcmaster.ca

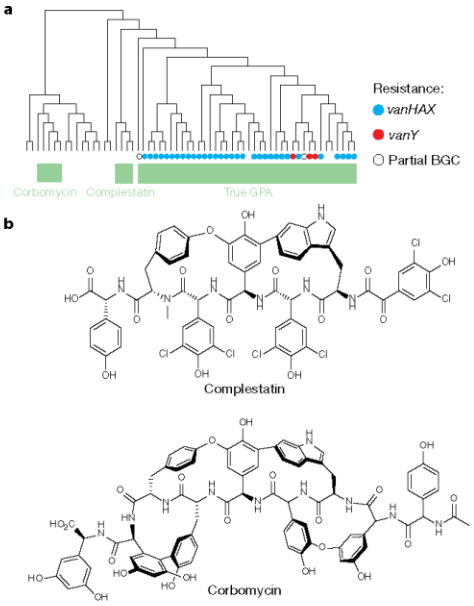


Fig. 1 | Phylogeny-guided discovery of complestatin and corbomycin. **a**, A maximum-likelihood phylogenetic tree of condensation domains at the C-2 module according to the labelling scheme shown in Extended Data Fig. 1a; the full tree is shown in Extended Data Fig. 1b. The presence of true GPA self-resistance genes—including *vanHAX* and *vanY*—within each corresponding BGC is marked by coloured circles. The tree is shown as a rectangular cladogram with a midpoint root. Labels indicate well-supported clades (Shimodaira–Hasegawa (SH)-like support values > 0.89) of condensation domains for the true GPAs, complestatin-type BGCs and corbomycin-type BGCs. Branches not marked by one of these BGC family labels are uncharacterized. Scaled and terminal-node-labelled trees are available at https://github.com/waglecn/GPA_evolution. **b**, The chemical structures of complestatin and corbomycin.

in BGCs with divergent evolutionary histories and a distinct common ancestor from 'true' GPAs, we noted two clades that lacked any known GPA resistance determinant, which indicates that they may also possess divergent biological activity.

One of these divergent clades contained a known compound, complestatin (Fig. 1b), as well as other members that were predicted to be structural variants. We purified complestatin from *Streptomyces* **sp. WAC01325 that contains this BGC[10] (Extended Data Fig. 1c). The second divergent clade contained BGCs with no characterized** members, and so we used our phylogenetic trees to make structural predictions and guide purification of the resulting metabolite from *Streptomyces* **sp. WAC01529 (Fig. 1b, Extended Data Fig. 1d, e, Supplementary Figs. 1–6). We named it corbomycin (after Nid de Corbeau—or Crow's Nest Pass—in Alberta, Canada, where** *Streptomyces* sp. WAC01529 was collected), with *crb* denoting the corresponding BGC (Extended Data Fig. 1d).

Corbomycin and complestatin belong to the type V family of GPAs[11], the other known members of which include chloropeptin I, neuroprotectin A/B and kistamicins[12,13]. Antibiotic activity of these compounds **has been noted but has not been investigated in depth[14,15]. The lack** of known resistant determinants in these BGCs therefore warranted further investigation of their mechanism of action.

Identification of a novel mode of action

Corbomycin and complestatin primarily show activity against Gram-positive bacteria, with minimum inhibitory concentrations (MICs) ranging from 0.5 to 4 μg ml^{-1}—a potency comparable to that of vancomycin (Extended Data Table 1). They are active against a range of laboratory strains and clinically relevant pathogens, including methicillin-resistant and daptomycin-resistant *Staphylococcus aureus* (MIC = 0.5–2 μg ml^{-1}). Notably, they are also active against vancomycin-resistant *Enterococcus* and vancomycin-intermediate *S. aureus*, which indicates that they have a different mode of action from that of typical GPAs. Indeed, isothermal titration calorimetry experiments failed to show binding to peptidoglycan stem pentapeptide under conditions in which vancomycin binding can be easily determined (K_d = 14.3 μM).

The activity of complestatin against vancomycin-resistant *Enterococcus* has been observed previously[16], but studies in vitro and in *S. aureus* **suggested that the mechanism involved the inhibition of fatty acid** synthesis by targeting the enoyl-ACP reductase FabI[15]. We questioned **whether this was the physiologically relevant target given that the size (molecular mass = 1,328 Da) and physiochemical properties of complestatin render it unlikely to penetrate the cell membrane and thus reach intracellular FabI. Target overexpression, knockout and** exogenous fatty acid supplementation experiments (Extended Data **Fig. 2) did not support fatty acid synthesis as the primary target of** complestatin or of corbomycin in *S. aureus* or *Bacillus subtilis*.

Given the activity of the glycopeptide antibiotics against the cell wall, we next tested whether corbomycin and complestatin affect peptidoglycan metabolism. The promoter of *ywaC* in *B. subtilis* is activated **almost exclusively by antibiotics that act on the cell envelope[17], and** was robustly activated by corbomycin and complestatin (Fig. 2a). P_{ywaC} **is also activated by compounds that target the cell membrane[17], but** permeabilization of the membrane was not observed using the voltage-sensitive fluorescent dye $DiSC_3$ (Fig. 2b). We then tested the effects of **these compounds on the biosynthesis of teichoic acid by using a** *B. subtilis* mutant deficient in TagO—this enzyme is involved in the early **stages of teichoic acid biosynthesis, and its knockout abolishes the** lethality that is observed upon the inhibition of this biosynthetic pathway[18]. Corbomycin and complestatin were equally active against Δ*tagO* and wild-type *B. subtilis* (MIC = 1 μg ml^{-1}, Extended Data Table 1), which **rules out an effect on teichoic acid synthesis. We therefore focused** on peptidoglycan metabolism as the potential target of corbomycin and complestatin.

To investigate which stage of peptidoglycan metabolism is blocked **by complestatin and corbomycin, we measured the build-up of the final cytoplasmic intermediate of peptidoglycan synthesis, uridine** 5′-diphosphate-*N*-acetylmuramic acid pentapeptide (UDP-MurNAc-PP). Treatment of *S. aureus* with antibiotics acting on lipid-linked steps **of peptidoglycan biosynthesis such as transglycosylation results in** the accumulation of UPD-MurNAc-PP (Fig. 2c). Corbomycin and com**plestatin had no effect (Fig. 2c), which indicates that peptidoglycan** metabolism is affected downstream of transglycosylation.

We next tested whether transpeptidation was inhibited by making use of fluorescent D-amino acids (FDAAs)—such as HCC-amino-D-alanine **(HADA)—that are incorporated into actively growing peptidoglycan by transpeptidases[19]. At concentrations twice that of the minimum inhibitory concentration (2 × MIC), the inhibition of peptidoglycan** synthesis by the β-lactam ampicillin was detected by decreased HADA incorporation, whereas corbomycin and complestatin caused no significant change over 1 h (Fig. 2d, Extended Data Fig. 3a–c). Furthermore, **whereas all antibiotics that block peptidoglycan synthesis result in** cell lysis through an imbalance between peptidoglycan synthesis and degradation, corbomycin and complestatin are bacteriostatic (Fig. 2e, Extended Data Fig. 3d, e). Complestatin and corbomycin were not synergistic with various cell-wall- and membrane-active antibiotics, but **were additive with each other (Extended Data Fig. 4a). Corbomycin**

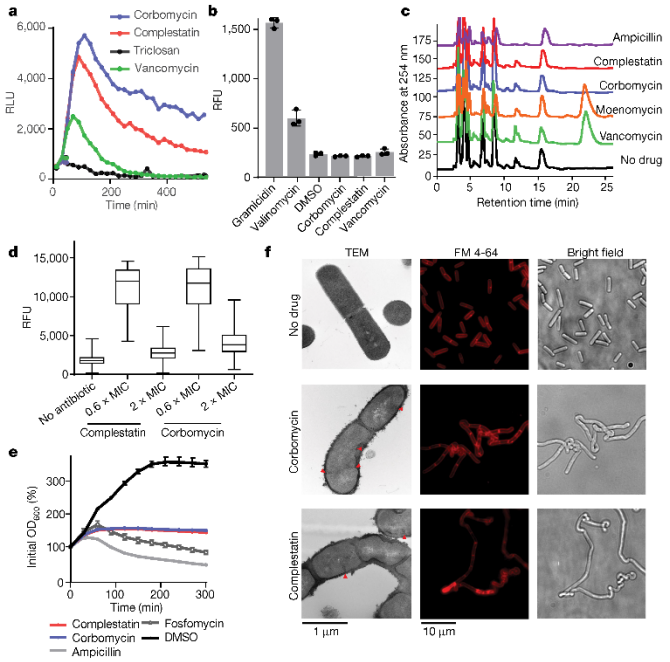


Fig. 2 | Corbomycin and complestatin affect peptidoglycan metabolism. **a**, Complestatin and corbomycin strongly activate P_{yuaG}, which controls the expression of *lux* genes in *B. subtilis*. RLU, relative luminescence units. **b**, Membrane disruption was measured using release of the voltage-sensitive fluorescent dye $DiSC_3$ from *B. subtilis* membranes, quantified as maximal relative fluorescence units (RFU). Gramicidin and valinomycin are membrane-active antibiotics and are used as positive controls for this assay. **c**, High-performance liquid chromatography (HPLC) chromatograms show accumulation of UDP-MurNAc-PP in *S. aureus* after treatment with 10 × MIC antibiotic, as seen by the peak at 23 min in the vancomycin and moenomycin samples. **d**, The incorporation of HADA was used to measure peptidoglycan synthesis and transpeptidase activity in antibiotic-treated *B. subtilis*. For samples shown left to right, the number of individual cells quantified was $n = 202, 81, 169, 138, 187$. The whiskers show the minimum and maximum values of the dataset, the box shows the upper and lower quartiles and the line shows the median value. **e**, Lysis of *B. subtilis* treated with various antibiotics above their MICs shows that corbomycin and complestatin are bacteriostatic, in contrast to the peptidoglycan-biosynthesis inhibitors ampicillin and fosfomycin. **f**, Microscopy of *B. subtilis* grown in 0.6 × MIC corbomycin or complestatin shows a characteristic twisted phenotype. Red triangles on TEM images mark sites of aberrant-division septa formation, thickened cell wall and granular formations. In **b** and **e**, the mean of three biological replicates is shown, with error bars showing the standard deviation. In **b**, individual data points are shown by dots. All experiments were repeated at least twice with similar results.

and complestatin therefore target peptidoglycan metabolism through a similar mechanism, which to our knowledge is distinct from that of any previously reported antibiotic.

Mode of action by autolysin inhibition

Given the unprecedented activity of complestatin and corbomycin on peptidoglycan metabolism, we examined the phenotype of *B. subtilis* grown in sub-MIC levels of antibiotic. At 0.6 × MIC, cells effectively formed septa but failed to divide, instead forming twisted and knotted chains of cells (Fig. 2f). Transmission electron microscopy (TEM) images showed aberrant septal structure, the absence of flagella, a 'shaggy' and thickened cell wall, and distortion of the cytoplasm. This distinctive phenotype was unlike that observed with several control antibiotics (Extended Data Fig. 4b), but matched that of *B. subtilis* strains defective in autolysins[20,21]. This diverse group of peptidoglycan-degrading enzymes is essential for normal peptidoglycan metabolism by enabling the insertion of new material into the existing cell wall and the cleavage of peptidoglycan at the division septa.

We attempted to select for spontaneous resistant mutants on 4 × MIC and 8 × MIC agar, but were unsuccessful for both *B. subtilis* 168 (frequency of resistance < 3×10^{-9}) and *S. aureus* ATCC 29213 (frequency of resistance < 10^{-10}). *B. subtilis* was instead serially passaged in sub-MIC levels of antibiotic for 25 days. Resistance was slow and difficult to develop, reaching a maximum MIC increase of only fourfold for both corbomycin and complestatin (final MIC = 4 μg ml^{-1}, Fig. 3a). We observed similar low levels of resistance development in serially passaged *S. aureus* (Extended Data Fig. 5a). Resistant mutants raised separately on corbomycin (strains COR14 and COR25) or complestatin (strains COM20 and COM25) displayed cross-resistance with each other but did not share cross-resistance to other cell-wall- or membrane-active antibiotics (Extended Data Fig. 5b, c, Extended Data Table 1); this is consistent with their activity being distinct from that of those antibiotics.

Whole-genome sequencing identified mutations in protein-coding regions of the *B. subtilis* resistant mutants (Fig. 3b). We identified mutations in various proteins that modify the cell wall directly (for example, the autolysin CwlO[20]), as well as those that are linked to the regulation

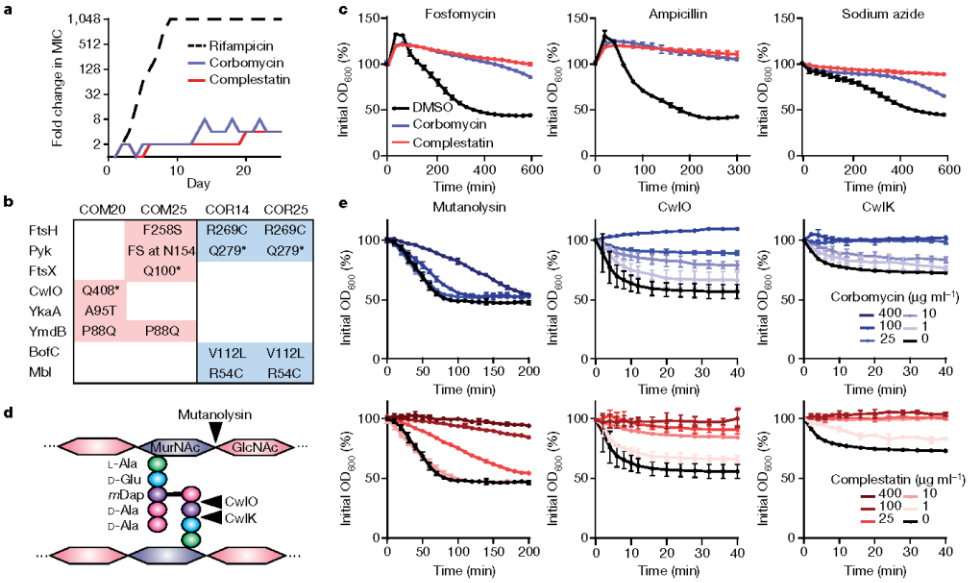


Fig. 3 | Corbomycin and complestatin inhibit autolysins. a, The MICs of *B. subtilis* serially passaged in sub-lethal concentrations of antibiotic over 25 days were tracked and plotted. As a comparison, the rapid development of high-level resistance to rifampicin was tracked. Serial passaging was completed on two independent occasions with similar results. **b**, Mutations identified in complestatin-raised isolates on day 20 (COM20) or 25 (COM25), or in corbomycin-raised mutants on day 14 (COR14) or 25 (COR25). Amino acid mutations are shown, with FS representing a frame shift. The CwlO Q408* mutation truncates the catalytic histidine (H431), resulting in an inactive enzyme. **c**, Corbomycin and complestatin protect *B. subtilis* from lysis. Cells were treated with agents to induce lysis in addition to corbomycin, complestatin or solvent control. **d**, Bond cleavage specificities of various autolysins. **e**, In vitro digestion of *B. subtilis* peptidoglycan by various hydrolases was monitored after preincubation with complestatin or corbomycin. For **c** and **e**, mean values and standard deviation from biological triplicates are plotted.

of autolysins (for example, the phosphodiesterase YmdB[22], CwlO regulator FtsX[20,23] and actin homologue Mbl[20]) or to cell division in general (for example, the metalloprotease FtsH[24] and pyruvate kinase Pyk[25]) (see Supplementary Discussion). The effects of single-gene deletions were characterized (see Supplementary Discussion and Extended Data Fig. 5d), but they conferred only up to twofold reduced susceptibility to corbomycin and complestatin (Extended Data Fig. 5e). This is within the accepted error for MICs and indicates a possible polygenic mechanism **by which they confer resistance and a non-protein or multisubunit** target for corbomycin and complestatin. On the basis of resistant mutations and phenotypic evidence, we hypothesized that corbomycin and complestatin block the function of autolysins.

To test whether corbomycin and complestatin inhibit autolysins in whole cells, we examined the effects of the compounds on *B. subtilis* lysis. By blocking peptidoglycan synthesis with ampicillin or fosfomycin, net peptidoglycan degradation by autolysins leads to lysis, but it could be antagonized by corbomycin or complestatin (Fig. 3c). Some bacteriostatic drugs—especially those that target the ribosome, such as chloramphenicol—are known to antagonize bactericidal antibiotics by preventing the synthesis of lytic enzymes[26] (Extended Data Fig. 6a). We therefore induced lysis using sodium azide to dissipate the proton motive force (Fig. 3c). Autolysin activity is inhibited by low pH next to the cell membrane, but this inhibition is reduced as this pH increases upon azide treatment[27,28]. Corbomycin and complestatin were found to protect cells from lysis induced by sodium azide, whereas chloramphenicol did not (Fig. 3c, Extended Data Fig. 7a). Similar results were **observed for the lysis of *S. aureus* induced by the inhibition of peptidoglycan synthesis (Extended Data Fig. 6b). These results support** the suggestion that corbomycin and complestatin inhibit autolysins in whole cells.

Next, we simplified our whole-cell system to an in vitro assay containing peptidoglycan and various peptidoglycan hydrolases with different bond specificities. Intact peptidoglycan is insoluble, so digestion can be monitored by tracking the decrease in optical density as peptidoglycan is solubilized. We first tested two muramidases—lysozyme from hen egg white and mutanolysin from *Streptomyces globisporus*. These enzymes hydrolyse β-1,4-glycosidic bonds between *N*-acetylmuramic acid (MurNAc) and *N*-acetyl-D-glucosamine (GlcNAc) of the glycan backbone of peptidoglycans (Fig. 3d). Complestatin strongly inhibited the digestion of peptidoglycan by both muramidases, whereas corbomycin blocked digestion only at higher concentrations—400 μg ml^{-1} corbomycin per 1 mg ml^{-1} peptidoglycan (Fig. 3e, Extended Data Fig. 6c). Next, we purified and tested two *B. subtilis* endopeptidases—CwlK (YcdD) and CwlO (YcvE)—that cleave the peptide stem of peptidoglycan either between L-Ala and D-Glu or between D-Glu and *m*DAP, respectively[29,30] (Fig. 3d). **We chose these enzymes as representative physiologically relevant** autolysins, as both are active during vegetative *B. subtilis* growth but target different bonds and have unrelated catalytic domains. The action **of both enzymes on peptidoglycan was strongly inhibited to similar** degrees by corbomycin and complestatin at concentrations as low as 10 μg ml^{-1} antibiotic per 1 mg ml^{-1} peptidoglycan (Fig. 3e). Corbomycin and complestatin therefore broadly block the activity of peptidoglycan hydrolases, irrespective of the enzyme family.

We next used FDAAs to measure the incorporation of peptidoglycan after the treatment of *B. subtilis* with 0.6 × MIC corbomycin. We found a significant increase in FDAA signal after treatment, in contrast

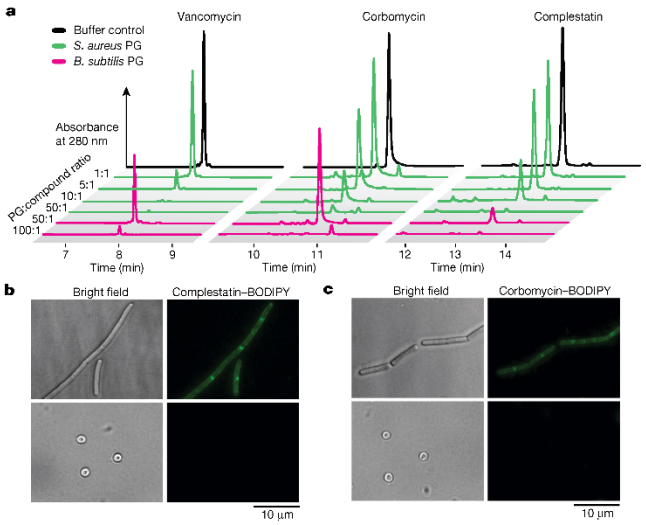


Fig. 4 | Corbomycin and complestatin bind to peptidoglycan. a, HPLC chromatograms of antibiotic left unbound after incubation with either *B. subtilis* or *S. aureus* peptidoglycan. Antibiotic and peptidoglycan were combined in various w/w ratios, and the shortening of a peak in comparison to the buffer control that lacked peptidoglycan represents compound binding and removal from solution. **b**, **c**, Imaging of *B. subtilis* rods (top) or protoplasts (bottom) stained with fluorescent complestatin (complestatin–BODIPY; **b**) or corbomycin (corbomycin–BODIPY; **c**). Identical staining conditions, exposure time and image adjustments were used for rod and protoplast images. All experiments were performed on two separate occasions with similar results.

to that from cells treated with 2 × MIC corbomycin (Fig. 2d, Extended Data Fig. 3b). A similar increase in nascent peptidoglycan labelling was observed using the fluorescent conjugate vancomycin–BODIPY FL (Extended Data Fig. 6d) and has been documented for multi-autolysin *B. subtilis* mutants[31]. This phenotype is consistent with a slowing of peptidoglycan hydrolysis without affecting peptidoglycan synthesis, which results in the accumulation of labelled peptidoglycan. At 2 × MIC, when growth is halted completely (Extended Data Fig. 3b), peptidoglycan synthesis may slow in balance with degradation and result in no significant change in HADA incorporation (Fig. 3c, Extended Data Fig. 3b). Combined with the results above, these findings suggest that complestatin and corbomycin inhibit autolysins both in vitro and in whole cells.

Corbomycin and complestatin bind peptidoglycan

The ability of corbomycin and complestatin to inhibit a broad range of structurally unrelated autolysins suggests that they bind to the common substrate, peptidoglycan. The inability to digest peptidoglycan preincubated with antibiotic then washed before treatment with CwlO further supported the suggestion that inhibition results from an interaction between antibiotic and peptidoglycan, not between antibiotic and enzyme (Extended Data Fig. 7a). We tested binding by incubating the antibiotics with insoluble peptidoglycan at a range of weight-to-weight ratios, collecting insoluble material including any bound antibiotic, and quantifying the remaining soluble antibiotic by HPLC. Validating our assay, vancomycin was shown to bind peptidoglycan whereas the negative controls daptomycin and rifampin did not (Fig. 4a, Extended Data Fig. 7b). Corbomycin and complestatin were bound and removed from solution by both *B. subtilis* and *S. aureus* peptidoglycan in a dose-dependent manner (Fig. 4a).

To further visualize peptidoglycan binding, we synthesized fluorescently labelled corbomycin and complestatin by derivatizing the carboxy terminus with BODIPY-FL-EDA to generate the conjugates corbomycin–BODIPY and complestatin–BODIPY (Extended Data Fig. 7c). These derivatives were shown to retain on-target activity (Methods, Extended Data Fig. 7d). We stained *B. subtilis* cells with 2 × MIC antibiotic and observed staining of the outside of the cell including the division septa, which is consistent with their binding to peptidoglycan (Fig. 4b, c). Generating protoplasts by removing the cell wall with lysozyme abolished staining by corbomycin–BODIPY or complestatin–BODIPY (Fig. 4b, c). Collectively, these results show a specific interaction between peptidoglycan and corbomycin or complestatin at a motif that is widely found in peptidoglycan from different species, including *Staphylococcus* and *Bacillus* spp.

To our knowledge, our results support an unprecedented mechanism of action whereby corbomycin and complestatin bind to peptidoglycan and block its access to autolysins. Although autolysins are a highly redundant family of enzymes and no single enzyme is usually essential, by binding the peptidoglycan itself corbomycin and complestatin block most—if not all—autolysin activity that is required for cell-wall expansion, thereby inhibiting growth.

In vivo efficacy in mouse

We investigated the in vivo efficacy of corbomycin and complestatin at treating an infection. Notably, both compounds are non-toxic to eukaryotic cells, including yeasts and human embryonic kidney (HEK) cells (Extended Data Table 1). Owing to the poor water solubility of the compounds, we chose to formulate the antibiotics as a topical lotion for our preliminary studies. We established a methicillin-resistant *Staphylococcus aureus* (MRSA; Rosenbach ATCC 33591) superficial skin infection in neutropenic mice and applied complestatin or corbomycin in a petroleum-jelly-based lotion. Fusidic acid, a common topically applied antibiotic, was used as a comparison. Both complestatin and corbomycin significantly reduced bacterial load by around 100-fold at 33 h post-infection, with a similar efficacy to fusidic acid (Extended Data

Fig. 8a) and in line with previous studies[32–34]. The weight loss exhibited by the mice during infection was significantly reduced, and wound scabbing and necrosis were noticeably improved, when compared with vehicle controls (Extended Data Fig. 8b, c).

Discussion

Although actinomycetes contain a wealth of potential chemical diversity, prioritization of BGCs for follow-up remains a bottleneck. By combining phylogenetic divergence with the absence of dedicated self-resistance determinants to prioritize BGCs, we describe the discovery of a new functional class of GPAs with a novel mechanism of action. This approach is similar to those that use phylogenetic analysis of sequence tags that have been amplified from metagenomic samples by PCR[3,4,35], but it incorporates information from multiple genes in an intact BGC rather than from single sequence tags. Tracking the phylogenetic history of complete GPA BGCs enables the identification of divergent BGCs, because not all genes within a cluster follow the same evolutionary lineage. This strategy also provided information about the grouping and location of domains in non-ribosomal peptide synthase, as well as important structural predictions for purifying corbomycin. Importantly, in contrast to previous methods, we also focused on BGCs that expressly lack a certain resistance mechanism. Although the biological activity of identified compounds cannot be predicted a priori, this strategy prioritizes compounds that are more likely to have new modes of action. The selection of BGCs that lack known resistance genes is generalizable to any class of antibiotic with a specific resistant mechanism (for example, target modification).

Antibiotics blocking nearly every step in peptidoglycan synthesis have been described, but—to our knowledge—complestatin and corbomycin are the first to inhibit peptidoglycan remodelling. Different members of the family may have slightly different peptidoglycan-binding sites, which could explain subtle differences in activity between corbomycin and complestatin. By acting as chemical probes for peptidoglycan binding and autolysin inhibition, corbomycin and complestatin—along with their fluorescent derivatives—will be useful tools for the study of peptidoglycan hydrolases.

To avoid cross-resistance with existing antibiotics, the development of compounds with novel targets is a coveted but elusive goal. Although others have explored the idea of inhibiting autolysins in the context of virulence[36,37] and β-lactam potentiation[38], the lack of an essential single protein prevented autolysins from emerging as a primary antibiotic target. The finding that corbomycin and complestatin inhibit most or all autolysins by binding peptidoglycan thus represents the discovery of antibiotics with a long-sought-after mechanism of action. Corbomycin and complestatin are active against multidrug-resistant clinical isolates, display low resistance development and are effective in vivo in a mouse model of skin infection, making them an exciting avenue for future development.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-1990-9.

1. Laxminarayan, R. et al. Antibiotic resistance—the need for global solutions. *Lancet Infect. Dis.* **13**, 1057–1098 (2013).
2. Wright, G. D. Something old, something new: revisiting natural products in antibiotic drug discovery. *Can. J. Microbiol.* **60**, 147–154 (2014).
3. Kang, H. S. & Brady, S. F. Arixanthomycins A–C: phylogeny-guided discovery of biologically active eDNA-derived pentangular polyphenols. *ACS Chem. Biol.* **9**, 1267–1272 (2014).
4. Peek, J. et al. Rifamycin congeners kanglemycins are active against rifampicin-resistant bacteria via a distinct mechanism. *Nat. Commun.* **9**, 4147 (2018).
5. Hover, B. M. et al. Culture-independent discovery of the malacidins as calcium-dependent antibiotics with activity against multidrug-resistant Gram-positive pathogens. *Nat. Microbiol.* **3**, 415–422 (2018).
6. Thaker, M. N. et al. Identifying producers of antibacterial compounds by screening for antibiotic resistance. *Nat. Biotechnol.* **31**, 922–927 (2013).
7. Yan, Y. et al. Resistance-gene-directed discovery of a natural-product herbicide with a new mode of action. *Nature* **559**, 415–418 (2018).
8. Tang, X. et al. Identification of thiotetronic acid antibiotic biosynthetic pathways by target-directed genome mining. *ACS Chem. Biol.* **10**, 2841–2849 (2015).
9. Waglechner, N., McArthur, A. G. & Wright, G. D. Phylogenetic reconciliation reveals the natural history of glycopeptide antibiotic biosynthesis and resistance. *Nat. Microbiol.* **4**, 1862–1871 (2019).
10. Chiu, H.-T. et al. Molecular cloning and sequence analysis of the complestatin biosynthetic gene cluster. *Proc. Natl Acad. Sci. USA* **98**, 8548–8553 (2001).
11. Nicolaou, K. C., Boddy, C. N. C., Bräse, S. & Winssinger, N. Chemistry, biology, and medicine of the glycopeptide antibiotics. *Angew. Chem. Int. Ed.* **38**, 2096–2152 (1999).
12. Breazzano, S. P. & Boger, D. L. Synthesis and stereochemical determination of complestatin A and B (neuroprotectin A and B). *J. Am. Chem. Soc.* **133**, 18495–18502 (2011).
13. Nazari, B. et al. *Nonomuraea* sp. ATCC 55076 harbours the largest actinomycete chromosome to date and the kistamicin biosynthetic gene cluster. *MedChemComm* **8**, 780–788 (2017).
14. Naruse, N. et al. New antiviral antibiotics, kistamicins A and B. I. Taxonomy, production, isolation, physico-chemical properties and biological activities. *J. Antibiot.* **46**, 1804–1811 (1993).
15. Kwon, Y. J., Kim, H. J. & Kim, W. G. Complestatin exerts antibacterial activity by the inhibition of fatty acid synthesis. *Biol. Pharm. Bull.* **38**, 715–721 (2015).
16. Park, O. K., Choi, H. Y., Kim, G. W. & Kim, W. G. Generation of new complestatin analogues by heterologous expression of the complestatin biosynthetic gene cluster from *Streptomyces chartreusis* AN1542. *Chem BioChem* **17**, 1725–1731 (2016).
17. Czarny, T. L., Perri, A. L., French, S. & Brown, E. D. Discovery of novel cell wall-active compounds using P_{vanG}, a sensitive reporter of cell wall stress, in the model Gram-positive bacterium *Bacillus subtilis*. *Antimicrob. Agents Chemother.* **58**, 3261–3269 (2014).
18. D'Elia, M. A., Millar, K. E., Beveridge, T. J. & Brown, E. D. Wall teichoic acid polymers are dispensable for cell viability in *Bacillus subtilis*. *J. Bacteriol.* **188**, 8313–8316 (2006).
19. Kuru, E., Tekkam, S., Hall, E., Brun, Y. V. & Van Nieuwenhze, M. S. Synthesis of fluorescent D-amino acids and their use for probing peptidoglycan synthesis and bacterial growth in situ. *Nat. Protoc.* **10**, 33–52 (2015).
20. Dominguez-Cuevas, P., Porcelli, I., Daniel, R. A. & Errington, J. Differentiated roles for MreB-actin isologues and autolytic enzymes in *Bacillus subtilis* morphogenesis. *Mol. Microbiol.* **89**, 1084–1098 (2013).
21. Blackman, S. A., Smith, T. J. & Foster, S. J. The role of autolysins during vegetative growth of *Bacillus subtilis* 168. *Microbiology* **144**, 73–82 (1998).
22. Diethmaier, C. et al. A novel factor controlling bistability in *Bacillus subtilis*: the YmdB protein affects flagellin expression and biofilm formation. *J. Bacteriol.* **193**, 5997–6007 (2011).
23. Meisner, J. et al. FtsEX is required for CwlO peptidoglycan hydrolase activity during cell wall elongation in *Bacillus subtilis*. *Mol. Microbiol.* **89**, 1069–1083 (2013).
24. Yepes, A. et al. The biofilm formation defect of a *Bacillus subtilis* flotillin-defective mutant involves the protease FtsH. *Mol. Microbiol.* **86**, 457–471 (2012).
25. Monahan, L. G., Hajduk, I. V., Blaber, S. P., Charles, I. G. & Harry, E. J. Coordinating bacterial cell division with nutrient availability: a role for glycolysis. *MBio* **5**, e00935-14 (2014).
26. Kudrin, P. et al. Subinhibitory concentrations of bacteriostatic antibiotics induce *relA*-dependent and *relA*-independent tolerance to β-lactams. *Antimicrob. Agents Chemother.* **61**, e02173-16 (2017).
27. Jolliffe, L. K., Doyle, R. J. & Streips, U. N. The energized membrane and cellular autolysis in *Bacillus subtilis*. *Cell* **25**, 753–763 (1981).
28. Calamita, H. G., Ehringer, W. D., Koch, A. L. & Doyle, R. J. Evidence that the cell wall of *Bacillus subtilis* is protonated during respiration. *Proc. Natl Acad. Sci. USA* **98**, 15260–15263 (2001).
29. Yamaguchi, H., Furuhata, K., Fukushima, T., Yamamoto, H. & Sekiguchi, J. Characterization of a new *Bacillus subtilis* peptidoglycan hydrolase gene, *yvcE* (named *cwlO*), and the enzymatic properties of its encoded protein. *J. Biosci. Bioeng.* **98**, 174–181 (2004).
30. Fukushima, T., Yao, Y., Kitajima, T., Yamamoto, H. & Sekiguchi, J. Characterization of new L,D-endopeptidase gene product CwlK (previous YcdD) that hydrolyzes peptidoglycan in *Bacillus subtilis*. *Mol. Genet. Genomics* **278**, 371–383 (2007).
31. Davies, A. Peptidoglycan Architecture and Dynamics in *Bacillus subtilis*. PhD thesis, Univ. Sheffield (2014).
32. Pletzer, D., Mansour, S. C., Wuerth, K., Rahanjam, N. & Hancock, R. E. W. New mouse model for chronic infections by Gram-negative bacteria enabling the study of anti-infective efficacy and host–microbe interactions. *MBio* **8**, e00140-17 (2017).
33. Chiang, N. et al. Infection regulates pro-resolving mediators that lower antibiotic requirements. *Nature* **484**, 524–528 (2012).
34. Kugelberg, E. et al. Establishment of a superficial skin infection model in mice by using *Staphylococcus aureus* and *Streptococcus pyogenes*. *Antimicrob. Agents Chemother.* **49**, 3435–3441 (2005).
35. Guo, J., Ran, H., Zeng, J., Liu, D. & Xin, Z. Tafuketide, a phylogeny-guided discovery of a new polyketide from *Talaromyces funiculosus* Salicorn 58. *Appl. Microbiol. Biotechnol.* **100**, 5323–5338 (2016).
36. Atilano, M. L. et al. Bacterial autolysins trim cell surface peptidoglycan to prevent detection by the *Drosophila* innate immune system. *eLife* **3**, e02277 (2014).
37. Humann, J. & Lenz, L. L. Bacterial peptidoglycan degrading enzymes and their impact on host muropeptide detection. *J. Innate Immun.* **1**, 88–97 (2009).
38. Skalweit, M. J. & Li, M. Bulgecin A as a β-lactam enhancer for carbapenem-resistant *Pseudomonas aeruginosa* and carbapenem-resistant *Acinetobacter baumannii* clinical isolates containing various resistance mechanisms. *Drug Des. Devel. Ther.* **10**, 3013–3020 (2016).

Article

Methods

Data reporting

No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

Strains and culture conditions

All strains, sourced from this study and previous studies[17,39–41], are listed in Extended Data Table 2. Strains were grown under standard culturing conditions at 37 °C in either Mueller Hinton Broth (MHB; BD Biosciences) or Lennox Broth (LB; Bioshop). Antibiotics were supplemented as necessary for *B. subtilis* strains (7.5 μg ml^{-1} kanamycin for BKK knockout collection strains, 1 μg ml^{-1} erythromycin and 12 μg ml^{-1} lincomycin for BKE knockout collection strains, 10 μg ml^{-1} chloramphenicol for pSWEET integrated strains). Antibiotics were purchased from Sigma, except for moenomycin (Cayman Chemicals), ampicillin and kanamycin (Bioshop).

Cluster identification

We sourced GPA BGC clusters consisting of previously published sequences (31 BGCs), and used GPA fingerprint sequences[6] and BLASTp followed by antiSMASH[42] to identify new BGCs from GenBank (18 BGCs) and our in-house collection (22 BGCs). Any sequence with a hit (*e*-value < 1 × 10^{-5}) was subsequently run through antiSMASH (v.4.2, options "–smcogs–knowncluster blast–full-hmmer"), and the results were searched for a GPA or GPA-like cluster, which we defined as a non-ribosomal peptide synthase (NRPS) cluster with 4-hydroxyphenylglycine and/or 3,5-dihydroxyphenylglycine biosynthesis, and optionally tailoring enzymes (halogenase, glycosyltransferase, acylation, methyltransferase, etc). This generated a list of 71 whole or partial clusters from whole or partial genome sequences, and included known published and putative unknown BGCs. A more complete analysis of these clusters is found in a previous study[9].

Tree construction

The NRPS detection rules in antiSMASH identify well-known functional domains in the large multimodular synthase genes to facilitate structural predictions of the products of the NRPS enzymes. These domains include the peptidyl-carrier protein or thiolation domains (T-), adenylation (A-) domains that activate precursor amino acids before they can be incorporated into the peptide scaffold, epimerization (E-) domains that catalyse amino acid conversion from D to L epimers, X-domains that facilitate oxidative crosslinking[43], and condensation (C-) domains that catalyse the formation of peptide bonds. The sequences identified as C-domains were aligned using MUSCLE (default parameters), and manually inspected for the presence of gaps and misaligned regions. This alignment was used to compute a maximum-likelihood (ML) tree using FastTree 2, using the WAG substitution model and CAT approximation with the default number of rate categories. Each block of C-domains labelled in the phylogeny for the numbered positions in GPA scaffolds is robust, with SH-like support values > 0.949 except for one (0.891 for C+2 corbomycin). Further bioinformatic methods are available in a previous study[9].

Owing to variation in the number of NRPS-encoded condensation domains in each BGC, we applied a labelling system for condensation domains on the basis of location with respect to the shared centrally encoded 4-hydroxyphenylglycine, position 0, to enable us to visualize trends within this tree (Extended Data Fig. 1a). Using this labelling system, the relationship between condensation domains of different modules and between strains within an individual module can be observed (Extended Data Fig. 1b). The location C-2 in this domain structure labelling scheme is shown as an example in Fig. 1a.

Antibiotic susceptibility testing

Antibiotic susceptibility was tested using standard procedures in MHB for most strains. *Enterococcus* strains were grown in brain heart infusion medium (BD Biosciences). *Streptomyces* MICs were determined in tryptic soy broth (TSB; BD Biosciences) with 0.5% yeast extract. *Neisseria gonorrhoeae* inoculum for MIC determination was prepared from strains streaked on enriched chocolate agar and resuspended in GW liquid medium[44], and microtitre plates were incubated with shaking at 500 rpm, 37 °C, 5% CO_2 and 90% humidity for 12 h. For MIC determination, daptomycin and bacitracin were supplemented with 1.25 mM $CaCl_2$ or 40 μg ml^{-1} $ZnCl_2$ respectively. Two-dimensional checkerboards were set up according to standard protocol.

Fermentation and purification of corbomycin

Structural predictions for corbomycin were informed using phylogenetic trees of NRPS adenylation domains, including 3,5-dihydroxyphenylglycine specificity, which is notoriously difficult to predict for this amino acid[45]. The NPRS was predicted to encode a nonapeptide, in contrast to the heptapeptide backbone of most GPA antibiotics. These structural predictions guided the initial purification of corbomycin, independent of antibiotic activity.

WAC01529 mycelium from 50 ml TSB seed culture was inoculated into each of eighteen 2.8-litre flasks containing 600 ml Bennett's medium (1% potato starch, 0.2% casamino acids, 0.18% yeast extract, 0.02% KCl, 0.02% $MgSO_4 \cdot 7H_2O$, 0.024% $NaNO_3$, 0.0004% $FeSO_4 \cdot 7H_2O$). After 4 days, fermentations were fed with 0.2 mM each of cysteine, histidine, glutamine and tyrosine. Amino acid supplements were prepared as 100× stocks in water (Cys, His, Glu) or 10 mM HCl (Tyr) and neutralized with two equivalents of $NaHCO_3$ after addition to fermentations.

Spent medium was extracted with 8% (w/v) HP-20 (Diaion) resin. Cell pellets were extracted twice with 500 ml MeOH and concentrated under vacuum with 100 g HP-20 (Diaion) resin. These resins were combined and eluted with H_2O (2 l), 20% MeOH (2 l), 40% MeOH (2 l) and 100% MeOH (4 l). Analysis of fractions by HPLC and liquid chromatography coupled with mass spectrometry (LC–MS) identified a peak with a molecular mass (around 1,300–1,600 Da) and a UV-absorption profile (maxima at 220 nm and 280 nm) that were consistent with the predicted structure. This fraction (100% MeOH) was extracted with ethyl acetate, MeOH/H_2O (1:4) and DMSO. The DMSO subfraction was found to contain the predicted glycopeptide and was applied to reverse-phase CombiFlash ISCO (RediSep Rf C18, Teledyne) and eluted with a linear gradient system (5–100% water/acetonitrile, 0.1% formic acid) to give 136 fractions. Fractions containing the predicted glycopeptide were combined and subjected to a Sephadex LH-20 column (400 ml), eluting with MeOH/acetonitrile/H_2O (1:2:1), to yield 36 subfractions. Identified subfractions were combined and further purified with an Agilent Eclipse XDB-C8 column (5 μm, 9.4 × 250 mm), to yield 23.6 mg of corbomycin.

Fermentation and purification of complestatin

Streptomyces sp. WAC01325 was fermented using conditions identical to those used for *Streptomyces* sp. WAC01529, except for amino acid feeding. After 3 days growth at 30 °C, 250 rpm, fermentations were fed with 0.2 mM each of 4-hydroxyphenylglycine, tryptophan and tyrosine. Amino acids were prepared as a 100× stock solution in 10 mM HCl and neutralized with two equivalents of $NaHCO_3$ after addition to fermentations. Growth was allowed to continue for a total of 8 days and complestatin was purified from spent medium and cell pellet in a similar manner to corbomycin. High-resolution electrospray ionization mass spectrometry (HR-ESI-MS) and ^{1}H nuclear magnetic resonance (NMR) confirmed the compound as complestatin.

Structural characterization of corbomycin

To elucidate its structure, corbomycin was subjected to 1D and 2D NMR and HR-ESI-MS experiments; all structural data are shown in Supplementary Figs. 1–6 and Supplementary Table 1. One- and two-dimensional NMR experiments were performed using a Bruker AVIII 700 MHz instrument equipped with a cryoprobe in deuterated DMSO. Chemical shifts are reported in parts per million (ppm) relative to tetramethyl

silane using the residual solvent signals at 2.50 ppm in proton NMR and 39.5 ppm in carbon NMR as internal signals. HR-ESI-MS data were acquired using an Agilent 1290 UPLC separation module and qTOF G6550A mass detector in negative-ion mode. Interpretation of MS and NMR data for structural determination are provided in the Supplementary Discussion.

fab and *cwlO* overexpression and knockout in *B. subtilis* 168

Full length *fabI*, *fabL*, *cwlO* and truncated $cwlO_{1-407}$ were cloned from *B. subtilis* 168 gDNA into pSWEET[46] for overexpression under a xylose-inducible promoter[46]. Both pSWEET-*bgaB* and PCR amplicons were digested with the restriction enzymes PacI and BamHI before ligation and electroporation into *E. coli* Top10 and selection with 100 μg ml^{-1} ampicillin. For transformation into *B. subtilis* 168, 1 μg of plasmid DNA was digested with PstI and transformed as previously described using selection with 10 μg ml^{-1} chloramphenicol[41]. As *B. subtilis* Δ*cwlO* displays loss of genetic competence[47], *B. subtilis* Δ*cwlO* pSWEET combination strains were generated by subsequently transforming 2 μg BKE34800 gDNA into respective pSWEET-containing strains and selected with appropriate antibiotics. For antibiotic-susceptibility testing, overexpression of genes was induced using 3% w/v xylose in LB medium.

P_{ywaC} luminescence testing

Activation of P_{ywaC} was tested in *B. subtilis* EB1385 as previously described with minor modifications[17]. EB1385 was grown overnight in MHB with 1 μg ml^{-1} erythromycin, then diluted to an optical density at 600 nm (OD_{600}) of 0.15 in fresh MHB. 100 μl of inoculum was dispensed in a white 96-well plate with a clear bottom. Antibiotics were added in triplicate in a twofold serial dilution from 16 μg ml^{-1} to 0.125 μg ml^{-1}. The plate was covered with an optically clear film and incubated at 37 °C with shaking, with OD_{600} and luminescence (0.1-s integration time) monitored using an EnVision Multilabel plate reader (PerkinElmer) at 10-min intervals. Luminescence is reported for the lowest concentration that fully inhibited growth at this inoculum (2 μg ml^{-1} vancomycin, 4 μg ml^{-1} corbomycin, 8 μg ml^{-1} complestatin, 1 μg ml^{-1} triclosan).

$DiSC_3$ dye fluorescence assay

Membrane permeability was tested using the voltage-sensitive dye $DiSC_3$. Mid-log phase *B. subtilis* 168 grown in LB ($OD_{600} \approx 0.6$) was washed twice in 5 mM HEPES (pH 7.3), and finally resuspended in 5 mM HEPES with 20 mM glucose (pH 7.3), adjusting the OD_{600} to 0.1. A portion of the cell suspension (200 μl) was dispensed in a black 96-well plate and $DiSC_3$ dye in DMSO was added to a final concentration of 1 μM. Cells were incubated at room temperature for 30 min with shaking to allow for dye uptake, and fluorescence quenching was monitored using a Synergy H1 microplate reader (excitation/emission 600 nm/660 nm). When fluorescence stabilized, test compounds were added to a final concentration of 10–12 μM (18 μg ml^{-1} gramicidin, 11 μg ml^{-1} valinomycin, 16 μg ml^{-1} vancomycin, corbomycin and complestatin) and fluorescence was monitored. Maximal fluorescence, observed 250 s after compound addition for all antibiotics, is reported. The experiment was performed in triplicate on two independent occasions.

UDP-MurNAc-PP accumulation in *S. aureus*

UDP-MurNAc-PP accumulation was measured in *S. aureus* as previously described[48]. In brief, mid-log phase *S. aureus* ATCC 29213—pretreated with 130 μg ml^{-1} chloramphenicol—was split into 10-ml aliquots and exposed to 10× MIC test antibiotic (10 μg ml^{-1} corbomycin, 20 μg ml^{-1} complestatin, 1.25 μg ml^{-1} ampicillin, 10 μg ml^{-1} vancomycin, 1.25 moenomycin, 160 μg ml^{-1} bacitracin zinc, 20 μg ml^{-1} fosfomycin, 160 μg ml^{-1} carbenicillin, 80 μg ml^{-1} mecillinam, 640 μg ml^{-1} aztreonam). After 1 h incubation at 37 °C, 250 rpm, cells were pelleted and extracted with boiling water. Soluble extracts were lyophilized, then resuspended in 100 μl water and 20 μl was analysed on an Inertsil ODS-4 4 × 150 mm column run with isocratic elution (0.5 ml min^{-1}, 50 mM sodium phosphate buffer, pH 5.2) at 37 °C. The results reported are representative of two independent experiments.

Determination of colony-forming units for bacteriostatic drugs

Mid-log phase *B. subtilis* 168 (OD_{600} = 0.25) in LB was dispensed into 1.5 ml aliquots and treated with 4 × MIC (4 μg ml^{-1}) or 8 × MIC (8 μg ml^{-1}) complestatin, corbomycin or chloramphenicol, or left untreated. Cells were incubated at 37 °C for 6 h, then cells were washed twice with LB to remove antibiotic that could inhibit growth, OD_{600} was adjusted to 0.4 and colony-forming units (cfu) were enumerated in triplicate. For growth curves, these cells were inoculated 1:200 into fresh LB and dispensed into a microtitre plate. The plate was covered with a clear, breathable film and OD_{600} was monitored on a Tecan Sunrise microplate reader at 37 °C with shaking.

Single-step selection of spontaneous resistant mutants

Single-step selection of resistant mutants in *B. subtilis* 168 and *S. aureus* ATCC29213 on solid medium was performed according to standard protocol. Mueller Hinton medium with 1% agarose and 4× or 8× solid MIC complestatin or corbomycin (32–64 μg ml^{-1}) were plated with 10^9–10^{10} cfu from an overnight culture. Cfu were enumerated directly from this overnight culture. Plates were incubated for 48 h at 37 °C and no resistant colonies were detected. The experiment was performed on two independent occasions.

Raising resistant mutants through serial passaging

Complestatin- and corbomycin-resistant mutants were raised beginning with the laboratory strain *B. subtilis* 168 or *S. aureus* ATCC 29213. To begin with, a single colony was inoculated into 1 ml MHB in a sterile test tube with 0.25 × MIC, 0.5 × MIC, 1 × MIC and 2 × MIC (where MIC = 1 μg ml^{-1} for both complestatin and corbomycin, and 0.0625 μg ml^{-1} for rifampicin). After 24 h growth with shaking, the lowest concentration with no growth was taken as the new MIC, and cells were subcultured into fresh tubes 1 in 100 from the highest concentration that supported growth. This process was continued for 25 days, and glycerol stocks were taken whenever there was a shift in MIC. At the end of 25 days, glycerol stocks were streaked on non-selective medium (Mueller Hinton agar) and single colonies were isolated for two generations. The MIC of purified strains was measured by microbroth dilution. Serial passaging was performed in biological duplicate using two independent lines.

For one line of the serially passaged cells, whole-genome sequencing was performed on the strain on the final day (COM25 and COR25), or at the earliest time point that the highest MIC was reached. For corbomycin, this strain arose at day 14 (*B. subtilis* COR14), and for complestatin, day 20 (*B. subtilis* COM20). Whole-genome sequencing on resistant mutants, as well as our laboratory *B. subtilis* 168, was performed with Illumina MiSeq (300 bp, paired-end reads) by the Farncombe Genomics Facility. To identify mutations unique to our evolved mutants compared with the wild type, each of the three sequenced strains were compared to the published *B. subtilis* 168 reference genome (GenBank accession number AL009126.3) using breseq (v.0.33.1)[49] to generate a list of differences. Changes in protein-coding regions that were unique to resistant mutants and not present in our laboratory *B. subtilis* 168 strain were identified for follow-up. Sequencing two individual colonies isolated from day 25 gave identical genotypes.

Cell lysis assay

Early exponential phase *B. subtilis* 168 or *S. aureus* ATCC 29213 ($OD_{600} \approx$ 0.25) grown in LB medium was dispensed (100 μl per well) into a round-bottom 96-well plate. Antibiotics or $MgSO_4$ were supplemented to the appropriate concentration, and appropriate lytic agents were added (50 μg ml^{-1} fosfomycin, 100 μg ml^{-1} ampicillin, 75 mM sodium azide). Excess Mg^{2+} inhibits azide-induced lysis by a mechanism that is not well-understood but is thought to be partially through the modulation of autolysin activity[50,51]. The plate was covered with a clear, breathable

film and OD_{600} was monitored on a Tecan Sunrise microplate reader for 10 h at 37 °C with shaking.

Autolysin overexpression and purification from *E. coli*

CwlK and the catalytic CwlO domain with signal peptides removed were cloned into pET28a (EMD Biosciences) using primers CwlK_Full-F (GTCACCCATGGGCCATGAATGGCATTCTCAAAA) and CwlK_Full-R (GTCACCTCGAGGTTAGGAATCATCTCCAAGTG), or CwlO_Cat-F (GTCACCCATGGGCACTGTTATCAGCAACTCTGG) and CwlO-R (GTCACCTCGAGTTGAACAACACGTCTTACAAC), respectively, and NcoI and XhoI restriction sites for introduction of C-terminal 6×His. Cloning was performed in *E. coli* Top10 followed by transformation into *E. coli* BL21(DE3) pLysS for expression. For expression, 1 l of LB medium supplemented with 50 μg ml^{-1} kanamycin and 35 μg ml^{-1} chloramphenicol was inoculated 1:50 from an overnight culture and grown at 37 °C, 250 rpm until OD_{600} reached 0.6. Cells were induced with 2 mM isopropyl β-D-1-thiogalactopyranoside (IPTG) at 37 °C for 3 h (CwlO), or 18 h at 18 °C with 1 mM IPTG (CwlK). Cell pellets were resuspended in 20 ml lysis buffer (25 mM HEPES (pH 8), 300 mM NaCl, 10 mM imidazole) and frozen at −80 °C. For purification, pellets were thawed, treated with Pierce protease inhibitor tablets (Thermo Fisher Scientific), 10 mg ml^{-1} lysozyme and 5 μg ml^{-1} DNase, and lysed on ice by sonication. Clarified lysates were loaded onto equilibrated 2 ml Ni-NTA agarose (Qiagen), washed (25 mM HEPES (pH 8), 300 mM NaCl, 25 mM imidazole) and eluted (25 mM HEPES (pH 8), 300 mM NaCl, 250 mM imidazole). Elutions were dialysed overnight in buffer A for downstream anion or cation exchange. CwlO was purified by cation exchange with a HiTrap SP HP 5 ml column (GE Healthcare) on an AKTA Explorer system (buffer A: 25 mM HEPES (pH 8); buffer B: 25 mM HEPES (pH 8), 1 M NaCl). CwlK was similarly purified by anion exchange on a HiTrap Q HP 5 ml column (GE Healthcare) (buffer A: 25 mM Tris-HCl (pH 8); buffer B: 25 mM HEPES (pH 8), 1 M NaCl). Buffer exchange and concentration of fractions containing pure protein was performed using Amicon Ultra centrifugal filters with a 3,000 Da cut-off. Protein purity was >95% as assessed by SDS–PAGE analysis.

Specific activity of the purified enzymes was confirmed by performing peptidoglycan digestion, as described below, with inactivated enzyme. CwlK—a Zn^{2+}-dependent metalloprotease[30]—could be inhibited by addition of 0.5 mM EDTA, whereas CwlO—a NplC/P60 cysteine peptidase[29]—could be inhibited by preincubation with 4 mM maleimide.

Peptidoglycan isolation and digestion

To prepare peptidoglycan as a substrate for binding or digestion experiments, *B. subtilis* 168 or *S. aureus* ATCC 29213 was grown to OD_{600} 0.6–0.7 in LB medium and peptidoglycan was purified as previously described with minor modification[52]. In brief, cell pellets were boiled in 4% SDS, then washed, sonicated to break up sacculi, treated with α-amylase and DNase, and finally digested with pronase overnight. The peptidoglycan was again boiled in 2% SDS, washed, and teichoic acids were hydrolysed with 1 M HCl. The pure peptidoglycan was washed with water until the pH reached 6 and lyophilized.

For enzymatic digestion of peptidoglycan, conditions for each enzyme were optimized as follows: mutanolysin (20 μg ml^{-1}, 1 mg ml^{-1} peptidoglycan, 20 mM sodium acetate, pH 6.5), lysozyme (100 μg ml^{-1}, 1 mg ml^{-1} peptidoglycan, 20 mM sodium acetate, pH 6.5), CwlO (20 μg ml^{-1}, 0.5 mg ml^{-1} peptidoglycan, 20 mM MES-NaOH, pH 5.5), CwlK (16 μg ml^{-1}, 1 mg ml^{-1} peptidoglycan, 10 μM $ZnSO_4$, 20 mM sodium acetate, pH 6.5). To test the effect of corbomycin and complestatin on digestion, 100 μl of peptidoglycan in buffer, as indicated, was preincubated with either antibiotic or a corresponding volume of DMSO with shaking for 30 min, then dispensed into a round-bottom 96-well plate and enzyme added. The plate was covered with a clear, breathable film to prevent evaporation, and OD_{600} was tracked on a Tecan Sunrise microplate reader shaking at 37 °C. Results show average and standard deviation of triplicate wells.

FDAA imaging

B. subtilis 3610 (wild type) was grown in LB medium at 37 °C to exponential phase (OD_{600} 0.1–0.2), then diluted with fresh LB medium 1:10 and allowed to grow to an OD_{600} of 0.2. Complestatin and corbomycin were added to exponentially growing cell cultures (0.3 ml) to a final concentration of 0.6× or 2× MIC and incubated for 5 min at 37 °C. Alternatively, ampicillin was added at 0.5×, 1×, 2× and 5× MIC and incubated for 1 min at 37 °C. FDAAs (100 mM DMSO solution) were added to the cultures to a final concentration of 0.5 mM and incubated at 37 °C with shaking for 1 h (complestatin and corbomycin). A shorter incubation time (5 min) was used for ampicillin-treated cells to prevent lysis. Cells were fixed by adding pure ethanol to the cultures to a final concentration of 70% (v/v), incubating on ice for 1 h, and finally washing twice with PBS. Failure to label cells with the L-isomer HCC-amino-L-alanine (HALA) confirmed that incorporation was conducted by transpeptidases (Extended Data Fig. 3c).

For cell imaging, cells were applied to a coverslip (24 × 50 mm; 1.5) and covered with an 8 × 8-mm wide, 2-mm thick PBS-agarose pad (SeaKem LE Agarose). The coverslip–pad combination was placed onto a customized slide holder on the microscope with the pad facing upwards. Phase-contrast and fluorescence images were acquired using a Nikon Ti-E inverted microscope equipped with a 1.4 numerical aperture Plan Apo 60× oil objective and Andor iXon EMCCD camera. NIS-Element AR (v.5.02) was used for image acquisition. Identical conditions (exposure, light source power and electron-multiplying gain) were applied to all the samples for valid comparisons. Filters for HADA imaging were as follows: excitation 395/25 nm (DAPI) and emission 435/26 nm.

Image processing was performed in FIJI (v.1.51). Images were scaled without interpolation, cropped and rotated. Linear adjustment was performed to optimize contrast and brightness of the images. Quantitative measurement of FDAA labelling intensity was achieved using a FIJI plugin, MicrobeJ, in which cells were identified in the phase contrast channel with width limit from 0.3 μm to 2 μm and length greater than 1 μm. FDAA labelling intensity was then quantified and averaged so that $n > 100$ for each condition.

Peptidoglycan binding assay

To assess peptidoglycan binding, 1.6 mg ml^{-1} antibiotic dissolved in DMSO was diluted to 0.1 mg ml^{-1} in 20 mM phosphate buffer (pH 7.1) with 0 mg ml^{-1}, 0.1 mg ml^{-1}, 0.5 mg ml^{-1}, 1 mg ml^{-1}, 5 mg ml^{-1} or 10 mg ml^{-1} peptidoglycan for 0,1:1, 5:1, 10:1, 50:1 and 100:1 w/w ratios, respectively. *B. subtilis* 168 and *S. aureus* ATCC 29213 peptidoglycan was tested similarly. Reaction volumes were adjusted so that a minimum of 24 μg peptidoglycan was used in the lowest ratio condition to enable a visible pellet of peptidoglycan to form. Mixtures were incubated for 1–2 h at 37 °C, then centrifuged for 10 min to thoroughly pellet insoluble material. 50 μl of supernatant containing unbound antibiotic (maximum 5 μg) was injected and detected by HPLC using a WAT094269 Symmetry Shield RP 8 3.5 μM 4.6 × 150 mm column and the following gradient at 0.8 ml min^{-1}, 40 °C: 0–2 min 0% B, 2–12 min 0–95% B, 12–18 min 95% B, 18–20 min 95–0% B (buffer A: H_2O + 0.1% formic acid, buffer B: acetonitrile + 0.1% formic acid).

Antibiotic–BODIPY semisynthesis

Vancomycin–BODIPY FL was synthesized as described previously[53]. Complestatin and corbomycin were derivatized by combining equimolar amounts of antibiotic, BODIPY FL ethylenediamine (Invitrogen), *N*,*N*-diisopropylethylamine and coupling reagents. For complestatin, 2-(1*H*-benzotriazol-1-yl)-1,1,3,3-tetramethyluronium hexafluorophosphate (HBTU) plus ethyl cyanohydroxyiminoacetate (oxyma) was used, and for corbomycin *N*,*N*′-diisopropylcarbodiimide (DIC) plus oxyma was used. All components were dissolved in dimethylformamide (DMF), except for corbomycin and complestatin, which were dissolved in DMSO. The reaction was carried out at room temperature for 24–48 h

then purified by semi-preparative HPLC using an XSelect CSH Prep C18 5 µm 10 × 100 mm column with a gradient of H_2O + 0.1% formic acid and acetonitrile + 0.1% formic acid. The derivatives were verified by HR-ESI-MS using an Agilent 1290 Infinity II Series HPLC system and 6550 iFunnel QTOF LC–ESI/MS as follows: corbomycin–BODIPY $[M + 2H]^{2+}$: calculated 912.3092, found 912.3094; complestatin–BODIPY $[M + H]^+$: calculated 1,642.2856, found 1,642.2832.

Complestatin–BODIPY and corbomycin–BODIPY had MICs against *B. subtilis* 168 of 16 µg ml^{-1}, and showed a similar fourfold increase in MIC against their respective resistant mutants (*B. subtilis* COM20 and COR14 = 64 µg ml^{-1}). Growth at sub-MIC levels produced twisted chains of cells (Extended Data Fig. 7d), similar to underivatized antibiotics. Therefore these derivatives have the same mechanism as the parent compounds.

Fluorescence microscopy

To label active peptidoglycan synthesis in antibiotic-treated cells, mid-log phase *B. subtilis* 168 (OD_{600} around 0.5) grown in LB medium was first pretreated with 10 × MIC for 30 min at 37 °C (10 µg ml^{-1} corbomycin, 20 µg ml^{-1} complestatin, 0.039 µg ml^{-1} ampicillin, 640 µg ml^{-1} aztreonam). Vancomycin–BODIPY was then added along with unlabelled vancomycin each at 1 µg ml^{-1}, as previously reported[53]. Cells were incubated at room temperature for 15 min with shaking, then washed three times with 20 mM HEPES buffer (pH 8) to remove unbound compound.

For staining with complestatin–BODIPY and corbomycin–BODIPY, *B. subtilis* 168 was similarly grown to mid-log phase in LB, then collected by centrifugation and resuspended in 20 mM HEPES buffer (pH 8) for rod staining, or MSM buffer (20 mM $MgCl_2$, 0.5 M sucrose, 20 mM maleic acid, pH 7) for protoplast generation. For protoplasts, cells in osmotically protective MSM buffer were treated with 4 mg ml^{-1} lysozyme for 1 h at 37 °C, then centrifuged and resuspended in fresh MSM buffer before staining. Rods or protoplasts were stained with 2 × MIC complestatin–BODIPY or corbomycin–BODIPY (32 µg ml^{-1}) for 15 min with shaking, then washed four times in their respective buffers to remove unbound compound.

Prepared cells were mounted on polylysine-treated glass and imaged using a Nikon Eclipse Ti inverted microscope with a fluorescein isothiocyanate fluorescence filter for BODIPY (excitation/emission 470/520 nm). Equivalent exposure time and image adjustments were used for rod- and protoplast-stained cells. FM 4-64FX (Invitrogen) was used where indicated with the corresponding fluorescence filter.

Transmission electron microscopy

To prepare cells for imaging, *B. subtilis* 168 was grown in LB medium with 0.6 µg ml^{-2} corbomycin or 1 µg ml^{-2} complestatin to OD_{600} 0.5–0.7, then collected by centrifugation, resuspended in 2% glutaraldehyde (2% v/v) in 0.1 M phosphate buffer (pH 7.4) and allowed to fix overnight at 4 °C. The samples were prepared for TEM as previously described[54] and viewed in a JEOL JEM 1200 EX TEMSCAN transmission electron microscope (JEOL) operating at an accelerating voltage of 80 kV. Images were acquired with an AMT 4-megapixel digital camera (Advanced Microscopy Techniques).

Mouse skin infection model

All mouse experiments were performed in the Central Animal Facility at McMaster University under animal use protocol 17-03-10 as approved by the Animal Research Ethics Board. We have complied with all relevant ethical regulations. Six- to ten-week-old female BALB/c mice (Charles River, 028) were used for all experiments. The wounded area was prepared as previously described[34]. In brief, anaesthesia was induced with 5% isoflurane and maintained at 2.5%. An area of approximately 2 cm^2 on the dorsal region of the neck was stripped using 25 autoclave tape strips (3M) until the skin was visibly shiny but not bleeding. For bacterial preparation, *S. aureus* ATCC 33591 was grown overnight in TSB medium, and subcultured to $OD_{600} \approx 0.4$. Bacteria were adjusted to 5×10^6 cfu ml^{-1} and a 20 µl volume (10^5 cfu) was applied to the wounded area. Mice in groups of two, three or four were randomly relocated from housing cages into singly housed treatment and control cages. Test antibiotics were formulated in petroleum jelly (Vaseline) at 1% (w/w) and animals were treated 8 times over a 33-h period at the following times post-infection: 1, 4, 8, 12, 20, 24, 28, 32 h. Each treatment was 30 mg petroleum jelly formulation by weight, except for the treatment at 12 h which was 50 mg. A dose of fusidic acid was chosen to give similar molar concentrations to complestatin and corbomycin and to mimic their bacteriostatic effect, as fusidic acid is mainly bacteriostatic but becomes bactericidal at high concentrations[55]. Control mice received 10% DMSO in petroleum jelly. Mice were euthanized at 33 h post-infection. At the end-point, an approximately 2 cm^2 area of skin was excised and homogenized in 1 ml PBS for 15 min at 30 rps (Retsch MM400). Homogenates were serially diluted and plated on tryptic soy agar containing oxacillin (2 µg ml^{-1}) for cfu determination. Six animals (technical replicates) were used in each treatment group, divided among two experiments performed on independent occasions (biological replicates). Nine vehicle control mice were used with at least two control mice in each batch of biological replicates performed.

Statistical analysis

Statistical analysis of the mouse skin infection model was performed using GraphPad v.6. Significance was tested by one-way ANOVA on ranks with Kruskal–Wallis test and Dunn's post hoc analysis for multiple comparisons ($n = 6$ for complestatin, corbomycin and fusidic acid, $n = 9$ for vehicle). Mean ranks are as follows: vehicle 27.78, 1% complestatin 10.33, 1% corbomycin 11.67, 0.25% fusidic acid 6.5.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

Phylogenetic trees, including those of 84 GPA genes and gene segments from 71 BGCs that were analysed in this study, are available at https://github.com/waglecn/GPA_evolution. *Streptomyces* sp. WAC01529 and *Streptomyces* sp. WAC01325 genome sequences are available in GenBank with accession numbers NZ_CP029617.1 and NZ_QHKK00000000.1, respectively. Source Data for the animal experiments shown in Extended Data Fig. 8 are included online.

39. King, A. M. et al. Aspergillomarasmine A overcomes metallo-β-lactamase antibiotic resistance. *Nature* **510**, 503–506 (2014).
40. Unemo, M. et al. The novel 2016 WHO *Neisseria gonorrhoeae* reference strains for global quality assurance of laboratory investigations: phenotypic, genetic and reference genome characterization. *J. Antimicrob. Chemother.* **71**, 3096–3108 (2016).
41. Koo, B.-M. et al. Construction and analysis of two genome-scale deletion libraries for *Bacillus subtilis*. *Cell Syst.* **4**, 291–305 (2017).
42. Blin, K. et al. antiSMASH 4.0-improvements in chemistry prediction and gene cluster boundary identification. *Nucleic Acids Res.* **45**, W36–W41 (2017).
43. Haslinger, K., Peschke, M., Brieke, C., Maximowitsch, E. & Cryle, M. J. X-domain of peptide synthetases recruits oxygenases crucial for glycopeptide biosynthesis. *Nature* **521**, 105–109 (2015).
44. Wade, J. J. & Graver, M. A. A fully defined, clear and protein-free liquid medium permitting dense growth of *Neisseria gonorrhoeae* from very low inocula. *FEMS Microbiol. Lett.* **273**, 35–37 (2007).
45. Gonsior, M. et al. Biosynthesis of the peptide antibiotic feglymycin by a linear nonribosomal peptide synthetase mechanism. *ChemBioChem* **16**, 2610–2614 (2015).
46. Bhavsar, A. P., Zhao, X. & Brown, E. D. Development and characterization of a xylose-dependent system for expression of cloned genes in *Bacillus subtilis*: conditional complementation of a teichoic acid mutant. *Appl. Environ. Microbiol.* **67**, 403–410 (2001).
47. Liu, T.-Y., Chu, S.-H. & Shaw, G.-C. Deletion of the cell wall peptidoglycan hydrolase gene *cwlO* or *lytE* severely impairs transformation efficiency in *Bacillus subtilis*. *J. Gen. Appl. Microbiol.* **64**, 139–144 (2018).
48. Ling, L. L. et al. A new antibiotic kills pathogens without detectable resistance. *Nature* **517**, 455–459 (2015).
49. Deatherage, D. E. & Barrick, J. E. Identification of mutations in laboratory-evolved microbes from next-generation sequencing data using *breseq*. *Methods Mol. Biol.* **1151**, 165–188 (2014).

50. Dajkovic, A. et al. Hydrolysis of peptidoglycan is modulated by amidation of meso-diaminopimelic acid and Mg^{2+} in *Bacillus subtilis*. *Mol. Microbiol.* **104**, 972–988 (2017).
51. Formstone, A. & Errington, J. A magnesium-dependent *mreB* null mutant: implications for the role of *mreB* in *Bacillus subtilis*. *Mol. Microbiol.* **55**, 1646–1657 (2005).
52. Schaub, R. E. & Dillard, J. P. Digestion of peptidoglycan and analysis of soluble fragments. *Bio Protoc.* **7**, 2438 (2017).
53. Tiyanont, K. et al. Imaging peptidoglycan biosynthesis in *Bacillus subtilis* with fluorescent antibiotics. *Proc. Natl Acad. Sci. USA* **103**, 11033–11038 (2006).
54. Yan, A. et al. Transformation of the anticancer drug doxorubicin in the human gut microbiome. *ACS Infect. Dis.* **4**, 68–76 (2018).
55. Verbist, L. The antimicrobial activity of fusidic acid. *J. Antimicrob. Chemother.* **25** (Suppl. B), 1–5 (1990).

Acknowledgements We thank S. French for help with the imaging of vancomycin–BODIPY stained cells, and V. Rao and V. Yarlagadda for MIC measurement of *N. gonorrhoeae*. This research was funded by a Canadian Institutes of Health Research grant (FRN-148463), the Ontario Research Fund, and by a Canada Research Chair to G.D.W.; the work was also supported by National Institute of Health grants R35GM122556 to Y.V.B. and 5R01GM113172 to M.S.V. and Y.V.B., and a Canada 150 Research Chair in Bacterial Cell Biology to Y.V.B. E.J.C. was supported by a CIHR Vanier Canada Graduate Scholarship. N.W. was supported by a CIHR Canada Graduate Scholarship Doctoral Award.

Author contributions E.J.C., N.W. and G.D.W. conceived the study and designed experiments. N.W. performed phylogenetic analysis, structural predictions and resistant mutant genome analysis. W.W. performed compound purification and structural elucidation. E.J.C., A.A.F.-C. and B.K.C. designed animal studies and A.A.F.-C. performed animal experiments. E.J.C. and D.S. performed peptidoglycan and enzyme purification. E.J.C. and K.K. performed the synthesis of BODIPY derivatives. Y.-P.H. and Y.V.B. designed FDAA studies, M.S.V. provided FDAAs and Y.-P.H. performed experiments. E.J.C. performed all other experiments. E.J.C. and G.D.W. prepared the manuscript.

Competing interests E.J.C., N.W., W.W. and G.D.W. are inventors on a provisional patent application that covers the use of complestatin and corbomycin as antimicrobial therapies.

Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-1990-9.
Correspondence and requests for materials should be addressed to G.D.W.
Reprints and permissions information is available at http://www.nature.com/reprints.