Developing genome mining tools for the discovery of bioactive secondary metabolites

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät
der Eberhard Karls Universität Tübingen
zur Erlangung des Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat.)

vorgelegt von
Mohammad Mahdi Alanjary
aus San Diego, USA

Tübingen
2018

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation:	10.10.2018
Dekan:	Prof. Dr. Wolfgang Rosenstiel
1. Berichterstatter:	Prof. Dr. Nadine Ziemert
2. Berichterstatter:	Prof. Dr. Daniel Huson
3. Berichterstatter:	Prof. Dr. Mohamed Donia

Abstract

With the rise of Multi-resistant strains of previously treatable pathogenic microorganisms, some of which immune to all known antibiotics, we face a public health crisis that threatens the lives of anyone prone to infection. This challenge needs to be faced on many fronts and an important step to finding a solution is to replenish our antibiotic arsenals with new drugs that evade current antibiotic resistance strategies. The majority of these compounds have traditionally been sourced from, or inspired by, natural products – compounds produced by living things. This continues to be a valuable resource as the millennia of development through natural selection has made for precisely adapted molecules with desired antibiotic properties. Unfortunately natural products research has experienced stagnation due to high rates of rediscovery and low returns on research investment. Fortunately the widespread use of cheap sequencing technologies, influx of complete whole genomes, and tools used to process them have simultaneously been on the rise. These "genome mining" tools have only begun to highlight chemical potential that has been hidden from traditional approaches from a diverse set of genera. As the detection of various classes of Biosynthetic Gene Clusters (BGCs), areas of the genome responsible for production of these compounds, has matured there are now more leads generated than can be experimentally verified. The problem now is to prioritize these leads for those that have the highest potential for downstream experiments. Common prioritization schemes include: using comparative genomics to highlight unique or shared BGCs, focusing on novel genera besides the traditional prolific producing organisms, and highlighting BGCs that imply antibiotic activity via antibiotic resistance determinates. This research is focused on providing automated and accessible tools to preform these analyses in high-throughput. In addition to the prioritization and de-replication of potential BGCs, applications to enrich for novel leads via resistance determinant and target screening are also presented. As the number of genomes from different taxa begins to rise, shifting from a single genome analysis to a comparative pan-genome approach also shows promise to reinvigorate natural products research. The tools in this research that leverage these approaches will be continually maintained on free public servers for the furthered research and discovery of new antibiotic and anti-infective compounds to ensure the threat of antibiotic resistance is controlled.

Zusammenfassung

Der Anstieg von multiresistenten Mikroorganismen, die zuvor behandelbar waren und von denen einige immun gegenüber allen bekannten Antibiotika sind, hat zu einer öffentlichen Gesundheitskrise geführt. Diese Herausforderung gilt es an verschiedenen Fronten anzugehen. Dabei besteht ein wichtiger Schritt darin, unsere Antibiotika-Arsenale mit neuen Medikamenten aufzustocken, welche es schaffen die derzeit bekannten Resistenzstrategien zu umgehen. Die meisten dieser chemischen Verbindungen wurden traditionell aus Naturstoffen (d.h. von Lebewesen produzierten Substanzen) gewonnen, oder durch diese inspiriert. Nach Jahrtausende währender Entwicklung durch natürliche Selektion weisen diese Naturstoffe sehr präzise Eigenschaften, auf und stellen somit eine wertvolle Ressource für z.B neue Antibiotika dar. Aufgrund hoher Wiederentdeckungsraten und niedriger Forschungsinvestitionen hat die Erforschung von Antibiotika aus natürlichen Produzenten in den letzten Jahrzehnten stagniert. Neue Sequenziertechnologien ermöglichen es nun riesige Datenmengen zu erheben. Parallel dazu wurde eine Fülle an Anwendungen entwickelt die es ermöglichen diese Genomdaten zu verarbeiten. Mit Hilfe dieser "Genom-Mining" Anwendungen wird es möglich das Potenzial zur Produktion von Naturstoffen aufzuzeigen, das verschiedenen Gattungen innewohnt, und das durch traditionelle Aufarbeitungsmethoden bislang nicht zutage kam. Auch Methoden zur Identifikation von Biosnythese-Genclustern (BGCs), d. h. Regionen im Genom, die für die Produktion bestimmter Substanzen verantwortlich sind, werden ständig weiterentwickelt. Das führt dazu, dass immer mehr Biosynthese-Gencluster entdeckt werden, die verantwortlich sein könnten für die Produktion neuer Antibiotika. Das Problem besteht nun darin, diejenigen Gencluster zu priorisieren, die das höchste Potenzial zur Produktion von Antibiotika haben, um den Aufwand für weiterführende Laborexperimente zu minimieren.

Zu den gängigen Methoden bei der Priorisierung von BGCs gehören: 1) Vergleichsmethoden die einzigartige oder häufig vorkommende BGCs unterscheiden; 2) die Fokussierung auf neue Gattungen neben den traditionellen Produzenten und 3) die Identifikation von BGCs mit antibiotischer Aktivität über Antibiotika-Resistenzbestimmungen.

Im Rahmen der vorliegenden Arbeit wurden automatisierte und benutzerfreundliche Computeranwendungen entwickelt, welche die o.g. Priorisierungsmethoden unterstützen und es ermöglichen diese im Hochdurchsatz durchzuführen. Neben der Priorisierung potenzieller BGCs ist es mit Hilfe dieser Anwendungen auch möglich neue, unbekannte BGCs zu identifizieren. Aufgrund der stetigen Zunahme von Genomesequenzen aus unterschiedlichen Taxa und den damit verbunden Möglichkeiten, sind wir überzeugt, dass der Wechsel im Fokus von Einzelgenomanalysen hin zu vergleichenden Genomanalysen einen wiederbelebenden Effekt auf die Naturstoffforschung haben wird.

Es ist unser Ziel die hier entwickelten Anwendungen mit freiem Zugang auf öffentlichen Servern bereitzustellen und instand zu halten um sie für die zukünftige Erforschung und Entwicklung neuer Antibiotika und Antiinfektiva zur Verfügung zu stellen.

Acknowledgments

I would like to thank my advisors Prof. Dr. Nadine Ziemert and Prof. Dr. Daniel Huson for their support and guidance, which has made this work possible. The invaluable discussions for when I was faced with obstacles were always promptly provided with a smile, making for a remarkable learning experience. I am also very grateful to all my colleagues and co-authors for their time and discussions, and of course for making the research environment a memorable one.

To my parents and brothers who have always been supportive and encouraging, I give special thanks. My parents, Jalal and Beverly, taught me the meaning of hard work as they strived for a better future for their children, for which I am eternally grateful. And to Dr. Michelle Schorn, my co-author in academia and in life, I give thanks for the intellectual and emotional support through all the obstacles faced during our research.

To my committee members, thank you for your time and valued feedback. I have had nothing but a welcoming experience at the University of Tuebingen and will always be grateful for the hospitality and vibrant culture here.

I would also like to thank the German Center for Infectious Biology [DZIF 9.704] for their financial support and the University of Tuebingen for the resources provided.

I dedicate this work in the name of that which sets my dividing strands of DNA into motion – it is truly a wonder that deserves recognition.

In accordance with the standard scientific protocol, I will use the personal pronoun we to indicate the world population in general or my scientific collaborators and myself.

Contents

1 Introduction

Antibiotics are commonplace in today's world and are a cornerstone of modern medicine. From prevention in post-surgery applications to subduing a fatal pathogen, it is easy to say that without these medical wonders, countless lives would be lost. Additionally, these compounds have been used as invaluable tools for scientific research that have furthered our understanding of microorganisms. Unfortunately, the efficacies of these discoveries are beginning to wear off as antibiotic resistance continues to rise. Due to widespread use and misuse of these compounds we have seen growing selection for resistant variants of once treatable pathogens. Some of these "superbugs" have even shown to be resistant against all known antibiotic treatments (1). Furthermore, this threat is accelerated by the fact that discovery and development of new antibiotics have slowed (2). The use of the variety of discovered compounds with bactericidal or bacteriostatic properties is roughly a century in the making, a mere blip on the time scale of species evolution. This historical context helps to highlight the integral role these compounds play in human health and in understanding the rise in antibiotic resistance. The historical sourcing of compounds from natural sources, "natural products," also shows how the main avenue of discovery has evolved and where it can be improved. Unfortunately stagnation due to high rates of rediscovery has impeded progress in natural products. While we risk falling into another pre-antibiotic era, we are simultaneously at a unique time for leveraging advances in genomics to reinvigorate discovery. With the ever-increasing number of fully sequenced genomes, enabling large-scale comparative analysis, and the addition of automated methods to process this data, we have the resources to remain in a world with effective antibiotics.

The pre-antibiotic era and arrival of antibiotics

The historical accounts of the "pre-antibiotic" era refer mainly to the pre-World War II period before widespread use of antimicrobial agents. Indeed antibiotics have been around for much longer as life evolved strategies to compete. In human history, there is evidence showing traces of the antibiotic tetracycline in bone marrow from north African fossils during the roman period (3, 4). Fungus-growing ants in the amazon basin have been in symbiosis with antifungal producing bacteria for at least 50 million years (5). The label of

pre-antibiotic merely aims to highlight the sharp contrast of medical practices to those that followed the discovery of targeted anti-microbial treatments.

Microbial infections posed a serious health risk prior to and during the early part of the 20th century. Cholera, tuberculosis, typhus, and syphilis, among other diseases, plagued populations of all ages with staggering death tolls. Tuberculosis (TB), or "white plague", was a leading cause of mortality during the turn of the century in western countries and earned the title "Capitan of the men of death" (6). Caused by an infection of the lungs by *Mycobacterium tuberculosis,* this disease has loomed over mankind throughout history with evidence of infection in a 9000-year-old fossil (7). Likewise, cholera has caused millions of lost lives worldwide with several recorded pandemics before the turn of the century (8). Accounts have even shown that microbial infections eclipsed the number of combat fatalities in wartime, with one study showing "spotted fever" resulted in the death of 600,000 Turkish soldiers (9). Indeed pathogens were a serious threat to public health with little means of defense at the time.

Prior to using targeted treatments against microbial infections, various general antiseptics and methods were employed such as the use of zinc, silver nitrate, and mercury. Injections and oral tablets of mercury were standard treatments for syphilis and later included other heavy metal agents such as arsenic and bismuth. For surviving patients there was risk of lifelong damage from toxicity (10). Of the most widely used and relatively effective tactics were those that caused no harm to the patient. Rest and fresh air were often the major prescription for tuberculosis sufferers during this time but was shown to have high rates of relapse (11). Other tuberculosis treatments included the pneumothorax or plombage surgical procedures, which involved the collapsing of an infected lung to allow it to heal (12). Besides treatment of disease causing infections, surgical wards in general carried the high risk of fatalities before antiseptics and antibiotics were in use. Mortality of amputees was as high as 60% in the early 1900s largely due to gangrene and sepsis (13). Despite the causes of these microbial borne disease being known, with the discovery of microorganisms dating back to the mid-17th century (14), there was a dearth of treatment options that could selectively eradicate them.

While research of various chemotherapies against microbes had already proven to be effective, with the synthetic discoveries of "Salvarsan" (15) and Sulfanilamide (16), their impact remained limited in the medical community due to reported side effects and issues

regarding storage. It wasn't until the arrival of penicillin that widespread antibiotic use began to take hold. Alexandar Fleming discovered the compound in 1928 from the fungus *Penicillium chrysogenum*, and further application by Howard Florey and Ernst Chain earned them the Nobel Prize in 1945 due to its enormous impact on public health (17, 18). This discovery launched natural products research into its own scientific field resulting in the golden age of antibiotics. Penicillin was the first bactericidal compound in the beta-lactam class of antibiotics, which remains the largest class of antibiotics to date. This compound was effective due to the signature characteristic of attacking the production of peptidoglycan, a crucial element of cell walls in gram-positive bacteria. The efficacy of this targeted wonder-drug lead to further natural products research and several other classes of antibiotics from a variety of natural sources soon followed. By 1944 Albert Schatz and Selman Waksman isolated streptomycin from the bacterium *Streptomyces greiseus*, which proved to be the first antibiotic treatment of tuberculosis (19). By halting protein production via binding to the universally shared 30S ribosomal subunit in bacteria, this compound was effective against both gram-positive and gram-negative bacteria, making it one of the first broad-spectrum treatments against a variety of infections (20). The expansion of new molecules and new classes of molecules rapidly proceeded with the addition of chloramphenicols, tetracyclines, and macrolide-lincosamide-streptogramins (MLS) under a decade later (21). With this new set of medical tools, countless lives have been saved, on the order of millions, not only by eradication of pathogenic infections but also through increased post-surgery survival (22). Throughout this period, however, mass production of these compounds and uses in industries such as agriculture, animal husbandry, and aquatic farming have been met with decreasing effectiveness due to the rise in antibiotic resistance (23).

The rise of antibiotic resistance and decline of discovery

Antibiotic resistance is a broad term used to describe a variety of means that help an organism survive antibiotic treatments. It is a natural evolutionary phenomenon whereby resistance is favored when under appropriate selection. Resistance is nearly as old as antibiotic production itself as illustrated with the example of beta-lactamases, enzymes that inactivate drugs such as penicillin by cleaving the amide bond of the beta-lactam ring; these genes have been shown to originate as far back as two billion years ago (24). Different mechanisms can confer resistance, which include general intrinsic factors such as lacking the

antibiotic target to sophisticated methods of cell protection. For instance, a gram-negative bacteria lacks a cell wall and thus would be unaffected by antibiotics targeting peptidoglycan, a key cell wall component. Preventing entry of an antibiotic or making enzymes that export them from the cell have been an effective tactic of defense; these export systems are also notable in their high export efficiency and broad spectrum of substrate specificities (25). For example, the *Escherichia coli* housekeeping efflux system AcrAB-TolC are shown to effectively export 8 different classes of antibiotics (26).

Target replacement is another strategy that involves using a mutated variant that is insusceptible to attack while retaining its function. This has been seen with resistant ribosomal polymerase (*rpoB*), a target of the antibiotic rifampicin in the genus Salinispora (27). As these modified versions may not have the same fitness under normal conditions, other species have been seen to harbor a resistant copy alongside the original (28). Similar to target replacement, target modification or target protection utilizes factors to either chemically alter or block the mechanisms of action on the target to render the antibiotic insusceptible. Examples of methyltransferases that act to alter ribosomal subunits have shown to render the target insusceptible to once fatal antibiotics (29). The protection of DNA gyrase and topoisomerase against Fluoroquinolone antibiotics has also been shown with pentapeptide repeat proteins such as Qnr (30).

Inactivation or degradation of the antibiotic itself is another method that has shown to be prevalent (31). And lastly, global adaption describes cellular or population responses to antibiotic attack that evade potency. For example, bypassing of a susceptible pathway, as is the case with daptomycin resistant enterococci (32), or forming biofilms to protect the population (33). Many of these systems evolved naturally before widespread antibiotic use and in some cases have bifunctional roles that could be involved in other ecological processes (34). Also, random mutations have shown to spawn resistance, even during a short time scale (35). However widespread selection for these traits via dissemination of antibiotics have shown to be a factor leading to the rise in resistant strains (36).

Resistance acquisition though mutational changes and selection over generations has been an important source for the rise of these traits as well. Bacteria also have an accelerated mechanism for acquiring resistance besides mutation - horizontal gene transfer (HGT). The ability to acquire DNA, even directly from the environment (37), has given bacteria an additional edge for rapid adoption of these defense tactics. Usually this involves the

exchange of DNA by a process of conjugation, whereby physical contact and some method of trans-membrane donation occur; additionally, DNA uptake directly from the environment (transformation) and exchange via a vector such as a bacteriophage (transduction) can result in HGT. In some cases, segments of chromosomal DNA can be directly transferred, as seen in *Mycobacterium* species, which can span across species and produce a variety of mosaic variants of recipient cells (38). HGT has been proven to facilitate the spread of resistance and pathogenicity (39, 40), with transfers demonstrated to take only a few hours to fully disseminate into a population as illustrated with conjugal F-plasmids in *E. coli* (41). This further compounds the threat of resistance as pathogenic bacteria can acquire these traits from widespread non-pathogenic sources in a relatively short timeframe (42). HGT has therefore helped to explain the swift response to historical antibiotic use.

When given the proper selection conditions, these traits can be highly advantageous, as illustrated with the rise of resistance in the last few decades. The threat of antibiotic resistance has been recognized early on, as Fleming noticed insufficient doses lead to a population of bacteria that were trained to tolerate penicillin. As hypothesized, the emergence of penicillin resistant staphylococci were documented in the early 1950s and resulted in many countries limiting use to prescription only (43). Methicillin, another beta-lactam antibiotic, was soon used to combat resistant strains, however, the pattern of resistance continued resulting in Methicillin Resistant *Staphylococci aureus* (MRSA). Many pathogens have acquired resistance over the following years including *E. coli*, *Salmonella enterica*, and *Klebsiella pneumoniae,* especially against the widely used beta-lactam class of antibiotics. By 2010, nearly 1000 resistance related beta-lactamases had been discovered (44). As time has progressed, we have seen the rise of strains with combined resistance to multiple classes of antibiotics, also known as Multi Drug Resistant (MDR) bacteria. According to the World Health Organization (WHO), some of these "superbugs" are capable of evading nearly all known compounds such as carbapenem-resistant enterobacteriaceae (CRE), named one of the critical threats to human health (45). In 2014, an estimated two million people were infected with some form of MDR in the United States resulting in 23,000 fatalities, roughly equivalent to that in the European Union (46). The spread of resistance can be far-reaching and rapid, with examples spanning across the European Union and significant increases in as little as 2 years (Figure 1.1)

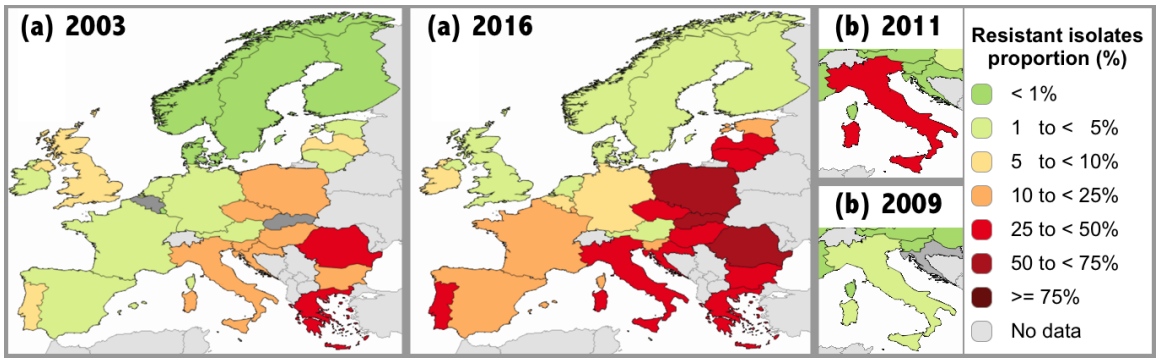

Figure 1.1: (a) Spread of *E. coli* strains resistant to third-generation cephalosporin 2003-2016. (b) Increase in *Klebsiella pneumoniae* strains resistant to carbapenems 2009-2011. Data sourced from European Centre for Disease prevention and Control (ECDC), Atlas of Infections Diseases. http://atlas.ecdc.europa.eu/ accessed May 12 2018.

Today, we are again facing the scourge of tuberculosis with the increase in MDR *Mycobacterium tuberculosis* and extensively drug resistant (XDR) strains, which are resistant to expensive second-line fluoroquinolones (47). The Center for Disease Control (CDC) estimates nearly one-forth of the world population is infected with latent TB with 1.7 million recorded deaths resulting in 2016 (48). Fortunately, the percentage of resistant strains has remained stable over the last 20 years, with the proportion of tested strains at 8.7% and 1.4% for isoniazid resistance and MDR resistance respectively (49). The direct impact of fully resistant variants would be devastating to modern medicine. Losing these invaluable tools could also indirectly negate medical advancements by having a patient survive a procedure but not the infection. To ensure the rise of resistance does not overcome our current medical tools, it is therefore critical to engage this threat on a variety of fronts from legislative safeguarding to development of new treatments.

Unfortunately, there has been a decline in new drugs to market after the pre-1970 "golden age" when most of the currently known classes of antibiotics had been discovered. The last 30 years have seen a drop in the number of approved antibiotics with significant stagnation over the last decade (Figure 1.2).

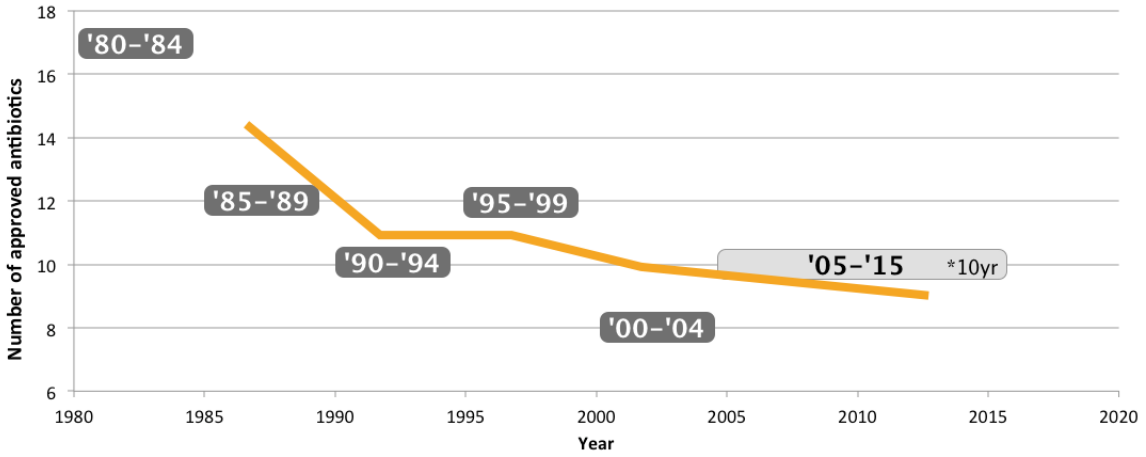

Figure 1.2: Decline in approved antibiotics from 1980 – 2015, boxes depict total for the time span. Data adapted from *Bassetti, et al.* (50), with updated approvals from *Deak, et al.* (51) depicted in a cumulative 10 year timespan. Orange trend line represents the moving average.

Additionally, roughly two-thirds of these known antibiotic classes target the cell-envelope or ribosome only, and nearly all new post-millennial approvals are from known classes (51–53). Although new approvals have been made over recent years, including additions from fully synthetic efforts, there still remains a lack of innovative new discoveries. This situation further compounds the risk of resistance, as resistant strains may also be insusceptible to new drugs that attack the same target (cross resistance). The decline in new drugs to market is due, in part, to an expensive and timely development pipeline, coupled with lower financial returns for developers. Compared to other classes of drugs, antibiotics may only yield a fraction of research and development efforts, with some examples earning up to 30 times less revenue of a non-antibiotic drug (54). This threatens future development in the private sector, as fewer resources will be devoted to research with weak returns. Besides the high cost of development, there exists stagnation in the discovery phase of drug development.

Currently, the main route for discovery has been sourcing from nature. These natural products are specialized compounds, also known as Secondary Metabolites (SMs), that are extracted from plants, animals, and various microorganisms. Today, the majority of antibiotics are either natural products or synthetics derived or inspired from natural products (55). One reason for this dominance is that natural products are the results of millions of years of evolution and thereby overcome difficult development hurdles such as efficient target binding and cell penetration (56). Because of this fact, natural products remain an

important field of research in the hunt for novel antibiotics and anti-infectives. However, the problem of rediscovery from these sources has been a big contributor to the stagnation in the drug development process. The time and effort used in the discovery of known compounds with the same activities impedes progress. This problem is partially due to heavy reliance on traditional discovery and screening methods that may not capture entire chemical space or simply limited due to focusing on historically familiar sources (56, 57). To reinvigorate the discovery pipeline we must therefore utilize new technologies and techniques, as well as expand the search to include promising underexplored sources.

Thesis Outline

This thesis is organized into two background sections and three research chapters that address improvements in natural product discovery using computational tools. In Chapter 3 we discuss the development of a new tool for prioritizing natural products with emphasis on novel drug target identification. In Chapter 2 we discuss the use of a tool developed to aid in the prioritization of new bacterial sources via a high-resolution phylogenetic workflow. Chapter 1 demonstrates the application of gene similarity methods to aid in the determination of known natural products and the diversity assessment of secondary metabolites. Details about the objectives of each research project can be found in section three. Finally, a discussion of all efforts and future outlook are addressed in section six.

2 Background

2.1 Drug Discovery from Natural Sources

Beginning with the discovery of penicillin, antibiotics have predominantly come from fungal, plant, or bacterial sources. These natural products have continued to fuel the discovery pipeline, as most known antibiotics are either inspired by natural products or are direct compounds (58). These producers have occupied a wide variety of environments and evolved under various selective pressures over the course of natural history, resulting in a wealth of compound diversity. Terrestrial bacteria have been extensively exploited, as the rich soil ecosystem affords an array of competitors and resources (59). It was noticed early on by Waksman and colleagues that particular soil-dwelling microbes, which formed complex networks and associations with surrounding plants and fungi, were found to kill other bacteria in co-cultures (60). It wasn't until later that these microbes were demystified and classified as bacteria belonging to the prolific Actinobacteria class. Members of this lineage, such as the genera *Streptomyces*, are historically the richest known sources of antibiotics (61). Focus on these bountiful sources is largely attributed to the observed chemical diversity they harbor. A few examples of this genus alone illustrate a spread of small to large secondary metabolites produced (Figure 2.1).

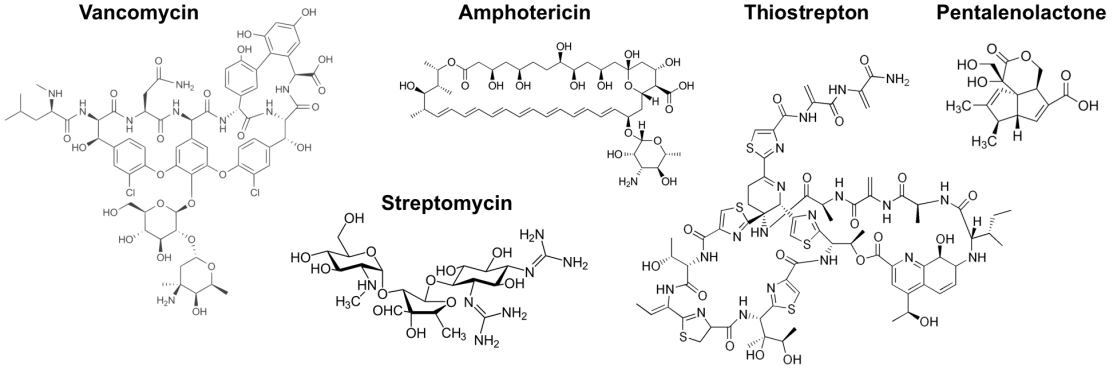

Figure 2.1: Variety of structures from antibiotic natural products from the genus *Streptomyces* from multiple classes: Vancomycin, Amphotericin, Streptomycin, Thiostrepton, and Pentalenolactone

This diversity has been the prime reason why natural products have been such a fruitful reservoir of useful compounds, with over 200,000 known molecules (62). Over the course of

natural selection, these secondary metabolites have been pre-screened for those that have interesting biological activities, such as ones with bactericidal properties. Leveraging the "research and development" of millions of years of evolution has been beneficial, but unfortunately there has been stagnation in the traditional discovery pipeline largely due to rediscovery of the same compounds (63). This problem is a major issue for this traditional "top-down" discovery model as it relies on a reasonable amount of luck to justify resources for prospecting (64). Fortunately, it has been shown that there are still many untapped areas of chemical diversity in new organisms and environments from which the traditional methods can benefit, such as underexplored myxobacteria and cyanobacteria genera (65). Additionally, with the advent of "bottom-up" approaches using gene sequencing we can systematically improve and expand the search for new compounds.

2.1.1 Traditional discovery pipeline

The basic workflow for natural products discovery over the last century has revolved around screening chemical extracts from collected organisms and natural sources (Figure 2.2). These extracts can be taken directly from the environment, prepared from raw biomass, such as solubilizing plant material, or by first cultivating a sample. For microorganisms, this is largely done via isolation of a producing strain and culturing to allow it to manufacture the potentially useful compounds. Culturing can also act as a screening step, as seen with Alexander Fleming's penicillin producing fungi, by growing in a co-culture assay.

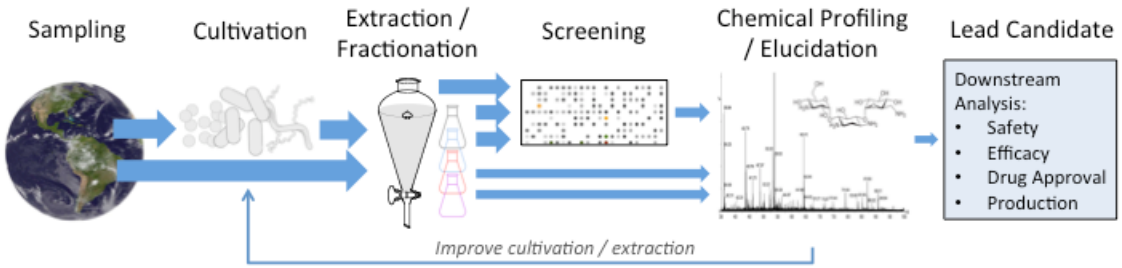

Figure 2.2: Traditional natural product discovery workflow. Extracts can be made directly from environment or through cultured microbes. Time intensive screening, chemical profiling, and fractionation are done before a lead can be further tested downstream. Multiple extraction and screening methods may be required to optimize detection.

A more common approach is to test the extracts directly using biological and toxicity assays, for example, placing either raw extracts or chemical fractions on a lawn of target bacteria to

determine antibiotic activity or use other multiplexing screens that look for various activities. While straightforward and robust, these steps can be labor intensive and time consuming, especially for organisms that need up to several months to accumulate enough biomass for testing. Once activity is established, accumulating enough material for downstream analysis is needed. This can be a problem if the source is from a slow growing organism or if it does not produce significant amounts of product. This is especially important for manufacturing, as many of these natural products are complex compounds that would require several multi-synthesis steps to be produced synthetically. Often this is too expensive and without the core structure production from the source, there would be no feasible way to produce the potential drug at a practical scale. After this barrier is overcome, the compound can then move on past the discovery stage to undergo further interrogation, toxicity screening, and structure elucidation before becoming a viable antibiotic. Despite these throughput bottlenecks, most of our known antibiotics have resulted from this process and hundreds of thousands of compounds have been isolated to date (66). Issues with this model, besides time and energy expense, are the inherent bias toward organisms that can be cultured in the lab in a timely manner and toward compounds that are produced in large enough quantities. This is a big drawback considering the majority of known microbes are "uncultivable" – extremely difficult to grow with standard techniques (67). Furthermore, certain products are only produced under specific conditions or stimuli and some products are governed by complex regulation systems, for example pristinamycin biosynthesis in *Streptomyces* (68). Unfortunately this makes it unfeasible to test every permutation for each organism manually, further restricting sampled compounds. This pipeline is therefore currently selecting for a subpopulation of chemical space and is at least in part responsible for the high rate of compound rediscovery. This problem is not a significant hindrance to development if caught early in the pipeline, however, the issue is that significant time and effort is usually already invested when an elucidated structure is known. This dissuades investment in discovery and further stuns progress.

Fortunately, efforts to improve this pipeline are underway. Using different media and growth conditions is one solution. Testing conditions of stress by manipulating carbon, nitrogen, phosphorous or iron supply is another tactic that has shown to produce a range of metabolites from the same sample (69). Rather than trying to mimic optimal growth conditions for "uncultivable" bacteria in the lab, another technique is to culture a sample *in*

situ – in the original habitat. Technologies such as the isolation chip (Ichip) have been recently developed to do this via permeable membranes so that clonal colonies can be grown directly where they were found (70). This technique has already lead to the discovery of a novel antibiotic, Teixobactin, that was produced by a new species of beta-proteobacteria, provisionally named *Eleftheria terrae* (71). Efforts to search for new sources in different environments such as marine sediment or an organism's microbiome are another expansion of discovery; one study showed over 25 thousand new compounds have already been discovered from marine sources around the globe (72). Even specific niches can harbor useful compounds; for example, isolated fungi from a microbial mat in an iron rich spring lead to the discovery of six new natural products (73). Symbiotic microbes also hold significant potential, as it has been observed that their hosts rely on the defenses that they produce (74). For example, marine Actinobacterial strains associated with the sponge *Halichondria panicea* have shown to produce 88 new putative compounds (75).

Identifying known molecules early on is another helpful endeavor. Computational methods such as Global Natural Product Social Molecular Networking (GNPS), are great tools for connecting elements in experimental mass spectra with those in a database (76). With this method, raw extracts or fractionated samples can be de-convoluted into a vector of chemical components. This is then displayed as an association network with spectra of known compounds. These methods are a great benefit to discovery, but still have a limited throughput requiring screening organisms under different conditions. These top-down approaches will continue to serve direct routes to discovery especially as this pipeline is expanded. Though to further solve the problem of rediscovery, we can also take a bottom-up approach and look beyond the apparent products toward the blueprints of these diverse compounds - their genes.

2.1.2 Genetic Research Unveils Natural Product Biosynthesis

Not only are *Streptomyces* one of the richest known sources of antibiotics, they also served as a model for understanding the biosynthesis of these valued compounds. With the study of organisms like *Streptomyces coelicolor,* new genomic tools were developed which helped to illuminate novel mechanisms of biosynthesis. David Hopwood used this strain to devise a genetic linkage map that established a means of associating genetic loci with observed phenotypes (77). This lead to the discovery of a plasmid responsible for methylenomycin A

production in *S. coelicolor*, which also included the resistance gene (78). Sanger sequencing and genetic manipulation techniques further enabled the correlation of specific genes responsible for expressed compounds. One major observation in microbes is that nearly every class of these production systems packages the required genes in close proximity to one another. These Biosynthetic Gene Clusters (BGCs) then form a cohesive recipe for the formation of these natural products that not only contain the core machinery for synthesis, but also accessory enzymes involved in tailoring, resistance, and export. These genetic tools exposed the origins of the vast chemical diversity of these specialized compounds and accelerated discovery by giving valuable insight into the genomic basis of natural product production.

Various systems of secondary metabolite production were later discovered including the ribosomally synthesized and post-translationally modified peptides (RiPPs), terpene synthases (TPS), and other derivatives of primary metabolism. The large macro enzyme complexes or collections, Non-Ribosomal Peptide Synthetases (NRPS) and Polyketide Synthases (PKS), were of particular interest due to the combinatory "assembly line" design, which helped explain the variety of observed chemical structures. The organization of PKSs showed similarities with products from Fatty Acid Synthases (FAS) with its composition of acyl subunits. Likewise, the genes responsible showed similar architecture to eukaryotic FAS type I as well as bacterial FAS II genes. In general NRPS and PKS BGCs were seen to have various domains that correlated to the number of subunits used to produce the end product, the so called co-linearity rule (79). Modules of the PKS type I consist of core catalytic domains that propagate and condense various acyl monomers: acyl transferase (AT), ketosynthase (KS), acyl carrier protein (ACP), and thioesterase (TE); Additional domains involved in modifications include ketoreductase (KR), dehydratase (DH), enoylreductase (ER), methyltransferase (MT), and others involved in special tailoring such as cyclization. The key observation that these multi-domain modules correlated with each unit of the molecule gave a standard recipe for the assembly of such complex structures. A specific order of these modules thus forms a macro-enzyme factory for the production of a particular metabolite. This is accomplished via a stepwise set of reactions where monomers propagate through a starter module, elongation modules, and termination module to form a growing chain that is then released. The AT domain catalyzes the loading of an acetyl-CoA, or one of its derivatives, to the ACP domain by a thioester linkage. The growing chain is

then passed to the KS domain of the next module, via catalysis by the KS domain. The chain bound to the KS domain and monomer bound to the current ACP domain, usually a malony-CoA or methylmalony-CoA, condense via the KS domain N-terminus activity. This results in an elongation of one unit as the chain passes from module to module (Figure 2.3).

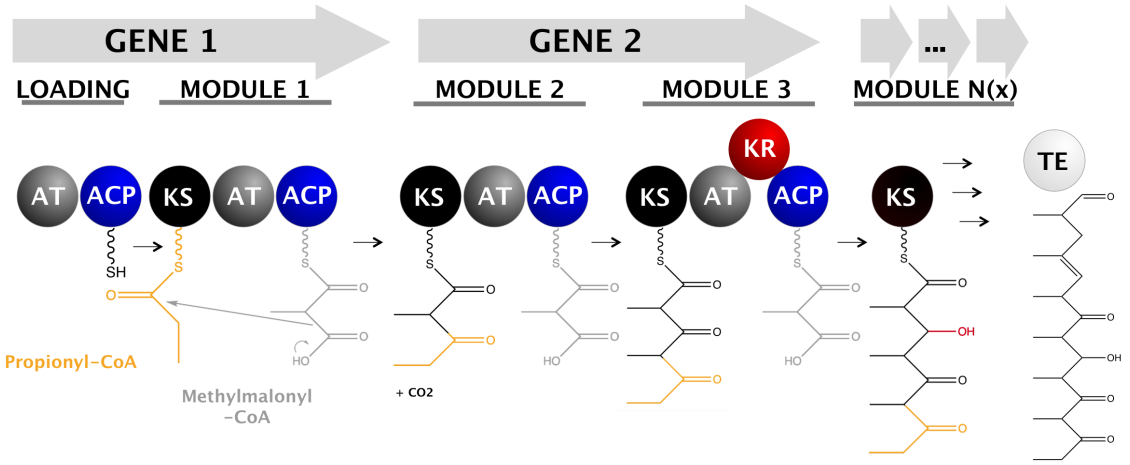

Figure 2.3: Conceptual example of modular PKS type I chain extension. AT domains select for the propionyl-CoA starter unit and Methylmalonyl-CoA extender units. Additional dehydratase (DH) and enoylreductase (ER) domains not shown here.

Each module may also contain additional domains that perform modifications to the previous unit in the chain, for example, the reduction of a double bond to a single by the ER domain or conversion of a ketone to an alcohol by the KR domain. The final step in the process releases the compound via the TE domain, which might involve a cyclization step. In addition to the macro enzymatic modular PKS, further divisions of iterative type I, trans-AT, type II, and type III systems have been described (80).

Similar to PKS systems, NRPS biosynthesis is analogous in modular design, with amino acids including nonproteinogenic variants used as monomers instead of acyl units (81). Additionally, due to the similar design to PKS systems, hybrid systems have been observed whereby acyl units may be used or where whole PKS production schemes are fused (82). NRPS modules contain three required catalytic domains – Adenylation (A), Peptide Carrier Protein (PCP), and Condensation (C). The first step in the process is activation of the starter or extender units by the A domain which uses ATP to form the reactive aminoacyl-adenylate. The PCP domain then catalyzes the loading to its 4'-phospho-pantethine (4'PP)

cofactor via a thioester bond. This unit then delivers the attached monomer to a nucleophile acceptor position in the C domain where it condenses with another unit in the electrophile donor position. This newly extended peptide is then moved to the downstream C domain, where further cycles can continue chain elongation. Alternatively, a C domain may be replaced with a Cy domain, which also catalyzes cyclization in addition to condensation of the upstream peptide. A variety of other tailoring functions can be included in a module as well, including: epimerization, formylation, methylation, oxidation, and reduction domains. Final cleavage of the completed chain can either be accomplished with a TE domain as with PKS, or a reduction domain that reduces the thioester bond to an alcohol or aldehyde. Post processing of the completed peptide is also seen using other tailoring enzymes; this can include halogenation, hydroxylation, glycosylation, and acylation.

With these two systems alone it is clear that a vast diversity of chemical novelty can be achieved with small changes in the order of modules or addition of the many possible tailoring steps involved. To add to this chemical potential, terpene biosysnthesis likewise has produced a large variety of compounds with approximately 60,000 known structures (83). Enzymes required for terpene synthesis can be found in bacteria, fungi, or plants, with the later being more common. Enzymes of this class all utilize five carbon building blocks, specifically isopentenyl diphosphate (IDP) and dimethylallyl diphosphate (DMADP). These precursors are generated from primary metabolism either from the mevalonate (MVA) pathway or the 2-C-methyl-D-erythritol-4-phosphate (MEP) pathway present in most bacteria. These precursors are then condensed into a variety of intermediates of variable chain length with enzymes such as farnesyl diphosphate and geranyl diphosphate among others. Different classes of TPSs then utilize these intermediates to produce more diverse structures. One key feature of TPS biosysnthesis is the formation of reactive carbocation intermediates that can lead to a mixture of products from a single pathway (84). With this feature and their ability to work as a multi-substrate enzymes (85), terpenes possess a great potential for chemical diversity.

The structural diversity of ribosomally synthesized molecules is clearly evidenced by the array of proteins that preform the necessaries of life, which RiPP pathways have exploited. RiPPs are a broad class of compounds that span an umbrella of modifications and cleavage of peptides produced from transcribed DNA; consequentially, RiPPs are responsible for several classes of compounds: lanthipeptides, cyanobactins, and thiopeptides, among others.

This class has shown applications in a number of commercial uses and is host to diverse biological activities. Although smaller molecule antibiotics have been sourced from RiPPs, these enzymes often produce large structures, typically over 1000 daltons, with macrocyclic rings (86). Biosynthesis involves an N-terminal leader strand of the precursor peptide that helps in recognition and guiding the various enzymes that act on the core segment; optionally, there may also be a C-terminal recognition site to aid in cyclization or excision. Modification reactions occur on the core peptide and depending on the class of RiPP, these modifications can include dehydration, prenylation, and cyclization, among others. The newly modified core peptide is then excised via a proteolysis step and a mature compound is produced.

Another class deriving form primary metabolism is the saccharides, which are responsible for a number of key cellular functions such as cell wall biosynthesis in gram-positive bacteria. Including hybrids of other classes, these pathways have helped to produce a significant amount of useful secondary metabolites including the life-saving antibiotics kanamycin and streptomycin. This class is also responsible for some virulence factors in pathogens that help protect against the host immune system such as the capsular polysaccharides (CPS) and lipopolysaccharides (LPS). CPSs are long chain repetitive glycan strings that can form differing protective layers around a bacterial cell helping to avoid recognition or attack by the host cells. General synthesis involves formation of CPS repeat units in the cytoplasm followed by export and polymerization into a fully formed capsule outside of the cell membrane. LPSs are also extracellular components that are comprised of an O-antigen, oligosaccharide, and a fatty-acid portion. These structures can serve intricate functions, such as modulating immune response, masquerading as a host cell, or altering its antigen profile diversity thereby evading immune detection (87). Saccharides are also a host to various endotoxins that can have serious human health impact. Given the varied biological activities, this class is shown to harbor many useful and harmful secondary metabolites and further expands the chemical profile of microorganisms.

Unveiling of the various production systems has helped in understanding the source of natural product chemical diversity. The key packaging of these systems in BGCs has also helped in determining the means of antibiotic resistance by investigating the surrounding genomic context. One example is the resistant *gyrB* gene included in the novobiocin BGC of *Streptomyces spheroids* (88). These "self-resistance" genes are a convenient way for an antibiotic

producer to avoid suicide by coupling the expression of both simultaneously; for example, modification of cell wall precursors via *vanX*, *vanH*, and *vanA* genes in the vancomycin-like producer *Amycolatopsis balhimycina* (89). Genomic context can also help to highlight which residues are responsible for resistance, thus giving clues to mechanism of action and revealing a compounds target; in the case of self-resistance via a modified target, as seen with *rpoB* genes in *Salinispora* species, genetic comparisons were used to highlight resistance conferring areas of the gene (27). This understanding has aided in structure prediction, increasing compound production, and discovery of new products. By focusing on the BGCs directly, rather than simply the source organism, this bottom-up approach is an attractive route to discovery. This also expands discovery, as the entire chemical potential of an organism can be accessed via its genetic potential despite undetectable production of compound or lack of gene expression.

2.1.3 Genomics Accelerates Natural Product Discovery

Gene manipulation tools and gene sequencing technologies have helped in the understanding of secondary metabolite production but also provided many useful applications for discovery. As seen with the plasmid-derived methylenomycin producer, transplantation of the entire production pathway into a new host is possible with a single plasmid transformation. It was also shown that the resistance gene accompanied this BGC which conferred methylenomycin resistance in the new host (78). Transplanting of chromosomal BGCs has also been widely utilized in discovery. Generation of F-plasmids containing sheared chromosomal DNA usually up to 40 killobase-pair (kb) produces a fosmid library, which can then be used to transform and screen recipients for the expected phenotype (90). This heterologous expression of an entire BGC can be very helpful in identifying, studying, or increasing yields of a natural product. Not only does this help confirm the hypothesis of a responsible cluster but also it makes for an easier platform to unveil key aspects of the product's formation. Transplanting BGS into model hosts can also enable gene manipulation tools that may not be established in the native host. For example, if production shows yields below detection, or the cluster is not expressed under laboratory conditions, then the insertion of a strong inducible promoter can activate this "silent cluster"; This protocol was shown to lead to the discovery of several novel metabolites (91). Likewise, gene deletion studies have helped to identify the genes that contribute to

production of a particular compound. By interrogating these gene "knock-outs" we can compare different stages in the synthesis process, which can ultimately help in structure elucidation or in describing the biosynthesis pathway. Another benefit of heterologous expression is safer investigation of products from dangerous pathogens by introducing them into a non-pathogenic organism. Heterologous expression of BGCs has been shown to be a valuable avenue for accessing uncultivable sources using BGCs from environmental DNA as well (92). PCR amplification of PKS related genomic regions were shown to produce recombinant environmental libraries, which lead to the discovery of eight novel PKS BGCs from diverse phylogenetic origins (93). Sequencing-free technologies such as DNA microarrays have also contributed to discovery by allowing for gene expression data to be studied in a rapid testing format (94). Sequence-guided discovery has also lead to the targeted expression of specific BGCs in a genome. Instead of generating a random fosmid library, technologies such as Transformation-associated recombination (TAR) cloning can be used, which has shown to capture specific regions of chromosomal DNA as high as 300kb (95). This can be used to design and capture specific areas of a genome or environmental sample and systematically discover their products. Genomic context has also helped in synthetic biology techniques, which has applications in optimization of known natural product yields and also in discovery of unknown or unnatural BGC derivatives (96).

Simply having the gene sequences alone has given research a step forward by helping to predict or answer key questions of natural product formation. For example by studying the variations in A-domains sequences of NRPS systems, we now know of ten residues in the binding pocket that are largely responsible for amino acid specificity (97). This "nonribosomal code" gives the ability to predict building blocks involved in biosynthesis from sequence, albeit not directly analogous to definitive translation as with DNA codon triplets. More sophisticated computational techniques of substrate prediction using these sequences have been developed with high accuracy however (98, 99). Of the main benefits of having the blueprints in hand is the ability to screen for natural products via their identified BGCs without the need to culture or perform biological screening. This has also unveiled hidden secondary metabolite potential by exposing silent gene clusters even in heavily studied organisms such *Streptomyces coelicolor* (100). Over the last decade tens of thousands of complete, or near complete, bacterial genomes have become publicly available

which has further exposed potential for secondary metabolite production using BGC prediction software (Figure 2.4).

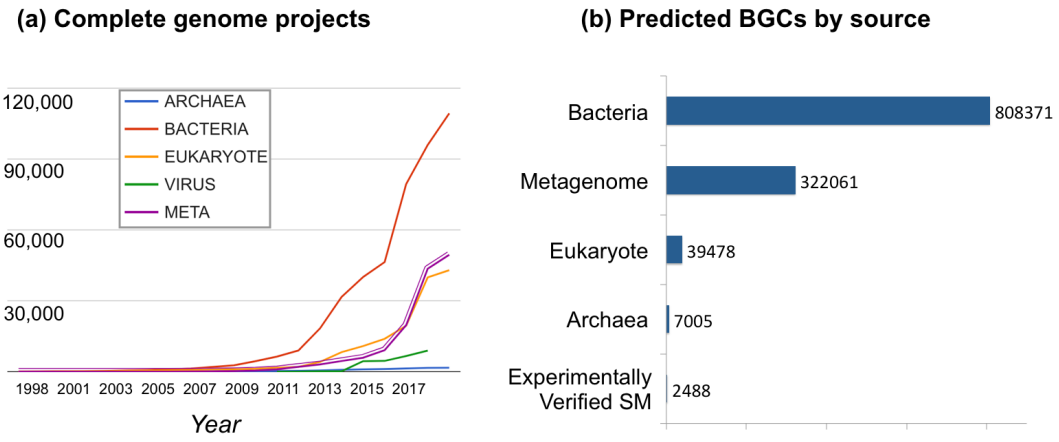

Figure 2.4: (a) JGI GOLDS completed genomes (http://gold.jgi.doe.gov). (b) Cluster prediction statistics from the Atlas of Biosynthetic gene Clusters (ABC) (101) sourced form (http://img.jgi.doe.gov) accessed May 15 2018

In addition to the increase in available genomic data the throughput and cost of sequencing has improved. With increases in throughput approaching 900 gigabase-pairs with the Illumina HighSeq X, this gives enough depth to compile multiple genomes in one run from a mixed environmental sample known as a meta-genome (102). This volume of data will require automated methods to be effectively leveraged. Thus with the maturation of analysis tools over the past years, and drop in sequencing cost, there is a new avenue for high-throughput screening available that can help overcome stagnation in the discovery pipeline.

2.2 Computational Methods and Automated Genome Mining

Downstream computational analysis of data is always reliant on the condition of the initial data. With the advancement of sequencing technologies it is therefore important to understand the limitations and future prospects for the generation of sequencing data. This introduction is meant to briefly describe various technologies in use to generate whole-genome data, a more in depth comparison of sequencing and assembly methods are referenced here (102–104). The increase in public genomes and drop in sequencing cost is largely due to the high-throughput "next-gen" sequencing technologies such as 454,

Illumina, and Ion Torrent sequencers; These methods all rely on a "shotgun" approach whereby DNA is randomly fragmented into shorter pieces, sequenced in parallel, and then computationally reassembled based on overlapping portions of sequenced ends. Each of these technologies generates reads through a process of sequencing by synthesis where a successful incorporation of nucleotide is read by the instrument. In practice, an initial sequencing may not be enough for a complete assembly and results in a "draft genome" - a set of unlinked contiguous sequences called contigs. This poses a problem as initial versions providing short read lengths are difficult to assemble, particularly for the repetitive sequences often seen in various BGCs. This problem of BGC's split on separated contigs could be solved through more re-sequencing efforts or using phylogenetic classifiers, such as NaPDoS (105), to connect them. Fortunately new entries into the market such as PacBio, and Oxford Nanopore sequencers have given a solution by using much longer read lengths via a single-molecule approach (102). With reads lengths around 10-15kb (102) smaller BGCs might be contained in a single read while larger BGCs will have a higher chance of full assembly. While read length is improving so is the throughput of data, which is allowing for higher resolution metagenomic surveying of sources. As demonstrated with programs such as MEGAN, we can pre-screen metagenomic data directly for their taxonomic and metabolic potential without the need of assembly (106). In addition to genomic data RNA sequencing has become a means to probe the expression of genes, which also provides another avenue for discovery (107). As these technologies improve and increasing volumes of data are seen there is a continually growing opportunity available for genome mining. Furthermore, with the increase in amount and diversity of whole genomes, a powerful approach using comparative genomics is on the horizon.

2.2.1 Natural Product Genome Mining

Genome mining is an attractive route to drug discovery because of the ease of obtaining genomes, its high-throughput capacity, and it requires relatively inexpensive *in-silico* analysis compared with experimental screening. The inaccessible "dark-matter" of uncultivable genomes can also be probed via meta-genome sequencing. This pipeline predominately involves searching for known BGC signature genes and analyzing the surrounding region. This motif-based technique has been the main route for manual and automated genome mining and with the addition of motif-free methods a wealth of predictions have been

found. Automated methods have sense matured and are able to reliably predict many classes of BGCs in fungal, plant, and bacterial genomes. Although there is still headway to be made, particularly with the more complicated fungal genomes (108), these methods are able to predict the NRPS, PKS, RiPP, terpene, saccharide and hybrid classes effectively. Besides BGC identification these tools also provide information such as accessory gene annotation, genomic context, and structure prediction of a product. A community-updated list of these cluster-mining tools can be found at the "The Secondary Metabolite Bioinformatics Portal" (109) (http://secondarymetabolites.org) and detailed method overviews have been reviewed in publication 4 as well as elsewhere (110, 111). Among these programs, antiSMASH (112), PRISM (113), SMURF (114), and CLUSEAN (115), achieve rapid searches of seed signature genes via high fidelity Hidden Markov Models (HMMs) and the updated HMMER3 algorithm (116). These models are based on sequence alignments of known signatures, for example KS domains for PKS clusters, which can outperform BLAST (117) searches in terms of speed and sensitivity. Once these anchors are identified a cluster is then defined based on a drop in signature gene density to identify boundaries. In addition to signature gene methods, probabilistic approaches such as clusterfinder (118), which is also integrated into antiSMASH, uses a technique of assigning every gene a likelihood of being apart of a BGC based on a large training set of currently known BGCs. Another interesting approach is the screening of regulation binding motifs for known up-regulation of a natural product. This method, INBEKT, was shown to identify the zincophore ethylene diamine disuccinic acid ([S,S]-EDDS) in *Amycolatopsis japonicum* (119); This was also shown to be generally applicable to other ionophores as publication 6 demonstrates the detection of aminopolycarboxylic acid siderophores (120).

To complement these motif-dependent and probabilistic methods, several motif-independent methods have been developed. As these approaches do not rely on previously known architectures of BGCs there is the potential to discover unknown systems of secondary metabolite production. One example, EvoMining (121), is based on the observation that many biosynthesis schemes have evolutionary roots in primary metabolism and so by looking for these "gene expansions" - duplicated and repurposed genes, we can identify regions of natural product biosynthesis. This method was shown to identify two previously uncharacterized enzymes, an argininosuccinate lyase used in the biosysnthesis of leupeptin, and a divergent AroA family enzyme, which lead to the discovery of a novel

arseno-organic compound (121). Likewise an approach that uses gene expression data to identify co-expressed clusters, MIDDAS-M (122), was shown to identify a BGC responsible for ustiloxin B production in fungi (123). RiPP detection in fungi remains a challenge using motif-dependent methods however this proof of principle highlights the possibility of finding classes that might elude motif methods. While this requires transcriptomic data, it is a valuable addition with an advantage to connect disparate genetic loci involved in biosynthesis as seen in plant genomes. Due to the difficulty to obtain a completed plant genome, this has helped to identify and connect contigs with hard to detect BGCs (124, 125). A third method, MIPS-CG (126), utilizes a comparative approach that analyzes co-localization of gene clusters in non-syntenic regions. By comparing to other genomes it highlights areas where these clusters have been evolutionarily maintained and possibly horizontally transferred to different organisms. This was shown to identify a kojic acid BGC (127).

With hundreds of thousands of processed antiSMASH jobs and over a million predicted clusters (with redundancies) from the IMG-ABC database (101), there are currently vast amounts of leads to investigate in wet-lab experiments. Therefore this poses a new challenge to prioritize and de-replicate these leads to efficiently interrogate them.

2.2.2 Prioritizing Natural Product Leads

The volume of predictions to investigate is infeasible to undertake using current biological screening methods. Even in an individual genome there can be many clusters identified, for example, *Streptomyces bingchenggensis* BCW-1 shows at least 47 BGCs for known and predicted secondary metabolites, with a large number of silent clusters (128). Being able to select which of these clusters have interesting biological activities or might be promising antibiotics is thus a high priority for genome mining. The first undertaking of de-replication, or identifying all clusters that are identical but found from different sources, has already begun in the last few years. This initial step of cataloging known clusters has amassed thousands of BGC sequences in currently updated databases. The MiBIG database (129) houses one of the largest collections of experimentally verified full and partial BGCs with over 1400 clusters. The Joint Genome Institute (JGI) has over a million clusterfinder results for all public isolates at the Atlas of Biosynthetic gene Clusters (ABC) (101). The recently released antiSMASH database maintains over 20,000 clusters from 3,900 unique isolates (130). In

order to effectively query these databases methods for gene similarity searching have been developed. MultiGeneBlast (MGB) allows for a cumulative search of entire operons or gene clusters by combining loci information and individual BLAST protein searches (131). Originally developed for the antiSMASH platform, this standalone program is able to query all sequenced public genomes from the NCBI GenBank database to help locate all similar clusters of a query. To use this method for de-replication however would be time consuming, as this would require a computationally expensive pairwise search of all gene cluster genes. Also the scoring scheme is great for sorting top likely hits, but because it is not a metric there would be difficulty in defining a similarity threshold to automate and would require manual inspection. To solve this issue, other gene cluster similarity metrics have been developed. One method uses protein family (Pfam) (132) annotation to calculate pairwise functional similarity of every cluster (Figure 2.5).

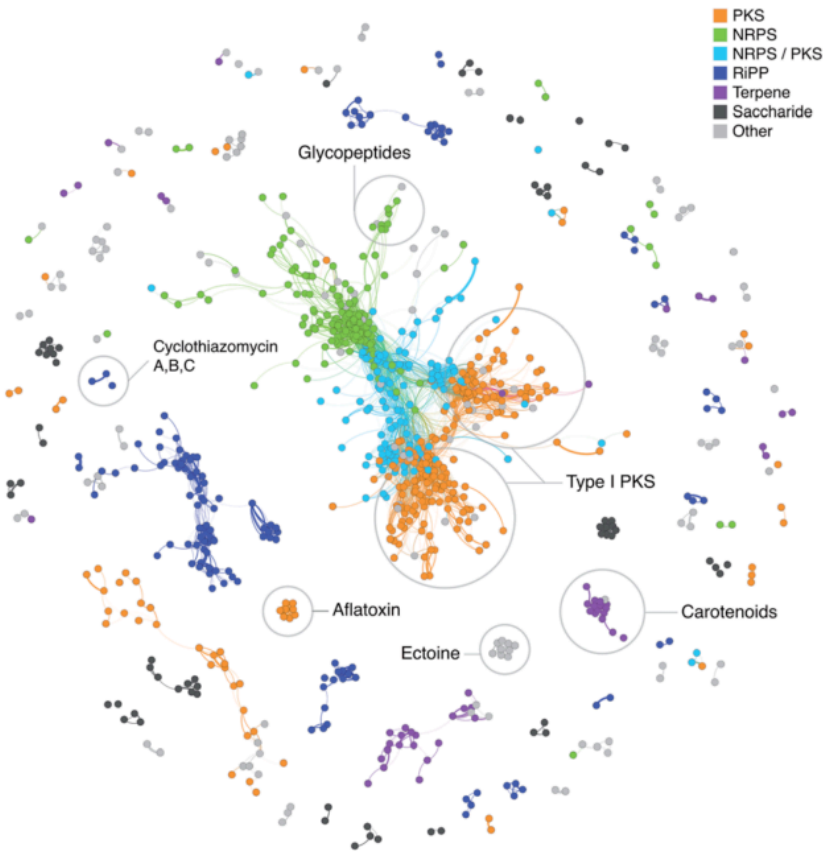

Figure 2.5: Example gene cluster network depicting Gene Cluster Families (GCFs) that produces like compounds. Data was generated using the MiBIG database (129) as detailed in publication 4, Ziemert et al(133). Colored regions depict gene cluster type with examples of known compounds circled based on MIBiG annotations.

This was used to develop a gene cluster similarity network of thousands of BGCs that showed three subfamilies for the large class of aryl-polyene BGCs (118). Despite the divergent sequence similarity for clusters of shared compounds, the method was able to associate these BGCs due to the higher order functional Pfam similarity measure. These networking techniques are therefore well suited for de-replication of like compounds. Furthermore the gene cluster network showed many groups of similar clusters in the network, gene cluster families (GCF), which had no described member. Thus by studying the diversity of these networks this method can also be used as a tool for prioritization as well as de-replication by exploring under-represented potential. These methods were shown to be effective as demonstrated in publication 2 and 3.

Many prioritization schemes have been adopted and the variety adds to the success rate by using multiple perspectives. One immediate method is prioritization by seeking novel chemical structures or interesting moieties. This has been widely employed through chemical spectra networking using platforms such as GNPS. For gene cluster data this tactic relies on gene product predictions. While this remains a difficult task recent advancements have improved upon current core structure perditions by interrogating mass spectrometry data as seen with PRISM (113) . As new structures inevitably lead to new chemical properties this is a valuable approach to identify compounds with new activates. Using this broad approach does not guarantee activity however. Other methods have taken a different approach to enrich for antibiotic activity. One such method, target directed genome mining (134), uses the fact that an antibiotic producer will encode a means of defense, which is then used as a marker for prioritization. Specifically this method aims to highlight a co-localized resistant target in a BGC. This helps save time with experimental screening by enriching for clusters with a higher probability of antibiotic activity; and it also gives a head start on elucidating the mechanism of action. This method can still lead to compounds with novel modes of action despite utilizing a known target. This has predominately been a manual screening process which publication 3, the Antibiotic Resistant Target Seeker (ARTS), aims to automate and address. Another popular approach is to focus on new taxa that harbor high levels of secondary metabolite production. Just as members of the Actinobacteria class have been exploited, new prolific clades have been discovered with promising resulting leads such as cyanobacteria and myxobacteria (135–137). The biosynthetic potential of these broad order

and family groupings can also differ for more specific taxonomic levels as seen with myxobacteria (138). Accurate phylogenetic placement of a particular species can thus serve as an indicator of the productive potential of a sample before experimental screening. This can also help to avoid resampling of closely related species and expand efforts more effectively. Phylogenetic classification has also shown to be a valuable tool to prioritize individual BGCs as seen with NaPDoS (105), a web tool to detect and classify conserved domains in NRPS and PKS BGCs. Because bacterial taxonomy is complicated by several factors, such as horizontal transfer, accurate phylogenetic classification is a non-trivial time consuming process. The Automated Multi-Locus Species Tree (autoMLST) in manuscript 1 was therefore developed to solve this problem and help to prioritize prolific sources.

3 Research Objectives

This thesis is focused on developing computational tools to aid the discovery of novel antibiotics from prokaryotes. The major goals of this research are to prioritize the wealth of leads form current genome mining methods and provide an orthogonal detection technique to complement current motif based strategies. Additional goals are to automate and simplify gold-standard species designation methods to identify promising new sources, and to apply these tools with comparative methods such as gene cluster networking to remove redundant leads to highlight novel antibiotic predictions.

Design requirements for distributed applications are to conform to these following criteria: Accessible to broad set of users, computationally feasible to host freely on in-house infrastructure, and open-source to enable widespread use and progress in natural product discovery. Specific objectives are detailed below.

Gene cluster networking:

- Preform comparative analysis of gene clusters for de-replication
- Highlight sources rich in diverse BGCs and aid exploration of other prioritization criteria

Automated Multi-Locus Species Tree (autoMLST):

- Provide accessible web interface for non-specialists to perform high-resolution species phylogenies
- Simplify workflow to accelerate species identification
- Help direct efforts on new promising sources of natural products

Antibiotic Resistant Target Seeker (ARTS):

- Automate target-directed genome mining by highlighting known resistance factors and cross-referencing with predicted BGCs
- Expand this process to include putative novel targets using several criteria associated with known resistant targets
- Prioritize gene cluster predictions with other resistance annotations
- Provided orthogonal approach to BGC detection methods

Chapter 1

4 Gene cluster networking

4.1 Introduction

Genome mining has shown to be a valuable avenue for the discovery of new anti-invectives from natural sources (139, 140). While several tools offer robust detection of nearly all known classes of Biosynthetic Gene Clusters (BGCs), including antiSMASH (112), RiPPMiner (141), clusterfinder (118), and PRISM (113), they do not help to prioritize which are the most likely candidates for novel activity. Because these predictions can lead to a large amount of potential clusters that require laborious experimental investigation there is a need for efficient de-replication and prioritization of these predictions. With over a million clusters predicted from the JGI Atlas of Biosynthetic gene Clusters (ABC) it is critical to eliminate identical clusters and define structurally similar classes in order to leverage this resource effectively. Methods for identifying homologous single genes have been in use for well over two decades using tools such as NCBI BLAST (142); however few methods have been developed for homology detection of whole clusters. MultiGeneBlast (131) is one such application that uses an "all vs. all" BLAST search of genes in a cluster and combines location information from the subject searches to get a cumulative score. This method has shown to work well and can identify similar clusters via a prioritized list of best hits as demonstrated in the included cluster blast feature in antiSMASH (143). However this method has some drawbacks that require manual investigation and does not lend itself well for automated de-replication. The non-symmetric scoring of the algorithm can result in a fragmented cluster producing highly identical hits with a full cluster for instance; also the highly repetitive genes in BGCs can lead to inflated similarity scores. Gene synteny is also weighted into the scoring of the algorithm, which is not always conserved for similar BGCs in other multiple taxa; for example, microcystin producers P. agardhii and M. aeruginosa have similar BGC composition for the same compound but with several rearrangements and inversions of genes (144). This issue is problematic for other synteny based cluster methods such as SYNS (145), SynBlast (146), and Synima (147). Additionally varying ranges of scores can be produced depending on the cluster size, and score thresholds would need to be

defined on an individual basis. Other similarity approaches that use phylogenetic clustering of key domains and alignments of profile HMMs have been demonstrated (105, 148). Scoring metrics have also been proposed to solve this problem and this has enabled an automatic approach to assign distances between clusters (118, 149). These distances can then be used for the de-replication and similarity networking of all clusters to help drive prioritization efforts. By grouping these like BGCs into Gene Cluster Families (GCFs) a secondary metabolite diversity map can then be inferred. The method described in Cimermancic et al. uses a functional similarity approach rather than a strict sequence homology score to define these GCFs. First all protein domains are annotated in query and subject and then compared using a symmetric composition metric via the Jaccard index combined with a domain duplication index. More recent applications of this method, such as BiG-SCAPE (150), have also included multiple sequence alignments to further distinguish domains but the experimentally confirmed weighting of factors remained over 85% reliant on these Jaccard and duplication indices (150). The end result showed that similar clusters could be associated despite having low sequence homology by using this protein similarity based approach.

Members of the Actinobacteria class have been one of the richest sources of antimicrobial compounds particularly from soil dwelling environments (151). Considering this, clusters form these taxa were used as the focus for exercising the gene cluster networking methods. One main goal in the design of this networking method is to be big-data capable and so we used the Joint Genome institute's Atlas of Biosynthetic gene Clusters database, JGI-ABC (101), which had over 4700 Actinobacteria genomes at the time. By using this set as well as other characterized sets, such as the experimentally confirmed catalog in the MiBIG database (129), it was possible to assess other potentially rich natural product environments by investigating the diversity and overlap with these references; as shown in publication 2, "Sequencing rare marine actinomycete genomes reveals high density of unique natural product biosynthetic gene clusters" (152), a map of cluster diversity could be built which helped to prioritize underexplored organisms from marine environments – "rare marine actinomycetes" (RMAs). This also identified which individual clusters were distinct from others in the reference that could potentially harbor novel chemistry. The same methods were also implemented in publication 3, "Comparative genome mining reveals phylogenetic distribution patterns of secondary metabolites in Amycolatopsis" (153), were differences in

natural product diversity could be seen at the genus level. This also helped to quickly identify potentially novel compounds and the use of high-resolution phylogeny added an extra layer of metadata to the networks, which helped in GCF prioritization.

4.2 Methods

4.2.1 Implementation of gene cluster networking

Workflow Overview

Cluster similarity scoring first relies on accurate protein family designation of all tested pairs of clusters. This step was achieved using pre-annotated clusters from the JGI-ABC dataset or with the Pfam-A (132) database and Hidden Markov Model (HMM) scans implemented in HMMER3 (116). The resulting domain table of hits was generated using the 'domtblout' option of the hmmsearch application. Distance values between all pairwise combinations of hits were generated using custom python scripts and a parsed list of domain tables as input (scripts available at: https://bitbucket.org/malanjary_ut/clustsimscore). Scores and cluster IDs were combined with annotations which produced an undirected graph file in GML format that could then be read by network exploration tools such as Gephi (154). The final graph is then visualized using the Yifan Hu (155) layout to spatially organize nodes based on edge weights in Gephi (Figure 6.1).

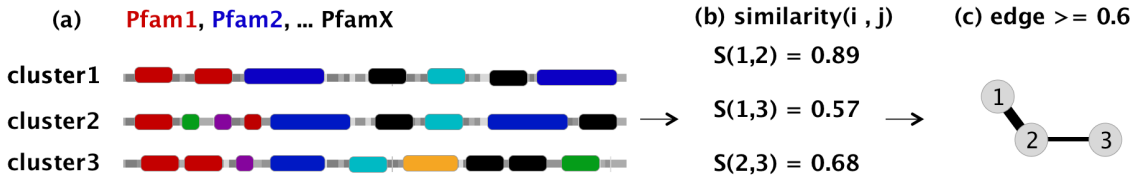

Figure 6.1: Workflow example of gene cluster networking steps (a) Protein domains are annotated for each BGC. (b) Accelerated calculation of similarity scores for all combinations of BGC pairs. (c) Parsing and visualization of final network

Similarity Scoring

Parsing of all HMM domain results yielded a set of unique Pfam domain IDs (x) for each cluster ID (i or j). The occurrence count of each Pfam domain was also stored in a count matrix (C):

$$x = (P_1, P_2, \dots, P_m) \qquad P = \{id_n \in [Pfam\_ids]\}$$

$$C_{m,n} = \begin{bmatrix} c_{11} & \cdots & c_{1n} \\ \vdots & \ddots & \vdots \\ c_{m1} & \cdots & c_{mn} \end{bmatrix} \quad c \in \mathbb{N}$$

The scoring uses two metrics to calculate total cluster distance - the Jaccard and domain duplication index. The Jaccard index, J(x_i, x_j), is a symmetric composition based index that measures the ratio of common Pfams and all Pfams present in both clusters:

$$J(x_i, x_j) = \frac{|\, x_i \cap x_j \,|}{|\, x_i \cup x_j \,|} \quad \in [0,1]$$

The domain duplication index, DDS(i,j), is calculated for a cluster pair (i,j) as the sum of all differences in corresponding Pfam count divided by the sum of the maximum count (excluding Pfams that are not present in either). The negative exponent is then taken as described in Lin et al.(149) resulting in a measure of equivalent repetition of domains.

$$D(i,j) = \sum_{p=1}^{n} |C_{pi} - C_{pj}| \qquad M(i,j) = \sum_{p=1}^{n} \max(C_{pi}, C_{pj})$$

$$DDS(i,j) = e^{-\frac{D(i,j)}{M(i,j)}} \quad \in [0,1]$$

The final similarity score is then defined as a weighted sum, S(i,j), of the Jaccard index and domain duplication index as demonstrated previously in Cimermancic et al. (118)

$$S(i,j) = 0.36 * J(x_i, x_j) + 0.64 * DDS(i,j)$$

Parallel Implementation

To accelerate processing, parallel computation of many pairs of (i,j) clusters could be achieved through matrix operations for the Jaccard and domain duplication steps. The resulting vectors could then be summed after scalar multiplication to give a vector of similarity scores that correspond to each (i,j) cluster pair. Sub-matrices of the count matrix $C_{m,n}$ (m cluster IDs and n Pfam IDs), A and B, are used as inputs to the final equation:

$$A = C[1, r; 1, n] \qquad B = C[1, r; 1, n] \;:\; r < m$$

$$\boldsymbol{Sp(A,B) = 0.36 * Jp(A,B) + 0.64 * DDSp(A,B)}$$

The parallel version of the Jaccard function, Jp, uses the Hadamard product of A and B to identify Pfams present. Any column with a zero will also result in zero for each row pair (resulting in set intersection effectively). This is calculated via a Boolean matrix, X, where all elements > 0 are converted to 1:

$$H(A,B) = (A \circ B)_{i,j} = (A)_{ij}(B)_{ij}$$

$$X_{r,n} = bool(\, H(A,B)\,) = (p_{rn}) \in \{0,1\}$$

Likewise the sum of A and B was used to define a Boolean matrix U (set union) of elements. A row-wise sum of these Boolean matrices was then taken to obtain a count vector of intersections and unions. These are then used to give a result vector of Jaccard indices:

$$U_{r,n} = bool(\, A + B\,) = (q_{rn}) \in \{0,1\}$$

$$\boldsymbol{J_p(A,B) = \frac{rsum(X)}{rsum(U)} = \left[\frac{\sum_{i=1}^{n} p_{1i}}{\sum_{i=1}^{n} q_{1i}}, \ldots, \frac{\sum_{i=1}^{n} p_{ri}}{\sum_{i=1}^{n} q_{ri}}\right]}$$

The parallel version of the domain duplication function (DDSp) also returns a solution vector by using the element-wise maximum function form NumPy (156) and matrix subtraction. The *Exp* function simply takes the exponent of each element in the vector:

$$|A - B| = C = (a_{ri}) \qquad \max(A,B) = D = (\max(p_{ri}, q_{ri})_{ri}) = (b_{ri})$$

$$D_p = rsum(C) = \left[\sum_{i=1}^{n} a_{1i}, \ldots, \sum_{i=1}^{n} a_{ri}\right]$$

$$M_p = rsum(D) = \left[\sum_{i=1}^{n} b_{1i}, \ldots, \sum_{i=1}^{n} b_{ri}\right]$$

$$\boldsymbol{DDS_p(A,B) = Exp(-\frac{D_p}{M_p})}$$

Cluster pairs, A and B, are generated such that all possible combinations of (i,j) clusters are represented with an iterative approach using m-1 pairs of (A,B) sub-matrices. The approach matches each cluster once with every other cluster without taking order into account (Figure 6.2). This simplistic method generates all combinations by matching the first (m-i) rows with the last (m-i) rows as illustrated below (m' accounts for -1 indexing) for i > 0 and i < m:

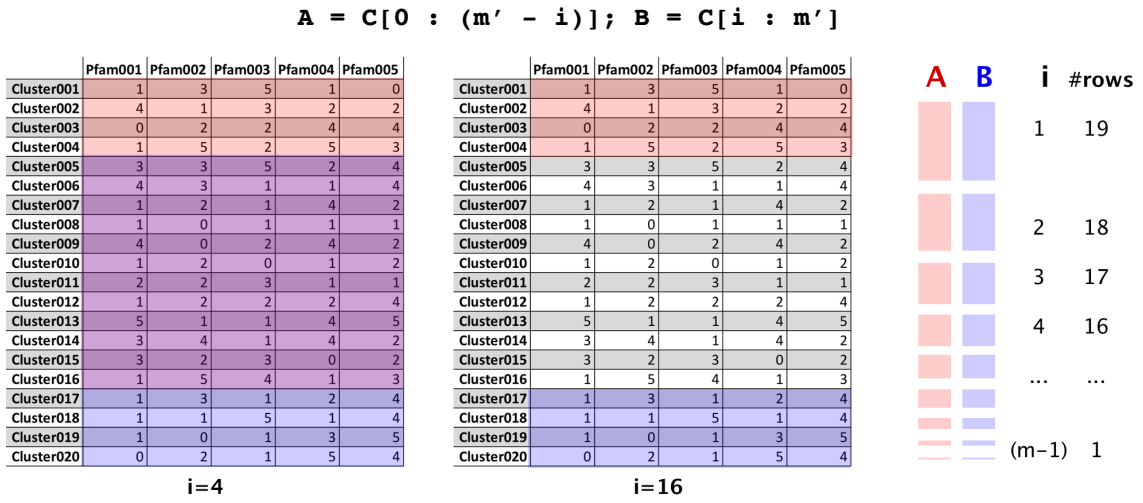

Figure 6.2: Illustration of sub matrix generation where top red sections produce the submatrix A, and bottom blue sections produce the submatrix B. The example count matrix C is a 20x5 set, thus m-1 (19) pairings are needed to produce all 190 combinations of cluster rows. This is equal to the number of combinations, C(n,2) = n!/(n-2)!2! = 190 = 1+2+3…+19.

Comparison of A and B pairs are further accelerated using the multiprocessing python toolkit to parallelize calls to $S_p(A,B)$. Results are finally collected and then written to a tabulated file of distances.

4.2.2 BGC Networking in Publication 2

Data Collection and Genome Sequencing

Data obtained from the JGI ABC database consisted of clusterfinder results of all public Actinobacteria isolates. The clusterfinder approach uses a probabilistic assignment of genomic regions to define BGCs and gives a prediction probability score for each cluster. FASTA sequences of all clusters with a score above 0.80 were then downloaded from: https://img.jgi.doe.gov/cgi-bin/abc along with Pfam (132) annotations. RMA genomes from the JGI database that were not included in the clusterfinder results were also

downloaded. Previously collected isolates from the Scripps Institute of Oceanography (SIO) were sequenced using 400bp Ion Torrent PGM sequencing as detailed in Schorn et al (152). BGCs for these genomes were then determined using antiSMASH 3.0 with clusterfinder enabled at a 0.8 threshold.

Gene Cluster Networking and Diversity Assessment

Pairwise similarities of all collected BGCs (detailed in section 6.2.1) were used to first de-replicate identical or nearly identical clusters by clustering those with similarities >= 0.99. The resulting list was used to identify all groups of nodes, Gene Cluster Families (GCFs), via Gephi's connected components function. One member was then kept per GCF and the number of de-replicated nodes was logged in the retained clusters's metadata. After all de-replicated nodes were removed from the initial similarity list a final network with a threshold greater than 0.6 was made using Gephi; The Yifan Hu (155) layout method was then used to visualize clusters. Gene cluster diversity was analyzed using common indices for species diversity such as the Shannon index (157). To normalize for different sample sizes of gene clusters the True Diversity index was calculated as a function of the Shannon index (158). Additionally the number of gene clusters that had at least one connection to any other cluster was compared to the total number of predicted clusters in a particular group to obtain that group's gene cluster uniqueness score.

4.2.3 BGC Networking in Publication 3

Data Collection and BGC Identification

Genomes from the genus Amycolatopsis were collected from the NCBI (142) and JGI-IMG databases (101) if they were high quality drafts with under 300 contigs and not from single cell sequencing. Additional strains from the Tuebingen collection, Amycolatopsis sp. H5 and KNN 50.9b, were sequenced using the Illumina HiSeq 1500 System as detailed in Adamek et al (153). Final assemblies were produced using the gsAssembler software (Newbler) v2.8 and submitted to the NCBI Prokaryotic Gene Annotation Pipeline for annotation (159). BGCs were then identified using antiSMASH v3.0 with default settings (143). A manual curation of all BGCs was done to improve border prediction accuracy as described in Adamek et al (160) using Artemis (161) to trim boundaries.

High-Resolution Phylogeny

A Multi-Locus Sequence Analysis (MLSA) was preformed to generate a high-resolution species tree of collected strains using several conserved housekeeping genes: atpD, clpB, gapA, gyrB, nuoD, pyrH, and rpoB. DNA alignments of each of the extracted genes were preformed using Clustal W (162). These were then concatenated to form a supermatrix of DNA sites. A Maximum Likelihood tree was then constructed in MEGA6 (163). Designation of major clades was also confirmed via pairwise ANIm values using JspeciesWS (164).

Manual GCF Identification and BGC Networking

Manual inspection using MultiGeneBlast (131) results and antiSMASH annotation were preformed on all clusters to define a curated set of GCFs. The following criterion was used to define a GCF: 1) BGCs share a similar genetic architecture showing a majority of genes that have the same function. 2) Genes required over a 50% BLAST amino acid similarity with at least 80% coverage to be considered similar. 3) Modular composition and BLAST similarity of KS and C domains for PKS and NRPS BGCs were also discriminated. These results were then recorded to a pairwise matrix for absence or presence and clustered using hierarchical cluster analysis in PAST (165). Rarefaction curves were produced from this manually curated set to determine BGC richness using EstimateS (166). Automated BGC networking was preformed (detailed in section 6.2.1) on all trimmed BGCs and a cutoff of 0.65 was chosen as it best reflected the manual curation set. Distance tables were then imported into cytoscape 3.4.0 for clustering and visualization (167).

4.3 Results

Performance Requirements

As the number of pairwise combinations for cluster comparisons would increase at roughly half an exponential rate a key design parameter was to implement a method that could accommodate the tens of thousands of clusters available in the JGI-ABC database. An initial implementation using an iterative approach was found to be limited in throughput despite using a multi-processing enabled workflow. Comparisons between the first iterative

approach and sub matrix parallel design were conducted by using a randomly sampled BGCs from the Actinobacteria dataset in publication 2. Tests using 1,000 to 5,000 clusters showed a dramatic speed difference of approximately 1860X (Figure 6.3).

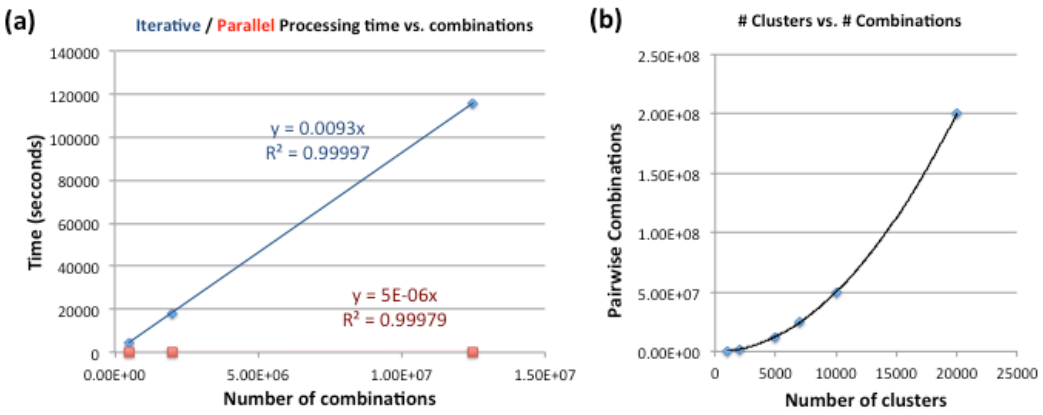

Figure 6.3: (a) Speed comparison of iterative approach (blue) vs parallel (red) shows an 1860x difference in combinations/second. (b) Amount of combinations needed for pairwise comparisons; equals n*(n-1)/2 combinations for n clusters.

The iterative approach finished in 32.3 hours for a 5000-cluster (~12.5 million combinations) test while the parallel version completed in 67 seconds (single CPU time). With this new implementation the large dataset of approximately 68 thousand clusters in publication 2 could then be computed and explored in a reasonable time. Memory requirements were higher for the parallel version but manageable at 6Gb for a 68K cluster comparison.

Networking Validation

BGCs de-replicated in the JGI ABC set showed to have identical annotations for known product where applicable. Annotated MiBIG clusters were also networked using this method to show that the similarity measures were effectively grouping identical compounds (Figure 6.4).

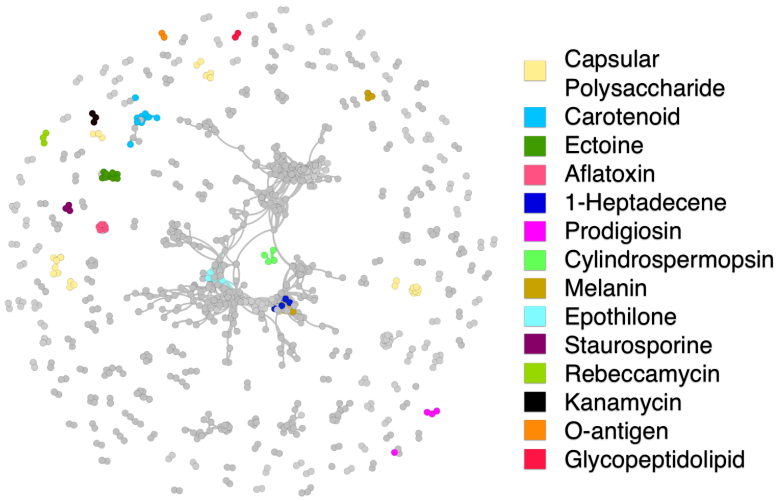

Figure 6.4: Networking of BGCs from MiBIG v1.3 that show groupings of frequent compound annotations. Nodes that did not show similarity above a 0.6 score are not displayed in the network.

As not all derivatives and related compounds are displayed the top identical product names were visualized in the network to confirm that they appear in the same GCF. This showed that the majority of those surveyed formed isolated groups while other clusters were found in subsections nested within a larger connected portion of the graph. General compound names, such as the capsular polysaccharides, form several isolated groups at the 0.6 similarity cutoff but were later seen to correlate to the assignment of compound based on their MiBIG structure annotation. These clusters were also apart of the "partial" sets in MiBIG so the possible incomplete clusters could affect the clustering efficiency. Likewise the separated carotenoid and prodigiosin nodes are marked as "partial" and only show similarity above a 0.5 threshold. Overall the majority of the compounds represent a single GCF illustrating that the network approach can be used to group identical and related compounds. BGC classes were also well organized in the network as seen in publication 4 with similar classes associating together in the network (Figure 2.4).

4.3.1 BGC networking in Publication 2: "Sequencing rare marine actinomycete genomes reveals high density of unique natural product biosynthetic gene clusters"

The collection of all Actinobacteria gene clusters filtered for those greater than a 0.8 clusterfinder probability resulted in 68,207 clusters from 4732 isolate genomes as of June

2016. After de-replication of nodes that showed a score of 0.99 and above, 22419 representative nodes remained. Many of these removed duplicates were seen to be identical clusters from sequencing projects of very similar species, and occasionally the same strain, as seen with *Mycobaterium Tuberculosis* (Over 2200 isolates sequenced as of May 2018). Only 1325 of the representative clusters were responsible for all redundancy with the top 32 de-replicated clusters accounting for half of the repetition. A histogram of all possible scores was used to determine the de-replication and final clustering thresholds (Figure 6.4); this choice was taken to obtain a conservative figure for cluster diversity while limiting the connectivity of nodes. Other thresholds were also tested and the metadata pertaining to BGC cluster type was analyzed to see how many edges had matching node type vs. mismatching types (Figure 6.5).

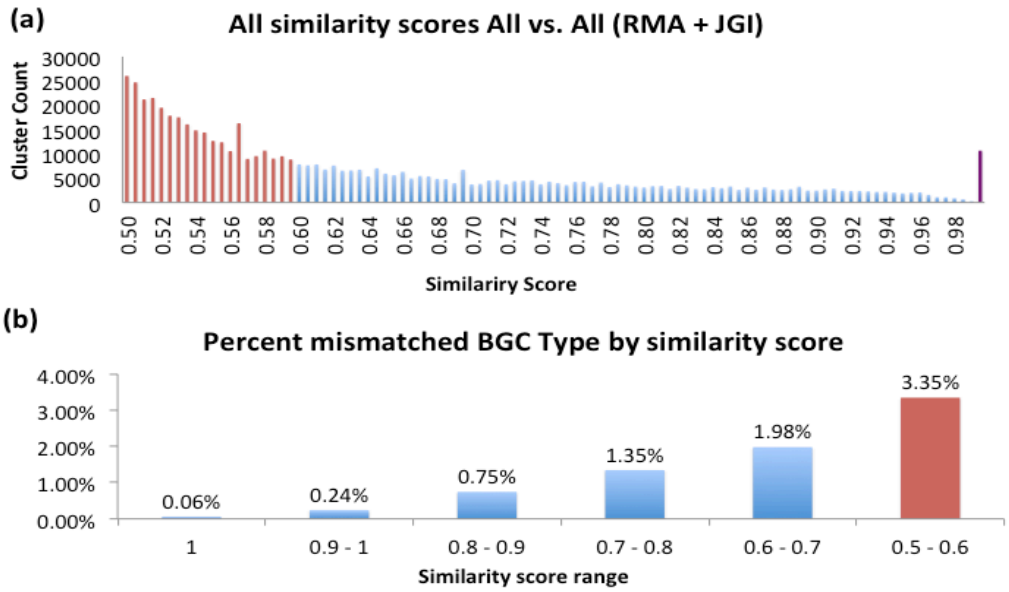

Figure 6.5: (a) Histogram depicting number of clusters binned at various edge values form 0.5 to 1.0. Selections greater than or equal to 0.99 were used to condense identical clusters (purple). The remaining edges above 0.6 were then used to define final connected GCF (blue). (b) Mismatched nodes as percentage of total edges at each binned similarity range. Red bar indicates edges that were omitted after cutoff selection.

As some annotations had multiple designations, ex: nrps-pks hybrid and nrps, a partial match with one or more annotations was considered a match. The number of mismatches was then calculated relative to the total number of edges for each binned similarity value. Because some discrepancies remained in the annotations, such as "null" or "putative"

designations, some mismatches (<2%) were tolerated. This results in a broader clustering by compound family and related core structures, which may under-represent compound diversity. Despite the underestimate of compound diversity 2970 GCFs were defined and split into two groups: one of highly connected nodes and the other with sparsely connected nodes to better visualize GCFs (Figure 6.6).

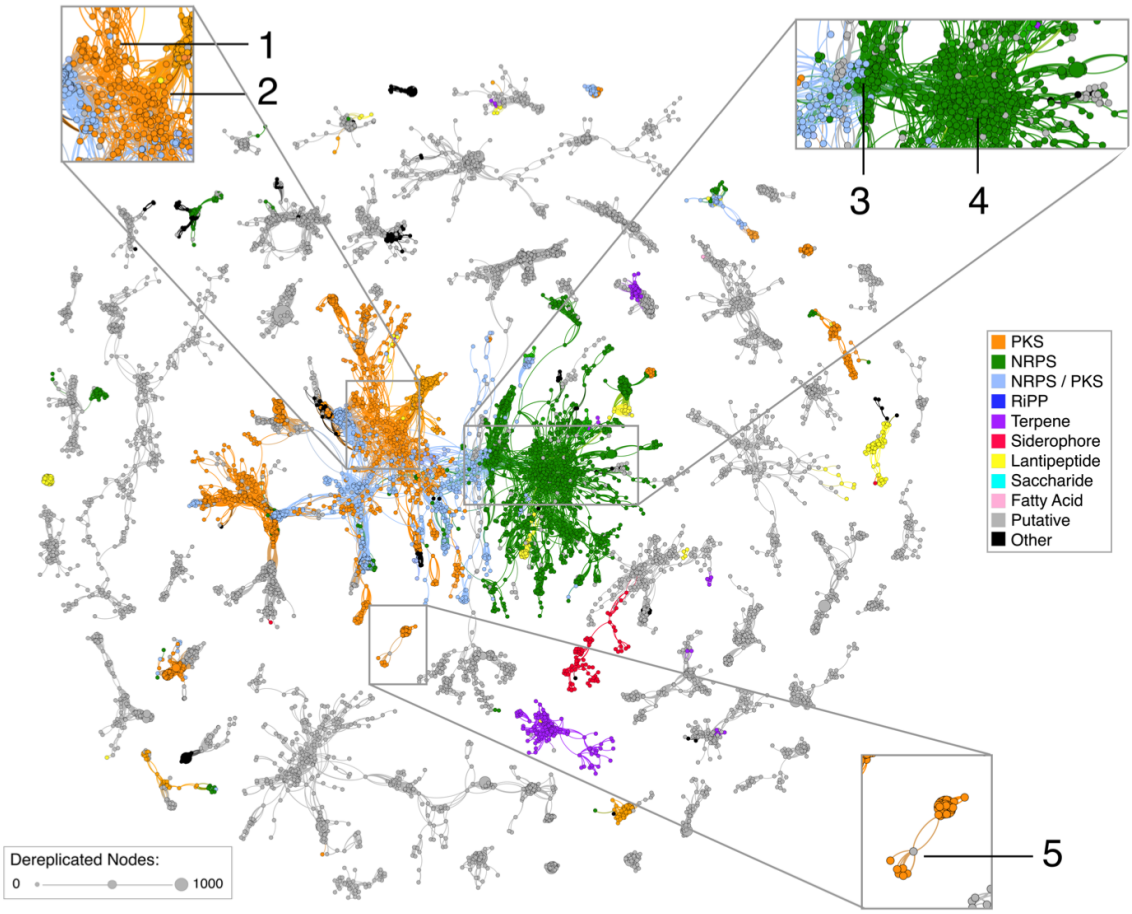

Figure 6.6: Highly connected portion of the network. Type I PKS macrolides such as oligomycin (1) and erythromycin (2) are within the larger PKS GCF. Siderophores, such as mycobactin (3) reside in the hybrid NRPS-PKS section. Cyclic depsipeptides, including homologues to pristinamycin (4) reside within the NRPS GCF. Rifamycin (5) and analogues form an isolated GCF. Figure adapted from Schorn et al. (152)

A useful method to correlate GCFs to compounds was to include MiBIG annotations. This too was a sparse measure of identifying compound groupings as only 3% of the total GCFs had at least one member from the MiBIG database. Those that did show annotations correlated with their respective BGC type, for example Type I PKS macrolides such as

oligomycin and erythromycin lie within this PKS GCF subsection. The large network shows many of the known classes of BGCs in isolated GCFs but the main connected component shows subsections that are not ideally resolved. For example PKS and NRPS clusters can be linked through the various NRPS/PKS hybrid clusters. The majority of the GCFs were seen in the second half of the network, which better illustrates diversity of RMA genomes (Figure 6.7).

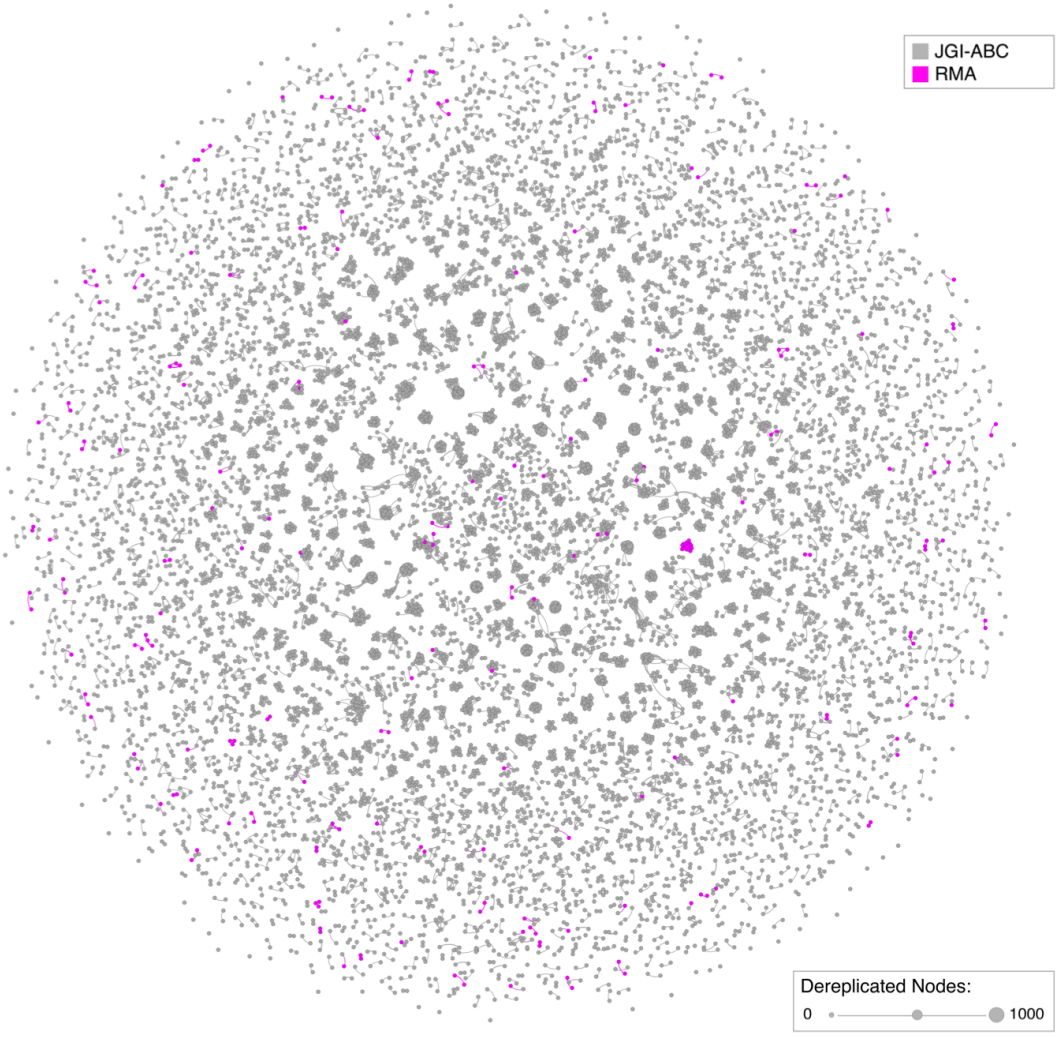

Figure 6.7: Majority of GCFs in 2nd network with BGCs from RMA genomes highlighted in pink. Figure adapted from Schorn et al. (152)

Here it can be seen that many of the RMA clusters form isolated, RMA only, communities with few that overlap with the other GCFs in the network. In addition to the lack of overlap in the resulting network, the number of BGCs that showed no significant similarity

(singletons) was 78% of the total BGCs from RMA genomes. This means a large portion of RMA secondary metabolite potential contained unique clusters not represented in the JGI Actinobacteria dataset. BGCs from marine *Streptomyces*, a promising source of natural products (168), were compared in order to assess the overlap. Additionally the diversity of GCFs, a measure of amount and size of GCFs, was also calculated and compared (Table 6.1).

	# BGCs	# Strains	# GCFs	% BGCs in network	True Diversity
RMA	1386	21	153	22.44	86.1595
Marine Streptomyces	1925	24	143	21.4	73.6128

Table 6.1: Network statistics for BGCs from RMA genomes compared with marine derived *Streptomyces*. Table adapted from Schorn et al (152)

The comparison shows very similar figures for GCF overlap despite slightly less isolate sampling in the RMA set. It is notable that despite the many sequenced terrestrial strains of *Streptomyces* in the JGI database the majority of marine derived clusters do not appear in the network (78.6%). Furthermore the shared clusters between RMA and the marine *Streptomyces* group only showed 4 GCFs (3% of RMA). This implies that there is still low sampling from both groups and more genomes from these sources would expand the known chemical space. Besides the low overlap and chance of rediscovery, the diversity of clusters also showed to be relatively high. True diversity, interpreted roughly as the effective number of GCFs if they were equally populated, showed similar potential to marine *Streptomyces*, which means many different chemical structures can be found in these sources.

4.3.2 BGC Networking in Publication 3: "Comparative genome mining reveals phylogenetic distribution patterns of secondary metabolites in Amycolatopsis"

The results from this study helped to compare manually defined GCFs with the same automated approach in publication 2. The majority of the GCFs at a threshold of 0.65 matched with the manually defined clusters but for some there was a merging of GCFs, which was expected using this global threshold as seen in publication 2. The example in Figure 6.8 shows the inclusion of clusters that had similar gene architectures but were

separated in the manual process based on C-domain similarity and module arrangement, which is not considered in the automated similarity method. The broader BGC class designation was shown to be consistent across all GCFs with the exception of one hybrid cluster included in an NRPS GCF.

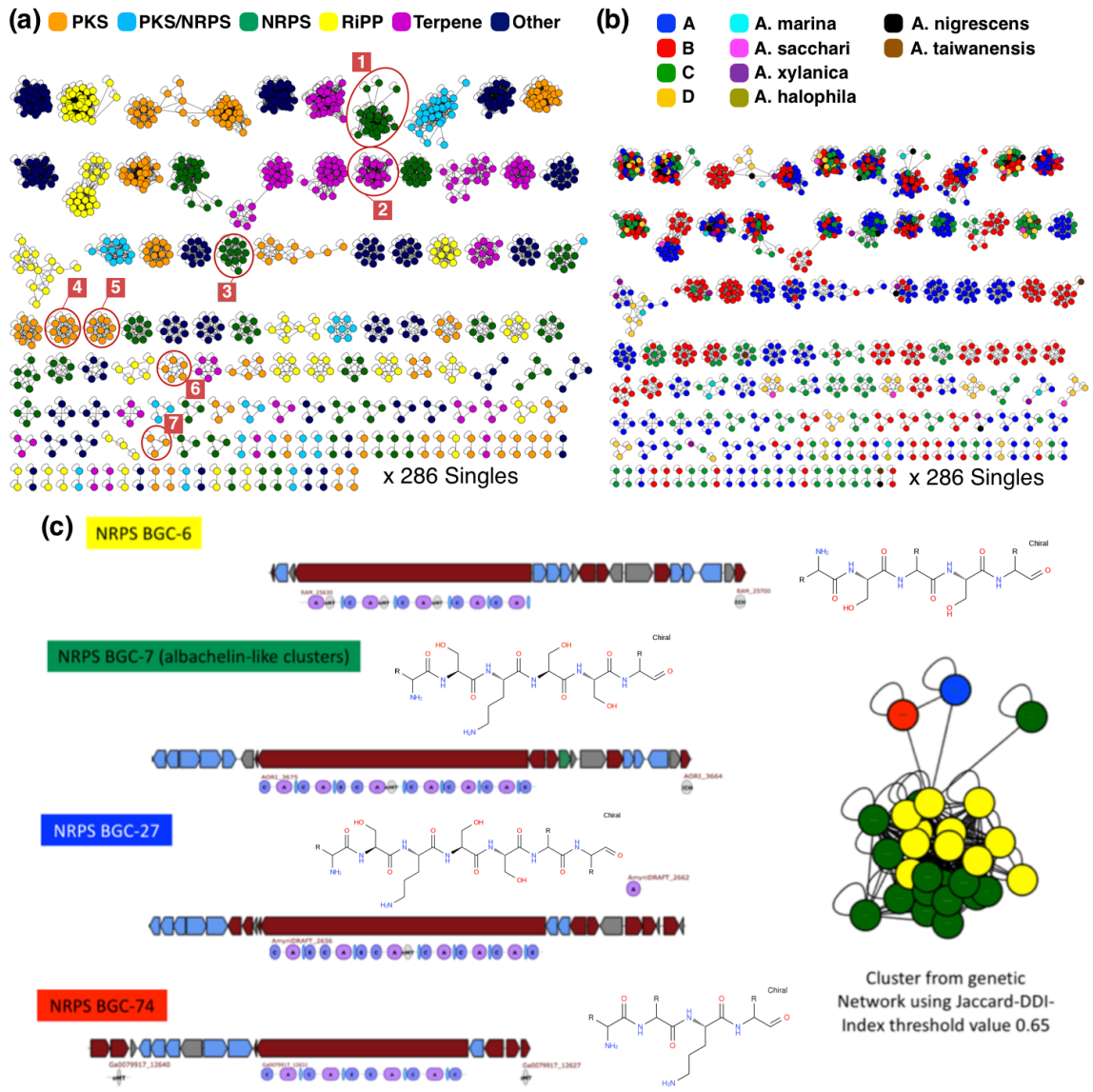

Figure 6.8: Gene cluster networking of *Amycolatopsis* strains. (a) Network colored by BGC class. 1. Albachelin-like NRPS clusters; 2. 2-methylisoborneol; 3. Glycopeptides; 4. Rifamycin; 5. ECO-0501; 6. macrotermycin-like PKS clusters; 7. Octacosamicin. (b) Networked colored by major phylogenetic groupings defined by MLSA (A,B,C,D) and other organisms. (c) Example of automated networking that include related NRPS BGCs (predicted structures shown) but were defined separately using manual criteria. Figure adapted from Adamek et al. (153)

Major phylogenetic clades were found via MLSA in this study and were used to annotate the network in Figure 6.8b. Interestingly, these major clades were shown to correlate to certain GCFs forming isolated clade specific communities. This clade specific metabolite potential also corroborated the heat-map analysis using manual GCFs performed in this study. The number of singleton clusters was another aspect that showed species specific potential with 21% of the total BGCs showing no similarity to each other; this implies an addition of about 6-7 unique BGCs on average for every strain that is sequenced. There were also some universally shared GCFs and others that were only shared between two clades. Notably, over half of the highlighted known compound groups happen to be multi-species GCFs suggesting a prioritization for shared clusters might be useful. These antibiotic clusters provide broad benefit and therefore perhaps many species have taken advantage of them, this idea has also been previously suggested (62). The combination of phylogenetic classification and network display thus provided a rapid survey of metabolite variety. Clades A, B, and C harbor the highest richness and diversity and so efforts spent on these species may have higher returns for new compounds; In particular, clade A shows to account for nearly half of the non-singleton GCF diversity (46%). In contrast, those from clade D have the majority of its non-singleton GCFs represented in the latter clades. This observation was also reflected in the average number of BGCs per isolate with 37, 34, and 30 clusters for clades A, B, and C respectively. An average of 18 for group D and related organisms, *A. sacchari* and *A. taiwanensis*, was shown to help prioritize efforts toward sources from the later clades.

Besides investigating diversity and uniqueness, identification of known compounds were used to highlight GCFs with potentially novel compounds. While several GCFs formed isolated known compounds, this compound grouping was not perfect at this threshold as it was shown that some similar but distinct BGCs were found in an automated GCF (Figure 6.8c). The difference between NRPS BGC-7 and NRPS BGC-27 for example were grouped because of the highly similar architecture but showed to have borderline C-domain BLAST similarity at around 51-60%. These inclusions were seen to only have a single connection to the group however so would likely be resolved at higher similarity score thresholds. The inclusion of a hybrid cluster with the only mismatched automated GCF was also only connected by a single edge. Given these were related BGCs it could be useful to have this

larger binning of compounds in order to prioritize those with drastically different structures from known BGCs however.

The compound examples shown were from manual identification via the included MultiGeneBlast analysis in antiSMASH. To test the automated identification of known compounds the same networking method was used with the inclusion of gene clusters from the MiBIG v1.3 database. This showed to not only highlight most of the previously identified compound groups but also several other GCFs with known members from diverse phylogenetic origins (Figure 6.9).

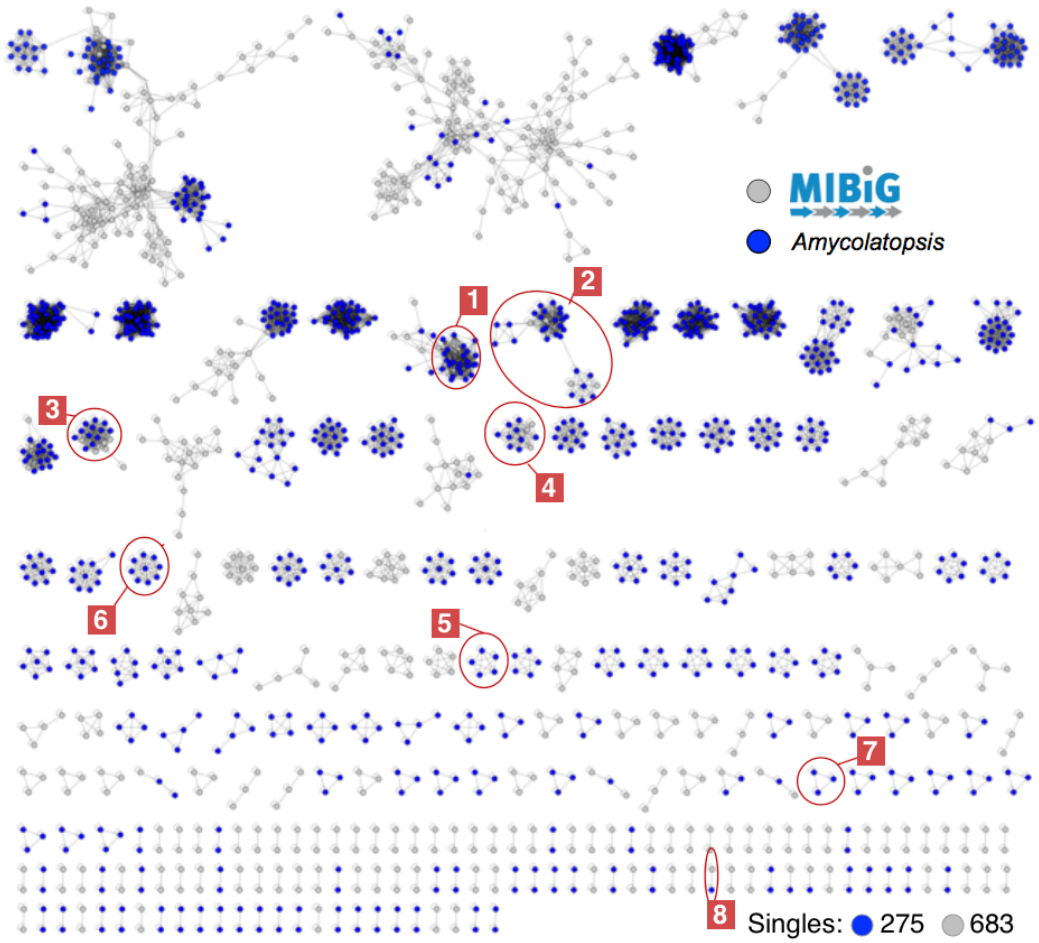

Figure 6.9: Amycolatopsis BGCs networked with known BGCs from the MiBIG database. 1. albachelin-like NRPS and similar clusters (see fig S6); 2. 2-methylisoborneol; 3. glycopeptides; 4. rifamycin; 5. ECO-0501; 6. macrotermycin-like PKS clusters; 7. octacosamicin; 8. chelocardin

With the added database unknown links could also be established as with the chelocardin cluster. A total of 11 unknown singletons could then be identified in this way. The larger networks also showed to span and connect previously isolated GCFs helping to establish related families. However the results mostly show that the majority of the *Amycolatopsis* GCFs are without significant similarity to what is in the MiBIG database. Compared to the original 133 GCFs 111 remain that do not have any connection to MiBIG nodes. These results show that the *Amycolatopsis* genomes are a promising source with little overlap to experimentally verified natural products.

4.4 Discussion

This implementation of gene cluster networking showed to be a rapid automated approach not only to de-replicate known compounds, but also to survey cluster diversity and prioritize uniqueness. Highly similar gene clusters and their compounds were shown to associate reliably with the simplistic Pfam composition metric in reasonable time after implementing the parallel vector method. Not shown here was the processing time for Pfam annotation, as this was previously calculated from the JGI-ABC database, however this step is rapidly achieved using the HMMER3 package (116) and is usually included for major genome annotation pipelines. With the processing time improvements this method is scalable for big-data applications shown to cope with tens of thousands of putative clusters. An extrapolation of 1 million clusters would result in a reasonable 29 days of single CPU time to calculate all comparisons. This would be infeasible using MultiGeneBlast or other BLAST similarity approaches, which might also require manual curation. As more genomes and meta-genomes are sequenced, and as sequencing throughput increases, this aspect of throughput is crucial to the downstream analysis of genomic data. As illustrated in publication 2, the initial clustering of identical compounds with the 0.99 thresholds resulted in a significant amount of de-replication. Nearly 67% of the data was shown to be redundant due to the multiple sequencing projects with identical species included in the Actinobacteria dataset. Although all combinations needed to be tested for similarity first this reduction helped to expedite network exploration and processing time for visualizing the final network.

Subsequently the new non-redundant set was used to identify known compounds and help determine which sources harbored the highest GCF diversity. As seen with the low overlap to previously sequenced Actinobacterial genomes in publication 2, the RMA

potential was shown to be a promising source for natural product discovery. Furthermore the distribution of these products into many isolated GCFs showed that higher chances of chemical diversity could be seen and that this group, along with other marine *Streptomyces*, warrants more sampling. This networking approach helped to solve two problems in the natural products pipeline: eliminating rediscovery of known compounds, and prioritization of sources based on potential novelty and diversity. An alternative prioritization prospect, as seen in publication 3, is to focus on GCFs that show heterogeneous phylogenetic composition, as several of the known antibiotic compounds were shared across different taxa. This idea that more ubiquitous ecologically advantageous compounds, such as antibiotics, have a higher chance of being shared across different genera has also been previously argued (62). Besides these prioritization possibilities the rapid screening of known compounds was demonstrated in publication 3 with the inclusion of the MiBIG database. This showed that several of the *Amacolatopsis* BGCs were unknown and harbor potentially novel compounds. This approach will only improve overtime as more contributions are made expanding and centralizing known compounds form BGCs. Currently this is somewhat limited as the majority of cataloged clusters are without an experimentally verified product, known as "orphan clusters". Once these links are established, as structure prediction and experimental evidence improves, this approach can lead to effective elimination of rediscovery of natural products.

In addition to the fast processing time the clustering accuracy was shown to be reliable by associated known compound groups. Validation using the MiBIG database of known clusters showed to group identical compounds into isolated sections; this was also seen in publications 2 and 3. In publication 3 there was clear separation of BGC classes with the exception of 1 hybrid cluster. With the larger JGI dataset more mixing was seen for the highly connected GCFs in publication 2. This cross connectivity was seen in the larger network due to the inclusion of more hybrids, which bridged PKS and NRPS classes. One factor could be the difference between trimmed vs. untrimmed clusters in the two publications. This difference was not tested but it highlights that automated cluster prediction may include some flanking genes that do not participate in the cluster which can lead to lower or higher scores depending on genomic context. Fortunately improvements to automated cluster boundary prediction have been developed in the new release of antiSMASH v4.0 (112), so this effect can be reduced. One solution to this mixing issue, if

BGC class is known beforehand, is to network each class separately to better distinguish GCFs from related groupings. This approach is used in a recent BGC networking program BiG-SCAPE (150).

Compared to the manually grouped approach in publication 3, which took several days of laborious work, the automated GFCs showed agreement for the majority of families but with some groupings occasionally including distinct but related BGCs. The example shown highlights this drawback but based on the compound predictions this discrepancy might be favorable as all of the structures shared a similar core scaffold which means GCFs will be conservatively defined leading to highlights for distinct core structures. Considering the time saved and throughput allowed from this method this shortcoming is acceptable, and the broader groupings may lead to more reliable prioritizations of novel compounds. Increasing the similarity thresholds is a solution to improving the resolving power however this comes at a cost of over estimation of compound diversity and higher singleton count. Other ideas for more resolved GCFs were explored but not implemented in this study such as using additional local cutoffs for large highly connected clusters or adaptive thresholds based on BGC type, size, or composition. Ultimately the simplistic approach was taken to ensure no bias toward a higher compound diversity was generated and to conservatively estimate novel prospects. A potential improvement to the GCF definitions could be made in future implementations by not only considering edge weight but also node connectivity. As seen in publication 3 some of the inaccurate cluster inclusions showed only a single connection to the GCF. This could be used for example to refine GCFs by examining all sparsely connected nodes via more strict cutoffs using a secondary threshold. Another interesting solution is to define GCFs using the Markov Cluster Algorithm (MCL), which can identify local groupings in large connected sections by simulating stochastic flow between connections in the graph (169). This method was not tested in these studies but can be applied using the current output format of the networking scripts.

Overall this BGC networking implementation was show to be effective and capable of de-replicating and prioritizing the wealth of publicly available BGC predictions as shown in publication 2 and 3. This demonstration has shown the basic application of combining databases of known natural products and prospecting GCF diversity but this method also enables a range of comparative analysis. As seen with the combination of phylogenetic classification metadata, it is possible to identify products shared among a broad range of

taxa, which might serve as a clue that the product is an advantageous compound. A variety of other metadata could also be used to cross-reference the network such as the inclusion of known resistance factors, known drug targets, or results from bioassays. This integration of bioassay data was demonstrated to work well for molecular network approaches (170). One interesting possibility is to cross reference gene cluster networking with molecular networks, which can help give an intersection of predictions for higher confidence leads; this can also be used to aid structure elucidation of an unknown cluster or establish a predictive link between orphan clusters and compounds. These comparative possibilities have only started in recent years and are likely to improve as the number of available genomes and meta-genomes continues to grow. With the increase in these diverse datasets, and expansion of known product databases, this gene cluster networking method is already showing promise to reinvigorate the natural products discovery pipeline (171).

Chapter 2

5 Automated high-resolution species trees (autoMLST)

5.1 Introduction

As demonstrated in chapter 1, natural product sources can vary significantly in secondary metabolite potential not only from a broad phylogenetic perspective but also within the same genus (153). Even within the same species there can be variation as seen in *Verrucosispora* and *Salinaspora* strains (172, 173). Accurate taxonomic classification of bacterial isolates is therefore an important tool to help identify viable sources and reduce natural product rediscovery (174). Understanding the true evolutionary phylogeny also has important applications to a variety of research. In natural products research it is often helpful to express the Biosynthetic Gene Clusters (BGCs) that encode these compounds of interest in a heterologous host due to issues with cultivation, expression in the native organism, or to avoid handling pathogenic strains. Choosing the closest relative of an isolate to increase compatibility of the transplanted BGC is a common idea as GC percent and codon usage will be similar. For example, a 100X increase in production was seen with tubulysin expression in a host more related to the native organism (175). Thus an accurate classification can identify a suitable host that has similar metabolic context as the source. This benefit is especially helpful when complicated pathways or unique precursor supply chains are involved (176). In addition to uncovering sets of related organisms for comparative analysis, a full species phylogeny can also help to highlight horizontal gene transfer via reconciliation of individual gene trees with the species tree (177). Phylogenetic background can also provide clues to BGC function as demonstrated in publication 5 where the acquisition of a new group of siderophores (salinichelins) coincided with the loss of known desferrioxamine clusters due to functional redundancy (178). Understanding the evolutionary background of a species is an informative tool but also simply providing a rigorous taxonomic identification can reduce errors in genome databases and help to guide experiments with proper comparative context. As current classification schemes still remain a challenge (179) systematic genomic classifications, such as using whole genome Average

Nucleotide Identity (ANI), are becoming a popular proposal to solve the taxonomy difficulties for prokaryotes (180).

Classical species delineation of bacterial isolates has utilized morphological and chemical properties, which remain an important factor in defining a type strain – an isolate that represents a particular species based on rigorous phenotypic and genomic criteria. Genomic data has long been in use though DNA-DNA hybridization techniques however due to labor intensive and time consuming processing this method has been largely supplanted by genomic sequencing of conserved areas, mainly the 16S ribosomal RNA sequences present in all bacteria (181). While this quick and inexpensive sequence to obtain has served to delineate much of the known genomes in public databases, there still remain some difficulties with using 16S sequencing alone. Using such a conserved region that maintains a strong purifying selection has its benefits as a taxonomic marker, however these same properties can also lead to low phylogenetic signal and reduced resolving power at the strain or species level as evidenced by the high similarity definition of species 98-99% (182). Additionally, complications when using partial 16S sequences (183) or selection of a sequence in an organism that contains multiple variants (184) can be a source of misleading designations. Despite these drawbacks 16S sequencing has remained the workhorse of taxonomic identification of submitted genomes due to large catalogues of these sequences that enable a rapid classification via similarity screens with tools such as BLAST (117, 185–187). This method continues to be the most widely used due to low sequencing costs and highly accessible rapid processing. While similarity serves as a useful heuristic to evolutionary history, it is important to note that the commonly used blast identity measure can lead to multiple species designations in genera with highly similar sequences (188).

An alternative to similarity scores is to model evolutionary history using phylogenetic methods that take into account parameters of evolution such as higher rates of transitions compared with transversions. Techniques to calculate a resulting tree of sequence evolution range from faster distance based methods, Unweighted Pair Group Method with Arithmetic Mean (UPGMA) and Neighbour-joining (NJ), to computationally rigorous character-based approaches: Maximum parsimony (MP), Maximum-likelihood (ML), and Bayesian Inference (BI). These later methods rely on an alignment of sequences where differences are interrogated to explain how they have evolved and are related. In general the character-based approaches can search and evaluate many hypothesis to arrive at a more accurate result

(189). These methods rely on evaluating many topologies, which require a computationally expensive search considering there are (2n - 3)!! (1*3*5*7…2n-3) possible rooted tree topologies (190); fortunately efficient non-exhaustive algorithms have roughly $O(n^2)$ complexity for n species using the Sub-tree Prune and Re-graft (SPR) approach of tree searching (191). Several models of evolution can be used which range from those with simplistic constant rates of evolution and few parameters to sophisticated models accounting for a variety of parameters. This extensive modeling can lead to a more accurate hypothesis of species relationship than from similarity alone (191). Unfortunately the variety of processing techniques discourages widespread use as best practices are not immediately apparent to non-specialists. Recently this barrier to use is being reduced through accessible web interfaces that utilize the computationally expensive ML approaches, such as IQ-TREE (192, 193) and RaxML (194). Additional measures such as model finding, included in IQ-TREE (195), automatically detects the most simplistic model that best explains the genetic data; this is a valuable step as the choice of model can give varied results using likelihood or Bayesian methods (196). These advancements have made it easier to perform a more rigorous analysis to identify the evolutionary relationship between a set of genes. However this process can often be insufficient in delineating confidant species splits using 16S data alone. The use of Multi-Locus Sequence Analysis (MLSA), a technique that integrates many genomic loci to increase phylogenetic signal, has shown improved resolving power and highlights the inaccuracies of relying on 16S data only (197). MLSA trees have primarily been processed by concatenating all aligned genes into a super-matrix as input for tree inference. Other approaches infer a species tree by first building gene trees separately and combining them using coalescent theory as this can be beneficial for recent or rapidly diverging lineages, or if other complications are present (198, 199). However the choice of which genomic loci is important as using genes subject to HGT can impair accurate estimation (200). Criteria such as using single copy ubiquitous housekeeping genes, restricted to genes with low synonymous vs. non-synonymous (dN/dS) mutations, help to focus on vertically inherited genes with low phylogenetic noise (201). Currently this has been a manual process that depends on the organisms in question but generally includes various ribosomal genes, and other ubiquitous essential genes such as DNA/RNA polymerase proteins as seen from the public database pubMLST (202).

To expedite this process and bring these methods into more widespread use we present the Automated Multi-Locus Species Tree (autoMLST). The goal of this project is to provide a simplistic "BLAST like" interface that can automate each step of this workflow and provide easy access to high-resolution methods of species inference. While existing sites aim to have similar accessibility to these methods such as EDGAR (203) and PATRIC (204), these only automate the tree inference step; additionally these methods do not include features such as model finding or use faster forms of tree inference such as NJ methods. From selection of appropriate reference organisms to final tree construction the autoMLST server aims to provide a quick means of obtaining an initial species designation by automating each step of the process. Also by providing annotations such as pairwise ANI estimates and BGC potential of a particular clade, autoMLST can help with discriminating query genomes for natural product prioritization. The application is presented in two pipelines: one that utilizes rapid placement on predefined trees, and a de-novo approach that handles costly computational time by limiting to the most relevant organisms.

5.2 Methods

Workflow Overview

This workflow is designed to automate MLSA tasks including selection of genomes to final tree construction (Figure 5.1).

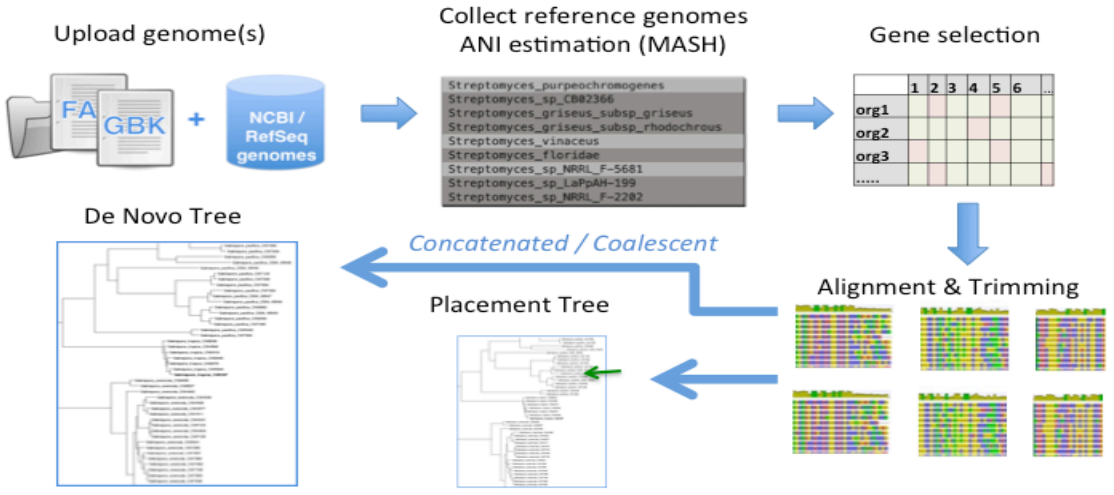

Figure 5.1: autoMLST workflow to automate reference genome and single copy gene selection. Alignment and trimming of each gene is first preformed then depending on user selection a placement tree or de novo tree is built. The user can also select to use a concatenated alignment or coalescent approach to build the final tree.

All front-end code and workflow scripts are open source and available at: https://bitbucket.org/ziemertlab/automlst. First, query genomes are screened against reference and type strains genomes using rapid ANI estimation via MASH (205). The nearest organisms are then selected and ubiquitous essential genes that appear in single copy are identified. Alignments of each gene are performed followed by automated refining using the "automated1" option in trimAL (206); this is designed to improve accuracy of maximum likelihood trees by removing aligned sites with a high proportion of gaps and variability. Finally a tree is produced using a rapid placement method or full inference from a maximum of 20 query genomes per run.

Reference Genomes and Build Process

Reference genomes were obtained from NCBI Refseq (207) in September 2017 and genbank files were then converted to a SQL database which includes all sequences and taxon metadata. To reduce redundant strains the top 10 highest quality genomes were retained for genomes that showed the same species taxid, as determined using the most complete "assembly level" metadata and lowest scaffold count; Any genomes marked as type strain or reference genome were retained and all that had ambiguous designations for genus were removed. The application begins by parsing all user submitted FASTA or GenBank sequences and uses chromosomal sequences to perform ANI estimation using MASH (205). Reference genomes with the highest ANI to each query and average ANI to the entire query group are then selected. Preference for type strains is given by allowing higher distances (+5% ANI) when selecting the nearest reference organisms. All sequences are then integrated with the query sequences into a final temporary SQL database. Searches for gene homologs are preformed using HMMER (116) and essential gene models. These models were collected from Pfam (132) and "equivologs", orthologous genes with confirmed conserved functions, from TIGRFAM (208); reference genomes have these searches pre-computed. The results are then added to the database where a gene matrix of all organisms is produced. The resulting matrix is screened for all single copy genes present in every genome and prioritized via pre-calculated dN/dS values. These values were determined based on codon alignments of reference organisms using Pal2Nal (209) and the PAML (210) application "yn00" and averaged. To reduce computation time a maximum of 100 genes are selected from the prioritized gene set. Nucleotide alignments of all genes are preformed

using MAFFT (211) which can perform in fast mode (FFT-NS-2) or local iterative (L-INS-i) as an option. Guide trees are also written during alignment with the "treeout" option; these are optionally used to further filter gene selection by removing trees with the highest median distance to all other trees (up to one standard deviation). Depending on the selection of workflow, trees are built with IQ-TREE (de novo workflow) or placed onto reference gene trees using the Evolutionary Placement Algorithm (EPA) in RaxML (212). The de novo workflow also has two modes: concatenated alignment inference or coalescent tree inference. This allows the comparison of both methods to identify areas of the tree that might be problematic. Each step is automated by default but can also be manually refined for gene selection and organism selection. A reanalyze button in the final results makes this process easier if a user wants to compare other gene sets or organisms.

For the rapid placement workflow, reference sets of families were built using type strains with identical NCBI family IDs. These sets included all families that had 10 or more members marked as type strain or reference genome and showed 10 or more single copy genes. Each set was built by running the command line version of autoMLST to obtain and align all unfiltered single copy genes. Gene trees were then built using IQ-TREE with 1000 bootstrap replicates and General Time Reversible (GTR) model. This was done using the ultra-fast bootstrap setting UFboot2 (213). Query genomes are matched with the applicable reference set by using the top MASH distances to find a reference set, or if none is found the user is notified. The single copy genes from the query organisms are then added to the reference alignments with the "--add" option in MAFFT. The updated alignments are used in conjunction with the precompiled reference tress and RaxML EPA to produce gene trees with placed query sequences. A final coalescent tree is then inferred from all gene trees with the ASTRAL-III v5.5 (214) application.

For the de novo approach, a total of 50 organisms including up to 20 query genomes are used as input. Alignment and refining are also done with MAFFT and trimAL. If the extra screening options are enabled the guide trees produced in the MAFFT alignment step are used to remove genes that show conflicting topologies. This is done by discriminating median pairwise Robinson-Foulds (RF) distances of each gene tree to the group of every other tree. This value represents the symmetric difference of splits in one tree but not the other. The highest values within one standard deviation of all median RF distances are then removed. After all single copy genes have been aligned they are either concatenated into a

partitioned alignment by default or optionally used directly to build a coalescent species tree using ASTRAL. The partitioned alignment is processed with the partition-aware features in IQ-TREE (215) which allows for gene specific parameters of evolution. Model selection and bootstrapping are also optionally performed in the same step. The final tree is then displayed in the browser, which allows for dynamic coloring schemes to depict type strains and organisms that belong to similar ANI groupings.

ANI Clans and Validation

As bacterial species definitions remain a challenge, with known misnomers and ambiguous assignment due to human error (216), we decided to use a systematic approach using ANI for validation. Definition of ANI "clans", groups of organisms with closely matching ANI values, were based on pairwise MASH distances of all reference genomes whereby all distances below various thresholds were used as input for Markov clustering using the MCL application (169). MASH distances were converted to yield a percent ANI to obtain an edge weight for clustering and three analyses were made at 97, 95, and 90 percent ANI thresholds. Below 90 percent was omitted due to higher rates of error in the ANI estimation as reported in MASH (205). Each group was given a unique clan ID and a final translation file associates each genome with the clan IDs at each threshold. These designations are directly used in final tree visualization by highlighting all non-singleton clans so that the user can quickly estimate clade distinctions or problems with the evolutionary hypothesis (Figure 5.2).

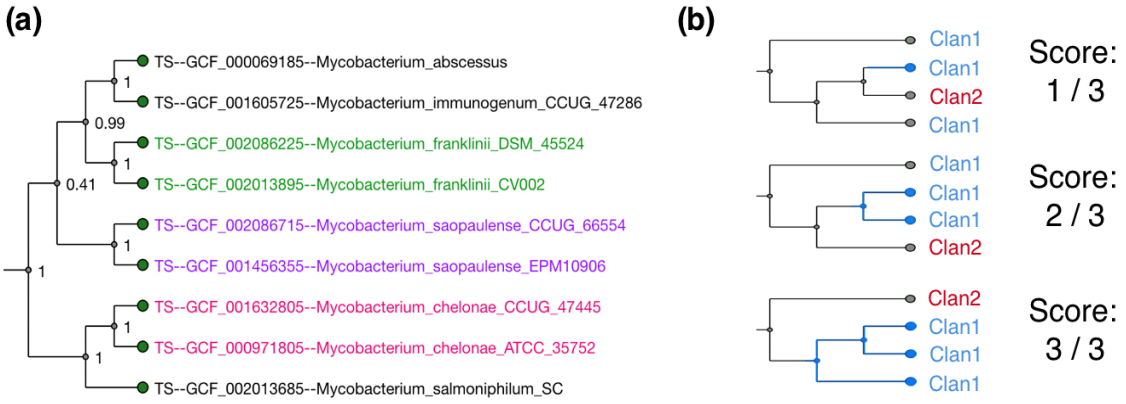

Figure 5.2: (a) Example of ANI clan visualization with *Mycobacterium* type strain (or representative genome) where a 97% threshold is shown. (b) Conceptual example of ANI scoring calculation using strict monophyletic branches as all descendants of the least common ancestor of clan. If multiple monophyletic groups exist the largest will be selected.

These groupings were also used to validate generated trees by checking if related genomes clade together on tree branches; this was done by using the Environment for Tree Exploration (ETE3) python library (217) to identify the largest monophyletic group (strictly homogeneous) for each ANI clan. The proportion of maximum monophyletic members to the total was then used to assess tree placement; a score of 100% would be given if all members appear in one branch with no other genomes included. This is done for every non-singleton ANI clan and the average is reported for each tree tested at various ANI clan definitions. All previously generated concatenated family trees composed of type strain organisms or reference genome were used for the validation.

To include more singleton ANI groupings and inspect branch length designations, a test of pairwise ANI values were correlated with pairwise tree distances between all nodes. A measurement of one leaf to another was done via multiple calls to "get_distance" in ETE3 and all pairs were matched with pairwise ANI values over 85%. The Pearson correlation coefficient and corresponding p-values were then automated using the Scipy library (218). Tree coverage, the ratio of leaves sampled to determine score, was also calculated by taking all leaves that had at least one data point divided by total leaves. Further validation of all branches was done by inspecting leaf and internal node bootstrap support values summarized via the average and ratio of well-supported splits in a given tree; this ratio was calculated as the number of supports equal to or greater than 80 divided by the total number of support values. Support values for the coalescent method uses a slightly different calculation of "local posterior probability" where quartet branches from all gene trees are used to define the probability of that topology (219). With the concatenated approach a random sampling (bootstrapping) of aligned sites are used to generate many resulting trees to infer support (213).

A comparison case study was also preformed using the manual high-resolution *Amycolatopsis* phylogeny generated in publication 3. These were submitted to the webserver after removing the restriction to number of genomes and ran using default settings with the de novo workflow in concatenated mode. Tree visualization was done using the tanglegram algorithm in Dendroscope (220) to compare topologies and similar groupings defined previously were highlighted.

Performance Testing

All speed tests were performed on a development webserver with 6 virtual cores equivalent of a 2.2GHz Xenon processor. Job submissions were made through the web interface to simulate real-world usage. A subsection of 40 genomes, ranging from 1-10 Mega-base-pair (Mbp), were taken at random from different families that were apart of the placement workflow. These were used so that we could also ensure proper placement of queries by relating with the corresponding reference genome in the tree. Tests of three workflows were made for each genome: Model Finder Plus (MFP) with bootstrapping enabled, de novo workflow with default options, and the placement workflow with default options. Start and complete times were taken from the run logs and used to generate average processing times and variance.

5.3 Results

5.3.1 Reference Generation and Performance

Over 22,000 genomes were integrated into the SQL database after limiting downloaded genomes to 10 per identical NCBI species ID. These genomes spanned 393 families and 1,995 different genera. The distribution of families was fairly even for the top 21 groups, which represented roughly half of the genomes in the database (Figure 5.3a).

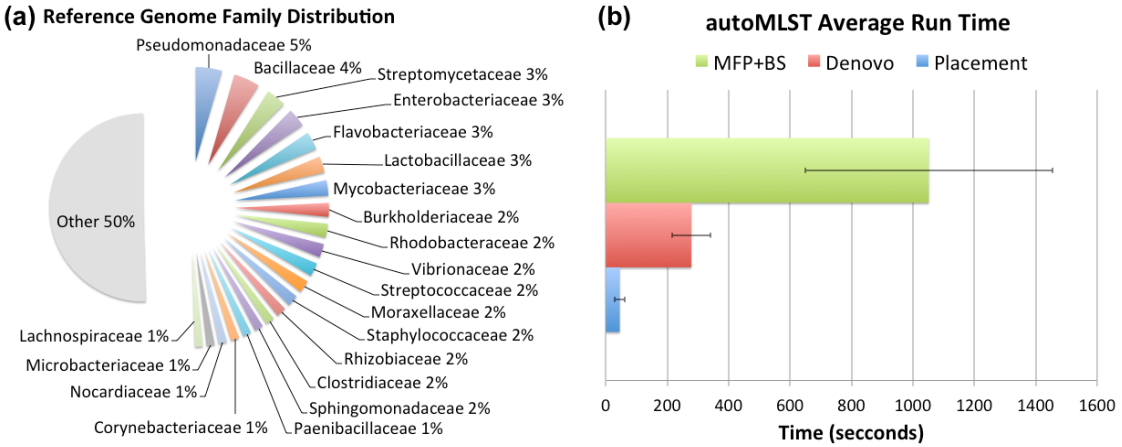

Figure 5.3: (a) Distribution of reference genomes by family showing the top 21 families comprising 50% of the database. (b) Average run times for 40 test runs using genomes from different families. Three separate runs were made per genome using model finding and bootstrap options (green), default de novo workflow (red), and default placement workflow (blue)

In contrast 119 families showed only one or two isolates to represent the entire family. With respect to genus the top 6 groups accounted for 18.1% of the reference genomes. *Pseudomonas*, *Streptomyces*, and *Bacillus* were the top sequenced isolates accounting for 10.7% of the reference, followed by *Mycobacterium*, *Lactobacillus*, *Streptococcus*, and *Staphylococcus*. The remaining top half of the database had fairly even coverage between 0.35 – 1.75% representing 50 genera in total. However, 1168 genera only contained 1 or 2 genomes as representatives. Genomes marked as type strain or representative genome showed 6060 isolates spanning 354 families. For the placement workflow a total of 128 families were found to have over 10 members, ranging from 11 to 313 members, and were used to generate reference trees.

Application performance was shown to be rapid with all 120 test-runs showing no reported errors. The placement workflow showed to have an average runtime of 45 seconds with a 6X and 23X increase in time for the de novo and MFP runs respectively (Figure 5.3b). The de novo workflow ranged from 4-5 minutes per run. However, with model finder and bootstrapping options enabled this increased to 10-27 minutes. This translates to an acceptable throughput of over 480 submissions a day using the de novo approach with defaults. These test runs were checked for placement and all methods positioned the query genomes alongside the corresponding reference nodes from which they were taken. A comparison of the alternate workflows to de novo showed very similar topologies with average RF distances of 8.1 (SD 7.3) and 7.3 (SD 7.2) for the placement and MFP workflows respectively. The differences were seen to mainly occur in closely related sections with short branch lengths and lower support values.

The development server was launched on April 10th 2018 and has aided in ongoing debugging efforts. As of the end of June 2018, 283 genomes were processed successfully. A total of 22 runs were logged as erroneous for three major reasons: Sequence parsing was not handling genbank records with missing sequences (annotations only), no family reference could be identified for some queries using the placement workflow, and multiple families were detected for one placement run. Measures have been taken restricting the use of genbank files without sequence and more error prompts given to the user so that they are notified to try a different set of organisms when using the placement workflow; also suggestions to limit a placement to one genome per run have been added.

5.3.2 Tree Validations

ANI clan validation

Family trees tested for validation were generated using the genome sets from the placement workflow except using a reconstructed tree with the default de novo workflow. Scoring via the monophyletic criteria showed the vast majority of trees had a perfect grouping of all ANI clans (>90% of applicable trees) with none below an average score of 0.75 for all three clan groupings tested (Figure 5.4). Similar results were seen for both the concatenated alignment and coalescent workflow with the exception of one tree with a score of 0.55 in the coalescent results. The topologies between the two workflows were also compared and seen to be similar with 25 families showing identical trees. Because tree sizes ranged widely the average of the RF distance ratio to maximum RF distance was taken and shown to be 0.09 (SD 0.07) showing limited differences in topology. Some trees could not be scored as they only formed singleton ANI clans (no other member found above ANI level), which are not considered in the average scoring; therefore these were assessed using the remaining validation methods.

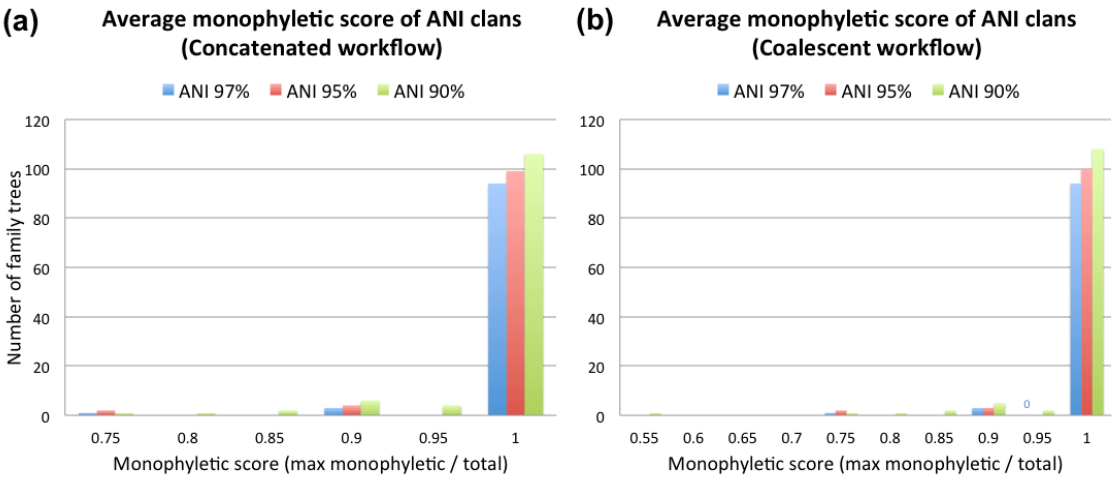

Figure 5.4: Histograms of monophyletic scoring of ANI clans at three clan definitions: 97, 95, and 90 percent ANI. (a) Scores from the concatenated workflow. (b) Scores from the coalescent workflow.

The ratio of leaves in multi-member ANI clans was calculated to determine what percent coverage of the trees were scored. This value ranged widely between each family with a minimum of 4% to 71% tree coverage. The average for the 97, 98, and 90 percent ANI groupings showed coverage percentages at 21, 23, and 33 respectively (with standard

deviations of 14, 14, and 18). Because many isolates in the tree could not be scored with this method we tested pairwise ANI scores at lower thresholds and calculated branch support values as an alternate validation for these singleton branches.

Bootstrap support validation

The bootstrap support values extracted from each of the validation trees were shown to be well supported for not only peripheral branches close to the leaves but also many internal branches as seen from the histograms of support values (Figure 5.5). This shows the majority of trees have nearly every split (over 90% of branches) confidently supported with a bootstrap support of 80 or higher. For the concatenated workflow over half of the trees have all branches well supported. Furthermore, all trees from both the concatenated and coalescent workflow showed average support values over 85; standard deviations for both workflows ranged form 0 to 21.3 with 90% of the trees under 15.

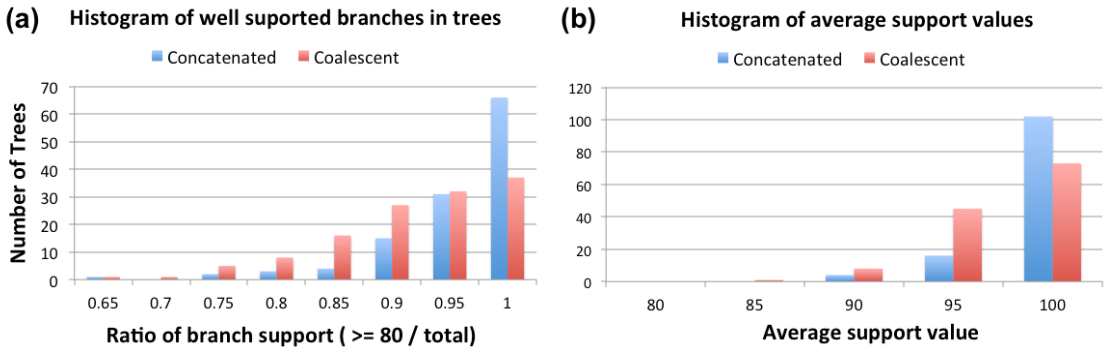

Figure 5.5: (a) Histogram of well-supported branches. This was represented as a ratio of branches with bootstrap support >= 80 divided by total support count for each tree. (b) Histogram of average support values.

By visual inspection the lower support values were usually seen near short branch length leaves of highly similar genomes or on internal branches where distant ancestral splits began to diverge. The comparison between coalescent and concatenated support values cannot be made as they represent two different definitions of support. Despite the lower values seen with the coalescent approach both illustrate confidant support for ancestral divergence.

Branch length validation

To assess if evolutionary distances in the trees (branch lengths) are estimated appropriately pairwise node distances were plotted against their corresponding ANI values. By using a

lower pairwise ANI threshold of 85% for the correlation a higher proportion of the tree could be scored compared with the ANI clan method. However ANI estimation is less precise below 90, this is apparent in the ANI vs. branch length plots (Figure 5.6).

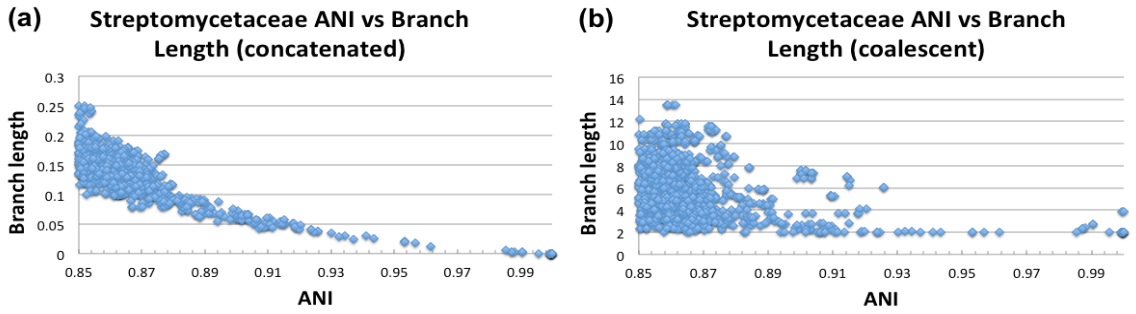

Figure 5.6: Example of pairwise node distances correlated with pairwise ANI values for the *Streptomycetacea* family. (a) Concatenated tree method. (b) Coalescent tree method.

For the concatenated alignment a clear linear relation was seen especially for values between 88-100% ANI. In contrast the coalescent distances, represented in coalescent units (221), did not correlate well with ANI. All trees from this method showed terminal leaf distances fixed at a distance of 1 thus skewing the comparison. This analysis was repeated for trees that had over five ANI connections and P-values less than 0.01. The resulting Pearson correlation coefficients were subsequently shown to have a strong relation for the concatenated method but not for the coalescent method (Table 5.1). Tree score coverage also showed an average of 57-64%.

	Concatenated			Coalescent		
	Coverage	**Pearson**	**P-value**	**Coverage**	**Pearson**	**P-value**
Min	17.6%	**-0.998**	2.03E-261	17.6%	**-0.998**	1.41E-55
Max	93.8%	**-0.730**	7.35E-03	93.8%	**-0.248**	9.90E-03
Average	57.7%	**-0.932**	2.25E-04	64.2%	**-0.623**	1.04E-03
Stdev.	18.6%	0.060	1.00E-03	18.6%	**0.175**	2.23E-03

Table 5.1: Tree coverage and correlation calculation statistics for concatenated and coalescent trees

The tested trees therefore illustrated that branch length designations were corroborated by evolutionary distances as defined by ANI. However the low correlation seen with the coalescent method confirmed that branch lengths were better estimated via the concatenated method and measures to adjust terminal branch lengths for coalescent leaves are needed.

Reference tree comparison

The manual MLSA analysis done in publication 3 highlighted four major clades within the *Amycolatopsis* genus, which were also represented in the autoMLST tree. The automatically generated tree used the default settings in de novo concatenated mode but with manual selection of the same out-group, *Nocardia farcinica* IMF 10152. Additionally 5 reference type strains were automatically added by autoMLST, which were the exact genomes of the corresponding queries. These added reference strains were removed when generating the final tanglegram comparison to make it easier to visualize (Figure 5.7). Besides having the same major clades the tree topologies were similar with a RF distance of 26 out of a maximum of 86. The differences in closely related strains accounted for most of the topology conflicts.

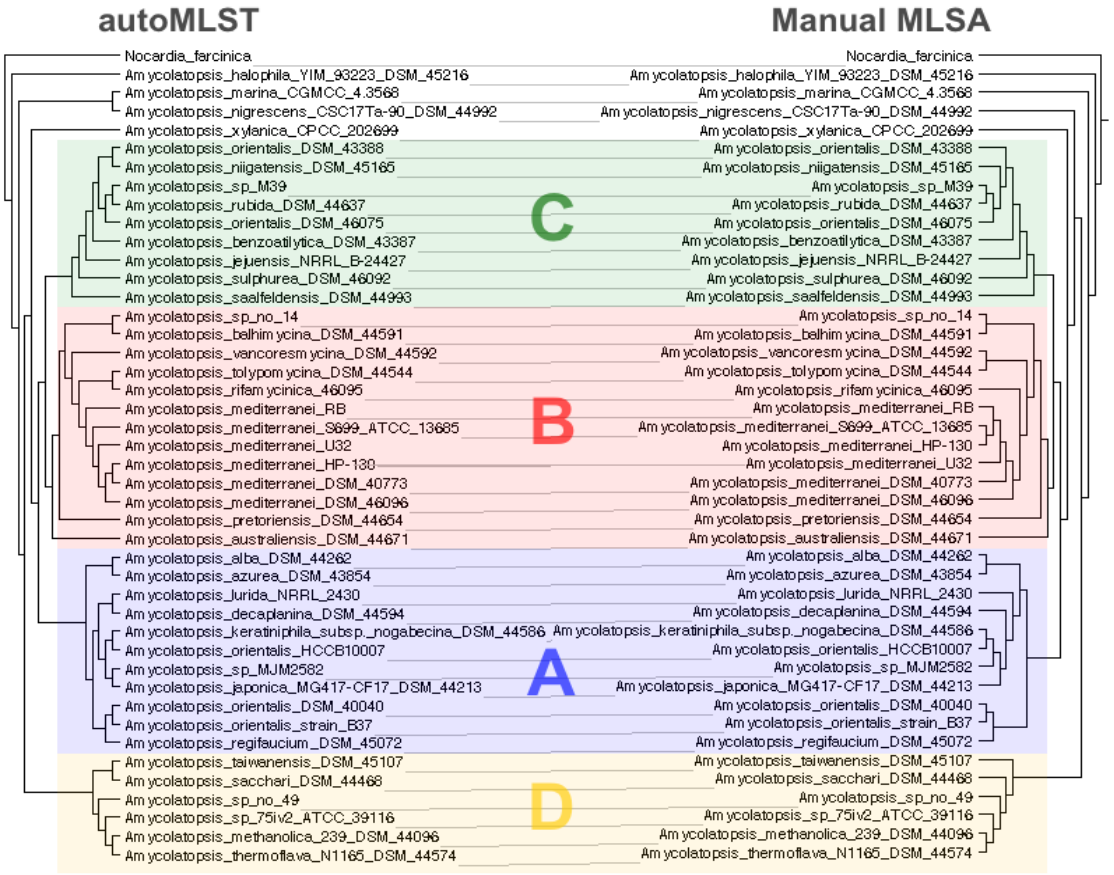

Figure 5.7: Tanglegram comparison of trees generated automatically with autoMLST (left) and using manual MLSA analysis (right). The manual MLSA tree file was provided by Dr. Adamek generated from publication 3 (153). Groups defined in this study are indicated using the same color scheme and labels as in Adamek et al.

For example the *A. mediterranei* clade, which had the shortest branch lengths and consequently the lowest support values (<65) in both trees, was responsible for 8 of the different splits in the trees. Although these strain level distinctions are less certain, one notable improvement by using more genes (85 selected) with the autoMLST method is that there were fewer polytomies (unresolved bifurcation) in the tree compared to the 7 gene manual MLSA; in contrast to the 5 out of 6 *A. mediterranei* strains in unresolved branches, only 2 of these strains were unresolved in the autoMLST tree. The genes automatically selected also overlapped with the majority of those used in the manual process including *atpD*, *gyrB*, and *pyrH*; two alternate subunits corresponding to *nuoD* and *rpoB* genes were also included. Furthermore, the automated selection overlapped with 19 of those found in the pubMLST database with many other commonly used MLSA genes including: ribosomal subunits, DNA / RNA maintenance, and DNA translation proteins. The support values for both trees were very similar with average values of 85 (SD 27) and 87 (SD 22) for the autoMLST and manual tree respectively; Also the ratio of well supported leaves equal to or greater than 80 were identical at 0.75.

5.4 Discussion

Classifying bacterial genomes by 16S sequencing is fast and accessible to many non-specialists but unfortunately this approach is limited in the resolution it can produce due to high sequence similarity. Fortunately MLSA analyses, which often include 16S sequences, have been employed over the years to solve this resolution problem (201); however the complications and lack of standardization have made them less accessible to a variety of research. Even the initial steps of finding appropriate species, outgroup organisms, and the best set of genes is non-trivial unless a user is familiar with the taxa. We therefore developed a tool that automates each step of the process and is accessible though an intuitive web interface. The aim of this project is to bring these high-resolution common best practices, such as using ML tree inference, model selection, and bootstrap analysis, into a rapid procedure that can be as widely used as 16S BLAST searching. Automation of gene and reference organism selection is a unique aspect of the project that allows for "one touch" processing of query genomes. This means anyone unfamiliar with the taxa in question can quickly process their genomes for an initial hypothesis.

While automation is the major goal, we heavily stress that inferring the evolutionary history from a snapshot of present sequences is inherently complicated and great care should be taken when concluding a species tree. As a disclaimer, all results and input quality should be closely scrutinized when concluding a final hypothesis; taking the autoMLST results for granted should be avoided. This quality control process is simplified as we have provided features for easy exploration of the tree and allow all raw data and alignments to be downloaded for inspection and confirmation. For example, by toggling the branch length view a user is able to identify erroneous branches that might be artificially long due to sequencing artifacts and thus potentially misplaced. The ANI grouping visualization also allows for a second sanity check by illustrating groups that should be closely related on the tree. We also provide alternative practices such as coalescent tree building to help encourage multiple perspectives to support a hypothesis; this is made easier as every job allows for rapid re-analysis in the results page. Finally, inspecting the alignments is a crucial step to identify problematic genes or organisms, which can be used to guide a reanalyzed job.

In spite of this disclaimer nearly all of the default automated trees generated for validation of each tested family showed to be well supported as defined by the two methods of bootstrap and coalescent support. These validation trees were analyzed as-is to get a sense of default performance. They also show that ANI clans are grouped effectively in the majority of all family trees. The few families that showed lower ANI grouping ratios further highlight the need to preform manual inspection and reanalysis of a result. Trees that were found to have lower ANI grouping values are currently being reanalyzed to provide a more accurate placement tree. Branch length accuracy was also verified in the concatenated workflow. The correlation with ANI shows both measures of evolutionary distance corroborate each other, however this was not seen with the coalescent approach. We are actively exploring methods to rectify this disadvantage of inaccurate branch lengths in this optional coalescent method. As the error seems to come from fixed terminal branch distances perhaps a hybrid definition of branch lengths using coalescent distances and average nucleotide substitutions for all genes can be used. Despite this, the topologies of the two methods were shown to be largely in agreement by RF distances. Likewise, similarity between the placement, MFP, and de novo workflows were in agreement as illustrated with the performance test genome set.

While the reference genomes used showed to span a robust variety of taxa the results indicated that our current publicly available sequenced genomes are still heavily skewed toward model taxa. Overtime the increase in whole genomes, and those from metagenomic data, will help to even out this bias and yield more complete databases. Because the de novo method does not rely on these references any unknown set of organisms can still be processed so long as they share enough conserved single copy genes in the entire query. The placement method however is limited in this respect; nevertheless many researchers are focused on model taxa and so would benefit from this rapid option. This workflow was shown to be a fast alternative with processing time under a minute using the performance test set. The accuracy of the placement was also confirmed as all genomes were placed with the corresponding reference leaves in these tests. More confirmations using variant strains are also actively being tested via the development sever to further validate this workflow. So far the user feedback has been positive overall despite being in the beta testing phase. With regard to throughput, the initial development server showed to be sufficient in supporting hundreds of submissions per day using the longer de novo workflow. Considering this, future redundant production servers will be capable for handing a high capacity of submissions. For more demanding needs users are able to download and install this server freely. Currently this process must be done from source but following the completion of a release candidate version we will package this using container solutions to make for easy deployment on any private server. We are actively working on this to help collaborators at the Fraunhofer institute setup a local version of autoMLST.

The comparison to the high-resolution *Amycolatopsis* tree in publication 3 is a clear example of how the manual MLSA process can be accelerated by autoMLST yielding comparable results. Major branches between strains were shown to be identical which highlighted the key subdivisions of BGC production within the genus, as defined in Ademek et al (222). The subtle differences in strain level topologies accounted for most of the discrepancies in the two trees, which is expected considering these branches have the lowest support values. This could simply be a consequence of variable placement in tree inference due to the highly identical genome sequences. The time saved by using this pipeline is a clear advantage over the manual process but also leveraging more phylogenetic signal via sampling a larger pool of conserved single copy genes helped increase tree resolution. Although this did not make major differences in the topology it resolved some strain level polytomies,

which may be important for research that requires discrimination of genomes of the same species; as seen with some natural product producers this variability within the same species can thus be identified. Overall these tests demonstrated that the tool has met the design goals for automating the laborious process of generating high-resolution species tress with results similar to manual methods. This example shows how a query genome can be classified into the different subdivisions of the genus and can therefore be used as a prioritization tool for discovering new natural products.

Although this initial version is available to the public and can be immediately used efforts to improve this application are ongoing. In addition to further validation with other comparative studies we aim to provide automatically updated reference genomes and assist the application of strain prioritization for natural products discovery. In its current state prior knowledge or literature research is required to assess if the organism is placed in a prolific clade rich in BGCs. We therefore are adding an extra visualization layer to the final tree to show basic counts for BGC detection. After cataloging of known BGCs in reference genomes present in public BGC databases, such as the antiSMASH database (130), these will be used to group prolific clades. Counts for BGCs present in the query genomes can also be included if these are present in the genbank file. In the meantime this prioritization process can be largely achieved using the ANI clan groupings and searching the secondary metabolite potentials of like group members. Besides the application of natural product discovery prioritization, this tool can currently be used to identify ideal related strains and provide an easy automated way to classify an unknown genome with higher fidelity than using 16S sequences alone. Although it is preferred to use at least draft quality genomes, the application with PCR fragment sequences is also possible if enough core genes are included. With the downloadable MLST alignments from a family or genus of interest this can be predefined and even help to generate primers for inexpensive PCR sequencing. Considering sequencing costs continue to fall this may be irrelevant in the future however.

Chapter 3

6 Targeted genome mining with ARTS

6.1 Introduction

The rise in resistant pathogens coupled with the decline of new antibiotics to market is a serious threat to human health that is accelerated via the dissemination though horizontal gene transfer (HGT) (223, 224). This health crisis must be addressed on several fronts to be effective. Efforts to replenish our antibiotic arsenals are a major factor to controlling this problem however stagnation in the drug discovery pipeline has hampered leads to new effective compounds (58). The majority of antibiotics have been and continues to be inspired by natural products – secondary metabolites (SM) produced by living organisms; often these are the first members of novel classes of compounds (225–227). Decades of exploiting rich resources, such as soil dwelling microbes from the Actinobacteria class, have proven to be fruitful using traditional cultivation and extraction techniques but lately these methods have experienced high rates of rediscovery (228). These techniques involve a process of collection, cultivation, and bio-activity screening that can lead to several time limiting bottlenecks and "filters" of chemical potential. For example organisms that take much longer to grow or do not thrive in laboratory environments go uninvestigated. It is estimated that the vast majority 90-99% of microbes are "uncultivable" in the lab (70, 225) which can be a serious contributor to the rediscovery of known chemical space. Additionally, hidden potential has even shown to be present in cultivated species with the presence of "silent gene clusters" (58) identified using new genomic techniques. These Biosynthetic Gene Clusters (BGCs) go unexpressed under laboratory conditions or are under complicated regulation so evade current screening methods. The antibiotic discovery phase is therefore in need of new methods to further the search for new compounds. Over the last few years reinvigoration with the application of "genome mining" methods has helped to expand this search (171). Mature applications such as antiSMASH (229), clusterfinder (118), and PRISM (113) can effectively detect several classes of SM systems and have resulted in large databases of putative BGCs to investigate; Another interesting detection method, EvoMining (121),

uses gene duplications from primary metabolism, "gene expansions", as a marker for BGC detection, which can potentially discover completely novel classes of BGCs. These methods have lead to known cluster databases, such as MiBIG (129), and those with a vast array of potentially viable antibiotics. For example, the "Atlas of Biosynthetic gene Clusters", a component of the "Integrated Microbial Genomes" Platform of the Joint Genome Institute (JGI IMG-ABC) (101) shows over a million putative BGCs as of June 2018. The vast majority of these predictions have no known compound associated with them (orphan clusters) as only 0.2 % of the database has experimentally verified products.

With no shortage of potential to investigate the main limiting factor to exploiting genome mining is now to prioritize these leads for laborious wet-lab experiments. One interesting approach uses the fact that many antibiotic producers also include the corresponding resistance determinate so as not to commit suicide (230). This idea was exploited by Wright and colleagues to enrich microbial libraries for producers of selected antibiotic scaffolds by enriching for those with self-resistance mechanisms (231). Methods of antibiotic protection include genes that encode for: transporters to export the compound out of the cell (26), proteins that neutralize its activity (31), or a target protein with a resistant mutation (27). With the later case a second copy of the protein is maintained, possibly due to the fact that the unaltered version of the protein results in higher fitness when the antibiotic is not present (28); in addition to duplication these resistance targets are often found within the BGC and are potentially horizontally acquired (232, 233). This simple tactic of co-expression with the antibiotic allows the organism to protect itself during production. Based on these concepts, Moore and colleagues screened for duplicated essential genes that are co-localized within putative BGCs to identify potential new targets. Using this genome mining tactic they were able to identify a fatty acid synthase (FAS) resistance gene in a hybrid BGC (134); The expression of the cluster and structure elucidation revealed the product was a previously described group of compounds that inhibit the FAS-II system (234, 235). This shows that not only does this tactic provide a valuable way to enrich for antibiotic activity in putative clusters, but also it can give downstream experiments a head start by hinting at the mode of action. While this method is beneficial it has mainly been employed manually or requires computational expertise to automate.

Another helpful clue for prioritization is the detection of HGT, as BGCs or resistance determinates are known to be subject to HGT (236). Detecting HGT has traditionally been

accomplished by interrogating compositional features of the DNA that are incongruent with the whole of the organism. For example, if an area of low GC content in a high GC organism is found it its possible this region was acquired horizontally. Organisms also show preference for certain degenerative codons, which have also been exploited for HGT detection (237). However these markers are subject to mutation toward the preference of the host and so this method can be problematic when detecting distant HGT events. With the advent of many whole genomes, comparative phylogeny has provided another option that has shown to be a reliable inference for HGT (238). A phylogeny of sequences of the same gene derived from vertical inheritance, orthologous genes, can show if the gene is divergent from the species phylogeny. Therefore by interrogating discrepancies in the gene trees with the species tree one can find HGT candidates. However this process is complicated by the fact that not all incongruences are caused by HGT. A combination of gene duplication, resulting in paralogous genes, and loss events can lead to incongruent trees. This incomplete lineage sorting can be overcome through various model based or probabilistic approaches that assume a parsimonious or likelihood explanation (238). Thus more confidant detection of HGT can be made using these approaches even for ancient events that have mutated overtime.

Here the Antibiotic Resistant Target Seeker (ARTS) (239) was developed to examine these three criteria: essential gene duplication, co-localization within a BGC, and phylogenetic evidence of HGT. To bring these methods into widespread use for natural product discovery we built an intuitive web interface with dynamic output to help the user explore these results effectively. The major goal of this project is to automate the steps required to perform target directed genome mining, however ARTS is also useful as an orthogonal cluster prediction method. By using the criteria and explorative functions in the results known and putative resistance determinants can be used as markers to identify BGCs that elude motif based detection, similarly demonstrated with gene expansions searches in EvoMining. This process involves detection of duplications from a list of shared essential genes anywhere in the genome. ARTS also screens for experimentally verified resistant targets and other known resistance determinates to quickly prioritize by a particular target or any known resistance (for example, beta lactamases).

6.2 Methods

Workflow Overview

The ARTS workflow does three criteria analyses on submitted sequences to determine duplication, co-localization, and HGT for all genes in an essential gene set (Figure 6.1).

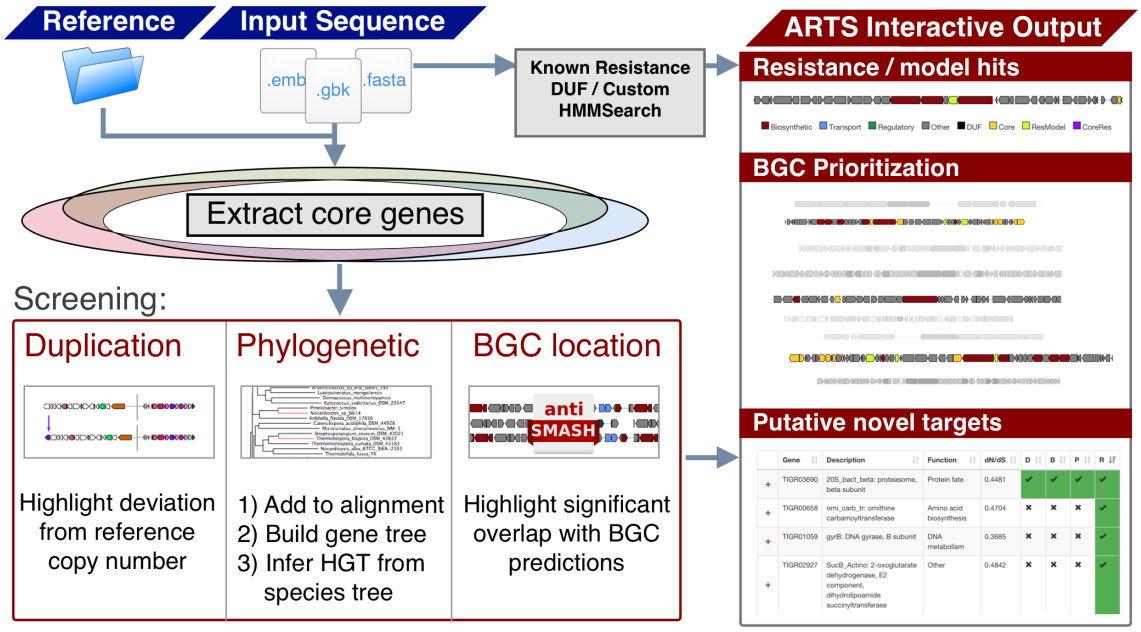

Figure 6.1: ARTS workflow overview. After extraction of know resistance and essential "core" genes criteria are crossed referenced. Duplication uses deviation from reference copy number, HGT is determined using phylogenetic reconciliation, and BGC co-localization is tested via overlapping gene and cluster boundaries. Finally results are presented to quickly identify prioritized clusters and to dynamically explore putative novel targets. Figure adapted from Alanjary et al (239)

Users submit whole genome or BGC sequences in Genbank, FASTA, or EMBL format as input; alternatively an NCBI accession number or antiSMASH job ID can be used to retrieve data automatically. The annotated Genbank is then parsed using Biopython (240) to identify all protein coding sequences, rRNAs, and cluster annotations. The first step includes BGC identification using antiSMASH (241) if this is not already provided in the genbank file. Afterward known resistance models are detected using those collected from the "The Comprehensive Antibiotic Resistance Database" (CARD) and ResFams (242–245). Essential gene sets, defined by conserved "core genes" from complete genomes (see reference and core gene section), are then detected in the query using HMMER3 (116). HMM domain

results are parsed and the best model hit for a gene is extracted if it passes 50% coverage length thresholds of both model and gene; this value was chosen to allow for missing domains and incomplete sequences while reducing fragmented hits. The identified core and known resistance genes are then screened for their location within BGCs by examining overlapping boundaries. Finally core genes are screened for HGT using the generated gene trees and species tree. Additional searches using Domain of Unknown Function (DUF) models from the Pfam (132) database and are used to highlight potential novel chemistry in a cluster; custom user submitted models can also be supplied in the advanced section of the program. All results are then summarized into an interactive output table to rapidly cross reference known and putative novel antibiotic targets.

Reference set and core gene selection

NCBI's RefSeq (246) database was used to download complete genomes to build the current *Actinobacteria* reference. By using complete genomes errors derived from missing genes in draft genomes can be avoided for the core gene calculation. Essential genes are inferred by a comparative genomics approach where ubiquitous "core genes" are those consistently found in reference organisms as detected using HMMER and Hidden Markov Models (HMMs) from the TIGRFAM (208) protein family database; In addition, predefined core genes from the TIGRFAM v15 "bacterial core gene set" (GenProp0799) are included. All TIGRFAM homologous proteins with emphasis on conserved function, "equivologs", and their hypothetical and domain variants are used for essential gene analysis. After HMM detection, counts for genes were recorded in a gene matrix consisting of all reference genomes. Family specific core genes are then defined as genes present in greater than 95% of genomes relative to each family based on the count matrix. Families with less than 10 genomes were combined and a lower ubiquity threshold of 90% is used instead to account for the more distant relationship. These core genes were then analyzed to build several gene trees used for accelerated gene tree creation. All core gene sequences are extracted into corresponding multi-record FASTA files and, where applicable, out-group sequences added using model matches from various sequences of Proteobacteria. Each core gene protein FASTA file is then aligned with MAFFT (211) followed by a codon alignment with Pal2nal (209). Trimmed copies of each codon alignment are made using trimAL (206) with the maximum likelihood optimized "automated1" setting. RaxML (194) is used to build each tree with 100 bootstrap

replicates using GTR-GAMMAI model selection. Pairwise selection (dN/dS) values were calculated for each alignment using the yn00 tool from the PAML (210) package and the median of all Nei-Gojobori dN/dS values were logged to the model metadata. Metadata for functional classification are taken from model descriptions and associated main categories in TIGRFAM Roles. Additional statistics such as global ubiquity percentages and how often a gene appears as a single copy are also recorded to allow the user to further prioritize selected genes.

Core gene filtering

To help identify viable targets and reduce false positives a filtered set of core genes is used in the default search mode. However the optional "exploration mode" omits this filtering to allow for searching of a broader set of core genes. Core genes associated with transport or regulatory functions were removed based on terms found in the model descriptions. We also noticed several common metabolic enzymes associated with biosynthesis in a BGC. These were removed if the protein sequences from the corresponding HMM seed alignments yielded positive hits for high frequency BGC Pfams. High frequency biosynthetic Pfams are determined using clusterfinder (118) Pfam frequency data where those above a frequency of 50 were used. This threshold was conservatively chosen based on the histogram of different Pfams (Figure 6.2).

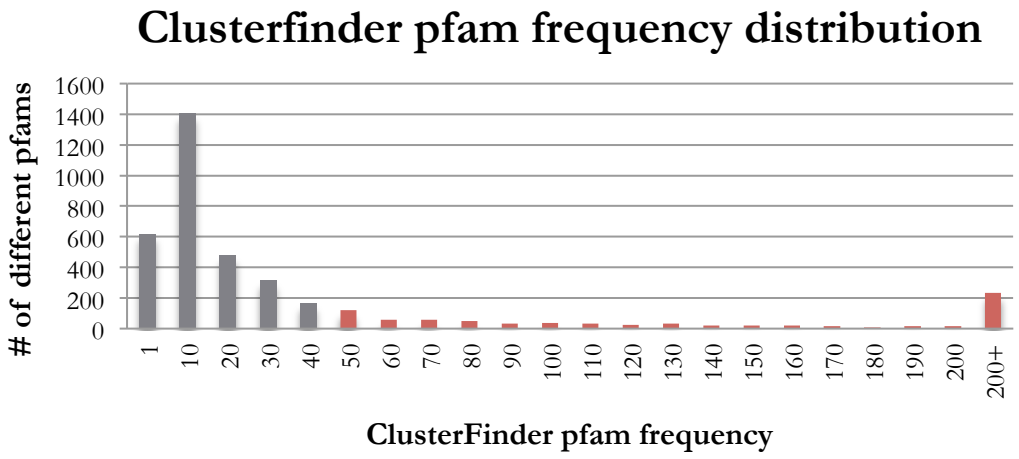

Figure 6.2: Clusterfinder frequencies defined from a manually curated set of 732 known BGCs (118). The common Pfams found in BGCs (red) were removed. Figure adapted from Alanjary et al (239)

6.2.1 Criteria Screening

Duplication screening

Duplication is determined comparatively to the reference set of organisms where median counts were calculated for each core gene. This was used to reduce the effect of outliers. A divergence from the norm is then defined as gene counts that are higher than these values plus their standard deviations in the reference set. Genes with counts greater than this baseline are recorded along with the bit-score, scaffold location, and reference count statistics provided for manual review.

BGC proximity screening

All detected core gene and known resistance genes that intersect with BGC boundaries on the same scaffold are marked for co-localization. Visualizations are appended to the antiSMASH generated graphics and colored by criteria to quickly identify the type of proximity hits in the "proximity" section of the results. Results from DUF and resistance model hits are also appended to cluster annotations in a similar manner where hits for both resistance and core models are marked indicating a known target; these are labeled "CoreRes" - a core model that is also in the known resistance set.

Phylogenetic screening

If the user sequences are compatible with the Actinobacteria reference set the phylogenetic screening can be used to detect HGT. Input sequences that do not have enough core genes, or sequences not part of the reference phyla will fail this screen or produce inaccurate results so the option to omit this screening is provided. The screening for HGT involves making sequence alignments of every detected core gene followed by gene and species tree inference. Alignments are accelerated using the reference set by adding the extracted nucleotide core sequences to the pre-trimmed reference codon alignments using the add method in MAFFT (mafft --add). Appended alignments are stored for user export and trimmed copies are made using the "automated1" method in trimAl. Trimmed alignments and existing reference trees are then used with the Evolutionary Placement Algorithm (EPA) (247) option available in RAxML to produce all gene trees. The species tree is then inferred from a coalescent of multi-locus gene trees using ASTRAL (248). The set of gene trees used are all single copy genes present in every reference and query organism with non-

synonymous vs. synonymous (dN/dS) mutation ratios less than 1; 16S rDNA sequences are also included if present. Each gene tree is then reconciled with the species tree to delineate incongruences due to duplications, transfers, and loss; this is determined using the parsimonious criteria defined functions in ranger-dtl-U (177) tool using default HGT cost values. All transfers involving the query organism are then parsed and sorted and an additional filtering to mask intra-genus transfers is applied based on the name of the query to highlight inter-genus transfers.

6.2.2 Performance and validation testing

Processing time was calculated using the 200 genomes from the validation testing. These jobs were spread evenly across two redundant servers running a 4-core and 8-core system equivalent to 2.2GHz Xenon processors per core. Errors were also logged based on failed runs that were recorded in the run queue database. Validation testing consisted of selecting known clusters and genomes that had experimentally verified self-resistance mechanisms via literature searches. Results for all searches were sourced from NCBI pubmed and the MiBIG database with additional findings reported from discussions with colleagues from the Scripps Institute of Oceanography (SIO). All known examples were screened thought the web interface to see if ARTS marked these genes as hits.

6.3 Results

To our knowledge, ARTS currently remains the only public web server that automates an extended target-directed genome mining that includes potentially novel targets. It demonstrates significant timesaving over manual methods by automating all criteria screens and helps with exploration of results via the interactive output. The webserver (https://arts.ziemertlab.com) and analysis scripts are open source and available freely to the public with source repository at: https://bitbucket.org/ziemertlab/arts

Usage and performance

From the public release in May 2017 over 2,000 jobs have been processed with positive feedback and feature requests from users as of June 2018. Overall less than 2% of the jobs showed to have errors, which were mainly related to lack of core genes discovered and file

format errors. The main request has been providing more reference sets, which we are actively working on. The collaboration with researchers at Novartis, to help setup and test a local beta version, also showed no error reports and positive feedback overall. Using the accelerated phylogenetic method, processing time was significantly reduced compared to a similar manual analysis. Phylogenetic analysis of all core genes, with alignments and Maximum-Likelihood tree construction, took over 89 hours of processing time alone on a 16-core machine. By leveraging the pre-computed reference, the analysis took 15 minutes using the same resources with ARTS for one run.

Output and interactive layout

The results are presented in various panels by criteria sections: core genes, resistance models, gene duplications, BGC proximity, and phylogeny. All sections can be searched and sorted by various additional properties to help identify potential targets (Figure 6.3).

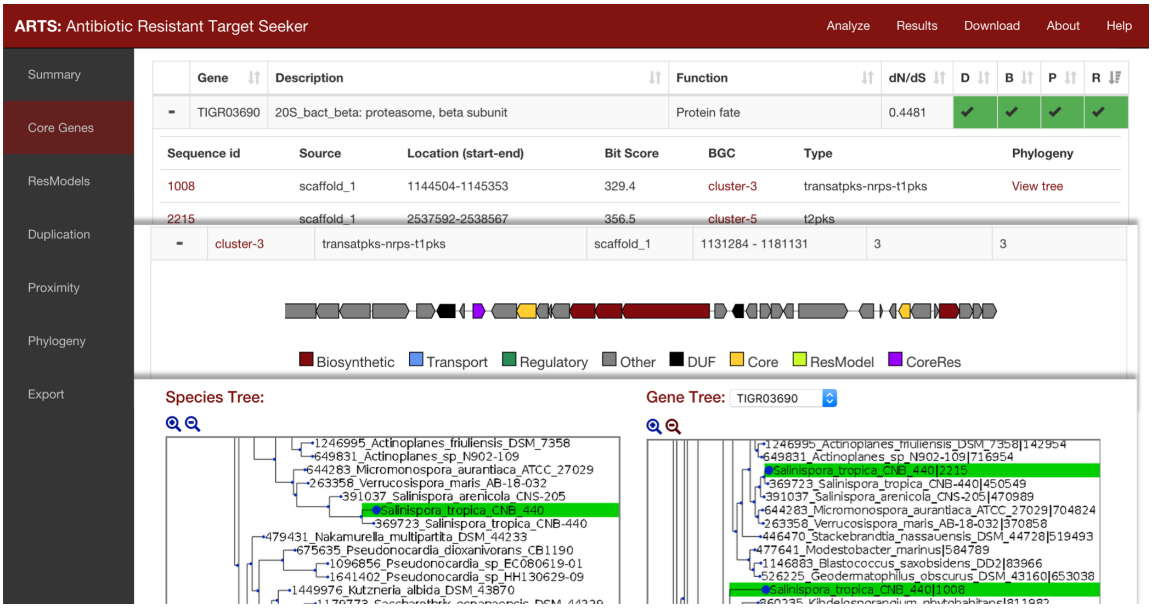

Figure 6.3: Screenshots of ARTS interactive layout sourced form the public example *Salinispora tropica* CNB-440. Shown here are the dynamic tables in the core gene section, BGC visualization in the proximity section, and tree comparisons in the phylogeny section. Figure adapted from Alanjary et al (239)

Sorting the core gene table by positive BGC hits and phylogeny will yield all core genes found in a BGC with HGT evidence. More details can be seen for each detected gene by expanding a row in these tables with links to other sections. Modified antiSMASH cluster

visualizations are augmented with ARTS hits which can also be sorted by hit type and count for rapid prioritization. A side-by-side visualization of phylogeny hits is also provided for user confirmation. All tables, trees and core gene alignments can be exported and saved for additional analysis from the export section. To help discriminate putative target hits, gene properties are also presented in the core gene section including: functional classification, average selection pressure (dN/dS) values, and how widespread the essential gene is relative to reference organisms (ubiquity); these can be used to help assess the viability of a target. Examples of all inputs and detailed tutorials have also been generated and are available on the help section: https://arts.ziemertlab.com/help.

Reference set and core genes

The complete Actinobacteria genomes as of September 2016 were comprised of those from 189 species representing 22 different families. Members from the genera Corynebacterium, Streptomyces, and Mycobacterium had the highest number of genomes with 14.8%, 9.5%, and 7.9% representation of the reference respectively. The remaining 83 genera showed relatively even representation between 0.5-4.2%. To verify that core genes inferred from reference organisms are essential, comparisons to experimentally verified essential genes were preformed. A total of 664 core gene models were identified using the union of family specific core genes. After gene filtering for those commonly involved in BGCs, 432 remained for the default search mode. A comparison to the Database of Essential Genes (DEG) v13 (249), a repository of genes found to be essential via experimental study, shows 538 genes in the unfiltered set match to one or more records. The functional classification of each was used to inspect the distribution of genes relative to the DEG set (Figure 6.4). All reference core genes compared to all ARTS hits showed enrichment for essential functions including: protein and amino acid synthesis, energy and metabolism, and transcription - Supplemental S4 (239). A variety of approaches to determine essential genes can be found in literature where ubiquity and conservation of sequence are properties frequently exploited (250–252). These properties were further interrogated to assess if the core genes represent essential genes. Nearly the entire core gene set shows dN/dS values less than 1, illustrating many are under purifying selection, with a mode of 0.35 and range of 0.05-1.05. Other ubiquity measures showed that many were shared with all species and usually present in single copy as seen in Supplemental S5 (239).

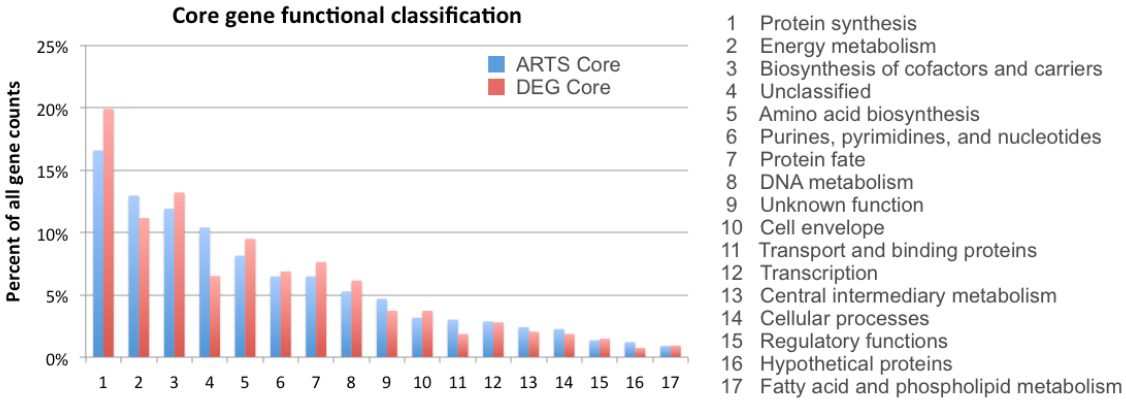

Figure 6.4: Functional classification of all core genes used in ARTS and those found in the Database of Essential Genes (DEG) (249). Figure adapted from Alanjary et al (239)

While more than half of the genes are present in over 90% of genomes, many are specific to certain genera. One example is the 20S proteasome, which is essential to some Actinobacteria genera but is lacking in others with only 64% global reference ubiquity. By defining core genes relative to family and then taking the union, this specific function is captured. All gene trees were saved and bootstrap analysis was preformed to assess the quality of each tree. Each tree was analyzed individually and the final species tree was then analyzed using the coalescent of 30 single copy genes (Figure 6.5).

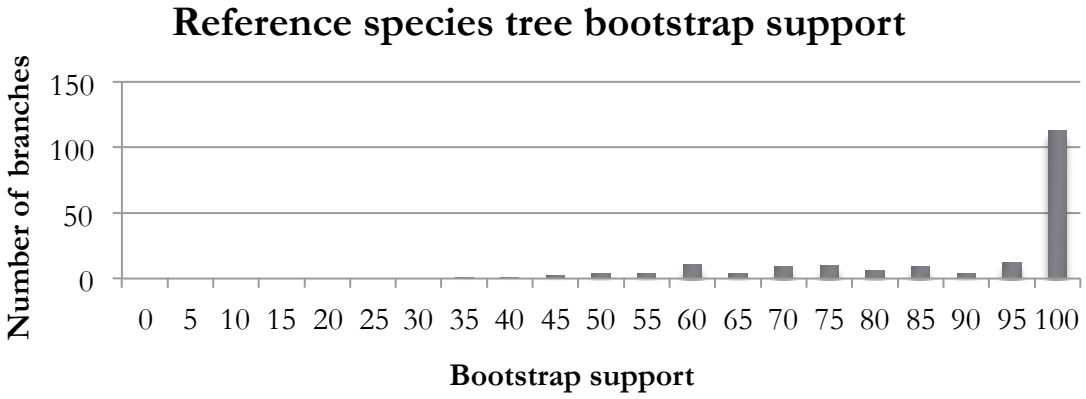

Figure 6.5: Histogram of bootstrap values from resulting coalescent species tree. Figure adapted from Alanjary et al (239)

In addition to the well-supported species tree the individual gene trees showed 520 gene trees had most of the branches (mode of all branches) over 95 with median values of 70-80, Supplemental S6 (239).

6.3.1 Self-resistance gene detection results

Positive examples from whole genomes

After extensive literature searching and discussions from colleagues 14 cases of self-resistance genes in genomes of at least draft quality were analyzed using ARTS (Table 6.1).

Product	Resistance gene	Organism	Gene Accession (ref)	ARTS hits	Criteria hits (>2, >3)	BGCs (total, core hit, res hit)	Genes (core, total)
Novobiocin	duplicated *gyrB*	*Streptomyces niveus* NCIMB 11891	WP_03123 2360 (88)	**D , B , R, *P**	25 , 5	30 , 11 , 8	383 , 7815
Clorobiocin	duplicated *gyrB*	*Streptomyces roseochromogenes* DS 12.976	AAN65247 (253)	**D , B , R, *P**	46 , 13	43 , 19 , 13	396 , 9055
Albicidin	pentapeptide repeat protein for GyrB	*Xanthomonas albilineans* GPE PC73	CBA16025 (254)	**B , R**	1 , 0	8 , 7 , 3	309 , 3208
Streptolydigin	mutated *rpoB*	*Streptomyces lydicus* NRRL2433	AAQ19729 (255)	**R**	28 , 2	35 , 10 , 13	384 , 8518
Rifamycin	mutated *rpoB*	*Amycolatopsis mediterranei* S699	AAS07760 (256)	**R**	28 , 5	30 , 11 , 8	379 , 9575
Rifampicin	duplicated *rpoB*	*Nocardia farcinica* IFM 10152	BAD59497. 1 (257)	**D, R**	21 , 3	17, 13, 7	550 , 5946
Thiocillin	duplicated ribosomal L11	*Bacillus cereus* ATCC 14579	AAP11944, AAP11947 (258)	**D , B**	2 , 0	10 , 5 , 2	310 , 5255
Erythromycin	duplicated 23S rRNA methyltransferase	*Saccharopolyspora erythraea* NRRL23338	WP_00995 0391 (259)	****D , B , P , R**	49 , 9	36 , 13 , 13	422 , 7198
Agrocin 84	duplicated Leu-tRNA synthase	*Agrobacterium radiobacter* K84	ACM31456 (260)	**D**	2 , 0	10 , 4 , 0	317 , 6684
Thiolactomycin	duplicated FabB/F	*Salinispora pacifica* DSM 45543	ALJ49913 (134)	**No hits**	18 , 3	25 , 10 , 9	365 , 4784
Salinospora-mide A	duplicated beta-proteasome subunit	*Salinispora tropica* CNB-440	ABP53490 (232)	**D , B , P , R**	16 , 2	19 , 9 , 7	362 , 4536
Vancomycin	Peptidoglycan remodeling	*Amycolatopsis orientalis* DSM 40040	CCD33128 (261)	**B, R ***	34, 7	39 , 16 , 17	381 , 8194
Cephamycin	duplicated beta-lactamase	*Streptomyces clavuligerus* ATCC 27064	AAF86620 (262)	**B, D, R**	26 , 3	45 , 20 , 15	546 , 7730
GE2270	duplicated elongation factor	*Planobispora rosea* ATCC 53733	AGY49599, AGY49600 (263)	**D , B , P**	26 , 4	26 , 13 , 9	372 , 8176
Notes: * Intra-genus phylogeny hits only seen; ** Other hits found using the advanced noise cutoff "E1" (90% model noise cutoff); *** Only *vanX* gene is detected							

Table 6.1: Positive examples of genomes with known self-resistance mechanisms analyzed with ARTS default mode. Hits to ARTS criteria are shown as; D: Duplication, B: BGC proximity, P: Phylogeny, R: Resistance model. Rows in grey indicate non-actinobacteria genomes, yellow indicate non-applicable BGC co-localization. Figure adapted from Alanjary et al (239)

In all but one case these genes showed hits for at least one criterion and in approximately two thirds of examples two or more criteria were highlighted. Three of these examples showed resistance targets outside of the cluster boundaries making the co-localization criteria inapplicable; likewise the phylogeny criteria was not applied for the three cases of non-actinobacteria organisms. For the missing case of FabB/F detection this was due to low homology and high confidence thresholds – by default ARTS uses the trusted cutoff values present in each HMM model. By using the optional exploration mode cutoffs these can be detected but with an increase in false positives. In the three cases of rpoB resistance these were only shown to highlight one criterion due to the fact that these do not appear to be co-localized in the cluster. Additionally only one of the three resistance genes were detected in *Amycolatopsis orientalis* DSM 40040 example due to strict similarity thresholds. Despite these cases over 80% of all genes were identified in the positive examples. This also resulted in a manageable average of 21 (SD 14.7) genes flagged for over one criterion out of 391 genes; for over two criteria this showed an average of 4 (SD 3.6) highlighted for this positive example set.

Positive examples from the MiBIG database

We identified 26 clusters from the MiBIG database with known self-resistance mechanisms to test with ARTS. Although the genomic context is missing, and so duplication or phylogeny could not be assessed, other criteria were able to highlight the known examples. The four missed examples were all duplicated resistance targets that were not detected due to the lack of genomic context and low scoring homology hit (Table 6.2).

Product	Resistance gene	Organism	Accession (ref)	BGC ID	ARTS Hit
Yatakemycin	Copy of DNA glycosylase	*Streptomyces sp. TP-A2060*	ADZ13541 (264)	BGC0000466	**No hit**
Azinomycin	Copy of DNA glycosylase	*Streptomyces sahachiroi*	ABY83174 (265)	BGC0000960	**No hit**
Avilamycin	duplicated 23S rRNA methyltransferase	*Streptomyces viridochromogenes Tue57*	AAG32067, AAG32066 (266)	BGC0000026	**No hit**
Kalimantacin	duplicated FabI	*Pseudomonas fluorescens BCCM_ID9359*	ADD82948 (267)	BGC0001099	**No hit**

Table 6.2: ARTS results showing missed cases of self-resistance genes from the MiBIG database. Figure adapted from Alanjary et al (239)

The remaining 22 cases showed to be flagged as core gene or resistance models (Table 6.3).

Product	Resistance gene	Organism	Accession (ref)	BGC ID	ARTS Hit
Griselimycin	Copy of *dnaN*	*Streptomyces sp. DSM 40835*	AKC91855 (268)	BGC0001414	**Core + Res.**
Coumermycin	Copy of *gyrB*	*Streptomyces rishiriensis DSM 40489*	AAO47226 (269)	BGC0000833	**Core + Res.**
Novobiocin	Copy of *gyrB*	*Streptomyces niveus NCIMB 9219*	AFI47646 (88)	BGC0000834	**Core + Res.**
Albicidin	pentapeptide repeat protein for GyrB	*Xanthomonas albilineans GPE PC73*	CBA16025 (254)	BGC0001088	**Res.**
Cystobactamide	pentapeptide repeat protein for GyrB	*Cystobacter sp. Cbv34*	AKP45389 (270)	BGC0001413	**Res.**
Rifamycin	mutated *rpoB*	*Amycolatopsis mediterranei*	AAS07760 (256)	BGC0000136	**Core + Res**
Rubradirin	two copies of Initiation factor	*Streptomyces achromogenes subsp. rubradiris*	CAI94679, CAI94684 (271)	BGC0000141	**Core,Core**
Thiocillin	two copies of Ribosomal protein L11	*Bacillus cereus ATCC 14579*	AAP11944, AAP11947 (258)	BGC0000612	**Core,Core**
GE2270	two copies of elongation factor	*Planobispora rosea ATCC 53733*	AGY49599, AGY49600 (263)	BGC0001155	**Core,Core**
Erythromycin	duplicated 23S rRNA methyltransferase	*Saccharopolyspora erythraea NRRL2338*	WP_009950391 (259)	BGC0000055	**Res.**
Pikromycin	duplicated 23S rRNA methyltransferase	*Streptomyces venezuelae ATCC 15439*	AAC69328, AAC69327 (272)	BGC0000094	**Res.**
Mupirocin	duplicated Ile-tRNA synthetase	*Pseudomonas fluorescens NCIMB 10586*	AAM12927 (273)	BGC0000182	**Core**
Borrelidin	duplicated Thr-tRNA synthetase	*Streptomyces parvulus*	CAE45679 (274)	BGC0000031	**Core**
Indolmycin	duplicated Trp-tRNA synthase	*Streptomyces griseus ATCC12648*	AJT38681 (275)	BGC0001206	**Res.**
Salinosporamide A	duplicated beta-proteasome subunit	*Salinospora tropica CNB-440*	ABP53490 (232)	BGC0001041	**Core + Res.**
Eponemycin	duplicated beta-proteasome subuint	*Streptomyces hygroscopicus ATCC 53709*	AHB38505 (276)	BGC0000345	**Core + Res.**
Vancomycin	Peptidoglycan remodeling	*Amycolatopsis orientalis* HCCB 10007	CCD33128, CCD33129, CCD33130 (261)	BGC0000455	**Res.**
Cephamycin	duplicated beta-lactamase	*Streptomyces clavuligerus ATCC 27064*	AAF86620 (262)	BGC0000319	**Res.**
Platencin	duplicated FabB/F	*Streptomyces platensis MA7339*	ACS13710 (277)	BGC0001156	***Core**
Thiolactomycin	duplicated FabB/F	*Salinispora pacifica DSM 45543*	ALJ49913 (134)	BGC0001237	***Core**
Thiotetroamide	two copies of FabB/F	*Streptomyces afghaniensis NRRL5621*	ALJ49924, ALJ49919 (134)	BGC0001236	***Core,*Core**
Andrimid	One copy of acetyl-CoA carboxyltransferase	*Pantoea agglomerans*	AAO39114 (278)	BGC0000956	**Res.**

Table 6.3: ARTS results showing positive examples of self-resistance form the MiBIG databse. Purple boxes show putative resistance, as in vitro experiment are not confirmed. Grey boxes and those marked with (*) show E1 exploration mode using 90% of normal bit-score thresholds. Figure adapted from Alanjary et al (239)

The 23S rRNA methyltransferase, was identified in clusters for Erythromycin and Pikromycin but not Avilamycin due to its significantly truncated sequence in this cluster. The results show that although these were single cluster submissions ARTS is still able to detect the resistance factors in the vast majority of cases. This illustrates that although using whole genomes is beneficial to take advantage of all criterion, draft or fragmented genomes can still be used to highlight self-resistance factors in clusters.

Positive examples from genomes outside reference phylogeny

ARTS so far includes one reference for the prolific Actinobacteria class however, as demonstrated by the positive results for all three examples from Firmicutes and Proteobacteria (Table 6.1), ARTS can still be applied to organisms that are not apart of the reference phylum. Because many of the core genes are ubiquitous to all of bacteria and in single copy the co-localization and duplication criteria served to identify these cases. One interesting example is the Agrocin 84 producer *Agrobacterium radiobacter* K84. This genome showed a duplicated Leu-tRNA synthetase, the target of Agrocin 84, located on the pAgK84 plasmid. This area did not show to be a BGC using default detection however, so only the duplication criteria highlighted the area. Indeed this was shown to be apart of the producing cluster (260) and only after a extended BGC search using clusterfinder was a putative 9kb segment identified. While more manual exploration is required to identify these types of hits some of the exploration functions made this process very simple. For example, by discriminating the duplicates by global ubiquity and single copy hits this example showed to be in the top three rows of the core gene table. Furthermore, the lower homology score seen with the resistant copy and clear origin from a plasmid helped to quickly show this gene was a likely potential resistant target. This demonstrates that not only can ARTS be used successfully without the phylogenetic criteria but also that it can function as an orthogonal cluster detection method, which can complement current methods.

Detection frequency

As the rate of false resistance targets is unknown in these genomes we used total detection frequency as a means of estimating worst-case false positive rates. All tests also used the optional exploratory mode to assess the upper range of hit frequencies. The complete

Actinobacteria set including positive example genomes from NCBI RefSeq comprised 200 complete isolates in total that were tested. All clusters from the MiBIG v1.3 set were also run though the ARTS web interface for testing. For the complete genome tests, each criterion except phylogeny showed a hit rate of 5% of core genes; this resulted in less than 20 hits on average for those with two or more criteria selected (Table 6.4).

Hit type	Average	SD	Min	Max
Core	489	79	271	653
Duplication	27	23	0	96
BGC proximity	27	24	0	140
Phylogeny	125	88	7	422
Two +	**16**	16	0	83
Three +	**2**	4	0	22

T**able 6.4:** Average ARTS detection counts for 200 tested genomes. Figure adapted from Alanjary et al (239)

The maximum values for these detections were rare. Despite these higher values the major discriminatory power is illustrated by the cross-referencing of criteria, where even in the most extreme case only 22 potential genes with three criteria are left to investigate. Although the figures for HGT seem high other reports have stated similar values of approximately 35% HGT for *Actinobacteria* (279). With the higher figures seen for phylogeny it is therefore recommended to use this measure in conjunction with other criteria however. Overall few average hits were seen for those with multiple criteria as seen with many of the positive case studies. For example *Planobispora rosea*, the producer of the thiazolyl peptide GE2270, shows positive hits for duplication, co-localization, and HGT for the resistant target elongation factor (EF-Tu); these modifications present in the domain II have also only been seen in thiazolyl resistant *Bacillus subtilis* (263) implying this may indeed have been shared horizontally.

The MiBIG clusters used totaled 1409 characterized BGCs. For comparison to the default ARTS analysis, using the filtered set of core genes, we examined the clusters for function and percent of clusters that showed ARTS hits. For single hits in a BGC roughly a third and a quarter of clusters showed hits to core genes or resistance models respectively (Figure 6.6). Compared to the default core gene mode this figure falls by more that half largely due to the reduction of transport proteins (cell envelope), protein synthesis, and energy and metabolism enzymes (Figure 6.6a). The functional classification of hits also

shows a significant amount of "unclassified" or "other" core gene hits, which underscores the possibility of finding resistance factors that are currently uncharacterized.

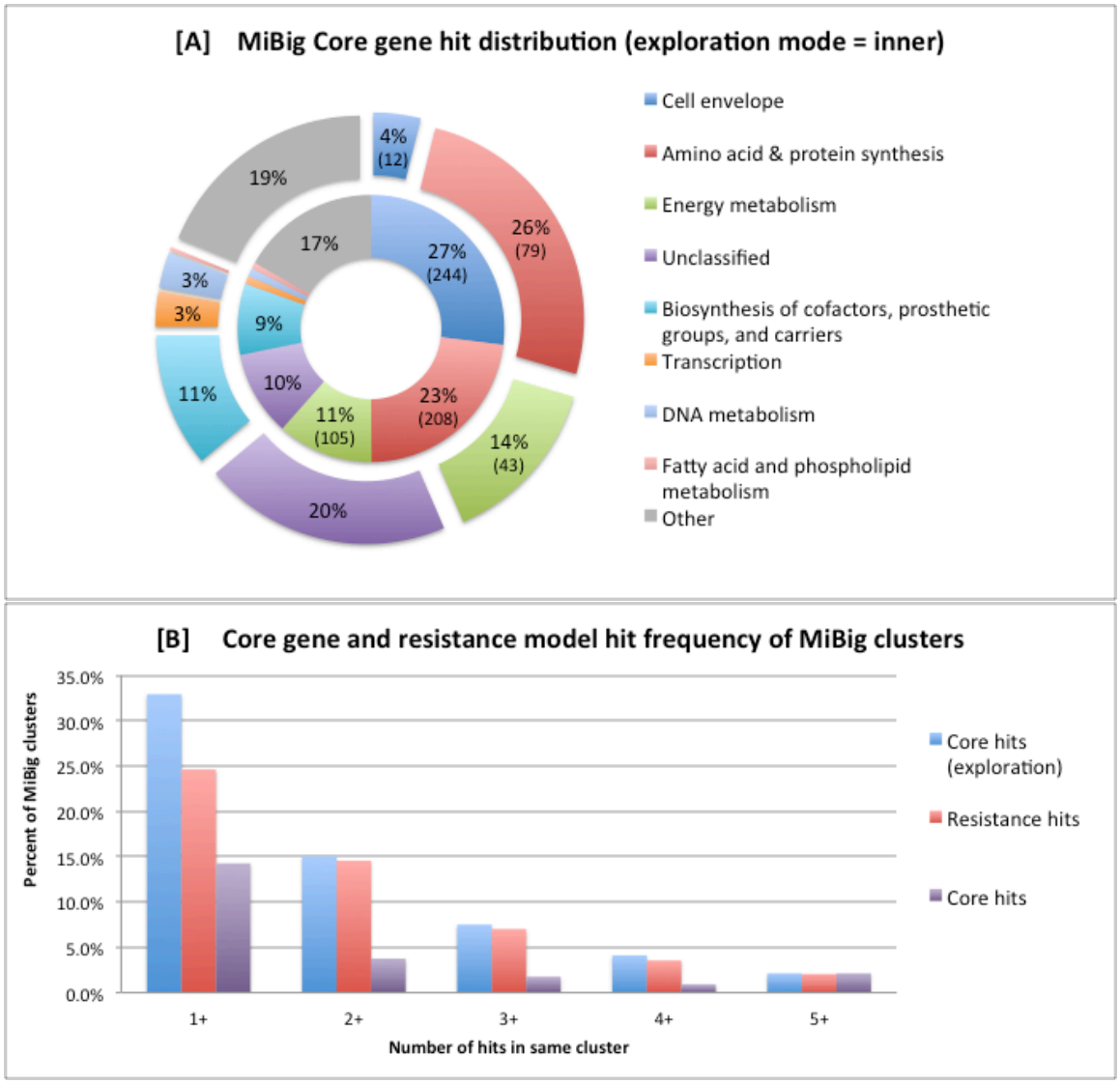

Figure 6.6: (A) Functional classification of all MiBIG hits from ARTS. The inner ring represents all core genes (exploration mode) with the outer showing the filtered set. (B) Core gene counts of ARTS hits relative to all BGCs analyzed. Purple shows the default ARTS search and blue shows the exploration mode search. Figure adapted from Alanjary et al (239)

Some multiple core gene hits are seen in a single cluster but with a sharp reduction in the filtered core list. For those with much higher values, some genes were found in peripheral areas of the cluster. Thus we believe clusters with high core gene hits are likely due to inaccurate cluster boundaries which happen to include neighboring areas of the genome that

contain core genes. As seen in Figure 6.7 the extra core gene annotation with ARTS might also help establish true cluster boundaries, which remains a difficult problem to be solved.

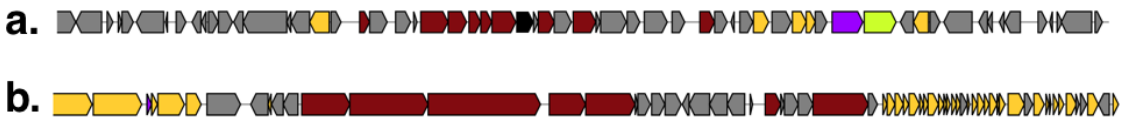

Figure 6.7: Comparison of positive example cluster from *Streptomyces roseochromogenes* DS 12.976 (a) where core genes are shown in yellow, known resistance in green and hits for both shown in purple. This is compared with another cluster in the genome (b) with a high number of core gene hits likely due to boundary issues. Figure adapted from Alanjary et al (239)

6.4 Discussion

With the many BGC predictions from current genome mining applications it is important to effectively enrich these prospects for those that will yield successful drug candidates without performing exhaustive experimental screening. Target directed genome mining affords an attractive approach to the prioritization of antibiotic clusters as it also helps to accelerate downstream experiments by providing clues to identify the target protein. This prioritization can be done using high confidence known resistance targets or instead be used to screen hundreds of putative essential genes to discover novel targets and antibiotics with new modes of action. This proof of concept was shown in other independent studies using duplication as a marker (134) as well as illustrated in the validation studies here. A key design goal for ARTS, besides the automation of this task, is to help make the many disparate results digestible to users. The integration of several dynamically presented tables is an easy to use solution that proved to rapidly identify each of the positive examples. Confirmation of each criterion could also be immediately visualized using the multiple sections in the final results. These functions proved beneficial when identifying the hits that showed few criteria. For example with the *Nocardia farcinica* IFM 10152 genome only duplication and known resistance was seen, as the resistant *rpoB* gene was not present in any of the BGCs. Filtering by function and checking the homology score quickly identified the resistant variant *rpoB2* as the lower score indicated a divergent version. The occasional high number of criteria hits in exploration mode also neglects to account for these post-filtering capabilities presented to the user. To illustrate this the *Streptomyces roseochromogenes* genome showed the highest number

of results for genes with two or more criteria, 73 using trusted cutoffs. By focusing on key categories such as DNA metabolism and transcription only 2 hits from this list are seen (including the positive control); filtering by protein and amino acid synthesis, another common target category, also shows a manageable hit count of 20. Moreover, although 94 duplications hits were found in this case, when sorted by single copy prevalence the resistant *gyr*B is ranked third in the list with only 11 hits over a 0.9 threshold. Due to this rapid filtering provided by the interactive navigation, potential false positives were intentionally retained to enable exploration and cross-examination with the other provided properties in the table. This way the user can explore the results and make more educated judgment calls to avoid excluding potential novel targets.

Although we have allowed the possibility of more false positives, results for total detection frequency showed to be manageable. The number of hits to curate per genome averaged between 10-30 genes with two criteria or more. Compared to the 664 possible essential genes, these highlights can be investigated relatively quickly. Alternatively, focus can be placed on high confidence known targets and known resistance factors which showed to highlight only 25% of the clusters in the MiBIG database. Likewise a third of the BGCs on average were highlighted with known resistance in the whole genome tests. The main source of high ARTS hit counts came form the phylogeny HGT screening. Because many of these were due to intra-genus transfers an additional filtering of inter-genus transfers is provided if the user has genus annotation in the genbank file. Here reported transfers of the same genus will be excluded to only highlight less common HGT events that might be more significantly related to resistance acquisition. This measure only helps to reduce the events seen in the highly represented genera in the reference such as those seen in Streptomyces and Corynebacterium however. The curation using the provided tree view in the phylogeny panel of the results is also helpful in reducing the number of potential hits. This view also helps to identify the potential source of HGT as the variant gene will be placed near a related genome from where it originated. This can aid further investigation if the source organism has a similar known resistance mechanism.

The Actinobacteria reference set showed to have higher representations of well studied genera such as *Streptomyces* but nevertheless showed modest coverage of 86 of the 130 known genera (280) in the class. With continued efforts and an increase in high quality genomes this should be expanded in the next iterations of ARTS along with representatives from other

promising natural product producers such as *Cyanobacteria* and *Myxobacteria* (135, 137). Expanded reference sets will help to improve HGT detection and provide more specific sets of essential genes but even without the use of the reference set the application can prove beneficial as seen with the Firmicutes and Proteobacteria positive examples. By using known resistance factors, known targets, duplication, and co-localization criteria, promising clusters can be highlighted independently of the reference phylogeny. These factors also illustrated that a potential BGC that is not identified can be highlighted by ARTS, as seen with the Agrocin 84 producing genome. Therefore this application has the potential to highlight novel BGCs for which we have no known biosynthesis motifs for detection from the duplication criterion alone.

Future and concurrent work to improve this pipeline nevertheless involves expanding the reference and improving HGT detection as this factor was shown to help discriminate potential leads; with the *Planobispora rosea* ATCC 53733 positive example, the probable HGT event helped to reduce the number of leads from 26 to 4 by using HGT as the third criteria. With the completion of the autoMLST application this process of reference generation can now be accelerated with the potential to provide family specific reference sets rather than using higher order taxonomy. These smaller taxa groupings can also allow alternative methods of species tree inference, such as concatenation, so long as the set of organisms remain low runtimes can remain practical. However the accelerated coalescent approach remains beneficial as it can support a larger span of organisms that can be computed in reasonable time. This was shown to handle close to 200 taxa with a speed increase of 350X compared to building all trees de novo. Besides the improvement of reference sets, updated models for new known resistance is in progress. In case there are models of interest that are not present in the current version of ARTS we have also included the ability for users to provide their own. In the advanced section a user can submit a valid HMM for either core genes or known resistance genes along with their genomes. Another useful feature would be the detection of homologous recombination events in addition to HGT. This remains a difficult process however as it would require screening several segments of one gene, which would dramatically increase processing time. However integration of other methods (281) is a possibility. The release of the antiSMASH API and their fast processing mode can also help to accelerate run times for jobs submitted to ARTS (112); this new version also improves cluster boundary prediction and so we hope to include this with the next release.

In addition to these updates we are actively using the software for BGC prioritization. By taking advantage of the recently released antiSMASH database (130) we are currently identifying candidates with reoccurring ARTS hits to highlight interesting leads. Preliminary results show common hits that are duplicated and co-localized in a BGC include various ribosomal subunits, a common target area for antibiotics. Also several known targets are seen and co-localized in the initial list including: *gyrB*, *EF-Tu*, and *dnaN*. As these data are not yet complete we hope to interrogate the results shortly with the hope of finding associated clusters for new targets or known targets with new susceptibilities.

Overall this first version of ARTS showed to accomplish the goals of automating the target directed genome mining approach with nearly every positive example showing detection. With the upper bound of hit frequencies remaining manageable, especially when using the post-processing curation features, this study showed to validate ARTS as a useful prioritization and exploration tool. Having worked with our collaborators at Novartis we were also able to provide easy installation and deployment solutions for this application via a simple container image; additional documentation to install from source on various Linux operating systems is also included. Novartis has also contributed a Virtual Private Server (VPS) cloud image though Amazon Web Services (AWS), which can help with quick deployments. To encourage more widespread use of these prioritization methods we also will continue to maintain the public server as we provide new versions. Ultimately we hope that this application will encourage broader use of these prioritization schemes for the natural product community. As we search for new antibiotics to combat the threat of clinical drug resistance it is crucial that we utilize comparative approaches to fully leverage the growing set of genomic data and enhance the impact of each downstream experiment.

7 Discussion and Conclusions

7.1 Overview and ongoing efforts

The danger of regressing to a time without useful antibiotics is a paramount concern in this research. While we are focused on replenishing our defenses against harmful microorganisms via new effective antibiotics, we acknowledge that additional measures will be needed to fully solve this problem. With proper legislative safeguards for our antibiotics and development of new innovative techniques, such as combination therapies (282) and phage treatments (283), we hope to remain in an era where we can continue to treat microbial infections. The comparative techniques developed here have been shown to help with the discovery of antibiotics but can also serve a broader impact to a variety of research. From industrial compounds such as biodegradable surfactants (284) to a variety of anti-invectives, natural products have shown to be a great source for many beneficial compounds. For example, a candidate anti-cancer compound salinisporamide A, a potent proteasome inhibitor (285), is also highlighted in ARTS. These techniques have come at a critical time as our traditional methods have been plagued with rediscovery of previously known compounds. Fortunately whole genomic data has continued to increase and with the improvement of sequencing technologies it is projected to increase at an accelerated rate. Shotgun metagenomic sequencing is also emerging as a means to obtain a large amount of whole genomes from taxa that previously could not be cultured and analyzed (286, 287). Not only does the volume of data hold more promise for investigating a wider space of chemical potential, but also the high diversity of whole genomes can enable more powerful comparative analysis. By shifting from single genome mining techniques to a multi-genome perspective we can start to fully leverage these data. The demonstrations in this thesis have showed some basic applications such as de-replication of known compounds and prioritization of likely leads but further development and integration of these techniques can help to highlight the "dark matter" of genomics for discovery.

In Chapter 1 the application of large-scale genome networking was demonstrated to identify areas of known chemical potential by using databases such as MiBIG. This application is of immediate help as it can mask those clusters that will lead to rediscovered compounds and save time for investigating other likely leads. Currently our databases remain sparse but continue to grow, as more experimental data is added. To complement the

confirmation via experimentally verified products, computational efforts to associate orphan clusters to other known compounds are also underway. Improvements in structure prediction from sequence (113) have shown to be a promising method for achieving these ends as well. Furthermore, the prospect of cross-referencing these gene cluster networks with GNPS techniques is an attractive possibility to expanding orphan clusters into this de-replication process. This rapid grouping of similar BGCs also hinted at other perspectives to help guide discovery. For example many of the effective compounds in the Amycolatopsis network from publication 3 showed to be apart of multi-species GCFs. With this extra annotation of high-resolution phylogeny, one possible application could be selecting for multi-species GCFs that have no known product. This idea of searching for common rather than unique BGCs has been proposed elsewhere (62), as the hypothesis is that antibiotics could be universally beneficial in contrast to unique secondary metabolites which might only have a benefit to a specific ecological niche. This example briefly illustrates the potential for cross-referencing different metadata in this comparative perspective by using a variety of results from bioassays and demonstrates the discriminatory power of using a multi-genome comparative approach. With the development of the tools presented in this thesis we now are actively looking to integrate these methods. In particular, classifying organisms by high-resolution species trees using autoMLST can achieve a similar analysis as in publication 3. Also the incorporation of results from ARTS is another prospect for annotating these gene cluster networks; if multiple organisms show the same low confidence putative target then perhaps it is worth investigating. For instance, a hypothetical essential protein that consistently shows up in similar clusters may be worth experimental effort. This can lead to more adventurous exploration with the ultimate goal of discovering novel compounds and drug targets.

Apart from providing more detailed metadata for comparative use, the high-resolution species trees in autoMLST (Chapter 2) aims to serve a demanding need for fast and fine-grained taxonomic identification. Although 16S data may not yield definitive results in some cases, it has a clear advantage of being fast and intuitive to a wide range of users. This aspect of being accessible is one of the major goals of the autoMLST server, which involves the integration of several workflows into a single pipeline. We have therefore generated a simple "BLAST like" web interface to achieve the multiple steps required to use modern MLSA methods. This application is distinct from similar web based phylogeny methods in that it

automates each step of the process, including non-trivial gene and organism selection. It provides advanced features such as model finding, bootstrap analysis, ANI estimation, and complementary methods of analysis so problems in inference can be easily identified. Automation of such a complicated process comes with the disclaimer that hypothesis should not be taken for granted. Despite this disclaimer, the several validation perspectives showed that the majority of the resulting trees were well supported by default processing. This was determined from branch support values, topological consistency with ANI values, and a comparison to manual analysis. While automation has shown to work well in the majority of cases, we have intentionally provided features to help in the quality control of results to encourage their use. With alternative methods of inference, easy to use reanalysis options, and bootstrap analysis any non-specialist can scrutinize and test their results easily. Efforts to provide useful metadata such as with the ANI clan annotations are also ongoing. For example, defining prolific BGC producing clans can help the user to quickly prioritize their samples for those that are potentially rich in natural products. Our known databases still require further expansion of representative genomes and verified type-strains, but for the de novo workflow this is not much of a disadvantage as query trees can be built independent of the reference organisms. However we are committed to providing continually updated reference genomes for the rapid placement workflow provided. Overall responses to this application have been positive and continued feedback from collaborators and public users involved in beta testing are helping to finalize a release candidate. In addition to providing the public web server we have made all source code freely available (https://bitbucket.org/ziemertlab/automlst) so that users can use this interface without limits to number of genomes. Setting up a private autoMLST server requires some expertise currently, however we are working with collaborators from the Fraunhofer institute to make this setup easier. We hope to provide a rapid deployment option using simple container technologies when a major release is finished, as is the case with the ARTS server.

ARTS (Chapter 3) has been live for over a year and remains the only public server to perform an extended target directed genome mining analysis. This application showed to highlight known clusters with associated self-resistance factors, which is an attractive approach to BGC prioritization. This not only helps by enriching predictions for likely antibiotic compounds but also guides mechanism of action studies, as the target is identified from the start. Utilizing the accelerated phylogeny approach for HGT determination was

shown to save several hours of processing time, as we were able to leverage pre-computation of over 664 individual genes. Performing a similar analysis manually would limit the amount of genomes screened and would be infeasible for large-scale genome mining. As seen in other studies (134), the duplication and co-localization criteria screening also offer valuable prospects for novel target identification. With the identification of duplicated known targets ARTS was applied to genera outside of the reference phylogeny and able to identify all positive examples form these genomes. Furthermore, examples using a single cluster without genome context were shown to be successful with most of the MiBIG clusters with known self-resistance being highlighted. This also demonstrates that fragmented draft genomes may also be successfully analyzed, although we encourage high quality complete genome to take full advantage of all criteria. The high positively identified cases were also matched with a low hit frequency, which on average highlighted 5% or less essential genes (except for HGT criteria). We used total hit frequency as a proxy for false positive rate considering the unknown status of most of the predictions. This was shown to only have high figures for the HGT criteria with approximately 25% of core genes highlighted on average. Although similar figures for HGT have been found elsewhere (279) this showed it is important to use this measure with multiple criteria for identifying confident hits. Predictions with two or more screening criteria showed to highlight a manageable amount of possible leads for most cases. The detection frequency study also fails to illustrate the explorative functions available to the user for cross-referencing other properties of viable targets, such as ubiquity. Considering these features we have allowed for the higher number of predictions to allow more educated discrimination by the user. Nevertheless the positive examples showed many examples where multiple criteria were highlighted, so quick predictions could also be made on this basis, as average hits for three or more criteria were around 1-4 core genes. Efforts to improve the HGT prediction and highlighting the most significant predictions, such as those originating from other genera, are still ongoing. Besides exploring other methods of tree reconciliation we also would like to add bootstrap support of branches into the automated prediction to limit any potential HGT inference due to poorly placed branches. A common request from users is also the addition of more reference sets, which is also in progress. Overall feedback from users has been positive thus far from both industry and public users and we hope to integrate these improvements with the next release.

7.2 Outlook and concluding remarks

Genome mining is an attractive avenue for natural products discovery as the cost of exploration is significantly reduced compared to traditional cultivation and screening practices. These methods also allow for capturing the full potential of chemical space indirectly through genomic potential. The detection of BGCs responsible for these compounds has matured over the last decade and resulted in efficient identification of a variety of classes of production mechanisms. This has resulted in an urgent need for prioritizing these clusters and, as demonstrated here, several approaches have been automated to accelerate this task. By comparing multiple BGCs using similarity networking, high-resolution taxonomic classification, and targeted genome mining we are able to enrich for leads with potentially lower likelihoods of rediscovery and a higher chance of desired activity. The subsequent expression of these leads can also be aided by improved identification of optimal hosts using high-resolution species or strain identification; these efforts are also complemented by recent advancements in genome manipulation technologies (288) and decreasing costs for synthesis of large DNA scaffolds (289). With the intersection of these technologies and the growing number of publicly available sequences enabling comparative analysis it seems likely that genome mining has yet to reach its full potential and can serve to provide insightful discoveries for natural products research.

Traditional methods still remain the workhorse of natural products discovery and we expect genome mining and comparative analysis will continue to support these endeavors. By focusing on new prolific genera for novel compounds to associating BGCs with desired products, genome mining methods can further improve discovery efforts, downstream analysis, and mode of action studies. This research has also kept in mind that it is important to provide tools that encourage widespread use to increase the chances of discovery. Considering this we have provided intuitive web interfaces that non-specialist can immediately take advantage of. We hope to continue to provide these tools publicly and utilize them for furthered antibiotic research so that we may never live in a world without effective antibiotics again.

8 List of publications and manuscripts

Research articles:

1. **Alanjary,M**., Kronmiller,B., Adamek,M., Blin,K., Weber,T., Huson,D., Philmus,B. and 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, W42–W48.

   Referred in text as publication 1

2. Schorn,M.A., **Alanjary,M.M**., Aguinaldo,K., Korobeynikov,A., Podell,S., Patin,N., Lincecum,T., Jensen,P.R., Ziemert,N. and Moore,B.S. (2016) Sequencing rare marine actinomycete genomes reveals high density of unique natural product biosynthetic gene clusters. *Microbiol.*, 162.

   Referred in text as publication 2

3. Adamek,M., **Alanjary,M**., Sales-Ortells,H., Goodfellow,M., Bull,A.T., Winkler,A., Wibberg,D., Kalinowski,J. and Ziemert,N. (2018) Comparative genomics reveals phylogenetic distribution patterns of secondary metabolites in Amycolatopsis species. *BMC Genomics*, 19, 426.

   Referred in text as publication 3

Reviews and relevant articles:

4. Ziemert,N., **Alanjary,M**. and Weber,T. (2016) The evolution of genome mining in microbes-a review. *Nat. Prod. Rep.*, 33.

   Referred in text as publication 4

5. Bruns,H., Crüsemann,M., Letzel,A.-C., **Alanjary,M**., McInerney,J.O., Jensen,P.R., Schulz,S., Moore,B.S. and Ziemert,N. (2017) Function-related replacement of bacterial siderophore pathways. *ISME J*., 10.1038/ismej.2017.137.

   Referred in text as publication 5

6. Spohn,M., Edenhart,S., **Alanjary,M**., Ziemert,N., Wibberg,D., Niedermeyer,T.H.J., Stegmann,E. and Wohlleben,W. (2018) Identification of a novel aminopolycarboxylic acid siderophore gene cluster encoding the biosynthesis of ethylenediaminesuccinic acid hydroxyarginine. *Metallomics*, 10.1039/C8MT00009C.

   Referred in text as publication 6

Manuscripts in progress:

7. **Alanjary M**., Steinke K., Adamek M., Huson D, Ziemert N. The Automated Multi-Locus Species Tree (autoMLST) enables rapid high-resolution bacterial species phylogenies.

   Referred in text as manuscript 1

9 Contributions

Publication 1

In the publication "The Antibiotic Resistant Target Seeker (ARTS), an exploration engine for antibiotic cluster prioritization and novel drug target discovery ", I performed all experiments pertaining to defining a computationally feasible pipeline and testing of various tools that can achieve sub-processes of the ARTS application. With the exception of open-source libraries, frameworks and external applications, I authored all code for the web interface, infrastructure maintenance scripts, and analysis pipeline. I also authored scripts for easy software distribution as well as the content of software documentation and help pages. All authors invested time into validation and bug testing with significant contributions from Martina Adamek and myself. We thank our collaborators Dr. Philmus and Dr. Kronmiller for valued discussions and additional Hidden Markov Models (HMMs) of known antibiotic resistance factors. Discussions with Dr. Weber and Dr. Blin of the antiSMASH project helped with a foundational component of BGC identification in ARTS. I authored the final manuscript with valued edits and feedback from my advisors, Dr. Ziemert and Dr. Huson. Contributions from my advisors on proper phylogeny construction, HGT testing, and concept were also integral to the realization of this project.

Publication 2

I contributed equally to the analysis of the results from experiments and sequencing conducted by Dr. Schorn in the publication "Sequencing rare marine actinomycete genomes reveals high density of unique natural product biosynthetic gene clusters". I aided in testing alternative genome assembly methods, quality control of genomes, and calculation of diversity indices. I preformed the collection of reference BGCs, implementation of high-throughput similarity scoring, and generation of gene cluster networks. Dr. Schorn wrote the manuscript with my contributions to methods.

Publication 3

Martina Adamek wrote the manuscript "Comparative genome mining reveals phylogenetic distribution patterns of secondary metabolites in Amycolatopsis" as well as performed all experiments. I aided with the generation of gene cluster similarity scoring and minor edits to the manuscript.

Publication 4

The review "The evolution of genome mining in microbes-a review" was written by Dr. Ziemert and Dr. Weber. I contributed with the generation of a gene cluster network of known BGCs.

Publication 5

In the paper "Function-related replacement of bacterial siderophore pathways" I preformed the Multi Locus Sequence Analysis (MLSA) and generation of a species tree, gene trees, and initial HGT assessment. All other experiments and analysis were performed by my co-authors. Dr. Burns and Dr. Ziemert wrote the manuscript with contributions to the methods from myself.

Publication 6

In the paper "Identification of a novel aminopolycarboxylic acid siderophore gene cluster encoding the biosynthesis of ethylenediaminesuccinic acid hydroxyarginine. Metallomics" I carried out a MLSA of related Streptomyces genomes to produce a high-resolution species tree. All other experiments and writing of the manuscript were performed by co-authors with additions to the methods from myself.

Manuscript 1

In the manuscript "The Automated Multi-Locus Species Tree (autoMLST) enables rapid high-resolution bacterial species phylogenies" I developed the workflow and oversaw web development. Katharine Steinke authored the front-end code base. I authored analysis and infrastructure scripts, excepting open-source libraries and frameworks. Katharine Steinke performed the initial validation followed by a large-scale validation by Martina Adamek and I. My advisors aided with the concept, resources, and direction of the project.

10 Abbreviations

4'PP	4'-phospho-pantethine
A	Adenylation
ABC	Atlas of Biosynthetic gene Clusters
ACP	acyl carrier protein
ANI	Average Nucleotide
antiSMASH	antibiotics & Secondary Metabolite AnalysisShell
ARTS	Antibiotic Resistant Target Seeker
AT	acyl transferase
ATP	Adenosine Triphosphate
autoMLST	Automated Multi-Locus Species Tree
AWS	Amazon Web Services
BGC	Biosynthetic Gene Cluster
BLAST	Basic Local Alignment Search Tool
C	Condensation
CDC	Center for disease control
CPS	capsular polysaccharides
CRE	Carbapenem-resistant Enterobacteriaceae
DH	dehydratase
DMADP	dimethylallyl diphosphate
DNA	Deoxyribonucleic acid
EDDS	ethylene diamine disuccinic acid
ER	enoylreductase
ETE3	Environment for Tree Exploration
FAS	Fatty Acid Synthase
GCF	Gene Cluster Family
GNPS	Global Natural Product Social Molecular Networking
HGT	Horizontal Gene Transfer
HMM	Hidden Markov Model
Ichip	Isolation chip
IDP	isopentenyl diphosphate
IMG	Integrated Microbial Genomes
JGI	Joint Genome Institute
KR	ketoreductase
KS	ketosynthase
LPS	lipopolysaccharides
Mbp	Mega-base-pair
MCL	Markov Cluster Algorithm
MDR	Multi-Drug Resistant
MEP	2-C-methyl-D-erythritol-4-phosphate

MGB	Multi-Gene-Blast
MiBIG	Minimum Information about a Biosynthetic Gene cluster
ML	Maximum Likelihood
MLS	Macrolide-Lincosamide-Streptogramins
MLSA	Multi Locous Sequence Analysis
MT	methyltransferase
MVA	mevalonate
NaPDoS	Natural Product Domain Seeker
NCBI	National Center for Biotechnology Information
NRPS	Non-Ribosomal Peptide Synthetase
PCP	Peptide Carrier Protein
PCR	Polymerase Chain Reaction
Pfam	Protein family
PKS	Polyketide Synthases
PRISM	Prediction Informatics for Secondary Metabolomes
RiPP	ribosomally synthesized and post-translationally modified peptides
RMA	Rare Marine Actinomyces
RNA	Ribonucleic acid
SIO	Scripps Institute of Oceanography
SM	Secondary Metabolite
TE	thioesterase
TPS	terpene synthases
VPS	Virtual Private Server
WHO	World health Organiszation
XDR	Extensively Drug Resistant

11 References

1. McCarthy,M. (2017) Woman dies after infection with bacteria resistant to all antibiotics available in US. *BMJ*, **356**.
2. Schäberle,T.F. and Hack,I.M. (2014) Overcoming the current deadlock in antibiotic research. *Trends Microbiol.*, **22**, 165–167.
3. Bassett,E., Keith,M., Armelagos,G., Martin,D. and Villanueva,A. (1980) Tetracycline-labeled human bone from ancient Sudanese Nubia (A.D. 350). *Science (80-.).*, **209**, 1532–1534.
4. Cook,M., Molto,E.L. and Anderson,C. (1989) Flourochrome labelling in Roman Period skeletons from Dakhleh Oaisis, Egypt. *Am. J. Phys. Anthr.*, **80**, 137–143.
5. Clardy,J., Fischbach,M.A. and Currie,C.R. (2009) The natural history of antibiotics. *Curr. Biol.*, **19**, 1–8.
6. René Dubos and Jean Dubos. The white plague: Tuberculosis, man, and society, 2nd ed. New Brunswick, N.J.: Rutgers University Press, 1987; originally published 1952. xxxviii + 277 pp. $2700 (cloth); $12.00 (paper) (2018) *J. Hist. Behav. Sci.*, **26**, 307.
7. Hershkovitz,I., Donoghue,H.D., Minnikin,D.E., Besra,G.S., Lee,O.Y.C., Gernaey,A.M., Galili,E., Eshed,V., Greenblatt,C.L., Lemma,E., *et al.* (2008) Detection and molecular characterization of 9000-year-old Mycobacterium tuberculosis from a neolithic settlement in the Eastern mediterranean. *PLoS One*, **3**, 1–6.
8. Hu,D., Liu,B., Feng,L., Ding,P., Guo,X., Wang,M., Cao,B., Reeves,P.R. and Wang,L. (2016) Origins of the current seventh cholera pandemic. *Proc. Natl. Acad. Sci.*, **113**, E7730–E7739.
9. Karagöz,E., Turhan,V., Hatipoğlu,M. and Ozkuzugudenli,B. (2017) Wartime infections and tragedies at the beginning of the 20thcentury in the eastern part of Turkey. *Infez. Med.*, **25**, 84–87.
10. Weismann,K. (1995) Neurosyphilis, or chronic heavy metal poisoning: Karen Blixen's lifelong disease. *Sex. Transm. Dis.*, **22**, 137–144.
11. DeFriez,A.I., Patton,W.E., Welch,E.J. and Badger,T.L. (1954) Bed Rest in the Treatment of Pulmonary Tuberculosis. *N. Engl. J. Med.*, **250**, 39–46.
12. Namana,V., Gupta,S.S., Sarasam,R. and Mathur,P. (2017) Historical TB treatment-Plombage. *Qjm*, **110**, 191–191.
13. Cohen,J. (1999) Hospital gangrene: the scourge of surgeons in the past. *Infect. Control Hosp. Epidemiol.*, **20**, 638–640.
14. Gest,H. (2004) The discovery of microorganisms by Robert Hooke and Antoni van Leeuwenhoek, Fellows of The Royal Society. *Notes Rec. R. Soc.*, **58**, 187–201.
15. Williams,K.J. (2009) The introduction of 'chemotherapy' using arsphenamine - The first magic bullet. *J. R. Soc. Med.*, **102**, 343–348.
16. Gelmo,P. (1908) Über Sulfamide der p-Amidobenzolsulfonsäure. *J. für Prakt. Chemie*.
17. Fleming,A. (2001) On the antibacterial action of cultures of a penicillium, with special reference to their use in the isolation of B. influenzae. 1929. *Bull. World Health Organ.*, **79**, 780–790.
18. Chain,E., Florey,H.W., Gardner,A.D., Heatley,N.G., Jennings,M.A., Orr-Ewing,J. and Sanders,A.G. (2005) THE CLASSIC: penicillin as a chemotherapeutic agent. 1940. *Clin. Orthop. Relat. Res.*, **439**, 23–26.
19. Comroe,J.H.J. (1978) Pay dirt: the story of streptomycin. Part I. From Waksman to Waksman. *Am. Rev. Respir. Dis.*, **117**, 773–781.

20. Barberis,I., Bragazzi,N.L., Galluzzo,L. and Martini,M. (2017) The history of tuberculosis: From the first historical records to the isolation of Koch's bacillus. *J. Prev. Med. Hyg.*, **58**, E9–E12.
21. Powers,J.H. (2004) Antimicrobial drug development – the past, the present, and the future. *Clin. Microbiol. Infect.*, **10**, 23–31.
22. Ventola,C.L. (2015) The Antibiotic Resistance Crisis: Part 1: Causes and Threats. *Pharm. Ther.*, **40**, 277–283.
23. Barbosa,T.M. and Levy,S.B. (2000) The impact of antibiotic use on resistance development and persistence. *Drug Resist. Updat.*, **3**, 303–311.
24. Hall,B., Salipante,S. and Barlow,M. (2004) Independent Origins of Subgroup Bl+B2 and Subgroup B3Metallo-?-Lactamases. *J. Mol. Evol.*, **59**, 133–141.
25. Sun,J., Deng,Z. and Yan,A. (2014) Bacterial multidrug efflux pumps: Mechanisms, physiology and pharmacological exploitations. *Biochem. Biophys. Res. Commun.*, **453**, 254–267.
26. Piddock,L.J. V (2006) Multidrug-resistance efflux pumps not just for resistance. *Nat. Rev. Microbiol.*, **4**, 629.
27. Freel,K.C., Millan-Aguinaga,N. and Jensen,P.R. (2013) Multilocus Sequence Typing Reveals Evidence of Homologous Recombination Linked to Antibiotic Resistance in the Genus Salinispora. *Appl. Environ. Microbiol.*, **79**, 5997–6005.
28. Ishikawa,J., Chiba,K., Kurita,H. and Satoh,H. (2006) Contribution of rpoB2 RNA polymerase beta-subunit gene to rifampin resistance in Nocardia species. *Antimicrob. Agents Chemother.*, **50**, 1342–1346.
29. Chow,C.S., Lamichhane,T.N. and Mahto,S.K. (2007) Expanding the nucleotide repertoire of the ribosome with post-transcriptional modifications. *ACS Chem. Biol.*, **2**, 610–619.
30. Vetting,M.W., Hegde,S.S., Fajardo,J.E., Fiser,A., Roderick,S.L., Takiff,H.E. and Blanchard,J.S. (2006) The Pentapeptide Repeat Proteins. *Biochemistry*, **45**, 1–10.
31. Davies,J. (1994) Inactivation of antibiotics and the dissemination of resistance genes. *Science*, **264**, 375–382.
32. Munita,J.M., Arias,C.A., Unit,A.R. and Santiago,A. De (2016) Mechanisms of Antibiotic Resistance. *Microbiol. Spectr.*, **4**, 1–37.
33. Stewart,P.S. (2002) Mechanisms of antibiotic resistance in bacterial biofilms. *Int. J. Med. Microbiol.*, **292**, 107–113.
34. Hollenbeck,B.L. and Rice,L.B. (2012) Intrinsic and acquired resistance mechanisms in enterococcus. *Virulence*, **3**, 421–433.
35. Stegmann,A.P., Honders,M.W., Hagemeijer,A., Hoebee,B., Willemze,R. and Landegent,J.E. (1995) In vitro-induced resistance to the deoxycytidine analogues cytarabine (AraC) and 5-aza-2'-deoxycytidine (DAC) in a rat model for acute myeloid leukemia is mediated by mutations in the deoxycytidine kinase (dck) gene. *Ann. Hematol.*, **71**, 41–47.
36. Sabtu,N., Enoch,D.A. and Brown,N.M. (2018) Antibiotic resistance : what , why , where , when and how ? 10.1093/bmb/ldv041.
37. Overballe-Petersen,S. and Willerslev,E. (2014) Horizontal transfer of short and degraded DNA has evolutionary implications for microbes and eukaryotic sexual reproduction. *Bioessays*, **36**, 1005–1010.
38. Derbyshire,K.M. and Gray,T.A. (2014) Distributive Conjugal Transfer: New Insights into Horizontal Gene Transfer and Genetic Exchange in Mycobacteria. *Microbiol Spectr.*,

2, 1–32.
39. Teuber,M. (1999) Spread of antibiotic resistance with food-borne pathogens. *Cell. Mol. Life Sci.*, **56**, 755–763.
40. Vogel,J.P. (2016) Bacterial type IV secretion: conjugation systems adapted to deliver effector molecules to host cells. *Trends Microbiol.*, **8**, 354–360.
41. Wan,Z., Varshavsky,J., Teegala,S., Mclawrence,J. and Goddard,N.L. (2011) Measuring the Rate of Conjugal Plasmid Transfer in a Bacterial Population Using Quantitative PCR D + λ X n X λ T n T X +. *Biophysj*, **101**, 237–244.
42. Jiang,X., Ellabaan,M.M.H., Charusanti,P., Munck,C., Blin,K., Tong,Y., Weber,T., Sommer,M.O.A. and Lee,S.Y. (2017) Dissemination of antibiotic resistance genes from antibiotic producers to pathogens. *Nat. Commun.*, **8**, 1–7.
43. ERIKSEN,K.R. and THERKELSEN,F. (1954) Infections, especially with penicillin-resistant staphylococci, following thoracic surgery treated with large doses of penicillin. *Acta Chir. Scand.*, **107**, 456–459.
44. Davies,J. and Davies,D. (2010) Origins and Evolution of Antibiotic Resistance. *Microbiol. Mol. Biol. Rev.*, **74**, 417–433.
45. Tacconelli,E., Carrara,E., Savoldi,A., Harbarth,S., Mendelson,M., Monnet,D.L., Pulcini,C., Kahlmeter,G., Kluytmans,J., Carmeli,Y., *et al.* (2018) Discovery, research, and development of new antibiotics: the WHO priority list of antibiotic-resistant bacteria and tuberculosis. *Lancet. Infect. Dis.*, **18**, 318–327.
46. Blair,J.M.A., Webber,M.A., Baylay,A.J., Ogbolu,D.O. and Piddock,L.J. V (2014) Molecular mechanisms of antibiotic resistance. *Nat. Publ. Gr.*, **13**, 42–51.
47. Nelson,K.N., Shah,N.S., Mathema,B., Ismail,N., Brust,J.C.M., Brown,T.S., Auld,S.C., Omar,S.V., Morris,N., Campbell,A., *et al.* (2018) Spatial Patterns of Extensively drug-resistant Tuberculosis (XDR-tuberculosis) transmission in KwaZulu-Natal, South Africa. *J. Infect. Dis.*, 10.1093/infdis/jiy394.
48. CDC (2018) Tuberculosis: Data and statistics. *www.cdc.gov*.
49. CDC (2016) Trends in Tuberculosis, 2016. *www.cdc.gov*.
50. Bassetti,M., Merelli,M., Temperoni,C. and Astilean,A. (2013) New antibiotics for bad bugs : where are we ? *Ann. Clin. Microbiol. Antimicrob.*, **12**, 1.
51. Dalia Deak, MPH; Kevin Outterson, LLM, JD; John H. Powers, MD; and Aaron S. Kesselheim, MD, JD,M. (2016) Progress in the Fight Against Multidrug-Resistant Bacteria? A Review of U.S. Food and Drug Administration–Approved Antibiotics, 2010–2015. *Ann. Intern. Med.*, 10.7326/M16-0291.
52. Fair,R.J. and Tor,Y. (2014) Antibiotics and Bacterial Resistance in the 21st Century. *Perspect. Medicin. Chem.*, 10.4137/PMC.S14459.Received.
53. Lewis,K. (2013) Platforms for antibiotic discovery. *Nat. Rev. Drug Discov.*, **12**, 371.
54. Fernandes,P. and Martens,E. (2017) Antibiotics in late clinical development. *Biochem. Pharmacol.*, **133**, 152–163.
55. Demain,A.L. (2009) Antibiotics : Natural Products Essential to Human Health. *Med. Res. Rev.*, **29**, 821–842.
56. Brown,E.D. and Wright,G.D. (2016) Antibacterial drug discovery in the resistance era. 10.1038/nature17042.
57. Wohlleben,W., Mast,Y., Stegmann,E. and Ziemert,N. (2016) Antibiotic drug discovery. *Microb. Biotechnol.*, **9**, 541–548.
58. Wright,G.D. (2017) Opportunities for natural products in 21 st century antibiotic discovery. *Nat. Prod. Rep.*, **34**, 694–701.

59. Woodruff,H.B. (2014) Selman A . Waksman , Winner of the 1952 Nobel Prize for Physiology or Medicine. **80**, 2–8.
60. WAKSMAN,S.A. and STARKEY,R.L. (1923) PARTIAL STERILIZATION OF SOIL, MICROBIOLOGICAL ACTIVITIES AND SOIL FERTILITY: III. *Soil Sci.*, **16**.
61. Chater,K.F. (2016) Recent advances in understanding Streptomyces. *F1000Research*, **5**, 1–16.
62. Tulp,M. and Bohlin,L. (2005) Rediscovery of known natural compounds: nuisance or goldmine? *Bioorg. Med. Chem.*, **13**, 5274–5282.
63. Kong,D.-X., Guo,M.-Y., Xiao,Z.-H., Chen,L.-L. and Zhang,H.-Y. (2011) Historical variation of structural novelty in a natural product library. *Chem. Biodivers.*, **8**, 1968–1977.
64. Jensen,P.R., Chavarria,K.L., Fenical,W., Moore,B.S. and Ziemert,N. (2014) Challenges and triumphs to genomics-based natural product discovery. *J. Ind. Microbiol. Biotechnol.*, **41**, 203–209.
65. Zhu,F., Ma,X.H., Qin,C., Tao,L., Liu,X., Shi,Z., Zhang,C.L., Tan,C.Y., Chen,Y.Z. and Jiang,Y.Y. (2012) Drug discovery prospect from untapped species: indications from approved natural product drugs. *PLoS One*, **7**, e39782.
66. Pye,C.R., Bertin,M.J., Lokey,R.S., Gerwick,W.H. and Linington,R.G. (2017) Retrospective analysis of natural products provides insights for future discovery trends. *Proc. Natl. Acad. Sci.*, **114**, 5601 LP-5606.
67. Rappe,M.S. and Giovannoni,S.J. (2003) The uncultured microbial majority. *Annu. Rev. Microbiol.*, **57**, 369–394.
68. Mast,Y., Guezguez,J., Handel,F. and Schinko,E. (2015) A Complex Signaling Cascade Governs Pristinamycin Biosynthesis in Streptomyces pristinaespiralis. *Appl. Environ. Microbiol.*, **81**, 6621–6636.
69. Sarkar,A., Funk,A.N., Scherlach,K., Horn,F., Schroeckh,V., Chankhamjon,P., Westermann,M., Roth,M., Brakhage,A.A., Hertweck,C., *et al.* (2012) Differential expression of silent polyketide biosynthesis gene clusters in chemostat cultures of Aspergillus nidulans. *J. Biotechnol.*, **160**, 64–71.
70. Nichols,D., Cahoon,N., Trakhtenberg,E.M., Pham,L., Mehta,A., Belanger,A., Kanigan,T., Lewis,K., Epstein,S.S. and Al,N.E.T. (2010) Use of Ichip for High-Throughput In Situ Cultivation of " Uncultivable " Microbial Species. *Appl. Environ. Microbiol.*, **76**, 2445–2450.
71. Ling,L.L., Schneider,T., Peoples,A.J., Spoering,A.L., Engels,I., Conlon,B.P., Mueller,A., Schäberle,T.F., Hughes,D.E., Epstein,S., *et al.* (2015) A new antibiotic kills pathogens without detectable resistance. *Nature*, **517**, 455.
72. Blunt,J.W., Copp,B.R., Keyzers,R.A., Munro,M.H.G. and Prinsep,M.R. (2016) Marine natural products. *Nat. Prod. Rep.*, **33**, 382–431.
73. Gerea,A.L., Branscum,K.M., King,J.B., You,J., Powell,D.R., Miller,A.N., Spear,J.R. and Cichewicz,R.H. (2012) Secondary metabolites produced by fungi derived from a microbial mat encountered in an iron-rich natural spring. *Tetrahedron Lett.*, **53**, 4202–4205.
74. Schmidt,E.W., Donia,M.S., McIntosh,J.A., Fricke,W.F. and Ravel,J. (2012) Origin and Variation of Tunicate Secondary Metabolites. *J. Nat. Prod.*, **75**, 295–304.
75. Schneemann,I., Nagel,K., Kajahn,I., Labes,A., Wiese,J. and Imhoff,J.F. (2010) Comprehensive Investigation of Marine Actinobacteria Associated with the Sponge Halichondria panicea. *Am. Soc. Microbiol.*, **76**, 3702–3714.
76. Wang,M., Carver,J.J., Phelan,V. V, Sanchez,L.M., Garg,N., Peng,Y., Nguyen,D.D.,

Watrous,J., Kapono,C.A., Luzzatto-Knaan,T., *et al.* (2016) Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nat. Biotechnol.*, **34**, 828–837.
77. Chater,K. (1999) David Hopwood and the emergence of Streptomyces genetics. *Int. Microbiol.*, **2**, 61–68.
78. Kirby,R., Wright,L.F. and Hopwood,D.A. (1975) Plasmid-determined antibiotic synthesis and resistance in Streptomyces coelicolor. *Nature*, **254**, 265–267.
79. Hopwood,D.A. (1997) Genetic Contributions to Understanding Polyketide Synthases. **2665**.
80. Fischbach,M.A. and Walsh,C.T. (2006) Assembly-Line Enzymology for Polyketide and Nonribosomal Peptide Antibiotics : Logic , Machinery , and Mechanisms. *Chem. Rev.*, **5**, 3468–3496.
81. Christopher T. Walsh, Robert V. OBrien,C.K. (2015) Nonproteinogenic Amino Acid Building Blocks for Nonribosomal Peptide and Hybrid Polyketide Scaffolds. *Angew Chem Int Ed Engl*, **52**, 7098–7124.
82. Du,L. and Shen,B. (2001) Biosynthesis of hybrid peptide-polyketide natural products. *Curr. Opin. Drug Discov. Devel.*, **4**, 215–228.
83. Xie,X., Kirby,J. and Keasling,J.D. (2012) Functional characterization of four sesquiterpene synthases from Ricinus communis (Castor bean). *Phytochemistry*, **78**, 20–28.
84. Keeling,C.I. (2008) Terpenoid biomaterials. *plant J.*, 10.1111/j.1365-313X.2008.03449.x.
85. Pazouki,L. and Niinemets,U. (2016) Multi-Substrate Terpene Synthases : Their Occurrence and Physiological Significance. *Front. Plant Sci.*, **7**, 1–16.
86. Arnison,P.G., Bibb,M.J., Bierbaum,G., Bowers,A.A., Bugni,T.S., Bulaj,G., Camarero,J.A., Campopiano,D.J., Challis,G.L., Clardy,J., *et al.* (2013) Ribosomally synthesized and post-translationally modified peptide natural products: overview and recommendations for a universal nomenclature. *Nat. Prod. Rep.*, **30**, 108–160.
87. Avci,F.Y., Li,X., Tsuji,M. and Kasper,D.L. (2013) Carbohydrates and T cells: A sweet twosome. *Semin. Immunol.*, **25**, 146–151.
88. Steffensky,M., Mühlenweg,A., Wang,Z.X., Li,S.M. and Heide,L. (2000) Identification of the novobiocin biosynthetic gene cluster of Streptomyces spheroides NCIB 11891. *Antimicrob. Agents Chemother.*, **44**, 1214–1222.
89. Schaberle,T.F., Vollmer,W., Frasch,H.-J., Huttel,S., Kulik,A., Rottgen,M., von Thaler,A.-K., Wohlleben,W. and Stegmann,E. (2011) Self-resistance and cell wall composition in the glycopeptide producer Amycolatopsis balhimycina. *Antimicrob. Agents Chemother.*, **55**, 4283–4289.
90. Kim,U.J., Shizuya,H., de Jong,P.J., Birren,B. and Simon,M.I. (1992) Stable propagation of cosmid sized human DNA inserts in an F factor based vector. *Nucleic Acids Res.*, **20**, 1083–1085.
91. Ahuja,M., Chiang,Y.-M., Chang,S.-L., Praseuth,M.B., Entwistle,R., Sanchez,J.F., Lo,H.-C., Yeh,H.-H., Oakley,B.R. and Wang,C.C.C. (2012) Illuminating the Diversity of Aromatic Polyketide Synthases in Aspergillus nidulans. *J. Am. Chem. Soc.*, **134**, 8212–8221.
92. Brady,S.F. and Clardy,J. (2005) Cloning and heterologous expression of isocyanide biosynthetic genes from environmental DNA. *Angew. Chem. Int. Ed. Engl.*, **44**, 7063–7065.
93. Courtois,S., Cappellano,C.M., Ball,M., Francou,F.-X., Normand,P., Helynck,G.,

Martinez,A., Kolvek,S.J., Hopke,J., Osburne,M.S., *et al.* (2003) Recombinant Environmental Libraries Provide Access to Microbial Diversity for Drug Discovery from Natural Products. *Appl. Environ. Microbiol.* , **69**, 49–55.
94. Debouck,C. and Goodfellow,P.N. (1999) DNA microarrays in drug discovery and development. **21**, 1–3.
95. Kouprina,N. and Larionov,V. (2016) Transformation-associated recombination (TAR) cloning for genomics studies and synthetic biology. *Chromosoma*, **125**, 621–632.
96. Awan,A.R., Shaw,W.M. and Ellis,T. (2016) Biosynthesis of therapeutic natural products using synthetic biology. *Adv. Drug Deliv. Rev.*, **105**, 96–106.
97. Döhren,H. Von, Dieckmann,R. and Pavela-vrancic,M. (1999) The nonribosomal code. *Chem. Biol.*, **6**, 273–279.
98. Rottig,M., Medema,M.H., Blin,K., Weber,T., Rausch,C. and Kohlbacher,O. (2011) NRPSpredictor2--a web server for predicting NRPS adenylation domain specificity. *Nucleic Acids Res.*, **39**, W362–W367.
99. Chevrette,M.G., Aicheler,F., Kohlbacher,O., Currie,C.R. and Medema,M.H. (2017) SANDPUMA: ensemble predictions of nonribosomal peptide chemistry reveal biosynthetic diversity across Actinobacteria. *Bioinformatics*, **33**, 3202–3210.
100. Pawlik,K., Kotowska,M., Chater,K.F., Kuczek,K. and Takano,E. (2007) A cryptic type I polyketide synthase (cpk) gene cluster in Streptomyces coelicolor A3(2). *Arch. Microbiol.*, **187**, 87–99.
101. Hadjithomas,M., Chen,I.-M.A., Chu,K., Ratner,A., Palaniappan,K., Szeto,E., Huang,J., Reddy,T.B.K., Cimermancic,P., Fischbach,M.A., *et al.* (2015) IMG-ABC: A Knowledge Base To Fuel Discovery of Biosynthetic Gene Clusters and Novel Secondary Metabolites. *MBio*, **6**, e00932.
102. Goodwin,S., Mcpherson,J.D. and Mccombie,W.R. (2016) Coming of age : ten years of next- generation sequencing technologies. *Nat. Publ. Gr.*, **17**, 333–351.
103. Simpson,J.T. and Pop,M. (2015) The Theory and Practice of Genome Sequence Assembly. *Annu. Rev. Genomics Hum. Genet.*, **16**, 153–172.
104. Alhakami,H., Mirebrahim,H. and Lonardi,S. (2017) A comparative evaluation of genome assembly reconciliation tools. *Genome Biol.*, **18**, 93.
105. Ziemert,N., Podell,S., Penn,K., Badger,J.H., Allen,E. and Jensen,P.R. (2012) The Natural Product Domain Seeker NaPDoS: A Phylogeny Based Bioinformatic Tool to Classify Secondary Metabolite Gene Diversity. *PLoS One*, **7**, e34064.
106. Huson,D.H., Auch,A.F., Qi,J. and Schuster,S.C. (2007) MEGAN analysis of metagenomic data. *Genome Res.*, 10.1101/gr.5969107.
107. Umemura,M., Koike,H., Nagano,N., Ishii,T., Kawano,J., Yamane,N., Kozone,I., Horimoto,K., Shin-ya,K., Asai,K., *et al.* (2013) MIDDAS-M: motif-independent de novo detection of secondary metabolite gene clusters through the integration of genome sequencing and transcriptome data. *PLoS One*, **8**, e84028.
108. Lee,T.A.J. Van Der and Medema,M.H. (2016) Computational strategies for genome-based natural product discovery and engineering in fungi. *Fungal Genet. Biol.*, **89**, 29–36.
109. Weber,T. and Kim,H.U. (2016) The secondary metabolite bioinformatics portal: Computational tools to facilitate synthetic biology of secondary metabolite production. *Synth. Syst. Biotechnol.*, **1**, 69–79.
110. Ziemert,N., Alanjary,M. and Weber,T. (2016) The evolution of genome mining in microbes – a review. *Nat. Prod. Rep.*, **33**, 988–1005.
111. Weber,T. (2014) In silico tools for the analysis of antibiotic biosynthetic pathways. *Int. J.*

Med. Microbiol., **304**, 230–235.
112. Blin,K., Wolf,T., Chevrette,M.G., Lu,X., Schwalen,C.J., Kautsar,S.A., Duran,H.G.S., Santos,E.L.C.D.L., Kim,U., Nave,M., *et al.* (2017) antiSMASH 4 . 0 — improvements in chemistry prediction and gene cluster boundary identification. **45**, 36–41.
113. Skinnider,M.A., Dejong,C.A., Rees,P.N., Johnston,C.W., Li,H., Webster,A.L.H., Wyatt,M.A. and Magarvey,N.A. (2015) Genomes to natural products PRediction Informatics for Secondary Metabolomes (PRISM). *Nucleic Acids Res.*, **9140**, gkv1012.
114. Khaldi,N., Seifuddin,F.T., Turner,G., Haft,D., Nierman,W.C., Wolfe,K.H. and Fedorova,N.D. (2010) SMURF: Genomic mapping of fungal secondary metabolite clusters. *Fungal Genet. Biol.*, **47**, 736–741.
115. Weber,T., Rausch,C., Lopez,P., Hoof,I., Gaykova,V., Huson,D.H. and Wohlleben,W. (2009) CLUSEAN: a computer-based framework for the automated analysis of bacterial secondary metabolite biosynthetic gene clusters. *J. Biotechnol.*, **140**, 13–17.
116. Finn,R.D., Clements,J. and Eddy,S.R. (2011) HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.*, **39**, W29-37.
117. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–10.
118. Cimermancic,P., Medema,M.H., Claesen,J., Kurita,K., Wieland Brown,L.C., Mavrommatis,K., Pati,A., Godfrey,P. a., Koehrsen,M., Clardy,J., *et al.* (2014) Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters. *Cell*, **158**, 412–421.
119. Spohn,M., Wohlleben,W. and Stegmann,E. (2016) Elucidation of the zinc-dependent regulation in Amycolatopsis japonicum enabled the identification of the ethylenediamine-disuccinate ([S,S]-EDDS) genes. *Environ. Microbiol.*, **18**, 1249–1263.
120. Spohn,M., Edenhart,S., Alanjary,M., Ziemert,N., Wibberg,D., Niedermeyer,T.H.J., Stegmann,E. and Wohlleben,W. (2018) Identification of a novel aminopolycarboxylic acid siderophore gene cluster encoding the biosynthesis of ethylenediaminesuccinic acid hydroxyarginine. *Metallomics*, 10.1039/C8MT00009C.
121. Cruz-Morales,P., Martínez-Guerrero,C.E., Morales-Escalante,M.A., Yáñez-Guerra,L.A., Kopp,J.F., Feldmann,J., Ramos-Aboites,H.E. and Barona-Gómez,F. (2015) Recapitulation of the evolution of biosynthetic gene clusters reveals hidden chemical diversity on bacterial genomes. *bioRxiv*, 10.1101/020503.
122. Umemura,M., Koike,H., Nagano,N., Ishii,T., Kawano,J., Yamane,N., Kozone,I., Horimoto,K., Shin-ya,K., Asai,K., *et al.* (2013) MIDDAS-M: Motif-Independent De Novo Detection of Secondary Metabolite Gene Clusters through the Integration of Genome Sequencing and Transcriptome Data. *PLoS One*, **8**, e84028.
123. Umemura,M., Koike,H. and Machida,M. (2015) Motif-independent de novo detection of secondary metabolite gene clusters—toward identification from filamentous fungi . *Front. Microbiol.* , **6**, 371.
124. Rajniak,J., Barco,B., Clay,N.K. and Sattely,E.S. (2015) A new cyanogenic metabolite in Arabidopsis required for inducible pathogen defence. *Nature*, **525**, 376.
125. Lau,W. and Sattely,E.S. (2015) Six enzymes from mayapple that complete the biosynthetic pathway to the etoposide aglycone. *Science (80-.).*, **349**, 1224 LP-1228.
126. Takeda,I., Umemura,M., Koike,H., Asai,K. and Machida,M. (2014) Motif-Independent Prediction of a Secondary Metabolism Gene Cluster Using Comparative Genomics: Application to Sequenced Genomes of Aspergillus and Ten Other Filamentous Fungal Species. *DNA Res.*, **21**, 447–457.
127. Terabayashi,Y., Sano,M., Yamane,N., Marui,J., Tamano,K., Sagara,J., Dohmoto,M.,

Oda,K., Ohshima,E., Tachibana,K., *et al.* (2010) Identification and characterization of genes responsible for biosynthesis of kojic acid, an industrially important compound from Aspergillus oryzae. *Fungal Genet. Biol.*, **47**, 953–961.
128. Wang,X.-J., Zhang,B., Yan,Y.-J., An,J., Zhang,J., Liu,C.-X. and Xiang,W.-S. (2013) Characterization and analysis of an industrial strain of Streptomyces bingchenggensis by genome sequencing and gene microarray. *Genome*, **56**, 677–689.
129. 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., *et al.* (2015) Minimum Information about a Biosynthetic Gene cluster. *Nat. Chem. Biol.*, **11**, 625–631.
130. Blin,K., Medema,M.H., Kottmann,R., Lee,S.Y. and Weber,T. (2017) The antiSMASH database, a comprehensive database of microbial secondary metabolite biosynthetic gene clusters. *Nucleic Acids Res.*, **45**, D555–D559.
131. Medema,M.H., Takano,E. and Breitling,R. (2013) Detecting sequence homology at the gene cluster level with MultiGeneBlast. *Mol. Biol. Evol.*, **30**, 1218–1223.
132. 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., *et al.* (2016) The Pfam protein families database: Towards a more sustainable future. *Nucleic Acids Res.*, **44**, D279–D285.
133. Ziemert,N., Alanjary,M. and Weber,T. (2016) The evolution of genome mining in microbes-a review. *Nat. Prod. Rep.*, **33**.
134. Tang,X., Li,J., Millán-Aguiñaga,N., Zhang,J.J., O'Neill,E.C., Ugalde,J.A., Jensen,P.R., Mantovani,S.M. and Moore,B.S. (2015) Identification of Thiotetronic Acid Antibiotic Biosynthetic Pathways by Target-directed Genome Mining. *ACS Chem. Biol.*, **10**, 2841–2849.
135. Dittmann,E., Gugger,M., Sivonen,K. and Fewer,D.P. (2015) Natural Product Biosynthetic Diversity and Comparative Genomics of the Cyanobacteria. *Trends Microbiol.*, **23**, 642–652.
136. Asmat,S., Shah,A., Akhter,N., Auckloo,B.N., Khan,I., Lu,Y., Wang,K., Wu,B. and Guo,Y. Structural Diversity , Biological Properties and Applications of Natural Products from Cyanobacteria . 10.3390/md15110354.
137. Herrmann,J., Fayad,A.A. and Muller,R. (2017) Natural products from myxobacteria: novel metabolites and bioactivities. *Nat. Prod. Rep.*, **34**, 135–160.
138. Hoffmann,T., Krug,D., Bozkurt,N., Duddela,S., Jansen,R., Garcia,R., Gerth,K., Steinmetz,H. and Müller,R. Correlating chemical diversity with taxonomic distance for discovery of natural products in myxobacteria. *Nat. Commun.*, 10.1038/s41467-018-03184-1.
139. Ikeda,H., Ishikawa,J., Hanamoto,A., Shinose,M., Kikuchi,H., Shiba,T., Sakaki,Y., Hattori,M. and Omura,S. (2003) Complete genome sequence and comparative analysis of the industrial microorganism Streptomyces avermitilis. *Nat. Biotechnol.*, **21**, 526–531.
140. Udwary,D.W., Zeigler,L., Asolkar,R.N., Singan,V., Lapidus,A., Fenical,W., Jensen,P.R. and Moore,B.S. (2007) Genome sequencing reveals complex secondary metabolome in the marine actinomycete Salinispora tropica. *Proc. Natl. Acad. Sci. U. S. A.*, **104**, 10376–10381.
141. Agrawal,P., Khater,S., Gupta,M., Sain,N. and Mohanty,D. (2017) RiPPMiner: a bioinformatics resource for deciphering chemical structures of RiPPs based on prediction of cleavage and cross-links. *Nucleic Acids Res.*, **45**, W80–W88.
142. Acland A, Agarwala R, Barrett T, Beck J, Benson DA, Bollin C, Bolton E, Bryant SH, Canese K, Church DM, Clark K, DiCuccio M, Dondoshansky I, Federhen S, Feolo M, Geer LY, Gorelenkov V, Hoeppner M, Johnson M, Kelly C, Khotomlianski V, Kimchi

A, Kimelman M,Z.K. (2013) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **41**, D8–D20.
143. Medema,M.H., Blin,K., Cimermancic,P., De Jager,V., Zakrzewski,P., Fischbach,M.A., Weber,T., Takano,E. and 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**, 339–346.
144. Christiansen,G., Fastner,J., Erhard,M., Börner,T. and Dittmann,E. (2003) Microcystin Biosynthesis in Planktothrix: Genes, Evolution, and Manipulation. *J. Bacteriol.*, **185**, 564–572.
145. Sarkar,A., Soueidan,H. and Nikolski,M. (2011) Identification of conserved gene clusters in multiple genomes based on synteny and homology. *BMC Bioinformatics*, **12**, S18–S18.
146. Lehmann,J., Stadler,P.F. and Prohaska,S.J. (2008) SynBlast: Assisting the analysis of conserved synteny information. *BMC Bioinformatics*, **9**, 351.
147. Farrer,R.A. (2017) Synima: a Synteny imaging tool for annotated genome assemblies. *BMC Bioinformatics*, **18**, 507.
148. Doroghazi,J.R. and Metcalf,W.W. (2013) Comparative genomics of actinomycetes with a focus on natural product biosynthetic genes. *BMC Genomics*, **14**, 611.
149. Lin,K., Zhu,L. and Zhang,D.Y. (2006) An initial strategy for comparing proteins at the domain architecture level. *Bioinformatics*, **22**, 2081–2086.
150. Yeong,M. (2016) BiG-SCAPE: exploring biosynthetic diversity through gene cluster similarity networks.
151. Monciardini,P., Iorio,M., Maffioli,S., Sosio,M. and Donadio,S. (2014) Discovering new bioactive molecules from microbial sources. *Microb. Biotechnol.*, **7**, 209–220.
152. Schorn,M.A., Alanjary,M.M., Aguinaldo,K., Korobeynikov,A., Podell,S., Patin,N., Lincecum,T., Jensen,P.R., Ziemert,N. and Moore,B.S. (2016) Sequencing rare marine actinomycete genomes reveals high density of unique natural product biosynthetic gene clusters. *Microbiol. (United Kingdom)*, **162**, 2075–2086.
153. Adamek,M., Alanjary,M., Sales-Ortells,H., Goodfellow,M., Bull,A.T., Winkler,A., Wibberg,D., Kalinowski,J. and Ziemert,N. (2018) Comparative genomics reveals phylogenetic distribution patterns of secondary metabolites in Amycolatopsis species. *BMC Genomics*, **19**, 426.
154. Bastian,M., Heymann,S. and Jacomy,M. (2009) Gephi: An Open Source Software for Exploring and Manipulating Networks. *Third Int. AAAI Conf. Weblogs Soc. Media*, 10.1136/qshc.2004.010033.
155. Hu,Y. (2005) Efficient and High Quality Force-Directed Graph Drawing. *Math. J.*, **10**, 37–71.
156. Oliphant,T.E. (2006) A guide to NumPy Trelgol Publishing.
157. Tuomisto,H. (2010) A consistent terminology for quantifying species diversity? Yes, it does exist. *Oecologia*, **164**, 853–860.
158. Lou,J. (2006) Entropy and diversity. *Oikos*, **113**, 363–375.
159. Tatusova,T., DiCuccio,M., Badretdin,A., Chetvernin,V., Nawrocki,E.P., Zaslavsky,L., Lomsadze,A., Pruitt,K.D., Borodovsky,M. and Ostell,J. (2016) NCBI prokaryotic genome annotation pipeline. *Nucleic Acids Res.*, **44**, 6614–6624.
160. Adamek,M., Spohn,M., Stegmann,E. and Ziemert,N. (2017) Mining Bacterial Genomes for Secondary Metabolite Gene Clusters. *Methods Mol. Biol.*, **1520**, 23–47.
161. Carver,T., Berriman,M., Tivey,A., Patel,C., Bohme,U., Barrell,B.G., Parkhill,J. and Rajandream,M.-A. (2008) Artemis and ACT: viewing, annotating and comparing sequences stored in a relational database. *Bioinformatics*, **24**, 2672–2676.

162. Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.
163. Tamura,K., Stecher,G., Peterson,D., Filipski,A. and Kumar,S. (2013) MEGA6: Molecular evolutionary genetics analysis version 6.0. *Mol. Biol. Evol.*, **30**, 2725–2729.
164. Richter,M., Rossello-Mora,R., Oliver Glockner,F. and Peplies,J. (2016) JSpeciesWS: a web server for prokaryotic species circumscription based on pairwise genome comparison. *Bioinformatics*, **32**, 929–931.
165. Hammer,Ø., Harper,D.A.T. and Ryan,P.D. (2001) PAST-Palaeontological statistics.
166. K.,C.R. and E.,E.J. (2014) EstimateS turns 20: statistical estimation of species richness and shared species from samples, with non-parametric extrapolation. *Ecography (Cop.).*, **37**, 609–613.
167. Su,G., Morris,J.H., Demchak,B. and Bader,G.D. (2014) BIOLOGICAL NETWORK EXPLORATION WITH CYTOSCAPE 3. *Curr. Protoc. Bioinformatics*, **47**, 8.13.1-8.13.24.
168. Dalisay,D.S., Williams,D.E., Wang,X.L., Centko,R., Chen,J. and Andersen,R.J. (2013) Marine Sediment-Derived Streptomyces Bacteria from British Columbia, Canada Are a Promising Microbiota Resource for the Discovery of Antimicrobial Natural Products. *PLoS One*, **8**, e77078.
169. van Dongen,S. and Abreu-Goodger,C. (2012) Using MCL to extract clusters from networks. *Methods Mol. Biol.*, **804**, 281–295.
170. Naman,C.B., Rattan,R., Nikoulina,S.E., Lee,J., Miller,B.W., Moss,N.A., Armstrong,L., Boudreau,P.D., Debonsi,H.M., Valeriote,F.A., *et al.* (2017) Integrating Molecular Networking and Biological Assays To Target the Isolation of a Cytotoxic Cyclic Octapeptide, Samoamide A, from an American Samoan Marine Cyanobacterium. *J. Nat. Prod.*, **80**, 625–633.
171. Bachmann,B.O., Van Lanen,S.G. and Baltz,R.H. (2014) Microbial genome mining for accelerated natural products discovery: is a renaissance in the making? *J. Ind. Microbiol. Biotechnol.*, **41**, 175–184.
172. Hou,Y., Braun,D.R., Michel,C.R., Klassen,J.L., Adnani,N., Wyche,T.P. and Bugni,T.S. (2012) Microbial Strain Prioritization Using Metabolomics Tools for the Discovery of Natural Products. *Anal. Chem.*, **84**, 4277–4283.
173. Jensen,P.R., Williams,P.G., Oh,D.-C., Zeigler,L. and Fenical,W. (2007) Species-specific secondary metabolite production in marine actinomycetes of the genus Salinispora. *Appl. Environ. Microbiol.*, **73**, 1146–1152.
174. Pidot,S.J., Coyne,S., Kloss,F. and Hertweck,C. (2014) Antibiotics from neglected bacterial sources. *Int. J. Med. Microbiol.*, **304**, 14–22.
175. Chai,Y., Shan,S., Weissman,K.J., Hu,S., Zhang,Y. and Müller,R. (2012) Heterologous Expression and Genetic Engineering of the Tubulysin Biosynthetic Gene Cluster Using Red/ET Recombineering and Inactivation Mutagenesis. *Chem. Biol.*, **19**, 361–371.
176. Davy,A.M., Kildegaard,H.F. and Andersen,M.R. (2017) Cell Factory Engineering. *Cell Syst.*, **4**, 262–275.
177. Bansal,M.S., Alm,E.J. and Kellis,M. (2012) Efficient algorithms for the reconciliation problem with gene duplication, horizontal transfer and loss. *Bioinformatics*, **28**, 283–291.
178. Bruns,H., Crüsemann,M., Letzel,A.-C., Alanjary,M., McInerney,J.O., Jensen,P.R., Schulz,S., Moore,B.S. and Ziemert,N. (2017) Function-related replacement of bacterial siderophore pathways. *ISME J.*, 10.1038/ismej.2017.137.

179. Oren,A. (2004) Prokaryote diversity and taxonomy: current status and future challenges. *Philos. Trans. R. Soc. B Biol. Sci.*, **359**, 623–638.
180. Garrity,G.M. (2016) A New Genomics-Driven Taxonomy of Bacteria and Archaea: Are We There Yet? *J. Clin. Microbiol.* , **54**, 1956–1963.
181. Woese,C.R., Kandler,O. and Wheelis,M.L. (1990) Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proc. Natl. Acad. Sci.*, **87**, 4576–4579.
182. Kim,M., Oh,H.S., Park,S.C. and Chun,J. (2014) Towards a taxonomic coherence between average nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of prokaryotes. *Int. J. Syst. Evol. Microbiol.*, **64**, 346–351.
183. Peplies,J., Kottmann,R., Ludwig,W. and Glöckner,F.O. (2008) A standard operating procedure for phylogenetic inference (SOPPI) using (rRNA) marker genes. *Syst. Appl. Microbiol.*, **31**, 251–257.
184. Louca,S., Doebeli,M. and Parfrey,L.W. (2018) Correcting for 16S rRNA gene copy numbers in microbiome surveys remains an unsolved problem. *Microbiome*, **6**, 1–12.
185. Quast,C., Pruesse,E., Yilmaz,P., Gerken,J., Schweer,T., Yarza,P., Peplies,J. and Glöckner,F.O. (2013) The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. *Nucleic Acids Res.*, **41**, 590–596.
186. DeSantis,T.Z., Hugenholtz,P., Larsen,N., Rojas,M., Brodie,E.L., Keller,K., Huber,T., Dalevi,D., Hu,P. and Andersen,G.L. (2006) Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl. Environ. Microbiol.*, **72**, 5069–5072.
187. Cole,J.R., Wang,Q., Fish,J.A., Chai,B., McGarrell,D.M., Sun,Y., Brown,C.T., Porras-Alfaro,A., Kuske,C.R. and Tiedje,J.M. (2014) Ribosomal Database Project: Data and tools for high throughput rRNA analysis. *Nucleic Acids Res.*, **42**, 633–642.
188. Conville,P.S. and Witebsky,F.G. (2007) Analysis of Multiple Differing Copies of the 16S rRNA Gene in Five Clinical Isolates and Three Type Strains of Nocardia Species and Implications for Species Assignment . *J. Clin. Microbiol.*, **45**, 1146–1151.
189. Yang,Z. and Rannala,B. (2012) Molecular phylogenetics: principles and practice. *Nat Rev*, **13**, 303–314.
190. Felsenstein,J. (University of W. (2004) Inferring Phylogenies Sinauer Associates, Inc, Sunderland, Massachusetts.
191. Zhou,X., Shen,X.X., Hittinger,C.T. and Rokas,A. (2018) Evaluating fast maximum likelihood-based phylogenetic programs using empirical phylogenomic data sets. *Mol. Biol. Evol.*, **35**, 486–503.
192. Nguyen,L.T., Schmidt,H.A., Von Haeseler,A. and Minh,B.Q. (2015) IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.*, **32**, 268–274.
193. Trifinopoulos,J., Nguyen,L.T., von Haeseler,A. and Minh,B.Q. (2016) W-IQ-TREE: a fast online phylogenetic tool for maximum likelihood analysis. *Nucleic Acids Res.*, **44**, W232–W235.
194. Stamatakis,A. (2006) RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, **22**, 2688–2690.
195. Kalyaanamoorthy,S., Minh,B.Q., Wong,T.K.F., Von Haeseler,A. and Jermiin,L.S. (2017) ModelFinder: Fast model selection for accurate phylogenetic estimates. *Nat. Methods*, **14**, 587–589.
196. Ripplinger,J. and Sullivan,J. (2008) Does choice in model selection affect maximum likelihood analysis? *Syst. Biol.*, **57**, 76–85.

197. Guo,Y.P., Zheng,W., Rong,X.Y. and Huang,Y. (2008) A multilocus phylogeny of the Streptomyces griseus 16S rRNA gene clade: Use of multilocus sequence analysis for streptomycete systematics. *Int. J. Syst. Evol. Microbiol.*, **58**, 149–159.
198. Simmons,M.P. and Gatesy,J. (2015) Coalescence vs. concatenation: Sophisticated analyses vs. first principles applied to rooting the angiosperms. *Mol. Phylogenet. Evol.*, **91**, 98–122.
199. Liu,L., Yu,L., Kubatko,L., Pearl,D.K. and Edwards,S. V (2009) Coalescent methods for estimating phylogenetic trees. *Mol. Phylogenet. Evol.*, **53**, 320–328.
200. Doroghazi,J.R. and Buckley,D.H. (2010) Widespread homologous recombination within and between Streptomyces species. *ISME J.*, **4**, 1136–1143.
201. Glaeser,S.P. and Kämpfer,P. (2015) Multilocus sequence analysis (MLSA) in prokaryotic taxonomy. *Syst. Appl. Microbiol.*, **38**, 237–245.
202. Jolley,K.A. and Maiden,M.C. (2010) BIGSdb: Scalable analysis of bacterial genome variation at the population level. *BMC Bioinformatics*, **11**, 595.
203. Blom,J., Kreis,J., Spanig,S., Juhre,T., Bertelli,C., Ernst,C. and Goesmann,A. (2016) EDGAR 2.0: an enhanced software platform for comparative gene content analyses. *Nucleic Acids Res.*, **44**, W22-8.
204. Gillespie,J.J., Wattam,A.R., Cammer,S.A., Gabbard,J.L., Shukla,M.P., Dalay,O., Driscoll,T., Hix,D., Mane,S.P., Mao,C., *et al.* (2011) PATRIC: the Comprehensive Bacterial Bioinformatics Resource with a Focus on Human Pathogenic Species . *Infect. Immun.*, **79**, 4286–4298.
205. Ondov,B.D., Treangen,T.J., Melsted,P., Mallonee,A.B., Bergman,N.H., Koren,S. and Phillippy,A.M. (2016) Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol.*, **17**, 132.
206. Capella-Gutiérrez,S., Silla-Martínez,J.M. and Gabaldón,T. (2009) trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, **25**, 1972–1973.
207. Pruitt,K.D., Brown,G.R., Hiatt,S.M., Thibaud-Nissen,F., Astashyn,A., Ermolaeva,O., Farrell,C.M., Hart,J., Landrum,M.J., McGarvey,K.M., *et al.* (2014) RefSeq: an update on mammalian reference sequences. *Nucleic Acids Res.*, **42**, D756-63.
208. Haft,D.H., Selengut,J.D. and White,O. (2003) The TIGRFAMs database of protein families. *Nucleic Acids Res.*, **31**, 371–373.
209. Suyama,M., Torrents,D. and Bork,P. (2006) PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res.*, **34**, W609-12.
210. Yang,Z. (2007) PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.*, **24**, 1586–1591.
211. Katoh,K. and Standley,D.M. (2013) MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.*, **30**, 772–780.
212. Group,S.C. and Group,S.C. (2011) Performance , Accuracy , and Web Server for Evolutionary Placement of Short Sequence Reads under Maximum Likelihood. **60**, 291–302.
213. Hoang,D.T., Chernomor,O., Von Haeseler,A., Minh,B.Q. and Vinh,L.S. (2018) UFBoot2: Improving the ultrafast bootstrap approximation. *Mol. Biol. Evol.*, **35**, 518–522.
214. Zhang,C., Rabiee,M., Sayyari,E. and Mirarab,S. (2018) ASTRAL-III: Polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinformatics*, **19**, 15–30.

215. Chernomor,O., Von Haeseler,A. and Minh,B.Q. (2016) Terrace Aware Data Structure for Phylogenomic Inference from Supermatrices. *Syst. Biol.*, **65**, 997–1008.
216. Van Belkum,A., Welker,M., Dunne,W.M. and Girard,V. (2015) The infallible microbial identification test: Does it exist? *J. Clin. Microbiol.*, **53**, 1786.
217. Huerta-Cepas,J., Serra,F. and Bork,P. (2016) ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Mol. Biol. Evol.*, **33**, 1635–1638.
218. Oliphant,T.E. (2007) Python for Scientific Computing. *Comput. Sci. Eng.*, **9**, 10–20.
219. Sayyari,E. and Mirarab,S. (2016) Fast Coalescent-Based Computation of Local Branch Support from Quartet Frequencies. *Mol. Biol. Evol.*, **33**, 1654–1668.
220. Huson,D.H. and Scornavacca,C. (2012) Dendroscope 3: an interactive tool for rooted phylogenetic trees and networks. *Syst. Biol.*, **61**, 1061–1067.
221. Degnan,J.H. and Rosenberg,N.A. (2009) Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends Ecol. Evol.*, **24**, 332–340.
222. Adamek,M., Alanjary,M., Sales-Ortells,H., Goodfellow,M., Bull,A.T., Winkler,A., Wibberg,D., Kalinowski,J. and Ziemert,N. (2018) Comparative genomics reveals phylogenetic distribution patterns of secondary metabolites in Amycolatopsis species. *BMC Genomics*, **19**.
223. Von Wintersdorff,C.J.H., Penders,J., Van Niekerk,J.M., Mills,N.D., Majumder,S., Van Alphen,L.B., Savelkoul,P.H.M. and Wolffs,P.F.G. (2016) Dissemination of antimicrobial resistance in microbial ecosystems through horizontal gene transfer. *Front. Microbiol.*, **7**, 1–10.
224. Dzidic,S. and Bedekovic,V. (2003) Horizontal gene transfer-emerging multidrug resistance in hospital bacteria. *Acta Pharmacol. Sin.*, **24**, 519–526.
225. Cragg,G.M. and Newman,D.J. (2013) Natural products: A continuing source of novel drug leads. *Biochim. Biophys. Acta - Gen. Subj.*, **1830**, 3670–3695.
226. Harvey,A.L. (2008) Natural products in drug discovery. *Drug Discov. Today*, **13**, 894–901.
227. Bérdy,J. (2005) Bioactive microbial metabolites. *J. Antibiot. (Tokyo).*, **58**, 1–26.
228. Katz,L. and Baltz,R. (2016) Natural product discovery: past, present, and future.
229. 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., *et al.* (2017) antiSMASH 4.0-improvements in chemistry prediction and gene cluster boundary identification. *Nucleic Acids Res.*, **45**, W36–W41.
230. Perry,J.A., Westman,E.L. and Wright,G.D. (2014) The antibiotic resistome: what's new? *Curr. Opin. Microbiol.*, **21**, 45–50.
231. Thaker,M.N., Waglechner,N. and Wright,G.D. (2014) Antibiotic resistance–mediated isolation of scaffold-specific natural product producers. *Nat. Protoc.*, **9**, 1469–1479.
232. Kale,A.J., McGlinchey,R.P., Lechner,A. and Moore,B.S. (2011) Bacterial self-resistance to the natural proteasome inhibitor salinosporamide A. *ACS Chem. Biol.*, **6**, 1257–64.
233. Freel,K.C., Millán-Aguiñaga,N. and Jensen,P.R. (2013) Multilocus sequence typing reveals evidence of homologous recombination linked to antibiotic resistance in the genus salinispora. *Appl. Environ. Microbiol.*, **79**, 5997–6005.
234. Parsons,J.B., Yao,J., Frank,M.W. and Rock,C.O. (2015) FabH mutations confer resistance to FabF-directed antibiotics in Staphylococcus aureus. *Antimicrob. Agents Chemother.*, **59**, 849–858.
235. Hayashi,T., Yamamoto,O., Sasaki,H., Kawaguchi,A. and Okazaki,H. (1983) Mechanism of action of the antibiotic thiolactomycin inhibition of fatty acid synthesis of Escherichia coli. *Biochem. Biophys. Res. Commun.*, **115**, 1108–1113.
236. Warnes,S.L., Highmore,C.J. and Keevil,C.W. (2012) Horizontal Transfer of Antibiotic

Resistance Genes on Abiotic Touch Surfaces: Implications for Public Health. *mBio* , **3**.
237. Lawrence,J.G. and Ochman,H. (2002) Reconciling the many faces of lateral gene transfer. *Trends Microbiol.*, **10**, 1–4.
238. Ravenhall,M., Škunca,N., Lassalle,F. and Dessimoz,C. (2015) Inferring Horizontal Gene Transfer. *PLOS Comput. Biol.*, **11**, e1004095.
239. Alanjary,M., Kronmiller,B., Adamek,M., Blin,K., Weber,T., Huson,D., Philmus,B. and 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**, W42–W48.
240. Cock,P.J.A., Antao,T., Chang,J.T., Chapman,B.A., Cox,C.J., Dalke,A., Friedberg,I., Hamelryck,T., Kauff,F., Wilczynski,B., *et al.* (2009) Biopython: Freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, **25**, 1422–1423.
241. Weber,T., Blin,K., Duddela,S., Krug,D., Kim,H.U., Bruccoleri,R., Lee,S.Y., Fischbach,M.A., Muller,R., Wohlleben,W., *et al.* (2015) antiSMASH 3.0-a comprehensive resource for the genome mining of biosynthetic gene clusters. *Nucleic Acids Res.*, **43**, W237-43.
242. 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., *et al.* (2013) The Comprehensive Antibiotic Resistance Database. *Antimicrob. Agents Chemother.*, **57**, 3348–3357.
243. Thai,Q.K., Bös,F. and Pleiss,J. (2009) The Lactamase Engineering Database: a critical survey of TEM sequences in public databases. *BMC Genomics*, **10**, 390.
244. Bush,K. and Jacoby,G.A. (2010) Updated functional classification of ??-lactamases. *Antimicrob. Agents Chemother.*, **54**, 969–976.
245. Gibson,M.K., Forsberg,K.J. and Dantas,G. (2014) Improved annotation of antibiotic resistance determinants reveals microbial resistomes cluster by ecology. 10.1038/ismej.2014.106.
246. O'Leary,N.A., Wright,M.W., Brister,J.R., Ciufo,S., Haddad,D., McVeigh,R., Rajput,B., Robbertse,B., Smith-White,B., Ako-Adjei,D., *et al.* (2016) Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.*, **44**, D733–D745.
247. Berger,S.A., Krompass,D. and Stamatakis,A. (2011) Performance, accuracy, and web server for evolutionary placement of short sequence reads under maximum likelihood. *Syst. Biol.*, **60**, 291–302.
248. Mirarab,S., Reaz,R., Bayzid,M.S., Zimmermann,T., S. Swenson,M. and Warnow,T. (2014) ASTRAL: Genome-scale coalescent-based species tree estimation. *Bioinformatics*, **30**, 541–548.
249. Luo,H., Lin,Y., Gao,F., Zhang,C.T. and Zhang,R. (2014) DEG 10, an update of the database of essential genes that includes both protein-coding genes and noncoding genomic elements. *Nucleic Acids Res.*, **42**, 574–580.
250. Huang,C.H., Hsiang,T. and Trevors,J.T. (2013) Comparative bacterial genomics: defining the minimal core genome. *Antonie Van Leeuwenhoek*, **103**, 385–398.
251. Christensen,H., Kuhnert,P., Olsen,J.E. and Bisgaard,M. (2004) Comparative phylogenies of the housekeeping genes atpD, infB and rpoB and the 16S rRNA gene within the Pasteurellaceae. *Int. J. Syst. Evol. Microbiol.*, **54**, 1601–1609.
252. Alam,M.T., Medema,M.H., Takano,E. and Breitling,R. (2011) Comparative genome-scale metabolic modeling of actinomycetes: The topology of essential core metabolism.

FEBS Lett., **585**, 2389–2394.
253. Schmutz,E., Mühlenweg,A., Li,S.M. and Heide,L. (2003) Resistance genes of aminocoumarin producers: Two type II topoisomerase genes confer resistance against coumermycin A1 and clorobiocin. *Antimicrob. Agents Chemother.*, **47**, 869–877.
254. Hashimi,S.M., Huang,G., Maxwell,A. and Birch,R.G. (2008) DNA gyrase from the albicidin producer Xanthomonas albilineans has multiple-antibiotic-resistance and unusual enzymatic properties. *Antimicrob. Agents Chemother.*, **52**, 1382–1390.
255. Sánchez-Hidalgo,M., Núñez,L.E., Méndez,C. and Salas,J.A. (2010) Involvement of the beta subunit of RNA polymerase in resistance to streptolydigin and streptovaricin in the producer organisms Streptomyces lydicus and streptomyces spectabilis. *Antimicrob. Agents Chemother.*, **54**, 1684–1692.
256. Floss,H.G. and Yu,T.-W. (2005) RifamycinMode of Action, Resistance, and Biosynthesis. *Chem. Rev.*, **105**, 621–632.
257. Ishikawa,J., Yamashita,A., Mikami,Y., Hoshino,Y., Kurita,H., Hotta,K., Shiba,T. and Hattori,M. (2004) The complete genomic sequence of Nocardia farcinica IFM 10152. *Proc. Natl. Acad. Sci. U. S. A.*, **101**, 14925–30.
258. Wieland Brown,L.C., Acker,M.G., Clardy,J., Walsh,C.T. and Fischbach,M. a (2009) Thirteen posttranslational modifications convert a 14-residue peptide into the antibiotic thiocillin. *Proc. Natl. Acad. Sci. U. S. A.*, **106**, 2549–2553.
259. Bibb,M.J., White,J., Ward,J.M. and Janssen,G.R. (1994) The mRNA for the 23S rRNA methylase encoded by the ermE gene of Saccharopolyspora erythraea is translated in the absence of a conventional ribosome-binding site. *Mol. Microbiol.*, **14**, 533–545.
260. Ryder,M.H., Slota,J.E., Scarim,A. and Farrand,S.K. (1987) Genetic analysis of agrocin 84 production and immunity in Agrobacterium spp. *J. Bacteriol.*, **169**, 4184–4189.
261. Marshall,C.G., Lessard,I.A.D., Park,I.S. and Wright,G.D. (1998) Glycopeptide antibiotic resistance genes in glycopeptide-producing organisms. *Antimicrob. Agents Chemother.*, **42**, 2215–2220.
262. Liras,P. (1999) Biosynthesis and molecular genetics of cephamycins. Cephamycins produced by actinomycetes. *Antonie Van Leeuwenhoek*, **75**, 109–124.
263. Sosio,M., Amati,G., Cappellano,C., Sarubbi,E., Monti,F. and Donadio,S. (1996) An elongation factor Tu (EF-Tu) resistant to the EF-Tu inhibitor GE2270 in the producing organism Planobispora rosea. *Mol. Microbiol.*, **22**, 43–51.
264. Xu,H., Huang,W., He,Q.-L., Zhao,Z.-X., Zhang,F., Wang,R., Kang,J. and Tang,G.-L. (2012) Self-resistance to an antitumor antibiotic: a DNA glycosylase triggers the base-excision repair system in yatakemycin biosynthesis. *Angew. Chem. Int. Ed. Engl.*, **51**, 10532–10536.
265. Wang,S., Liu,K., Xiao,L., Yang,L., Li,H., Zhang,F., Lei,L., Li,S., Feng,X., Li,A., *et al.* (2016) Characterization of a novel DNA glycosylase from *S. sahachiroi* involved in the reduction and repair of azinomycin B induced DNA damage. *Nucleic Acids Res.*, **44**, 187–197.
266. Mosbacher,T.G., Bechthold,A. and Schulz,G.E. (2005) Structure and function of the antibiotic resistance-mediating methyltransferase AviRb from Streptomyces viridochromogenes. *J. Mol. Biol.*, **345**, 535–545.
267. Mattheus,W., Masschelein,J., Gao,L.-J., Herdewijn,P., Landuyt,B., Volckaert,G. and Lavigne,R. (2010) The kalimantacin/batumin biosynthesis operon encodes a self-resistance isoform of the FabI bacterial target. *Chem. Biol.*, **17**, 1067–1071.
268. Kling,A., Lukat,P., Almeida,D. V, Bauer,A., Fontaine,E., Sordello,S., Zaburannyi,N., Herrmann,J., Wenzel,S.C., Konig,C., *et al.* (2015) Antibiotics. Targeting DnaN for

tuberculosis therapy using novel griselimycins. *Science*, **348**, 1106–1112.
269. Wang,Z.-X., Li,S.-M. and Heide,L. (2000) Identification of the Coumermycin A1Biosynthetic Gene Cluster of Streptomyces rishiriensisDSM 40489. *Antimicrob. Agents Chemother.*, **44**, 3040–3048.
270. Baumann,S., Herrmann,J., Raju,R., Steinmetz,H., Mohr,K.I., Huttel,S., Harmrolfs,K., Stadler,M. and Muller,R. (2014) Cystobactamids: myxobacterial topoisomerase inhibitors exhibiting potent antibacterial activity. *Angew. Chem. Int. Ed. Engl.*, **53**, 14605–14609.
271. Kim,C.-G., Lamichhane,J., Song,K.-I., Nguyen,V.D., Kim,D.-H., Jeong,T.-S., Kang,S.-H., Kim,K.-W., Maharjan,J., Hong,Y.-S., *et al.* (2008) Biosynthesis of rubradirin as an ansamycin antibiotic from Streptomyces achromogenes var. rubradiris NRRL3061. *Arch. Microbiol.*, **189**, 463–473.
272. Almutairi,M.M., Park,S.R., Rose,S., Hansen,D.A., Vázquez-Laslop,N., Douthwaite,S., Sherman,D.H. and Mankin,A.S. (2015) Resistance to ketolide antibiotics by coordinated expression of rRNA methyltransferases in a bacterial producer of natural ketolides. *Proc. Natl. Acad. Sci.*, 10.1073/pnas.1512090112.
273. Thomas,C.M., Hothersall,J., Willis,C.L. and Simpson,T.J. (2010) Resistance to and synthesis of the antibiotic mupirocin. *Nat. Rev. Microbiol.*, **8**, 281–289.
274. Olano,C., Wilkinson,B., Sanchez,C., Moss,S.J., Sheridan,R., Math,V., Weston,A.J., Brana,A.F., Martin,C.J., Oliynyk,M., *et al.* (2004) Biosynthesis of the angiogenesis inhibitor borrelidin by Streptomyces parvulus Tu4055: cluster analysis and assignment of functions. *Chem. Biol.*, **11**, 87–97.
275. Vecchione,J.J. and Sello,J.K. (2009) A novel tryptophanyl-tRNA synthetase gene confers high-level resistance to indolmycin. *Antimicrob. Agents Chemother.*, **53**, 3972–3980.
276. Schorn,M., Zettler,J., Noel,J.P., Dorrestein,P.C., Moore,B.S. and Kaysser,L. (2014) Genetic basis for the biosynthesis of the pharmaceutically important class of epoxyketone proteasome inhibitors. *ACS Chem. Biol.*, **9**, 301–309.
277. Peterson,R.M., Huang,T., Rudolf,J.D., Smanski,M.J. and Shen,B. (2014) Mechanisms of self-resistance in the Platensimycin- and platencin-producing streptomyces platensis MA7327 and MA7339 strains. *Chem. Biol.*, **21**, 389–397.
278. Liu,X., Fortin,P.D. and Walsh,C.T. (2008) Andrimid producers encode an acetyl-CoA carboxyltransferase subunit resistant to the action of the antibiotic. *Proc. Natl Acad. Sci. USA*, **105**, 13321–13326.
279. Jeong,H., Sung,S., Kwon,T., Seo,M., Caetano-Anollés,K., Choi,S.H., Cho,S., Nasir,A. and Kim,H. (2015) HGTree: database of horizontally transferred genes determined by tree reconciliation. *Nucleic Acids Res.*, **44**, D610-619.
280. Barka,E.A., Vatsa,P., Sanchez,L., Gaveau-Vaillant,N., Jacquard,C., Klenk,H.-P., Clément,C., Ouhdouch,Y. and van Wezel,G.P. (2016) Taxonomy, Physiology, and Natural Products of Actinobacteria. *Microbiol. Mol. Biol. Rev.*, **80**, 1–43.
281. Martin,D.P., Murrell,B., Golden,M., Khoosal, a. and Muhire,B. (2015) RDP4: Detection and analysis of recombination patterns in virus genomes. *Virus Evol.*, **1**, vev003-vev003.
282. Claeys,K.C., Fiorvento,A.D. and Rybak,M.J. (2014) A Review of Novel Combinations of Colistin and Lipopeptide or Glycopeptide Antibiotics for the Treatment of Multidrug-Resistant Acinetobacter baumannii. *Infect. Dis. Ther.*, **3**, 69–81.
283. Lin,D.M., Koskella,B. and Lin,H.C. (2017) Phage therapy: An alternative to antibiotics in the age of multi-drug resistance. *World J. Gastrointest. Pharmacol. Ther.*, **8**, 162–173.
284. Jiang,J., Gao,L., Bie,X., Lu,Z., Liu,H., Zhang,C., Lu,F. and Zhao,H. (2016)

Identification of novel surfactin derivatives from NRPS modification of Bacillus subtilis and its antifungal activity against Fusarium moniliforme. *BMC Microbiol.*, **16**, 1–14.
285. Gulder,T.A.M. and Moore,B.S. (2010) Salinosporamide Natural Products: Potent 20S Proteasome Inhibitors as Promising Cancer Chemotherapeutics. *Angew. Chem. Int. Ed. Engl.*, **49**, 9346–9367.
286. Ma,J., Prince,A. and Aagaard,K.M. (2014) Use of whole genome shotgun metagenomics: a practical guide for the microbiome-minded physician scientist. *Semin. Reprod. Med.*, **32**, 5–13.
287. Quince,C., Walker,A.W., Simpson,J.T., Loman,N.J. and Segata,N. (2017) Shotgun metagenomics, from sampling to analysis. *Nat. Biotechnol.*, **35**, 833.
288. Luo,Y., Enghiad,B. and Zhao,H. (2016) New tools for reconstruction and heterologous expression of natural product biosynthetic gene clusters. *Nat. Prod. Rep.*, **33**, 174–182.
289. Hughes,R.A. and Ellington,A.D. (2017) Synthetic DNA Synthesis and Assembly: Putting the Synthetic in Synthetic Biology. *Cold Spring Harb. Perspect. Biol.*, **9**.