A MULTISCALE ANALYSIS OF DIVERSITY AND GENETIC EXCHANGE IN BACTERIA

by

Bradon R. McDonald

A dissertation submitted in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy

(Microbiology)

at the

UNIVERSITY OF WISCONSIN–MADISON

2016

Date of final oral examination: 10/14/16

The dissertation is approved by the following members of the Final Oral Committee:

Cameron R. Currie, Ira L. Baldwin Professor of Bacteriology
Garret Suen, Assistant Professor of Bacteriology
Katherine D. McMahon, Professor of Bacteriology, Civil and Environmental Engineering
Heidi Goodrich-Blair, Professor of Bacteriology
Michael R. Sussman, Professor of Biochemistry

A MULTISCALE ANALYSIS OF DIVERSITY AND GENETIC EXCHANGE IN BACTERIA

Bradon R. McDonald

Under the supervision of Professor Cameron R. Currie
At the University of Wisconsin-Madison

Lateral gene transfer, the ability of organisms to acquire genetic material from another, non-parent organism, has changed the fundamental evolutionary framework of microbiology. Greater appreciation for its impact on microbial evolution has led to debates about how evolutionary novelty arises in microbial lineages, the usefulness of phylogenetics to understand their evolution, and the nature or even existence of species. All of these changes to evolutionary theory governing microbes are dependent on the rate and phylogenetic scale of factors enabling and limiting genetic exchange. In this dissertation I investigate the dynamics and evolutionary consequences of genetic exchange in bacterial lineages across large, intermediate, and small phylogenetic distances. In Chapter 1 I argue the importance of explicitly addressing the scale of evolutionary patterns and processes in microbiology. The scale of ecology and evolution makes them challenging to study, from their small size and rapid evolution to their vast diversity that evolved over the entire history of life on Earth. Experimental design, theoretical development, and data interpretation can all benefit from a deeper understanding of these scales. Chapter 2 provides an overview of long-term evolutionary dynamics in microbes, demonstrating the impact of contingency on the evolutionary trajectory of bacterial phyla. Genomes of these bacteria cluster by both metabolic potential and genome content dedicated to niche-specific functions, suggesting that barriers to LGT prevent disruption of overall phylum-specific genome characteristics.

In Chapter 3 I investigate the rate of lateral gene transfer in the genus *Streptomyces*, showing that the rate of successful gene acquisition and retention is low over evolutionary time. By placing these events in a temporal context via a molecular clock, I suggest a different interpretation of lateral gene transfer through the use of a scale-conscious approach. I also show that the observed rate of transfer is negatively correlated to the phylogenetic distance encompassed by a lineage. This is indicative of high neutral turnover of transferred genes, a key observation for both comparative analyses between lineages and inferring the potential evolutionary impact of lateral gene transfer. At a small phylogenetic scale, chapter 4 investigates the diversity of very closely related *Pseudonocardia* isolated from fungus growing ants. Here I show geographical and phylogenetic structuring of overall genome content, SNP diversity, and natural product biosynthesis potential. The distribution of these genes is likely impacted by vertical inheritance and selective pressure from the pathogenic fungus *Escovopsis*. I also show a limited amount of gene content diversity within these very closely related bacteria, despite the theoretical potential of lateral gene transfer to continuously introduce new gene families. Together, these studies provide insight into the impact of lateral gene transfer over a range of temporal scales in diverse microbial lineages.

ACKNOWLEDGMENTS

The work I present here would not be possible without the mentors and teachers who guided me on this path. In particular I would like to thank John McCutcheon, whose endless enthusiasm for science and patient mentorship provided me with both inspiration and a role model to follow early in my carrier. I am also indebted to Nancy Moran for giving me my first opportunity to work in a research lab. Thank you to Cameron Currie, who gave me the freedom to explore, and the guidance to finish, wide-ranging projects throughout my doctoral work. I would also like to thank my thesis committee for providing helpful suggestions and valuable discussions.

Graduate students and post-docs in the Currie lab also played an essential role in my success as a graduate student. Frank Aylward and Heidi Horn were an endless source of good discussions. Thanks to Jenny Bratburd for playing devils advocate for all of my ideas. Thank you also to Gina Lewin, who provided support and encouragement as we progressed through graduate school together.

Finally I would like to thank my parents, who raised me, supported me, and fostered my curiosity. Their encouragement and guidance have empowered me to explore the beauty of the biological world.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1

The Problem of Scale in Microbial Ecology and Evolution

Bradon R. McDonald and Cameron R. Currie.

1.1 Introduction

The astonishing diversity of microbial life on Earth has fascinated and perplexed biologists for hundreds of years. They play critical roles in large scale processes that affect all organisms on earth. Their activity drives global nutrient cycles, including the oxygenation of Earth's atmosphere. This byproduct of photosynthesis dramatically changed the evolutionary trajectory of life on earth, leading to a major shift in gene abundance from fermentative to oxidative phosphorylation related functions (David and Alm, 2011). The release of carbon dioxide by plant biomass degrading microbes (Book et al., 2016) plays a major role in the carbon cycle. Microbes are also intimately involved with the evolution and ecology of animals and plants. Multicellular organisms evolved in the constant presence of microbes, and their diverse interactions with microbes form a critical part of their biology (McFall-Ngai, 2015). These interactions range from the mutualisms found in sap feeding insects (Moran et al., 2008) and gut communities that enable herbivory (Hess, 2011) to devastating pathogens that have changed the course of human history (Bos et al., 2011). The

ubiquitous presence of microbes on and in eukaryotes means that even pathogens interact with both their eukaryotic host and its associated microbes as a community (Ahmer and Gunn, 2011).

Small ribosomal subunit RNA (16S) sequencing enabled taxonomic surveys of environmental microbes with unprecedented depth. This technique led to sequencing of environmental DNA from a wide range of habitats, from deep sea thermal vents (Moyer et al., 1994) and volcanic hot springs (Ferris et al., 1996; Hugenholtz et al., 1998) to agricultural soil (Jansen et al., 1996) and human skin (Gao et al., 2007; Grice et al., 2009). These studies revealed an entire new domain of life, the Archaea (Woese and Fox, 1977). It also enabled the first DNA-based phylogenetic trees encompassing all of life on Earth, including many micrbial phyla that had not been successfully cultured.

The rapid decrease in sequencing cost led to the sequencing of the first complete bacterial genome in 1995 (Fleischmann et al., 1995), enabling deeper analysis of microbial diversity than had been possible by laboratory experiments alone. Comparative studies between both closely and distantly related bacterial genomes greatly expanded the knowledge of evolutionary patterns in microbes. Genomic diversity within microbial species was found to be significantly higher than in eukaryotic species, which was attributed to frequent lateral gene transfer (LGT). Greater appreciation for the impact of LGT on microbial evolution has altered the entire evolutionary framework in microbes (Doolittle and Zhaxybayeva, 2009; Achtman and Wagner, 2008). Metagenomics has revolutionized the study of microbial communities, enabling investigation of community-level patterns of diversity and function across numerous environments (Handelsman, 2004). The massive amount of sequencing data currently being generated is used to infer a wide range of ecological and evolutionary

processes.

With rapidly growing datasets and investigation of a wide array of environmental functions and processes performed by microbial communities, there have been increasing efforts to integrate ecological and evolutionary theory into our understanding of microbes. 16S surveys and metagenomic sequencing have revealed biogeographical patterns in the distribution of taxonomic groups (Martiny et al., 2006), functions (Green et al., 2008), and the co-occurrence of microbial taxa (Ma et al., 2016). Seasonal variation in both marine (Gilbert et al., 2012) and soil (Copeland et al., 2015) systems provide insight into the dynamics of microbial communities during environmental fluctuations. Further experiments have assessed key properties of microbial communities, including resilience to disturbance and functional redundancy of community members (Allison and Martiny, 2008). These studies also inform crucial aspects of human health. Large scale surveys of human microbiome diversity in healthy (Huttenhower et al., 2012) and unhealthy (Turnbaugh et al., 2009) individuals has led to the development of models linking microbial community dynamics to human disease (Gilbert et al., 2016). Here, I discuss the influence of different temporal, spatial, and organizational scales on our understanding of biological processes.

1.2 Pattern, Process, and Scale in Ecology and Evolution

Ecological and evolutionary processes occur at different scales, as do the patterns that emerge due to these processes. In a seminal publication, Simon Levy provided a clear synthesis of the promise and challenges inherent in addressing ecological questions through an appreciation of scale (Levy, 1992). At the heart of this synthesis is the idea of "patchiness":

the distribution of organisms, resources, or interactions is often uneven. The degree of patchiness observed is heavily dependent on the scale of observation. At very small scales, the distribution of organisms or resources may appear stochastic. As the observation scale increases, consistent patterns often emerge. Finally, at the largest scales, smaller scale patches may no longer be visible, leading to the appearance of a uniform distribution. The ideal scale of observation for a particular phenomenon is therefore dependent on the question a study tries to answer and the nature of the phenomenon. The meaning of small, intermediate, and large scales is dependent on both the pattern being investigated and the process that generates it. Patterns often appear at larger scales than the processes that generate them.

As with all complex adaptive systems, biological entities are characterized by multiple levels of interaction with new emergent properties at progressively larger scales. At the local, population scale, genotypic diversity within populations provides the potential for natural selection. Population expansions and bottlenecks have significant impact on this standing diversity. Local interactions between organisms shape the evolutionary trajectory of all interacting partners, either directly through co-evolutionary interactions or indirectly via competition for resources. At the higher scale of species-level processes, niche specialization between related organisms and the evolutionary impact of sustained interactions drive adaptation and dictate longer term species divergence and diversification. By studying groups of species in a genus, the impacts of these longer term dynamics can be observed. These impacts include the rate of diversification, the diversity of ecological roles performed by closely related species, and the quantity and characteristics of interspecies interactions. With ever-increasing scale comes long term extinction and diversification

dynamics, evolutionary contingency of broader lineages, and the collective ecosystem properties generated by the interactions of many diverse organisms with their environment. Our ability to disentangle ecological patterns is greatly enhanced by posing questions, planning experiments, and interpreting data with a conscious appreciation of scale.

In microbial systems, the complexities resulting from differences in physical, temporal, and organizational scales play an even larger role. Understanding processes that occur at scales dramatically different from our normal experience always poses a challenge, and microbes interact and evolve at the smallest and largest extremes of any organisms on Earth. The evolutionary history of microbes spans the entire history of life, while multicellular organisms encompass only a small part. Microbes can reproduce, and therefore evolve, many times faster than macroorganisms. They can interact and disperse across extremely small physical scales (Vos et al., 2013), enabling them to exploit diverse niches in what appears to be uniform environments at the human scale. All of these factors make considerations of scale critical in microbial ecology and evolution.

The following sections provide several examples of approaching microbial ecology and evolution with an explicit focus on scale. Section 1.3 discusses the challenge of linking microbial taxonomic classification to their evolutionary diversity, while section 1.4 examines the issues surrounding species concepts in microbies. Section 1.5 describes the physical scale of microbial ecology and the complexities it causes, followed by a discussion of microbial communities in section 1.6. I conclude with a survey of emerging experimental and computational approaches that will enable more detailed examination of microbial ecology and evolution at extreme scales in section 1.7.

1.3 The history of microbial life - Phylogenetic diversity of microbes

The diversity of multicellular life spans approximately 600 million years of evolutionary history. Animals arose approximately 540 million years ago, followed 100 million years later by land plants. While staggering in its many forms, the diversity of multicellular life is dwarfed by microbes by orders of magnitude. Many of the major culturable bacterial phyla are estimated to have diverged over two billion years ago (Battistuzzi et al., 2004), and phyla entirely composed of microbes that have not been cultured make up the majority of bacterial lineages (Rinke et al., 2013). Each new technique for identifying microbial taxa has revealed greater and greater diversity. Classifying this incredible diversity, and placing the evolution of all life on earth in a single consistent framework, requires a clear appreciation of evolutionary timescale and a consistent way of classifying microbial and eukaryotic life.

Before the era of DNA sequencing, microbial diversity was often classified using the phenotypic characteristics available to microbiologists: morphology, cell wall structure, metabolic capabilities that could be tested in a laboratory, and interactions (generally pathogenic) with macro organisms. The characteristics for different microbes and their higher taxonomic groupings described in the classic Bergey's Manual of Systematic Bacteriology (Bergey, 1923) became the standard for identifying bacteria based on these properties. The combination of these phenotypic characterization techniques led to the delineation of approximately 5,000 bacterial species. The advent of DNA sequencing enabled the use of 16S sequencing, first developed by Carl Woese and colleagues, as a DNA based classifica-

tion method (Woese, 1987). This was found to largely recapitulate the taxonomic groups identified phenotypically if 97% 16S identity was used as the cutoff to determine whether two organisms were of the same species. Further DNA sequencing technology advances allowed the use of average nucleotide identity of sequenced genomes (ANI), which was found to largely recapitulate existing taxonomy with a cutoff of 94% ANI (Konstantinidis and Tiedje, 2005). Similar approaches, often using ANI of conserved housekeeping genes at various sequence thresholds, are the most frequently used metrics for assigning bacterial sequences into so-called "operational taxonomic units", or OTUs.

Our intuitive understanding of what a species or genus means in terms of diversity, and the processes and patterns we can expect to observe and study at this taxonomic scale, is largely based on what we see in animals and plants. These taxa can serve as useful benchmarks for understanding the scale of microbial diversity, as the relationship between macro-organism phylogenetic scale and the scale of eco-evolutionary processes that affect them is better understood. A large study of speciation rate across all plants and animals found that lineages split approximately every 2 million years, and suggested that this divergence occurs as a neutral process independent of ecological adaptations (Hedges et al., 2015). Although some evolutionary mechanisms differ in microbes, especially genetic exchange (discussed below), it is reasonable to use macro-organims species diversity within a given sequence similarity cutoff as an initial approximation for what we would expect to find in microbes as well.

However, trying to use the microbial taxonomy methods for drawing taxonomic boundaries in eukaryotes proves problematic. For example, applying the 97% identity cutoff to 18S sequences from animals results in the classification of humans and frogs as the

same species (McCallum and Maden, 1985). Although investigating tetrapods as a whole reveals a wealth of conserved characteristics, similar to what one would find in a microbial "species" delineated using the same metric, there are approximately 30,000 species of tetrapods. Because the phenotypic classification methods used for microbes were unable to differentiate subtle but ecologically relevant phenotypic differences, and the cutoffs for subsequent sequence-based metrics were chosen based on the existing taxonomy, a microbial species as currently defined taxonomically is not equivalent to an animal or plant species. This provides a clear example of the challenge posed by evolutionary scale in microbes.

The scale of microbial taxonomy does not align with that of macro-organisms, and therefore interpreting patterns of microbial diversity through the lens of equivalent macro-organism taxonomy can be misleading. Indeed, the amount of phylogenetic diversity that is grouped into a single species in bacteria would often be classified as an order or subphylum in animals. Even the stricter sequence identity cutoffs often used to identify OTUs in metagenomics studies still group organisms more broadly than most eukaryotic species. One percent divergence in core genes seems small relative to the diversity of microbial life on Earth, but actually indicates large absolute evolutionary distances due to their slow rate of evolution. For example, all mammals fall within 1% 18S identitiy of humans (Gonzalez and Schmickel, 1986). Therefore, eco-evolutionary processes or patterns that are expected to occur at the species or population level in eukaryotes may occur at finer phylogenetic scale than these sequenced-based approaches are able to distinguish in microbes.

Without a full appreciation for the huge evolutionary scale of even high percent identity

OTUs, our interpretation of microbial diversity and evolution can be easily misled. A recent commentary on 16S OTU diversity across many environments (Amann and Rosselló-Móra, 2016) suggested that the number of bacterial species may only be in the millions based on rarefaction analysis of OTU discovery (Schloss et al., 2016). For comparison, all of the estimated 5-60 million species of animals, from humans to sponges, would be one OTU at 92% 18S identity. The microbial diversity survey found 15,743 microbial clusters at 90% identity. Thus, if microbial lineages diversify approximately at the same rate per small ribosomal subunit RNA divergence as animals, one would predict between 78-944 billion species of microbes. In order for the number of microbial species to be in the millions, microbes would have to diversify more than four orders of magnitude more slowly than animals. Although sequence divergence metrics are the only available tool to assess the majority of microbial diversity, this diversity should be placed in context with better characterized groups of organisms to avoid over-interpretation. Many metagenomic analyses use the ANI whole genome identity approach to identify OTUs instead of 16S divergence. Even at this finer scale, organisms grouped into a single OTU via strict sequence identity cutoffs have been show to be paraphyletic and ecologically heterogeneous (Koeppel and Wu, 2013).

The disconnect between our intuitive understanding of the amount of diversity at a particular taxonomic level and the breadth of diversity encompassed in microbial groups at that level becomes even greater when looking at broader taxonomic groups. For example, molecular clock analysis of the genus *Streptomyces* suggests it is approximately 380 million years old, as old as land vertebrates (See Chapter 3). The estimated divergence time of *Escherichia* and *Salmonella*, considered to be very closely related microbial genera, is 100

million years (Ochman and Wilson, 1987). This divergence date indicates these genera are significantly older than placental mammals, an animal group containing 4,000 species. They also display 98% 16S identity, meaning that 1% 16S sequence divergence translates to 50 million years of time for these organisms. Statements about organisms like *Salmonella* and *Escherichia* with high core gene percent identity being very closely related need to be put in proper context. They are relatively closely related given the diversity of bacteria, but they are very distantly related in terms of generations or years. Few researchers would use the terms "very closely related" to describe two eukaryotic organisms separated by 100 million years. Additionally, microbial groups at a single taxonomic level can vary significantly from each other in terms of age or diversity. Without taking this into account, evolutionary patterns that vary between groups of the same taxonomic level may be attributed to biological differences rather than differences in scale. *Streptomyces* are estimated to be nearly four times older than the divergence of *Salmonella* and *Escherichia* for example. A direct comparison between genera may show more gene content diversity in *Streptomyces*, due to their greater age alone. If age is not taken into account however, a study could easily conclude that greater diversity in *Streptomyces* is due to more freqent gene acquisition rather than simply a difference in evolutionary timespan.

Understanding the rate of evolutionary processes in microbes, including phenotypic evolution, gene content differences and genetic exchange, lineage diversification and extinction, depends on a clear appreciation for the evolutionary scale at which we measure processes and patterns. For example the calculated rate of point mutations, even in synonymous sites, is influenced by the phylogenetic distance between organisms in a dataset (see Chapter 3). Mutations that are selected against are lost over evolutionary time, such

that older microbial groups may have lower observed mutation rates. Normalization of observed mutation rate by phylogenetic distance will enable better quantitative comparisons of mutation rates between microbial groups of different ages.

1.4 Units of evolution - Concepts and definitions of species and populations

One of the most contentious issues in microbial ecology and evolution revolves around the smallest but arguably most important phylogenetic levels: species and populations (Doolittle and Zhaxybayeva, 2009; Gogarten et al., 2002; Achtman and Wagner, 2008). It is critically important to understand what diversity of organisms comprise these units, because they are not merely taxonomic levels. Populations are the unit of natural selection and species are the smallest evolutionarily independent lineages (de Queiroz, 2005). Therefore, labeling a particular group of organisms as a population or species is making a statement about the biological processes they undergo, not just their taxonomy. Further, accurately identifying them is important for investigating the processes that affect them and the patterns that these processes create at the proper scale. One process in particular has generated the most controversy related to species and populations: genetic exchange (Nowell et al., 2014; Cordero and Hogeweg, 2009; Ge et al., 2005). This process includes both exchange and recombination of homologous sequences, analogous to what occurs in sexually reproducing eukaryotes (homologous recombination), and acquisition of novel DNA fragments from other organisms (lateral gene transfer, or LGT).

Both of these processes are distinct from what generally occurs in eukaryotes. Homolo-

gous recombination in microbes occurs not across whole genomes but in fragments, so that two organisms may recombine across only one part of their genome while being genetically isolated in another (Shapiro et al., 2012; Vetsigian and Goldenfeld, 2005). Lateral gene transfer, baring a few exceptions (Keeling and Palmer, 2008), does not occur in eukaryotes. The comparison of the first three sequenced *E. coli* genomes revealed that only 40% of genes where shared by all the genomes (Welch et al., 2002). This vast gene content diversity within the same taxonomic species revolutionized evolutionary microbiology. Is implications have led some to challenge fundamental aspects of microbial evolution, including the validity of evolutionary trees of organisms and the existence of species or populations in microbes (Syvanen, 2012).

This interpretation of genetic exchange and its impact on microbial evolution is largely based on the impression that it occurs very frequently with limited phylogenetic boundaries. Both of these observations are dependent on the scale of analysis, and further research has revealed a much more nuanced picture of genetic exchange. Homologous recombination is more promiscuous in microbes than in sexually reproducing eukaryotes, in large part because it is easier for exogenous DNA to interact with a microbial genome. The likelihood of successful recombination depends on the percent nucleotide similarity between donor and recipient, such that homologous recombination rate decreases logarithmically as sequence similarity decreases (Vulić et al., 1997). Some microbial taxa undergo homologous recombination much more than others (Vos and Didelot, 2008). Indeed, analysis of the two widespread lineages of *Listeria monocytogenes* found that the observed recombination rate in one lineage was six times higher than the other (Bakker et al., 2008). A number of large scale analyses have compared recombination rates across a range of bacterial

groups, but they generally treat different taxonomic groups of bacteria as equal in terms of phylogenetic depth. In general this would lead to under-estimates of recombination rate for older lineages due to loss of acquired alleles, either due to selection against recombinants or neutral processes.

Although barriers to genetic exchange are often seen as the gold standard for distinguishing species in multicellular eukaryotes, the difficulty of defining species boundaries remains a continuing issue in these organisms as well (de Queiroz, 2005, 2007). Many other criteria have been advocated as metrics for defining species boundaries including morphology, ecological differences, and phylogenetic clustering. The preferred metric often varies due to the limits of data available in different fields: for example an archeologist cannot determine the potential for inbreeding between specimens. In an attempt to bridge the gap between disparate species definitions, de Quieroz made a key distinction between a species concept, which is the evolutionary properties a group of organisms in a species are to have, and a species definition, which are the empirical properties used to distinguish organisms in one species versus another. He argues that the many disagreements over ways of classifying species in eukaryotes are debates over a preferred definition, all of which are trying to identify organisms that fit under the same theoretical species concept. He states this unifying species concept as "independently evolving metapopulation lineages"; a species is comprised of a group of populations that evolve independent of other populations through time. Divergence of one species into two is a gradual process, and different definitions are able to distinguish newly independent lineages at different time points as this process unfolds. Ecological differences due to changes in behavior may be apparent earlier in the divergence process for example, with morphological differences appearing later. Because

of this, the use of different species definitions may lead to disagreements about species boundaries for newly diverged lineages, but the majority of definitions attempt to identify groups of organisms that fit the same species concept.

There are several key aspects that determine how disruptive the effects of genetic exchange would be to the independently evolving metapopulation species concept in microbes. They largely revolve around the evolutionary and temporal scales of genetic exchange and divergence. Unlike sexually reproducing eukaryotes, even microbes with "frequent" homologous recombination are unlikely to recombine in every generation, and recombination does not encompass the entire genome. Therefore, the rate of recombination in microbes is more akin to migration of alleles from a distant population than sexual reproduction within a population. The rate of allelic migration per generation and the selective effect (positive, negative, or neutral) of the incoming genetic material are critical parameters that determine whether the evolutionary trajectory of a population will be altered by genetic exchange from another source (Cohan, 1994). Epistasis and differences in selective pressures for different populations may often lead to microbes with hybrid genotypes having lower fitness. This is analogous to eukaryotic hybrid sterility. Such microbial populations could continue to evolve independently despite gene flow between them at loci not under positive selection in either population. Due to this, they would still fit the de Quieroz species concept.

Lateral gene transfer represents an even larger departure from normal evolutionary dynamics in eukaryotes, because it involves the acquisition of completely novel genetic material and can occur across much longer phylogenetic distances. Although there are some examples of LGT occurring between distantly related organisms (Boucher et al., 2003),

indeed between bacteria and archaea or eukaryotes (Hilario and Gogarten, 1993), LGT events do not occur at random. So-called highways of LGT were identified relatively early in the genomics era, with some taxa more likely to acquire genes from specific groups (Beiko et al., 2005). Further research has demonstrated that organisms in the same environment (Smillie et al., 2011), or from the same phylogenetic group (Andam and Gogarten, 2011), are more likely to exchange genes. Intensive study of LGT in *E. coli* has shown high toxicity of many genes engineered into *E. coli* regardless of phylogenetic distance to the donor (Sorek et al., 2007). Combined analysis of protein production and gene expression in *E. coli* found that genes thought to be laterally acquired were rarely translated, despite being transcribed as much as vertically inherited genes (Taoka et al., 2004). This indicates that there are multiple potential barriers to successful acquisition and utilization of genes through LGT, and that the presence of a laterally acquired gene in a genome does not indicate selective benefit. Epistasis also plays a role in structuring LGT, as some genes are unlikely to be acquired unless the recipient genome already contains particular metabolic functions (Press et al., 2016). These patterns of bias appears at a wide range of phylogenetic scales, but further sampling is needed to identify the smaller scales at which LGT may occur in an unbiased manner. Unbiased LGT may serve as an indicator of species or population level grouping of microbes.

The frequency and selective effect of LGT also determines its impact on aspects of evolutionary theory, particularly evolutionary independence and the validity phylogenetic trees. Although the absolute number of LGT events identified in microbial genera may number in the thousands, the evolutionary impact of these transfers is predicated on their frequency relative to the generation time of the microbes and their phenotypic effect on

the recipient. As demonstrated above, percent sequence divergence of core genes is a very coarse metric for measuring evolutionary distance, and a single percent divergence may represent millions of years of evolution. Unfortunately, LGT is more difficult to study under realistic conditions in the lab than homologous recombination or point mutations. Naturally competent organisms may take up DNA in culture, but the amount of this DNA in the environment is unclear. The diversity of plasmids, phages, and other mobile genetic elements that facilitate LGT in nature is very difficult to replicate in the lab. Thus, transfer rates are generally calculated using comparative genomic approaches, often using sequence divergence rather than time or generations as the metric of evolutionary distance (Vos et al., 2015). Comparative approaches face an additional challenge, in that the high rate of turnover for laterally acquired genes makes it difficult to get accurate estimates of the neutral transfer rate (Lerat et al., 2005; Hao and Brian Golding, 2006). A better estimate of neutral gene acquisition and turnover rate would provide a basis for comparing observed gene distributions against a null model.

1.5 The world through a microbe's eye - Biogeography and environmental complexity

The influence of spatial distribution on the ecology and evolution of organisms has been a central pillar of ecology for decades, but it is perhaps the least understood aspect of microbial ecology (Martiny et al., 2006; Green and Bohannan, 2006; Green et al., 2008). Part of the difficulty in describing biogeography in microbes has its roots in the phylogenetic scale challenges mentioned above. Observed biogeographical patterns for groups of organisms

are highly dependent on both the geographic scale of observation and the phylogenetic scale of the organisms under study. For example, while individual species of mice may have patchy distributions across a continent, the distribution of rodents is much more uniform. Therefore characterizing the distribution of a microbial species must take into account both the size of the geographic area and the fact that microbial species are defined very broadly relative to their eukaryotic counterparts. Biogeographical patterns are likely to emerge at smaller phylogenetic scales.

A recent biogeography study of *Streptomyces* illustrates how changing phylogenetic scale can reveal previously obscured patterns (Andam et al., 2016). Isolates of these bacteria can be found across a wide geographic range, without a clear pattern of spatial distribution. Detailed analysis of many isolates from this species in the northern United States demonstrated a clear biogeographical pattern, including evidence of a migration northward following the retreating glaciers during the last ice age. Thus, not only do new patterns emerge, but they provide insight into the influence of large scale global change on the diversity of soil microbes. The other major challenge for studying biogeography in microbes is their most obvious trait: their small physical size. Because of their small size, what appears to be a homogenous environment at the scale observable through conventional methods can be a heterogeneous environment at the micro scale with patchy distributions of nutrients, moisture, and other organisms (Vos et al., 2013). This micro-scale, spatially structured environmental heterogeneity can support a vast diversity of microbes (MacLean, 2005). Classic studies using media left stationary on a benchtop have shown that a gradient of a single resource, oxygen, can lead to phenotypic divergence and the evolution of multiple ecological strategies on short timescales in *Pseudomonas* (Rainey and

Travisano, 1998). These same patterns are repeated on a vastly more complex scale in real world environments such as soil or marine environments.

The complex, heterogeneous structure of an environment such as soil has significant impacts on the microbial community living there. Using microscopy to inform further mathematical modeling, Raynaud and Nunan argued that individual bacterial cells are actually relatively isolated (Raynaud and Nunan, 2014). The average bacterium in their model only has, on average, 120 other bacteria within a 20um radius, which is 10 times the length of an *E.coli* cell. Therefore, despite the high estimated number of microbial cells per gram of soil (10^8), a particular organism only interacts with a fraction of this diversity directly. They also note that cell density and inter-cell distance is highly variable, such that some bacteria may have many interacting partners and others very few. Two organisms may be found in the same environment and sequenced in the same metagenome, and yet very rarely interact in nature. Interactions with more distant bacteria can occur indirectly through diffusion, but these interactions are also governed by spatial structure. At the bacterial scale, soil is a vast network of individual soil particles linked together by small water pockets and films (beautifully illustrated in Figure 2 of Vos et al. (2013)). Nutrients, migrating bacteria, phage, and predators all travel along these water films. Therefore, soil hydration becomes a critical component for driving interactions between bacterial microcolonies. Both modeling (Long and Or, 2005, 2009) and experimental soil microcosms (Treves et al., 2003) have shown that drier soils, i.e. soils where the soil particles and their associated microbes are more isolated from each other (Or et al., 2007), sustain more diversity by preventing direct competition between organisms. Limited diffusion also reduces potential competitors for public goods, facilitates quorum sensing, and leads to

hiFgher concentrations of excreted compounds and toxins than would be predicted if soil were a well-mixed environment. Even on a 2D surface, spatial structure of antibiotic producers and degraders combined with diffusion can lead to high microbial diversity (Kelsic et al., 2015).

Treating soil as a vast collection of individual particles linked by water films provides a direct analogy between these environments and aquatic ones. Marine and fresh water environments, from a microbe's perspective, are essentially the opposite extreme from dry soils: many widely dispersed particles linked by nearly universal water film connections. Some microbes spend most of their time swimming or floating in the large spaces between particles, while others colonize the particles directly and disperse through the water. The heterogeneity of these environments also enables closely related organisms to use different ecological strategies. In a series of papers, incipient speciation driven by ecological specialization has been identified in the marine bacterium *Vibrio cyclitrophicus* (Shapiro et al., 2012; Yawata et al., 2014). While one lineage is predominantly found on very small particles or free floating, the other is primarily found attached to particles. After many experiments showed no phenotypic difference between the two lineages, experiments utilizing microfluidic devices showed that the particle-associated strains were much slower to detach from a surface and chemotax towards a new nutrient source, but better at colonizing and forming biofilms on a surface once attached. Micro-scale heterogeneity enables the maintenance of ecological diversity in marine environments as well as terrestrial soil. Being able to investigate environmental at this small scale will increase our understanding of the forces driving and sustaining microbial diversity.

1.6 Microbiomes - Community ecology and interspecies interactions

Ultimately, the objective of many microbial ecology research programs is to understand the factors influencing the structure of microbial communities, and the effect that community structure has on community function. Addressing these questions requires a scale-conscious perspective on interactions between species and their broader effects on the community. All three of the previously mentioned aspects of microbial ecology and evolution are important components in these interactions.

Many of the interactions between microbes that mediate competition and competitive exclusion are driven by phenotypic traits that differ at small phylogenetic scales. These include the production of secondary metabolites that can have diverse effects from growth inhibition and signaling to protection from osmotic stress and iron sequestration. Predators such as phages also generally have narrow host ranges (Paez-Espino et al., 2016), sometimes at the sub-strain level. Polymorphic toxins (Zhang et al., 2012) and bacteriocins (Cotter et al., 2013) generally act to inhibit the growth of bacteria that are closely related to the organism that carries them. Since many of these functions are of interest for therapeutic purposes, understanding their evolution and diversification among closely related bacteria can provide new opportunities for drug discovery.

Research into the human gut microbiome provides an excellent example of the benefits of studying community interactions at a smaller scale. TMA, a byproduct of choline metabolism produced by some bacteria, is converted to TMAO by the liver. TMAO concentrations in the blood are correlated with atherosclerosis risk. Romano et al. (2015)

found strains from both the Firmicutes and Proteobacteria that are capable of metabolizing choline, but that strains from the same species vary in this trait. They also found a significant increase in TMAO levels in the blood even when TMA producing bacteria make up a very small proportion of the gut microbial community. Genus or even species level classification of microbial community composition would therefore be uninformative for this medically important microbial trait. Further, finding very low abundance of the relevant metabolic pathway would not indicate that the effect on the host is inconsequential. By investigating the variation in activity of very closely related bacteria and their effect on the host in the lab, studies such as this can inform metagenomic analyses and greatly improve our ability to infer biology from sequencing data.

One of the most cited correlations in human gut microbiome studies is the correlation between the relative amount of Firmicutes and Bacteriodes and obesity. However, more recent meta-analyses of raised questions about the statistical significance of this correlation. By reanalyzing ten studies and applying a variety of statistical tests, Sze and Schloss (2016) found no correlation between obesity and the relative abundance of Firmicutes, Bacteroidetes, or the ratio of the two. Although they did find some correlations with overall community diversity, they show that few studies had significant statistical power to identify such weak effect sizes. In fact the authors point out that many ecologists question the relevance of correlating broad diversity metrics with ecosystem-level attributes, given how little we understand the processes that generate these patterns. As one alternative explanation, they suggest that 16S-based taxonomic information may not provide sufficient information to understand microbiome involvement in obesity. Instead they suggest gene expression or metabolite production may signal microbe-host interactions related to obesity,

both of which are finer scale measures of microbial community activity.

1.7 Approaches for addressing scale in microbiology

Just as emerging sequencing technologies have enabled the explosion of microbial ecology research, the continued development of both computational and laboratory tools provide new opportunities for overcoming the challenges posed by scale in microbes. Single-cell sequencing (Blainey, 2013; Gawad et al., 2016) can provide information on the genes contained in a single genome in an environment, which is critical information lost using metagenomics approaches. Bacteria with very high core gene similarity can have different genome content or particular mutations, such that the full complement of genes found in composite genomes from metagenomics data may not exist in any actual microbe. Single cell sequencing can provide a wealth of information about the genomic potential of individual microbes in a community. It also enables the use of population genetic tools for assigning organisms to clusters and investigating processes that occur at the population scale, including changes in population size and selection on particular genes. This information presents a clearer picture of both a single organism's functional capabilities and fine scale diversity within a community.

Other methods will enable researchers to go beyond genomic potential and investigate the activity of microbes in a particular environment. Imaging mass-spectrometry can reveal the suite of molecules present in a sample at very small scale (Stubbendieck et al., 2016). This technique has a wide range of applications, from protein detection (Stoeckli et al., 2001) to studying pairwise interactions and microbial communities.(Watrous and Dorrestein,

2011). By providing an image of the sample with detected chemical species, it provides spatial as well as concentration data. This enables detection and quantification of protein or small molecule production by individual cells in a community without requiring genetic manipulation to add antibody-detectible tags or the creation of monoclonal antibodies to a particular molecule.

Single-cell transcriptomics also provides the opportunity to analyze microbial activity in real time. Organisms with the genomic potential for particular functions may nevertheless fail to perform these functions under many conditions, with important consequences for a microbial community. Several examples include the unknown conditions required to induce expression of most secondary metabolism biosynthetic clusters (Scherlach and Hertweck, 2009) and the lack of transcriptional response to the presence of cellulose in many strains of *Streptomyces* despite the presence of cellulases in their genome (Book et al., 2016). By assessing gene expression in individual cells which may experience different conditions in a patchy environment, we can gain a deeper understanding of phenotypic variation of microbes with similar genomic potential.

Perhaps the greatest potential for understanding the microbial world rests in the ability to perform experiments and manipulations at their scale. Microfluidics devices allow control of physical structure and chemical concentration gradients at the scale of individual cells or microcolonies, and these tools have already revealed new insights into a wide range of bacterial properties, from basic cell growth and physiology to motility and interspecies interactions. It has also been shown that *E.coli* are chemotactically sensitive to nanomolar concentrations of amino acids, a thousand times more sensitive than previously thought (Mao et al., 2003). Analyses of marine microbes have shown that they are able to respond

an order of magnitude faster than *E.coli* to chemoattracants, perhaps due to the selective pressure for pursuing short lived resource patches in the ocean (Stocker et al., 2008). Using microfluidics devices as tiny chemostats, *E.coli* have been shown to self-organize into multicell arrangements that enhance nutrient access, and therefore enable dense cell growth (Cho et al., 2007). Quorum sensing has been shown to occur with only a few cells in confined spaces, prompting reconsideration of its function in natural environments that contain small pockets with low diffusion rates (Boedicker et al., 2009). Further experiments have investigated nutrient competition between colonies (Keymer et al., 2008), predation (Park et al., 2011), bacterial cell aging (Stewart et al., 2005), and the evolution of antibiotic resistant persisters (Vega et al., 2012). In all of these studies, the ability to investigate microbial activity at the individual cell or small colony scale reveals new dynamics that are not apparent at larger scales.

The extremes of scale inherent in microbial ecology and evolution present an additional barrier to understanding their complex interactions and the patterns that these create, at both large and small scales. At large scales, putting the vast evolutionary history and diversity of microbes in clearer context is important for interpretation of evolutionary patterns and ecological diversity within an environment. Since evolutionary processes occur over time measured in generations, it is critical to connect our evolutionary distance metrics to some approximate sense of generations. One percent sequence divergence sounds small, but in fact often represents thousands of generations or more in microbes. At the small scale, greater appreciation for the heterogeneity of environments at a microbe's scale and phenotypic differences at very fine phylogenetic scales is also important. This appreciation will provide opportunities for addressing outstanding questions the evolution

and maintenance of diversity in seemingly simple environments. Understanding both small and large scale processes and patterns has great promise for investigating some of the most important questions facing microbiologists, from climate change to human health. The long history of ecological and evolutionary research in macro organisms provides a path for microbial research to follow.

Ultimately, the sequencing technologies that have enabled the revolution in microbial ecology and evolution over the past four decades are limited by our understanding of processes at the microbial scale. Sequencing provides a wealth of patterns, but we can only infer the processes that generate them. New technological tools and theoretical frameworks have great potential to unravel both the mechanisms unique to microbes and the guiding principles that structure microbial ecology and evolution.

In the following chapters of this dissertation, I investigate aspects of bacterial evolution and diveristy at a variety of phylogenetic scales. At a very large scale, Chapter 2 presents an assesment of the relative impact of evolutionary contingency and lateral gene transfer across bacterial phyla. In Chapter 3, I investigate the rate of lateral gene transfer and point mutation per million years in the genus *Streptomyces*. I show that the rate of transfer is low relative to the age of the lineage, and that observed point mutation and LGT rates are dependant on the evolutionary distance covered by the dataset. Chapter 4 analyzes a set of very closely related defensive mutualist *Pseudonocardia*, isolated from fungus-growing ants. I identify variable population dynamics and natural product biosynthetic potential in isolates from different geographic locations at kilometer scales. These studies demonstrate the importance of considering both the evolutioanry scale of genomic datasets and the timescales over which processes effecting their evolution occur.

Chapter 2

The Impact of Contingency on Fundamental Genomic Properties of Bacterial Phyla

Bradon R. McDonald, Garret Suen, and Cameron R. Currie

2.1 Abstract

Historical contingency, the limits placed on an organisms evolutionary potential due to inherited characteristics, is a critical factor impacting the diversification of life on earth. Its influence over large evolutionary timespans is still poorly understood, particularly in bacteria. Their ability to acquire new genetic material through lateral gene transfer may significantly reduce the impact of contingency. Here we investigate the influence of contingency on basic genomic characteristics across the history of life in bacteria. We find that global metabolic capabilities are more dependent on phylum classification than isolation environment. We also find that organisms from different phyla vary in their likelihood to gain specific gene classes as their genomes expand. Genomes with high 16S copy number, an indicator of rapid growth ecological strategies, are significantly clustered within two subclades of the Gammaproteobacteria and Firmicutes, and are rarely found in other phyla. Our analysis demonstrates the power of historical contingency to shape

the trajectory of lineages across billions of years of evolution despite continuing genetic exchange.

2.2 Introduction

Bacteria span an astonishing range of metabolic, ecological, and phylogenetic diversity. Beyond the thousands of microbes that have been characterized, recent research into the dozens of not yet cultured bacterial phyla suggests an even greater amount of microbial diversity previously unknown to science (Rinke et al., 2013). Characterizing this diversity and understanding the evolutionary processes that generate it has important implications for understanding the forces that determine the structure of microbial communities and interactions. The evolutionary history of each lineage is a combination of vertical inheritance, lateral gene transfer (LGT), and selection due to physical and ecological forces. This history shapes and constrains the evolutionary trajectory of extant lineages.

Growing appreciation for the role of LGT in bacterial evolution has raised questions about the phylogenetic structure of phenotypic and genomic diversity (Keeling, 2009). The ability of bacteria to acquire genes from distantly related organisms, if frequent enough over evolutionary time, could lead to a complete decoupling between metabolic or phenotypic traits and deep phylogenetic relationships. However, a number of studies have demonstrated strong effects of evolutionary contingency on the distribution of successful LGT events. Press et al. (2016) found a strong effect of epistasis between genes that influenced the ability of particular genes to be successfully acquired and retained. Other studies have found that many genes are toxic if artificially transferred into E. coli (Sorek et al., 2007), and

that most genes in E. coli suspected to have been acquired through LGT are not translated, even though they are transcribed (Taoka et al., 2004). These acquired genes also have high turnover rates (Lerat et al., 2005), due to either detrimental effects on the recipient or neutral gene loss. The impact of LGT over evolutionary time is therefore dependent on a range of processes limiting successful integration of acquired genes, in addition to the absolute transfer rate.

Here we investigate the degree to which several large scale metrics of bacterial diversity are structured by deep phylogenetic relationships. Using complete bacterial genomes from the KEGG database (Moriya et al., 2007; Kanehisa et al., 2014), we investigate overall metabolic capabilities, genome content bias, and 16S copy number distribution in eight phyla and subphyla. We identify a strong phylogenetic signal for each of these metrics, despite the billions of years of evolution and LGT that have affected these lineages.

2.3 Results

Our dataset consisted of 1,206 genomes from across the bacterial domain. Focusing on phyla with at least 40 genomes, we analyzed the Alpha, Beta, Gamma, and Delta Proteobacteria, Firmicutes, Actinobacteria, Bacteroidetes, and Cyanobacteria. We first sought to characterize overall metabolic diversity across this vast bacterial diversity and compare metabolic similarity between organisms in the same phylum versus those isolated from the same environment (Figure 2.1). We generated a metabolic similarity network, linking genomes in the network if they shared high similarity in enzymatic reaction substrate and product distributions. This clustering shows a clear pattern of clustering based on phylo-

genetic assignment. The three proteobacteria subphyla cluster together, and within that large network component bacteria from each subphylum are more likely to cluster together. Genomes from the gammaproteobacterial genera *Coxiella*, *Buchnera*, *Francisella*, and *Xylella* form their own small components. A range of strictly host-associated alphaproteobacteria form their own component, including *Wolbachia, Erlichia, Bartonella*, and *Liberibacter*. The Bacteroides show the best overall clustering by environment, with three components that are predominantly composed of strains biased by environment. Two of the components contain mostly host-associated strains, including the genera *Bacteroides, Porphyromonas, Odoribacter*, and *Prevotella*. The other host-associated component contains genomes of the obligate intracellular mutualist *Sulcia*.

The Actinobacteria form two components (Figure 2.2). The largest component is composed of two main clusters, connected by similarity between a few sets of genomes. *Micromonospora* sp. L5 and *Frankia* sp. EAN1pec share an edge, and *Micrococcus luteus*, *Corynebacterium urealyticum* DSM7109, and *Corynebacterium jeikeium* are each linked to *Mycobacterium leprae* Br4923 by an edge. The smaller component contains genomes from the *Bifidobacteria*, along with *Corynebacterium glutamicum* and *efficiens* strains. The Firmicutes show the most metabolic diversity, with 10 separate components. These components are consistent with subdivisions within the firmicutes.

Another key aspect that drives genome content is genome size; larger genomes tend to have a higher proportion of genes dedicated to signaling, membrane transport, and secondary metabolism (Konstantinidis and Tiedje, 2004). We therefore investigated the change in genomic proportion devoted to a range of biological functions as genome size increases in phyla/subphyla from our dataset. The analyzed phyla showed significant

differences in correlation between genome size and gene content in several KEGG functional categories (Figure 2.3, Supplemental figured A.1 and A.2). The Gammaproteobacteria and Firmicutes each diverge significantly from trends in gene content for all bacteria in a few KEGG categories. Each of these phyla lack correlations between genome size and Xenobiotic Biodegradation and Metabolism. Gene content in the Gammaproteobacteria does not correlate with genome size for the category Amino Acid Metabolism (R2: 0.083), in contrast to the positive correlations observed for this category in the other four phyla (Figure 2.3). The Firmicutes have a weak negative correlation for Carbohydrate Metabolism (R2: 0.122), whereas positive correlations are observed for the other four phyla. This may be explained by the specialized physiology of many Firmicutes with smaller genomes (1,500 - 3,000 genes), which lack the capacity for oxidative phosphorylation and instead use fermentation to generate ATP. This leads to a larger proportion of genes in the Carbohydrate Metabolism category for these genomes, and a lower proportion of genes in Energy Metabolism. Smaller Firmicute genomes are also enriched in Membrane Transport and deficient in Metabolism of Cofactors and Vitamins.

We found that the Alpha, Beta, and Gamma proteobacteria showed weak negative correlations with gene content in the categories Energy Metabolism (R2: 0.193, 0.157, and 0.393 respectively) and weak or insignificant correlations with Transcription (R2: 0.072, 0.008, 0.1506), whereas these were found to be positively correlated with genome size for both Firmicutes (R2: 0.107 and 0.365) and Actinobacteria (R2: 0.279 and 0.373). Genome size in the proteobacterial subphyla also correlated positively with gene content in the category Membrane Transport (R2: 0.331, 0.469, 0.260 respectively) and negatively with gene content in the category Cofactor/Vitamin Metabolism (R2: 0.304, 0.656, and 0.605

respectively); such correlations were not observed in the Firmicutes (R2: 003, 0.027) and Actinobacteria (R2: 0.001, 0.047).

Additionally, some functional categories show little or no correlation with genome size even when analyzed by phylum. For example, Energy Metabolism has a particularly low R2 values, with only Actinobacteria and Gammaproteobacteria showing R2 values > 0.2. As mentioned above, genome size in the Gram-positive phyla do not correlate with gene content in the category Membrane Transport, while genome size in the Proteobacteria does not correlate with gene content in the category Transcription. Genome size in the Gammaproteobacteria also does not correlate with gene content in the categories Carbohydrate Metabolism or Lipid Metabolism. Because gene content in these categories differs between genomes within these phyla, other factors independent of genome size, such as ecological niche, may drive differences in gene content in these categories.

We also investigated an indirect marker of metabolic strategy: 16S rRNA gene copy number. The number of 16S copies correlates with life history strategy: genomes with multiple 16S copies frequently use boom-bust strategies, analogous to R-selected organisms in eukaryotes, while genomes with few 16S copies predominantly use slow growth strategies (Klappenbach et al., 2000). The vast majority of bacterial genera with median 16S counts greater than four are found in the gammaproteobacteria and the firmicutes (Figure 2.4). Within the gammaproteobacteria, the enterobacteriaceae are particularly enriched in high-16S copy genera. Genera within the vibrionaceae also have large numbers of 16S genes, with as many as ten per genome. Within the firmicutes, most high 16S copy strains fall among the Bacilli. The other phyla have very few strains with more than four copies of 16S.

2.4 Discussion

Our analysis demonstrates phylogenetic coherence in several major properties of microbial genomes at deep phylogenetic scales. Despite the fact that LGT is more likely to occur between organisms in the same environment, isolation habitat is not a strong predictor of overall metabolic potential. The stability of overall metabolic capabilities within phyla and subphyla, despite their diverise habitat distribution, suggests more subtle adaptations to new environments. These bacteria may adopt ecological strategies that play to their existing strengths, dictated by billions of years of evolutionary history.

The evolution of rapid growth, boom and bust ecological strategies poses a particular challenge for microorganisms. It requires recalibration of nearly every part of cell physiology, from signaling and regulatory systems for rapid response to nutrient changes to ribosome production and DNA replication rates. Ecological strategies built around rapid growth, as inferred from 16S copy number, have predominantly arisen in particular subdivisions of the Gammaproteobacteria and Firmicutes. This derived trait appears to have evoloved multiple times independently, but is generally uncommon across the diversity of bacteria analyzed in this study. Further research into these lineages and comparisons to closely related low 16S genera will provide key insight into the physiological traits that enable the evolution of this complex and ecologically significant ecological strategy.

The diverse physiological capabilities of microbes are shaped by deep evolutionary relationships between taxa, augmented by additional raw material for adaptation and selection provided through LGT. Acquired genes that are maintained over evolutionary time tend to lay at the edges of an organism's overal metabolic network (Shapiro et al.,

2016), such that core metabolism evolves primarily through vertical inheritence over long evolutionary time periods. Deeper understanding of the predisposition to particular ecological strategies will provide a better framework to interpret large scale taxonomic distribution studies, where information on the ecological roles of different community members is limited.

2.5 Methods

Dataset. The initial dataset for these analyses comprised gene annotations for 1,206 complete bacterial genomes available from the Kyoto Encyclopedia of Genes and Genomes (KEGG) database (accessed: 04/04/2011). We analyzed genomes from the phyla or subphyla Gammaproteobacteria (269 genomes), Firmicutes (251), Alphaproteobacteria (136), Actinobacteria (122), Betaproteobacteria (90), Deltaproteobacteria (41), Bacteroidetes (51), and Cyanobacteria(40).

Genome clustering. Genome clustering by metabolic similarity used the predicted substrates for all KEGG annotated enzymes in each genome. The number of enzymes that either produced or consumed one of 14,783 metabolites in each genome was used to cluster all genomes via spearman correlation. Genomes linked by edges in Figure 2.1 and 2 share a spearman correlation of at least 0.85. Genome clustering by genome content in specific categories, Supplemental figures S1 and S2, used the number of each gene family within the category for spearman correlation analysis.

Genome size evolution. For this analysis, all genomes with less than 600 RNA and protein coding genes were removed. This eliminated 20 small endosymbiont genomes whose genomes are significantly reduced by gene loss specific to their host nutrient requirments. We analyzed the proportion of each genome dedicated to a particular functional categories and then compared these proportions across genome size. Pathway/category assignments are not mutually exclusive as the enzymatic reactions of some proteins are required for multiple pathways. The Secondary Metabolism functional category reported here combines the KEGG categories Polyketide/ terpenoid biosynthesis and Other Secondary Metabolism.

All other pathways were reported as defined by KEGG. Data compilation and regression analysis was conducted using the Python packages SciPy and pandas. For each functional category or pathway, the proportion of KEGG annotated genes within each genome classified in that category or pathway was plotted versus log10 total genes in the genome. R2 values and trendlines were calculated using the stats.linregress() function from SciPy.

16S copy number. We calculated the median number of 16S genes in genomes from each genus, mapped onto an ASTRALII (Mirarab and Warnow, 2015) generated phylogeny of all KEGG represented genera in our phyla of interest. The phylogeny was generated using a 94 conserved core genes from a representative member of each genus, identified using TIGRfam (Haft et al., 2013) HMMer (Eddy, 2011) models for TIGRfam's "core bacterial gene" protein set (GenProp0799). Protein sequences from each representative genome were aligned using MAFFT v7.221 (Katoh and Standley, 2013). 100 boostraped alignments were generated for each gene independently using RAxML-8.1.24 (Stamatakis, 2014) with the PROTGAMMABLOSUM62 substitution model, which were used to generate phylogenies using FastTree 2.0 (Price et al., 2010). These phylogenies were then used as the input for ASTRALII.

2.6 Figures

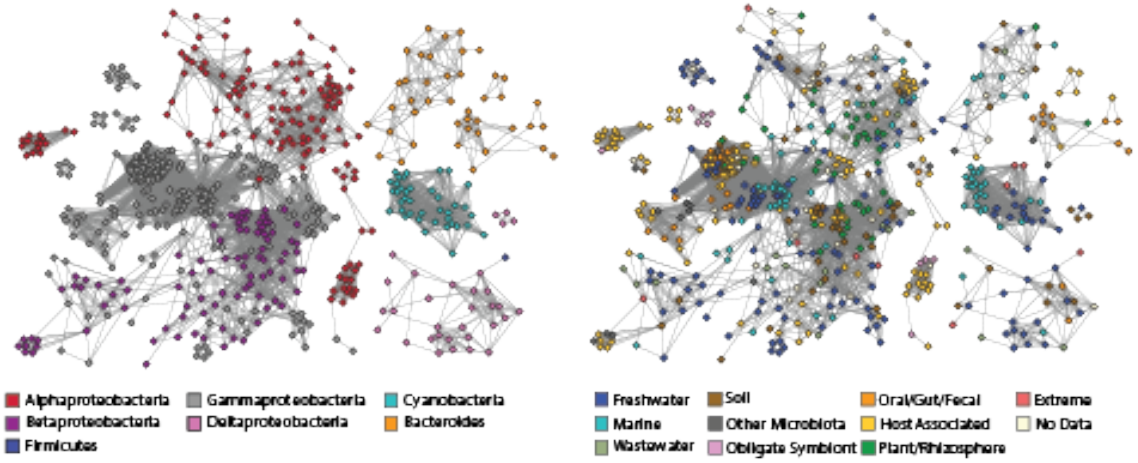

Figure 2.1: **Metabolic similarity across bacterial phyla and environments.** Nodes represent complete bacterial genomes. Edges indicate metabolic capability spearman correlations at or greater than 0.85 between pairs of genomes. Genomes are colored by phylum or isolation environment.

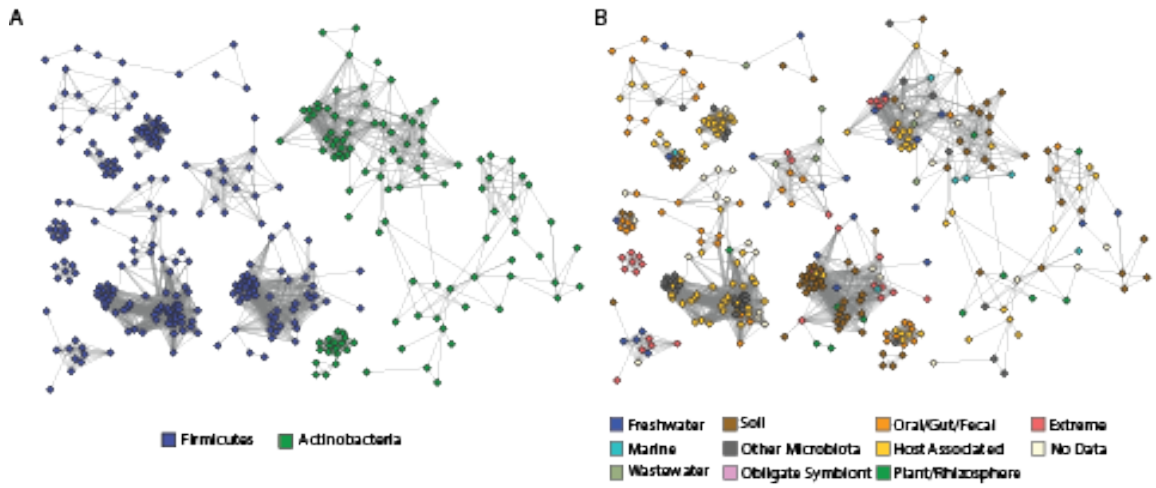

Figure 2.2: **Metabolic similarity across bacterial phyla and environments in Firmicutes and Actinobacteria.** Nodes represent complete bacterial genomes. Edges indicate metabolic capability spearman correlations at or greater than 0.85 between pairs of genomes. Genomes are colored by phylum or isolation environment.

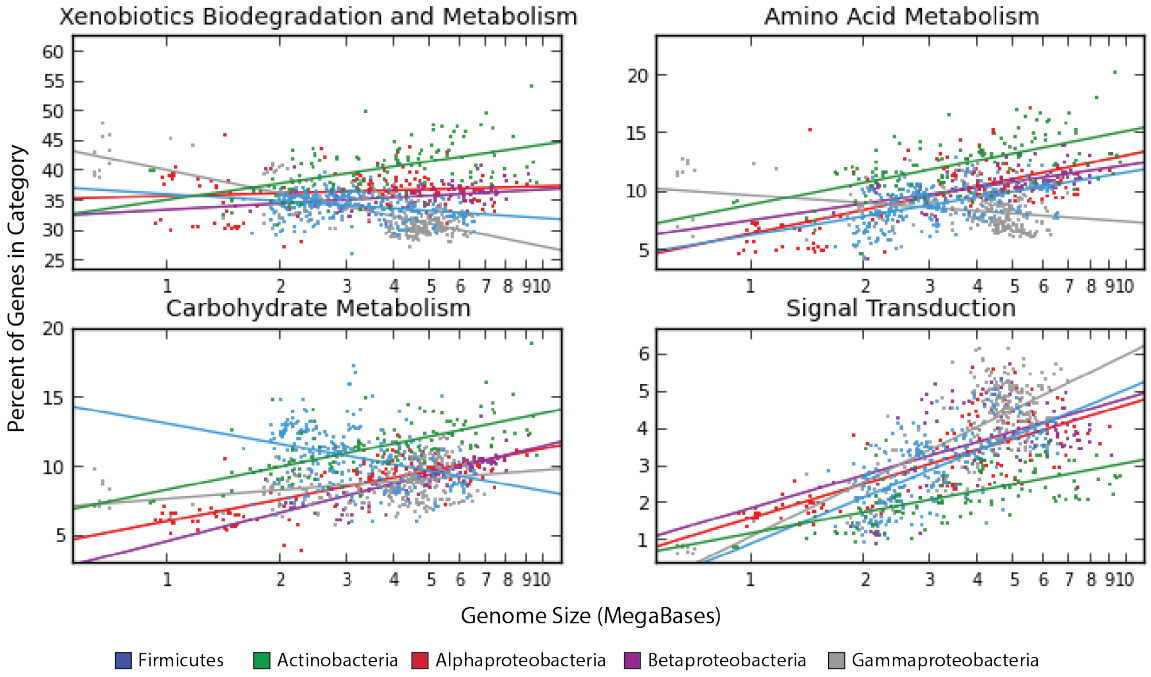

Figure 2.3: **Correlation between genome size and genome content across well sampled bacterial phyla.** In genomes with more than 600 genes, the genomic proportion dedicated to each KEGG gene category is plotted versus genome size.

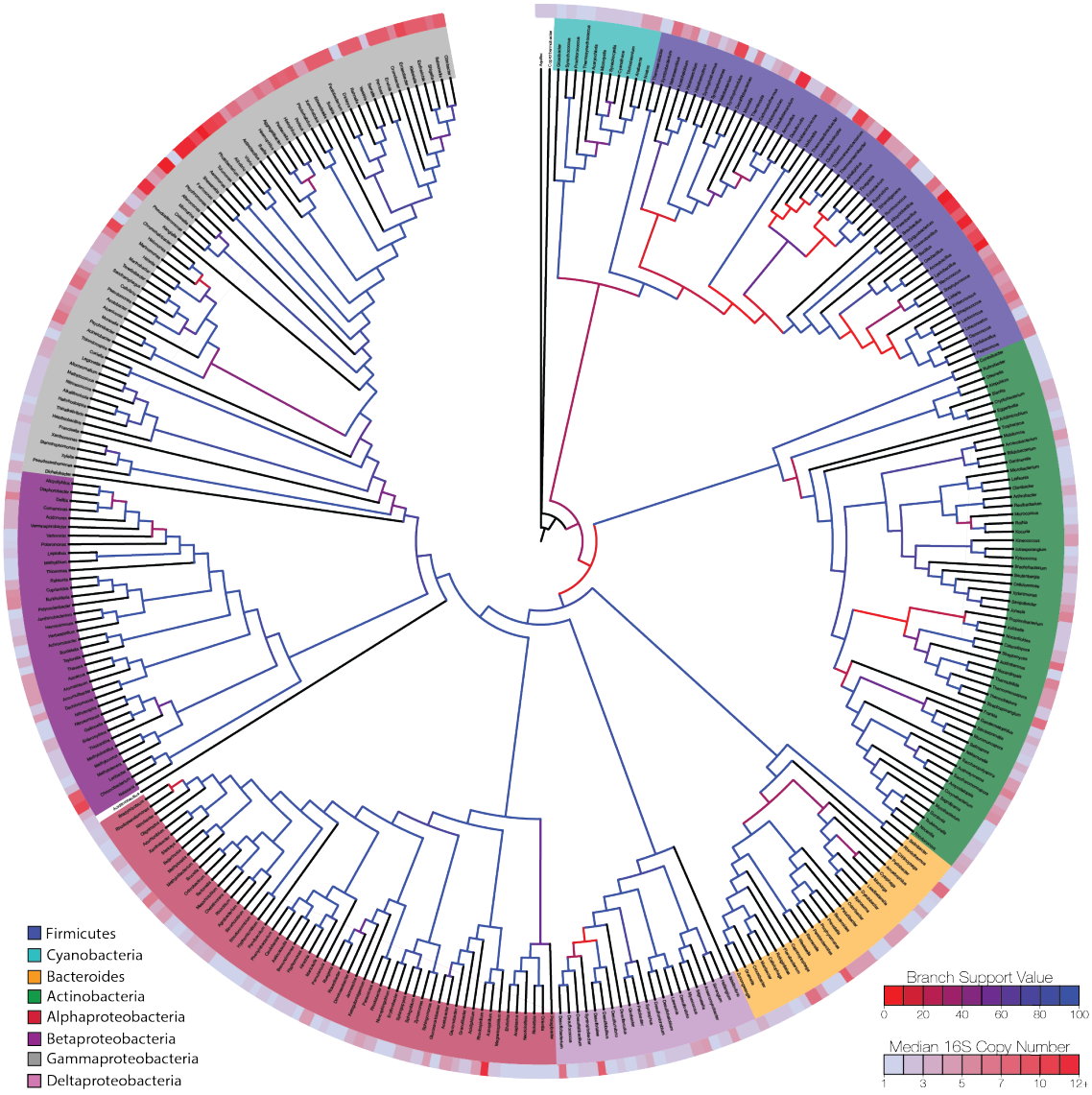

Figure 2.4: **16S copy number distribution in bacterial genera.** Each leaf represents a bacterial genus, colored by phylum. The outer ring shows the median 16S copy number among genomes in each genus. Branch colors indicate the ASTRAL support values for each branch.

Chapter 3

Evolution of the Ancient Bacterial Genus *Streptomyces*

Bradon R. McDonald and Cameron R. Currie.

3.1 Abstract

Lateral gene transfer (LGT) profoundly shapes the evolution of bacterial lineages. LGT across disparate phylogenetic groups and pan-genomic diversity suggest a model of bacterial evolution that views LGT as rampant and promiscuous. It has even driven the argument that species concepts and tree-based phylogenetics cannot be applied to bacteria. Here we show that LGT is an exceedingly rare event in the ubiquitous and biomedically important bacterial genus *Streptomyces* by calculating LGT rate versus time. We estimate that the *Streptomyces* are 380 million years old via molecular clock, as ancient as land vertebrates. Calibrating LGT rate to this geological timespan, we find that, per million years, on average only ten genes are acquired and subsequently maintained. Over that same timespan *Streptomyces* accumulate thousands of point mutations. By explicitly incorporating evolutionary timescale into our analyses, we provide a dramatically different view on the frequency of LGT and its impact on bacterial evolution.

3.2 Introduction

The bacterial domain encompasses prodigious diversity generated over billions of years of evolution (Rinke et al., 2013). Despite the critical role bacteria play in shaping nearly every aspect of life on Earth, understanding the complex evolutionary processes that generate this diversity remains a challenge. In contrast to sexual reproduction in eukaryotes, bacteria were originally thought to undergo strict clonal cell division with little to no genetic exchange. The discovery of conjugative plasmids (Lederberg and Tatum, 1946), and later transformation and transduction (Thomas and Nielsen, 2005), suggested that genetic exchange may play a role in bacterial evolution. The subsequent linking of non-homologous lateral gene transfer to important bacterial phenotypes, such as antibiotic resistance (Davies, 1994) and virulence (Ochman et al., 2000), resulted in the recognition of LGT as a driving force in bacterial evolution. With the advent of comparative genomics and the identification of significant gene content differences between related bacteria (Welch et al., 2002), the prevailing view of rampant exchange of genes across bacteria emerged. Expanding on this view, some have argued that genetic exchange is so rampant that bacterial species do not exist as discrete entities (Boucher et al., 2003) and their evolutionary histories fit a web of life model rather than a tree of life (Gogarten et al., 2002; Soucy et al., 2015; Ge et al., 2005; Retchless and Lawrence, 2010).

The disruptive impact of genetic largely depends on several factors, including the degree to which barriers to LGT structure exchange between distantly related lineages. At larger phylogenetic scales, gene transfers have been shown to be more common within than between phyla, and some phyla exchange genes more frequently than others (Beiko

et al., 2005; Andam and Gogarten, 2011). At the population level, both geographic distribution and sequence dissimilarity can lead to reduced rates of homologous recombination (Whitaker et al., 2003; Cadillo-Quiroz et al., 2012). Analyses that span the species level to intermediate, genus-level diversity provide opportunities to investigate the combined effects of LGT and mutation across physiologically similar organisms that are diverging due to selective pressures in the different ecological niches they occupy. This has the advantage of providing sufficient diversity to reliably detect LGT and identify trends that occur over evolutionary timescales, such as the loss of acquired genes that are selectively neutral or mildly deleterious (Lerat et al., 2005; Kuo and Ochman, 2009; Van Passel et al., 2008). By including sufficient sampling of closely related organisms, it also reduces the amount of variation due to differences in core physiology between organisms in the dataset. Therefore, it enables easier detection of rapidly evolving diversity and ecologically relevant variation.

The ubiquitously distributed bacterial genus *Streptomyces* provides an excellent model for intermediate scale analysis of genetic exchange and mutation. These diverse filamentous bacteria have been isolated from soil, marine, and host-associated environments, with ecological roles ranging from plant biomass degradation (Book et al., 2016; Chater et al., 2010) to defensive mutualisms with eukaryotes (Kroiss et al., 2010). Complex interactions between *Streptomyces* and other organisms are often driven by the production of natural products (Kelsic et al., 2015), which have been mined for drug discovery for decades (Hopwood, 2007). Here we utilize phylogenomic analyses combined with molecular clock dating to investigate the temporal scale of *Streptomyces* genomic and phylogenetic diversity, focusing on non-homologous LGT and point mutations.

3.3 Results

Based on our pan-genomic analyses of 122 *Streptomyces* genomes, 80 publicly available and 42 additional genomes sequenced for this study, we identified vast phylogenetic and genomic diversity. The 42 strains chosen for genome sequencing were selected to obtain genome coverage across the genus; we selected strains to fill in gaps based on the phylogenetic location of the publicly available genomes in a 16S rRNA gene phylogeny (Figure B.1). Using this expanded genomic dataset, multilocus phylogenies were generated using both a traditional multilocus approach based on 94 housekeeping genes (Figure 3.1) and an alternative gene-tree consensus based approach implemented in ASTRALII (Mirarab and Warnow, 2015) (Figure B.3). Both phylogenies are largely congruent, with two major clades of *Streptomyces* containing 88 genomes and a number of basal lineages containing the remaining 34 genomes (Figure 3.1, Figure B.2). These clades did not match the genome distribution on the 16S rRNA gene phylogeny, likely due to the poor resolution of 16S at finer phylogenetic scales. Most marine-isolated *Streptomyces* are found in basal lineages, suggesting a possible marine origin for the genus (Long and Xiao, 2016). Further, many strains in Clade I were isolated from insect-associated niches (Figure B.2). Support values were high across both phylogenies, with the exception of internal nodes in Clade II where tree topology differed in poorly supported nodes for both methods.

Exploring gene content, we identified a total of 39,893 gene families across the genus. Of these, 1,048 were conserved in 95% of *Streptomyces* genomes in the dataset (Figure 3.2). Each of the two major clades and the basal group contained 200 gene families that were conserved in the respective clade but not the others. Each *Streptomyces* clade also contained

a large number of unique genes, with 3,558 unique genes in Clade I and 8,140 in Clade II. Of these a small percentage (5% in Clade I and 2.5% in Clade II) were conserved across the subclade. Shared genome content was correlated with phylogenetic distance, as more closely related genomes shared a higher proportion of genes (Figure B.4b). Overall the results of our gene content analysis, without incorporating divergence times, are consistent with the prevailing view of bacterial evolution: high gene content diversity driven by frequent LGT.

Since mutation and LGT are dynamic processes that occur through time, identifying the rate of these events is critical for understanding their impact on the long-term evolution of a bacterial lineage. We estimated divergence times across the *Streptomyces* phylogeny using *Cyanobacterial* fossils (Brocks, 1999; Garvin et al., 2009), the estimated origin of life on Earth (Mojzsis et al., 1996; Rosing, 1999), and the *Escherichia-Salmonella* divergence time (Ochman and Wilson, 1987) as calibration points for relaxed molecular clock analysis (Figure B.4, Extended Data Table S1). Our analyses inferred that the genus *Streptomyces* diverged from *Kitasatospora* approximately 382 million years (my) ago, in the late Devonian, and the two major clades diverged approximately 123 my ago, in the early-mid cretaceous. These divergence times also allow us to approximate the timespan required for strains to diverge by 1% amino acid identity. Among 11 pairs of strains separated by 1% amino acid divergence in the core genes used for the phylogeny, the average divergence time is 8 ± 2.5 my.

Investigation of LGT dynamics in *Streptomyces* revealed both functional and phylogenetic biases in gene transfer events. In total, we identified 320,263 genes laterally acquired by *Streptomyces* lineages using a gene tree reconciliation approach implemented in AnGST

(David and Alm, 2011). Gene functional classes over-represented in LGT events consisted of secondary metabolism and xenobiotic metabolism (Table B2.). Core biological functions such as transcription and translation were under-represented (Table B.3). Combined with the molecular clock dating, we estimate that over all rates of detectable LGT events per node into Clade I or Clade II are 5.93 and 9.08 per my, respectively (Figure 3.3a, Table B.4). Nodes in Clade I and Clade II were significantly more likely to receive a gene transferred from a member of their own clade than from another source ($p < 1e\text{-}5$, Fisher's Exact Test), with rates of 3.82 and 7.07 per million years for Clade I and II, respectively. Estimated transfer rate per node from Clade I to Clade II is 1.08 transfers per my, and from Clade II to Clade I is 1.43 per my. Inferred transfers of genes from basal *Streptomyces* to a genome in one of the major clades are significantly less common, occurring at approximately once every two million years. Acquisition of genes from different actinobacterial genera occurred even less frequently, at once every 3 million years for Clade II and once every 5 million years for Clade I.

To investigate how the loss of neutral or deleterious genes acquired through LGT impacts estimates of gene transfer rates, we calculated transfer rate versus branch length across the *Streptomyces*. We found that the rate of detectable LGT events per million years is negatively correlated with branch length (Figure 3.3b), and the rate can be approximated relative to branch length using a power law function (α = -0.664 and R2 = 0.662). KEGG (Moriya et al., 2007; Kanehisa et al., 2014) genes with functions involved in replication and repair, translation, cell growth and death, and mobile elements make up a greater proportion of LGT events in short-branch-length nodes than long-branch-length nodes (2 sided T-test, p values 0.009, 0.018, 0.019, 0.027, respectively). This suggests that the lower number of

detected LGT events involving core genes is due to stronger selection against transferred core genes, not a reduced number of actual transfers in these categories. It also suggests that mobile elements are gained and lost more rapidly on evolutionary timescales relative to other gene classes, which is consistent with expectations based on their biology.

We also investigated the relative contribution of point mutation versus LGT to *Streptomyces* diversity over time by calculating the rate of synonymous and non-synonymous point mutations per million years in two different sets of conserved TIGRfam (Haft et al., 2013) gene families: 705 families conserved across all *Streptomyces* (Strept-conserved) and the 94 universally conserved genes used to generate our phylogenies (Bact-universal). Observed mutation rates differed between the two sets of genes we analyzed, particularly for synonymous mutations (Figure 3.4). Synonymous mutation rate was 1.5-2 fold higher in Strept-conserved genes than the Bact-universal genes. Similar to LGT rate, observed rates of both synonymous and non-synonymous mutations are also influenced by the evolutionary distance between genome pairs; synonymous mutation rate appear much higher in closely related genome pairs, while the difference in observed nonsynonymous mutation rate is less dramatic but still apparent. Using only pairwise comparisons of genomes separated by less than 100 my from the Strept-conserved dataset, the estimated median rate of synonymous mutations is 1.62×10^{-8} per site per year and median nonsynonymous mutation rate is 1.78×10^{-9} per site per year. Extrapolating the ratio of synonymous and nonsynonymous sites in the gene sequences we analyzed to total coding sequence length, we estimate a total of 13,714 synonymous and 10,429 non-synonymous mutations accumulate in *Streptomyces* lineages per million years.

Given that natural product biosynthetic gene clusters (BGCs) were over-represented in

LGT events, and the known role some of these play in producing small molecules that shape ecological interactions that are predicted to be under selection (Sit et al., 2015; Ziemert et al., 2014), we examined the distribution and exchange of these secondary metabolite producing pathways. We identified a total of 4,945 natural product BGCs in *Streptomyces*; 1,759 clusters could be classified into 405 BGC families based on PFAM (Finn et al., 2014) domain content, while the domain structure of the other 3,186 were too dissimilar to match another *Streptomyces* cluster. We found that nearly all BGC families were non-randomly distributed in *Streptomyces*, based on a branch-length permutation test: across the genus, 82.7% of BGC families were more phylogenetically restricted than expected by chance. This pattern also holds true at finer phylogenetic scales, as 79.2% and 76.1% of BCG families were more phylogenetically restricted than expected by chance in Clade I and Clade II, respectively. Interestingly, we also found very few cases of transfer and subsequent maintenance of complete BGC operons (Figure 3.5). Analyzing LGT events effecting each gene found within a BGC, our analysis suggests that, at least over long evolutionary time scales, the vast majority of BGCs appear to have been effected by LGT. Specifically, our findings infer that 93% of BGCs acquired at least one gene through LGT within the last 50 my. However, only 57 BGCs had been acquired intact from one source, while the other BGCs were composed of a mixture of genes from multiple sources, including vertically inherited genes. Of the BGCs entirely acquired from a single source, most have been acquired within the last 10 my.

3.4 Discussion

Using *Streptomyces* divergence times rather than sequence similarity provides a new perspective on the rate of LGT and its potential impact on bacterial evolution over ecological and evolutionary timescales. Although we inferred over 300,000 gene transfer events, a single successful gene transfer every hundred thousand years is sufficient to generate the observed number of events given the huge timespan encompassed by *Streptomyces* evolution. Further, gene transfers from distantly related *Streptomyces* or other Actinobacteria are orders of magnitude less frequent than transfers from closely related lineages, suggesting that distantly related bacterial lineages are typically genetically isolated from each other for tens- to hundreds-of-thousands of years at a time. We also find a strong effect of evolutionary distance between sampled genomes on the inferred rate of LGT. Our estimated transfer rate decreases steadily with branch length according to a 2/3rds power law, likely due to acquisition and subsequent loss of acquired genes with neutral or deleterious fitness effects (Lerat et al., 2005). Overall, these results stand in contrast to the prevailing view that successful LGT events are rampant over ecological time scales. Our analysis focused solely on the actinobacterial genus *Streptomyces*; further research is needed to determine how LGT dynamics vary over time in different lineages of bacteria. Nevertheless, given that our estimates of genomic diversity and total LGT events, as based on traditional pan-genome approaches, identify genomic diversity and total LGT events consistent with previous work across bacteria (e.g., ~300,000 LGT events), *Streptomyces* are unlikely to be outliers with respect to LGT rates.

We also show that approximately 23,000 point mutations accumulate per million years.

While positively selected LGT events provide a rare but important source of genetic diversity, these thousands of point mutations likely provide the vast majority of novel genetic diversity in *Streptomyces* over ecological timescales. Similar to the evolutionary distance effect on inferred LGT rate, pairwise comparisons between more distantly related organisms lead to lower estimated mutation rates, and estimated mutation rates are also lower in genes that are conserved across most bacteria than in genes conserved in all *Streptomyces*. The effects of evolutionary distance and gene dataset on inferred mutation rate are stronger in synonymous sites, suggesting widespread fitness effects of synonymous mutations in *Streptomyces* core genes as has been shown in a variety of bacteria and eukaryotes (Agashe et al., 2016; Knöppel et al., 2016; Akashi, 1994; Bailey et al., 2014). These results suggest that comparisons of evolutionary event rates between bacterial groups can be strengthened through normalizing rates based on evolutionary distance between samples within each group.

Genes involved in the biosynthesis of natural products, the compounds that are critical for modern medicine and for which *Streptomyces* are well known, are among the genes most frequently acquired through LGT. In contrast to previous work in the related actinobacterial genus *Salinispora* (Ziemert et al., 2014), we found that most biosynthesis clusters were composed of genes apparently acquired from multiple sources rather than a single full-operon transfer event. These differences could be explained by the *Salinaspora* dataset being comprised of much more closely related genomes than our *Streptomyces* dataset. Thus, many BCG transfer events in *Streptomyces* might in fact be transfers of full clusters, followed by gene shuffling with other clusters in the same genome over evolutionary time. This would result in the scattered distribution of transferred genes we observed in *Streptomyces*

BCGs, without requiring different transfer dynamics than those seen in *Salinaspora*.

Applying the temporal framework to *Streptomyces*, our results show that the biomedically and ecologically important genus *Streptomyces* is truly ancient; *Streptomyces* as a group are approximately as old as tetrapods and 60 my older than seed plants. The two major *Streptomyces* clades are approximately as old as flowering plants and older than the divergence of *Salmonella* and *Escherichia*. This molecular clock analysis is consistent with several other molecular clock analyses performed across all bacteria (Battistuzzi et al., 2004) and Actinomycetes (Embley, 1994), using different methods and datasets. The majority of comparable nodes in our molecular clock analysis fall within the credibility intervals of previous rigorous analysis in bacteria performed by Battistuzzi et al.(Tabl B.1). Older nodes have larger confidence intervals and are more variable. This variability has limited effect on our analyses of *Streptomyces*, because only recently diverged nodes, i.e. less than 400mya, were employed as reference points for the *Streptomyces* molecular clock that was used for downstream analyses. Although our molecular clock analysis is consistent with previous work, uncertainty in molecular clock dating generally means that the absolute rates of LGT we identified are estimates. However, the extremely low rate of LGT means that our general conclusions remain valid even if the *Streptomyces* lineage is significantly younger than our analysis indicates. The relative number of LGT events to point mutations is robust to uncertainty in the molecular clock, since both rates are calculated using the same divergence times.

Our results provide new insight into the paradox that, despite widespread LGT events, bacteria seemly form natural groups with coherent properties (Achtman and Wagner, 2008; Nowell et al., 2014). The argument that LGT significantly disrupts vertical evolution in

bacteria emerged from a combination of high pan-genomic diversity and examples of gene acquisitions from distantly related organisms (Cordero and Hogeweg, 2009). However, we find strong evidence of phylogenetic biases in *Streptomyces* LGT, even in the most frequently transferred gene classes such as secondary metabolism. We also find evidence that many transferred genes may be selectively neutral or deleterious, leading to rapid turnover of acquired genes. These results suggest that LGT dynamics can be highly structured by phylogenetic (Kloesges et al., 2011) and physiological (Lercher and Pál, 2008; Taoka et al., 2004) factors even within a genus, which limit its disruptive effect on bacterial evolution. Our incorporation of an absolute timescale reveals that the actual rate of successful transfer is many orders of magnitude lower than estimated bacterial generation times in nature (Lawrence and Ochman, 1998; Ochman et al., 1999), and we show that thousands of point mutations may accumulate over that same timescale. Viewed from this perspective, the high absolute number of transfers detected in large bacterial genomic datasets may not be the result of high transfer rates but of long evolutionary timespans separating sampled strains. Because even bacteria in the same genus can be separated by tens to hundreds of millions of years, placing LGT in a temporal context also potentially explains the variable gene content patterns observed among closely related genomes without the need to invoke rampant LGT at ecological time scales. Together, our results support a model of rare positively selected LGT events over millions of years driving gene content diversity in bacteria, with vertically inherited point mutations and homologous recombination dominating bacterial evolution over ecological timescales and finer phylogenetic levels.

3.5 Methods

Genome annotation. In order to ensure consistent annotations across all genomes, protein-coding genes were predicted *de-novo* using Prodigal (Hyatt et al., 2010). These were annotated using whole protein HMMer3 (Eddy, 2011) models generated from KEGG (Moriya et al., 2007; Kanehisa et al., 2014) database gene families and TIGRfam 13.0. They were also annotated using domain HMMer models from PFAM 27.0 and antiSMASH 2.0 (Blin et al., 2013). The TIGRfam noise score cutoff and antiSMASH score cutoffs were used to remove false positive hits, while KEGG hits were removed as false positives if their e-value was greater than 1e-5, or if the difference in length between the HMMer model consensus and the protein was greater than 50%. Ribosomal RNA genes were identified by performing BLAST v2.2.25 (Altschul et al., 1990) analyses using sequences from the SILVA database (Quast et al., 2013). Proteins were also classified into families de-novo based on sequence homology using ProteinorthoV2 (Lechner et al., 2011) with default parameters.

Genome-based phylogenetics. The *Streptomyces*/Actinobacteria multilocus phylogeny was generated using TIGRfam annotated proteins. The 94 TIGRfam proteins in the "core bacterial protein" set (GenProp0799) was used as the molecular dataset. The protein sequences with the top HMMer bitscore for each protein family in each genome were aligned using MAFFT (Katoh and Standley, 2013). These protein alignments were then converted to codon alignments and were concatenated. Recombinant regions were identified using BratNEXTGEN (Marttinen et al., 2012), and masked to remove the potential confounding influence of homologous recombination on species tree topology. This removal did not have a significant influence on the final topology. RAxML-7.2.6 (Stamatakis, 2006) was

used to generate the phylogeny using the GTRGAMMA substitution model and 100 rapid bootstraps on the final, recombination-free alignment. The phylogeny of all bacteria used for molecular clock calibration was generated using a similar workflow, except that protein sequences were used to generate the phylogeny using the PROTGAMMABLOSUM62 substitution model in RAxML. Since genomes in this phylogeny were generally very distantly related to each other, no correction for homologous recombination was performed. The gene-tree based phylogeny was generated using ASTRALII. 100 bootstrap alignments were generated for each of the core TIGRfam families using RAxML. Phylogenies for each of these alignments were generated with FastTree 2.0 (Price et al., 2010) and used as the input data for ASTRALII.

16S rRNA gene phylogeny. *Streptomyces* 16S gene sequences were obtained from RefSeq (Tatusova et al., 2014), along with 16S sequences extracted from the genomic dataset and three *Cyanobacterial* 16S sequences that were used as outgroups. These were aligned using MAFFT and then hand curated and trimmed to remove low quality 16S sequences. The curated set of sequences was realigned and used to generate a phylogeny with FastTree.

Molecular clock analyses. Reltime (Tamura et al., 2012) was used to approximate divergence times for two different phylogenies: the bacterial tree of life and the Actinobacterial phylogeny containing the full set of *Streptomyces* genomes. The all-bacteria phylogeny and protein alignment described above were used as the input for Reltime. The algorithm was set to use 'Many Clocks' and gamma distributed rates with invariant sites. Approximate time intervals for the evolution of *Cyanobacteria*, (2500-3500 million years ago)25,26, the divergence of *Salmonella* and *Escherichia* (50-150 million years ago)29, and the origin of bacteria (3500-3800 million years ago)27,28 were used to calibrate the molecular clock found in

Extended Data Figure 3.4. Using a single calibration point can correctly infer the divergence dates of the others with less than 20% error. The confidence intervals for the origin of Actinobacteria, *Streptomyces*, and the divergence of *Streptomyces* Clade I and Clade II were then used to calibrate a second molecular clock analysis of the Actinobacteria/Streptomyces phylogeny, i.e. Figure 3.1.

Lateral gene transfer analysis. For each protein or domain database, the protein sequences for all ProteinorthoV2 gene families with more than 3 genes were aligned using MAFFT and then converted to codon alignments. 10 bootstrapped alignments were generated using RAxML for each gene family, and FastTree was used to generate a phylogeny for each bootstrapped alignment. These bootstrap trees were then used as the gene tree inputs for AnGST. Default reconciliation event costs were used (LGT=3, DUP=2, LOS=1). The *Streptomyces*/Actinobacteria molecular clock tree was provided as the species tree, and AnGST was run in ultrametric mode to avoid biologically improbable LGT events from extant genomes to deep ancestral nodes. The rate of lateral gene transfer between or within subsets of *Streptomyces* was calculated as the number of genes acquired by genomes in the analyzed clade divided by the age of last common ancestor of the clade in millions of years. LGT events affecting secondary metabolite clusters that occurred within a time interval were identified by finding genes which first appeared in a *Streptomyces* lineage within that interval. There is not a statistically significant difference in the inferred percentage of genes transferred into draft versus complete genomes ($p = 0.21$, Mann-Whitney U), indicating that using predominantly draft genomes does not generate significant detection bias in our LGT analysis. The percent of inferred genes losses is somewhat higher and more variable in draft genomes, mean $15.30 \pm 8.91\%$, versus $10.41 \pm 4.81\%$ for complete genomes ($p = 0.06$,

Mann-Whitney U). This is likely an artifact of variable quality draft genome assemblies.

Mutation rate estimation. The concatenated codon alignment used to generate the *Streptomyces* phylogeny, along with concatenated codon alignments of all genes conserved in 95% of the *Streptomyces* genomes, were used to calculate point mutation rates. The number of synonymous and non-synonymous mutations and sites was identified by the codeML package in PAML (Yang, 2007), and molecular clock divergence dates were used to calculate mutation rates per million years. We estimated the total number of synonymous and non-synonymous sites across all coding regions by multiplying the number of sites in the TIGRfam protein coding sequence by the average total length of protein coding sequence. Total number of mutations per million years was calculated based on the mutation rate and the estimated total number of sites per genome.

Natural product biosynthesis cluster families. Natural product biosynthetic gene clusters were predicted using the ClusterFinder algorithm (Cimermancic et al., 2014) and PFAM annotations. Genes that were found on the end of predicted clusters and had less than 0.8 probability of being part of a cluster were removed. After this trimming, clusters with fewer than 3 genes and clusters lacking genes with an antiSMASH domain hit were removed. These curated clusters then were grouped into families using the modified Lin similarity metric (Lin et al., 2006) with a Jaccard weight of 0.36, GK-Gamma weight of 0.64, and overall similarity threshold of 0.7. We then processed the matches with the MCL algorithm, which was run with default parameters, to generate final cluster families. We used a subtree permutation approach to identify cluster families that were phylogenetically restricted, as in Cafaro et al. (2011). For each cluster family, we generated subtrees from the multilocus phylogeny containing only the genomes which possessed the cluster. We

then compared the total branch length of this subtree to the branch length distribution of 1,000 subtrees containing the same number of taxa, randomly sampled from the multilocus phylogeny. Cluster families were identified as phylogenetically restricted if their subtree total branch length was significantly less than the distribution mean by two sided T-test, with a p-value cutoff of 1e-5.

3.6 Figures

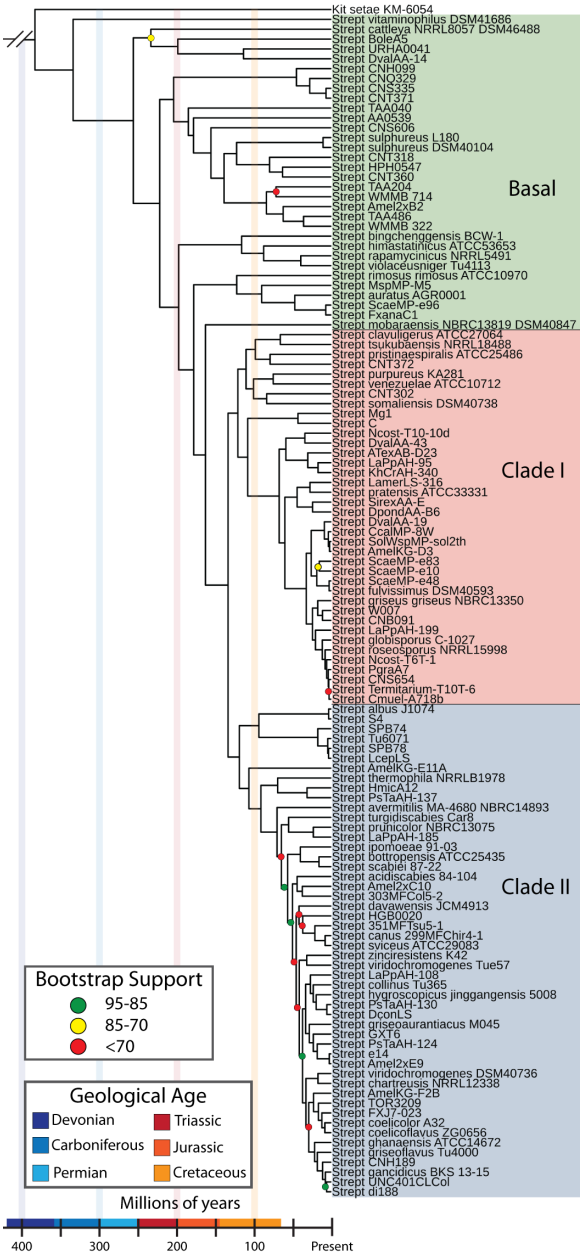

Figure 3.1: **Molecular clock phylogeny of the *Streptomyces*.** TIGRfam-based multilocus phylogeny of *Streptomyces* using 94 universally conserved housekeeping genes. Branch lengths indicate Reltime-estimated divergence times. Bootstrap values are shown by colored circles on all nodes with values $\leqslant 95$. *Streptomyces* is abbreviated as Strept, and *Kitasatospora* is abbreviated as Kit.

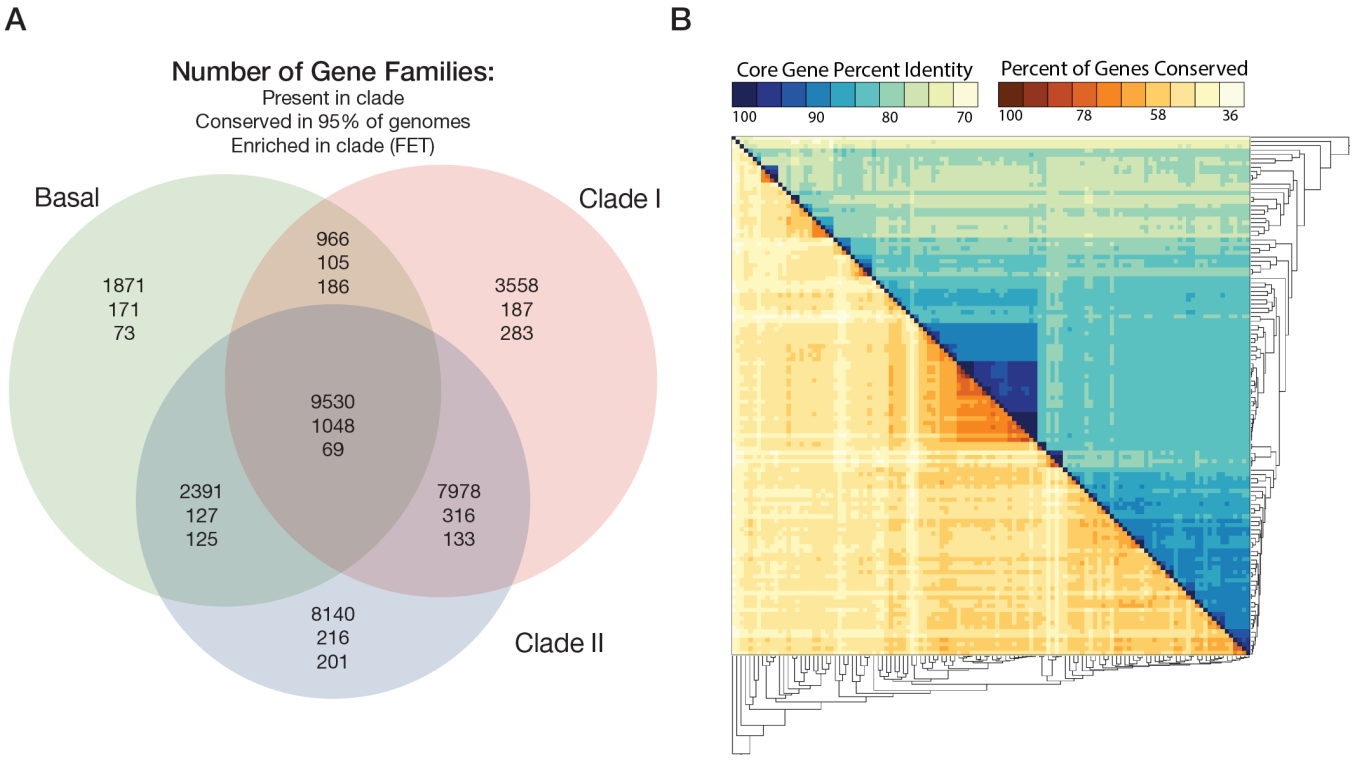

Figure 3.2: ***Streptomyces* genome content diversity.** A. Proteinortho gene families present, conserved, and enriched in the three main divisions of *Streptomyces*. Enrichment was determined by fisher's exact test. B. Pairwise comparison of conserved Proteinortho gene family percentage versus TIGRfam core gene percent identity.

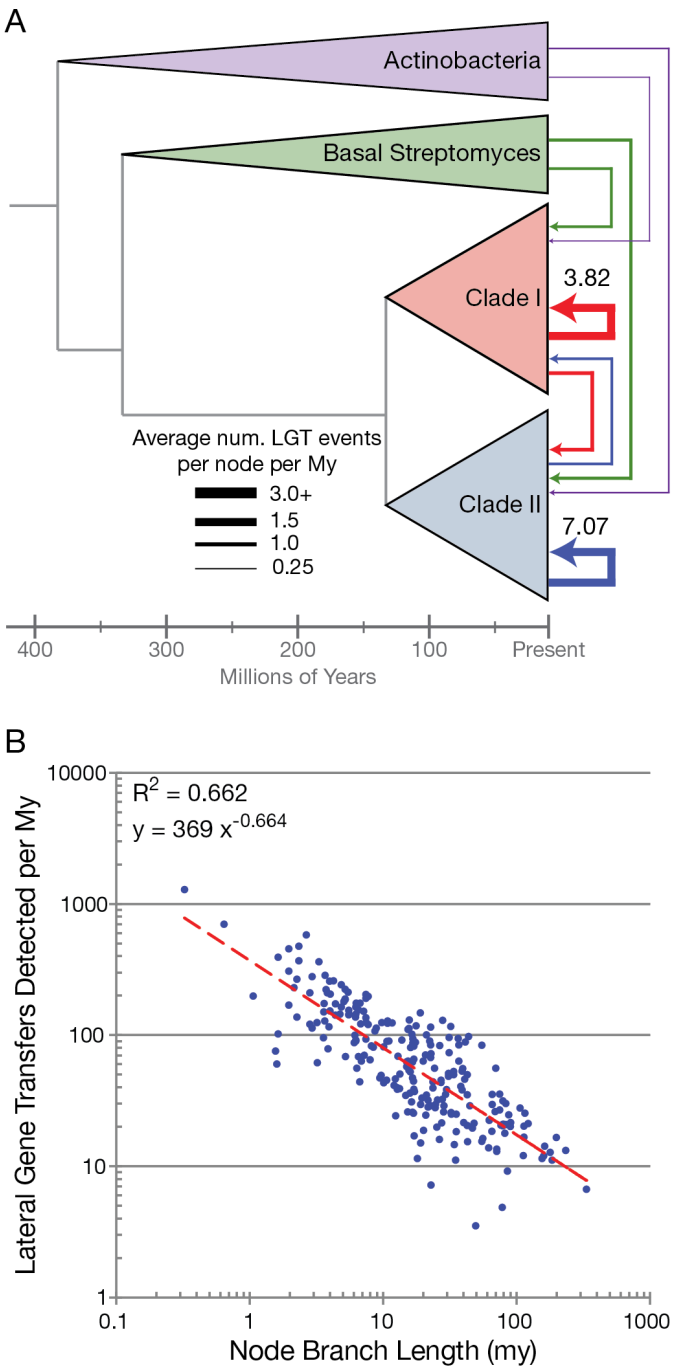

Figure 3.3: **The rate of lateral gene transfer in *Streptomyces*.** A. Average rate of LGT across the *Streptomyces* phylogeny. Line thickness indicates the average number of detected LGT events per genome (including ancestral reconstructions) per million years from each source. Rates greater than 3 are labeled, and all rates appear in Table S2. B. Detected rate of LGT on each branch of the phylogeny. Detected LGT rate is negatively correlated with branch length.

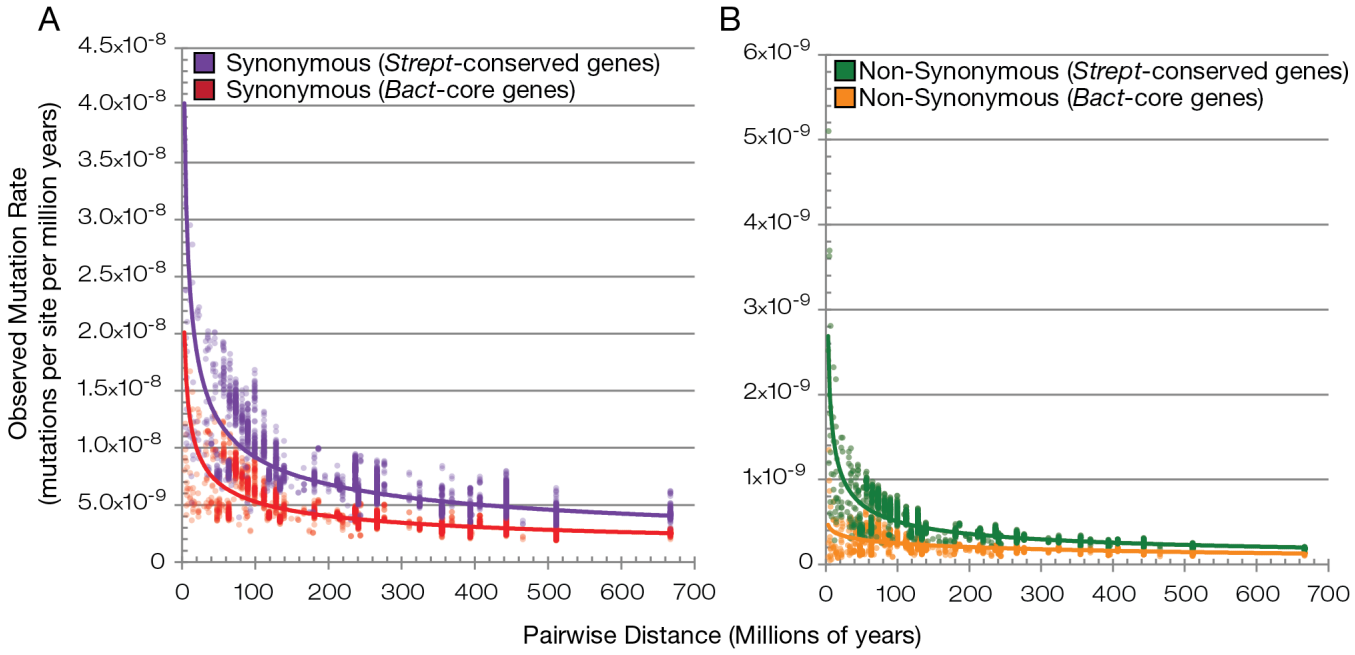

Figure 3.4: **The rate of point mutations in TIGRfam gene families.** Bact-core genes consists of 94 housekeeping TIGRfam gene families conserved across bacteria. Strept-conserved consists of 705 TIGRfam gene families that are found in 95% of our *Streptomyces* genome dataset. A. The observed rate of synonymous point mutations varies by gene dataset and pairwise distance, likely due to stabilizing selection of synonymous sites in highly conserved genes. B. The observed rate of non-synonymous sites also varies by dataset and pairwise distance, but to a lesser degree than synonymous sites.

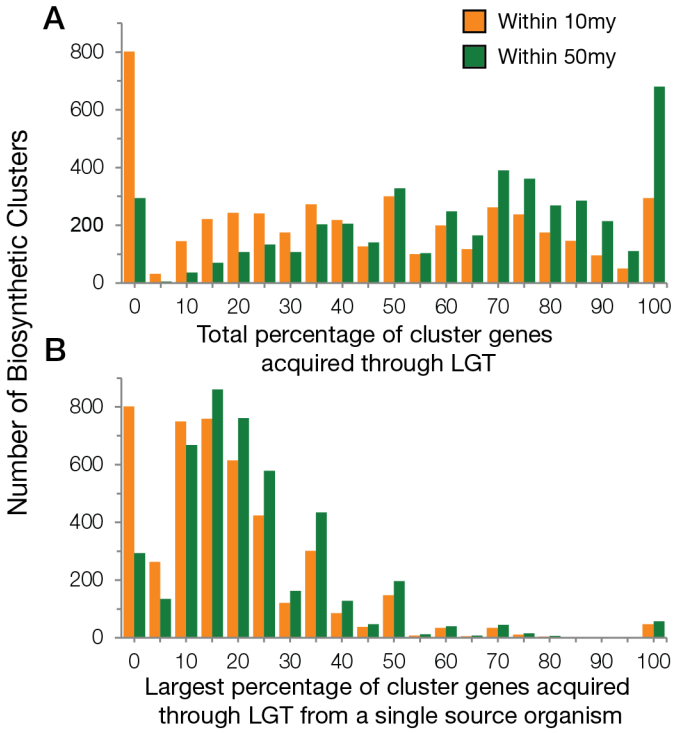

Figure 3.5: **Sources of laterally transferred secondary metabolite biosynthesis genes.** A. Most biosynthetic gene clusters have acquired some genes through LGT within the last 50my years. B. The vast majority of clusters appear to be a mix of genes from multiple sources, including vertical inheritance.

Chapter 4

Population Genomics of Fungus-growing Ant-associated *Pseudonocardia*

Bradon R. McDonald, Marc Chevrette, Jonathan L. Klassen, Heidi A. Horn, Eric J. Caldera, Evelyn Wendt-Pienkowski, Matias J. Cafaro, Michael G. Poulsen, Nicole M. Gerardo, Antonio C. Ruzzini, Ethan B. Van Arnam, George M. Weinstock, Jon Clardy, and Cameron R. Currie

4.1 Abstract

The geographic and phylogenetic scale of ecologically relevant microbial diversity is still poorly understood. Here we use a model mutualism, fungus-growing ants and their defensive bacterial associate *Pseudonocardia*, to investigate population-level diversity of bacteria across kilometer-scale geographic space. These bacteria grow on the ant cuticle and provide chemical defense against fungal pathogens via production of natural products. Our study includes 42 strains isolated from ant colonies in a 20km transect in and around Barro Colorado Island in Panama. Population genomic analysis revealed a strong geographic signal in SNP differences and genome content. We also identify several instances of migration

between the mainland and neighboring island. We find evidence of genetic exchange in these host-associated organisms. Gene content differences between our reference genome and other strains range from 621 to 1,532. The degree of genetic diversity, as determined by Tajima's D, varies between populations. This occurs despite the populations living in close proximity to each other and the potential for recombination. In addition, natural product biosynthesis clusters are also correlated with isolation location. These results demonstrate that ecologically relevant differences can be found between bacteria with extremely high core gene similarity within limited geographic ranges, despite continued gene flow.

4.2 Introduction

Understanding the phylogenetic and geographic scale of ecologically relevant diversity in microbes is a continuing challenge (Martiny et al., 2006). Due to both their small size and insufficient knowledge of ecological niches in complex communities, most comparative analyzes of microbial diversity have focused either on a few well characterized model organisms or microbial groups that span hundreds of millions of years of evolution. However, a growing body of evidence suggests that what are considered extremely fine-scale divisions between microbes approach the level of diversity that is ecologically relevant. Knowledge about the scale of ecological and phenotypic diversity in microbes is critical from both scientific knowledge and practical perspectives. Identifying ecological patterns at the correct scale provides deeper insight into the processes that generate these patterns, which influence the evolution of all microbial life on earth. Practically, sampling and experimental design are heavily influenced by our understanding of ecological processes

and the evolutionary scale at which they operate. Enzyme and small molecule discovery efforts can also be guided by this understanding (Lewin et al., 2016), making the knowledge of ecological diversity important for applied as well as basic science objectives.

Analysis of sub-strains of the same species in the genus *Streptomyces* revealed biogeographical patterns influenced by glacial movement in the last ice age using a selection of housekeeping genes (Andam et al., 2016). At an even finer phylogenetic scale, *Vibrio cyclotrophicus* strains isolated from different size organic particles in the ocean with nearly identical housekeeping gene sequences are ecologically distinct (Shapiro et al., 2012; Yawata et al., 2014). Whole genome analysis of these *Vibrio* demonstrated divergence in specific genes and, after significant experimental examination, subtle but ecologically significant phenotypic differences were identified.

Although the environmental forces driving the distribution of fine-scale microbial diversity are poorly understood for most taxa, those that are associated with extreme environmental conditions (Cadillo-Quiroz et al., 2012) or with eukaryotic hosts (Moran et al., 2009) can be used to address biogeographical and population-scale ecological questions more easily. Bacteria from the genus *Pseudonocardia*, which form a defensive mutualism with many fungus-growing ant species (Currie et al., 2003), provide a useful model system. The bacteria grow on the external surface of the ants, and are passed from one ant to another within the first hours of adult life (Marsh et al., 2014). Queens generally carry the bacteria with them when forming a new colony, but there is phylogenetic evidence of host switches over evolutionary time (Cafaro et al., 2011). The *Pseudonocardia* produce natural products that inhibit the growth of *Escovopsis*, a co-evolved pathogen of the ant's fungus garden. Several analyses of the biosynthetic gene clusters used to produce the natural

products suggest they are frequently found on plasmids (Sit et al., 2015; Van Arnam et al., 2015), leading to diverse chemical potential in very closely related organisms.

We sought to characterize the genomic diversity and chemical potential of these bacterial mutualists within a 20km transect on and around Barro Colorado Island in Panama. We used primarily sequenced genomes from population-scale sampling of *Pseudonocardia* strains isolated the ant species *Apterostigma dentigerum*, many of which have been used for a previous multilocus sequence analysis (Caldera and Currie, 2012). Queens of this ant species generally form new colonies within 400 meters of their parent colony, such that traveling across our study area would take 20 colony generations. Using a combination of comparative genomic and population genetic tools, we show a strong effect of geography on conserved gene diversity, gene content, and natural product potential in these microbes at the kilometer scale.

4.3 Results

Our analysis focused on 42 *Pseudoncoardia* genomes isolated from fungus-growing ant cuticles, primarily from the region around Barro Colorado Island (BCI). This island was formed from a hilltop that became isolated during the flooding that generated the Panama Canal. Several complete or near-complete genomes were generated using Pacific Biosciences sequencing to act as high quality references, while the rest of the genomes were sequenced using Illumina. These genomes were part of a clade of very closely related genomes in a multilocus phylogeny of *Pseudoncoardia*, along with a number of isolates from other fungus-growing ants from Panama and several other countries in Central and South America

(Figure 4.1). With an average core gene nucleotide sequence identity of 99.35%, these genomes are well within frequently used cutoffs to be considered ecologically equivalent organisms (Caro-Quintero and Konstantinidis, 2012).

Although core gene sequence identity was very high, we identified a number of strain clusters based on genome-wide SNP differences (Figure 4.1). We mapped all genomes in the clade to the PacBio genome for *Pseudoncoardia* sp EC080625-04 using nucmer (Kurtz et al., 2004) and identified 280,007 bi-allelic polymorphic positions that were covered by contigs from every genome. Putative population assignment inferred by fineSTRUCTURE (Lawson, 2012) using this SNP data matched the isolation locations for most strains (Figure 4.2a), with two groups of BCI isolated strains and a number of strains from the mainland around Pipeline Road (PLR). fineSTRUCTURE's subdivisions of BCI strains showed a geographic pattern on the island, with the less diverse subpopulation occupying the northern and western part of the island. PLR strains also clustered by geographic distribution, with one group clustering around the northern part of our sampling area and another in the southern portion. The PLR sampling area also contained sympatric subpopulations. Several strains isolated in the southern part of our sampling area clustered together, separately from the majority of PLR isolated strains. Finally, a further 3 strains of mainland-isolated *Pseudoncoardia* fell among populations largely consisting of *Pseudoncoardia* isolated from other ant species or other countries. Sampling of these populations is more limited in our dataset, with many strain clusters consisting of only one or two isolate genomes. There were 3 strains which did not fall into the same geographic area as the other strains they clustered with genetically. Two mainland-isolated strains clustered in one of the BCI populations (*Pseudonocardia* sp. EC080625-04 and EC080529-09) and a single BCI-isolated strain that

clustered with the mainland populations, *Pseudoncoardia* sp. EC080619-08.

The total number of polymorphic sites between genomes in reference-aligned regions ranges from 300 to more than 200,000, consistent with fineSTRUCTURE's population assignments (Figure C.1). Overall, core gene percent identity between isolates from the BCI populations and other genomes in our dataset correlates well with geographic distance (Figure 4.2b). BCI strains have very high sequence similarity with other isolates from the island, except for the isolate which falls among the mainland isolated strains, and lower core gene similarity to other mainland strains. Nearly all strains in the dataset from the area around BCI share core gene percent identity above 99.5%, except for three strains with a percent identity to the BCI strains of approximately 98.75%. Genome content diversity also follows a geographical pattern, with BCI-isolated strains generally sharing higher genome content similarity (Figure 4.2c). The pan-genome size of these closely related *Pseudoncoardia* strains is relatively small, with 6,617 actNOG gene families present in at least one genome (Figure 4.3). The core genome is approximately 3,238 actNOG gene families, which represents about half of the total genes found in most *Pseudoncoardia* genomes in our dataset.

Analysis of contig mapping to the reference genome provided insight into the source of gene content variation between these closely related genomes. Gene content varies by as much as 2,000 genes between strains. The vast majority of contigs either mapped to the reference across nearly their entire length, or failed to map almost entirely. Cases of only part of a contig mapping to the reference were very low, at only 3.5% of contigs with more than 10% and less than 90% mapping to the reference. Further, non-mapping contigs had a lower GC content than mapping contigs, at 72% versus 74% respectively

(T-test statistic -19.7, p-value 2.82E-85). Combining all genomes in the dataset, many KEGG (Moriya et al., 2007; Kanehisa et al., 2014) pathways associated with secondary metabolism were over-represented among genes that did not map to EC080625-04. Similarly, genes for the degradation of xenobiotics, degradation of a number of amino acids, and transposon/phage genes were also over-represented. Genes involved in many core biological functions and metabolic pathways are under-represented in these genes. Overall the function and sequence characteristics of DNA that does not map to the reference genome are consistent with the hypothesis that most population-level gene content diversity in ant-associated *Pseudoncoardia* is driven by acquisition and loss of mobile genetic elements and plasmids (Sit et al., 2015; Van Arnam et al., 2015), and that these mobile elements often contain secondary metabolite clusters.

We used two different approaches to identify diversity and selection within conserved genes diverging between the different populations, focusing on Tajima's D (TD) (Tajima, 1989) and gene tree topologies. TD measures the amount of sequence variation in a population relative to the expected amount of variation under evolutionarily neutral conditions, with genes below zero showing less than expected diversity and genes above zero showing greater diversity. We calculated TD values for all genes with greater than 1% polymorphic sites in various strain groups (Figure 4.4). The full population-scale dataset had a mean TD value of -0.444 across 4,225 genes. The BCI strain groups had much lower genetic diversity, with a maximum of 404 genes containing enough polymorphic sites to calculate TD and a mean value of -1.41. While the PLR-A (green) population also had a low TD value, PLR-B (dark red) had a median TD value of -0.03, and the PLR-B1 (red), and PLR-B2 (orange) had mean TD values of 1.28 and 0.28 respectively. High TD genome-wide TD

values suggest a recent population bottleneck in the PLR-B1 and 2 lineages. This suggest complex population dynamics that vary between clusters of very closely related strains within this small geographic region.

To identify genes under selection, we generated phylogenies of each gene in the EC080625-04 reference genome and identified genes in which each population or internal node in the fineSTRUCTURE dendrogram formed a well-supported monophyletic clade (Figure 4.5a). Given the high relatedness between genomes, genes diverging due to different selective pressures between populations would form monophyletic clades that matched population boundaries, while genes under purifying selection would be unlikely to form clades that matched population groupings unless the strains had diverged enough to become genetically isolated. In general, monophyletic genes were distributed across the chromosome in the population groups tested. The internal node shared by the BCI clade (purple) had the highest number of monophyletic genes, at 1,023, while subclades BCI-A (dark blue) and BCI-B (light blue) were distinguished by 344 and 43 genes respectively. We identified KEGG gene categories enriched in the monophyletic genes for each lineage using Fisher's Exact Test and the Benjamini-Hochberg procedure to limit the false discovery rate to 10%. Gene categories enriched among BCI monphyletic gene trees include xenobiotics degradation (odds ratio 1.69), tryptophan metabolism (odds ratio 2.18), ABC transporters (odds ratio 1.67), and nucleotide excision repair (odds ratio 4.64). KEGG gene categories enriched among the monophyletic genes in PLR-B2 (red) include secondary metabolism (odds ratio = 1.82) and aminobenzoate degradation (odds ratio = 2.21). No mobile elements were found to have monophyletic gene trees in this lineage, indicating continued movement of selfish genetic elements between these strains and other *Pseudoncoardia*.

We also investigated genes in the top and bottom 5% of TD values within each population. In most populations these genes were spread across the genome, except in clustering near the origin in PLR-B (dark red) and a cluster of genes with high TD in PLR-B2 (orange) and B2A(yellow) that appear to be part of a mobile element (Figure 4.5b). Across all populations, genes containing the PFAM (Finn et al., 2014) domain of unknown function DUF222 are significantly enriched among genes with unusually low TD values (Fisher's exact test, odds ratio 50.02, p value 1.2E-9). Eleven genes in EC080625-04 contain this domain, eight of which are in the lowest 5% of TD values in at least one population and seven of which are in the lowest 5% in at least three populations. Also known as 13e12 repeat proteins, these have been described homing endonucleases (Gibb and Edgell, 2007) and are an insertion target for phage in *Mycobacterium* (Cole, 1999). Genes containing this domain are also upregulated as part of the SOS regulon in *Corynebacterium* (Jochmann et al., 2009) and *Mycobacteria* (Davis et al., 2002). Two other PFAM domains are enriched in the low TD genes, one of which is transposase related (DDE_Tnp,) and another repeat domain of unknown function (RCC1_2). These results suggest recombination or selective sweeps affecting a few domain classes across *Pseudoncoardia* in the area around BCI, some of which are likely mobile elements.

When the full set of genomes is analyzed as a single dataset, a number of clusters of genes showed high TD values. These clusters included several transporters and signaling proteins, along with a cluster of genes involved in exopolysaccharide biosynthesis and cell envelope biosynthesis. Genes with low TD values are more scattered in the genome, and include a putative prophage and several transposases in addition to the aforementioned DUF222 domains. There are several clusters of low TD genes, including a number of genes

in a type-VI secretion system. A number of core genes also show abnormally low TD, including FtsQ and cytochrome C oxidase subunit I. In the BCI populations, low TD genes include a large number of hypothetical proteins along with two secretion system associated proteins, one type IV VirD4 family and one type VII EccB family. High TD genes include an MT0933-like antitoxin protein along with a range of hypothetical proteins. None of the genes with outlier TD values when both BCI populations are analyzed together have high TD in one population alone, suggesting that balancing selection between the two BCI populations may maintain genetic diversity at these loci.

Since production of natural products is thought to be the primary ecological role of ant mutualist *Pseudoncoardia*, we investigated the diversity and distribution of natural product biosynthetic gene cluster families (BGCs) among the sampled *Pseudoncoardia* populations (Fig 4.6). We identified 27 BCG families, 7 of which were widely distributed across strains in our dataset only within the southern BCI population. A siderophore BGC is found only in samples collected from BCI, including *Pseudoncoardia* sp. EC080619-08, which clusters well within the mainland population by SNP distribution. Likewise, a type 2 polyketide BGC is found only within the northern BCI population and *Pseudoncoardia* EC080619-08. Similarly, another type 2 polyketide BGC is found primarily among BCI isolates, along with *Pseudoncoardia* sp. EC080619-08 and *Pseudoncoardia* sp. EC080525-06. An aminoglycoside BGC also shows a large positive bias towards the island as 7 of 8 occurrences are from BCI populations. Mainland specific BGCs include a lassopeptide found only in isolates from the southern part of our mainland sampling area, and a nonribosomal peptide found only in the northern mainland population. Furthermore, multiple BGCs including an, ectoine, an oligosaccharide, a terpene, an an NRPS are present within mostly distant populations

and show a negative bias towards both PLR and BCI populations.

4.4 Discussion

Population genomic analysis of *Pseudonocardia* associated with fungus-growing ants demonstrates that extremely closely related bacteria can exhibit both gene content diversity and distinct population dynamics, and that these properties vary at kilometer scale geographic space. The acquisition of secondary metabolite clusters, and the recombination between homologous clusters, may be critical in arming *Pseudonocardia* with small molecules in their evolutionary race with the fungus garden pathogen *Escovopsis*. The distribution of secondary metabolites suggests that these external symbionts still undergo genetic exchange with other organinisms, unlike intra-cellular insect symbionts which are largely isolated from the environment (Moran et al., 2008). Our results indicate that secondary metabolite biosynthesis genes are particularly abundant on contigs that do not map to closely related reference genomes, and that gene acquisition and loss often affects multiple genes rather than single loci. Although previous research has suggested much of this gene acquisition is driven by plasmids, the relatively poor quality of our Illumina draft genomes makes it difficult to distinguish plasmid contigs from mobile DNA integrated in the chromosome.

Tajima's D analysis of different strain groups revealed striking differences in genetic diversity. The mainland isolated PLR-B1 and PLR-B2 populations had significantly different mean Tajima's D values, which may indicate recent population bottlenecks or balancing selection across the chromosome in the PLR-B1 population. The mean for all 42 genomes together may suggest purifying selection across the genome as a whole in these *Pseudono-*

cardia. The observation that Tajima's D values can differ significantly between subgroups of the dataset versus the entire dataset suggests that fine-scale sampling of microbial lineages can reveal population dynamics that are obscured when closely related organisms are treated as a single group.

The large number of monophyletic gene trees that separate the BCI lineage from the mainland lineage provide strong support for genetic isolation of many loci between these lineages. This observation is particularly important when investigating evolutionary independence between closely related bacterial lineages. As homologous recombination occurs in relatively small stretches of DNA rather than across the entire chromosome, bacterial populations can be genetically isolated at some loci while recombining at others. Genetic isolation at genes under selection in different environments could enable divergence of two populations, even as they continue to exchange alleles in genes not under differential selection. Identification and characterization of genes that diverge between closely related organisms can inform hypotheses about the selective pressures that differ between environments or ecological strategies.

Conserved gene sequence similarity and gene content diversity are largely consistent with isolation location. BCI-isolated *Pseudonocardia* are highly similar, forming a single lineage that is distinct from most mainland isolated strains. The only BCI-isolated strains that do not share very high sequence similarity to the others instead share high identity with some mainland isolates, suggesting continuing migration of ant hosts between the island and mainland. Similarly, two mainland isolated strains share high sequence similarity with the island strains. Gene content differences between strains are more variable, with strains from the BCI populations differing from mainland strains by around 15%. Between BCI

strains, gene content differences range from 188 to 2,063.

The distribution of secondary metabolite BGC families also follows isolation location at this fine geographic scale, as a number of BGC families we identified were found in BCI-isolated strains. This may be due to geographic diversity of fungus-growing ant pathogens. The acquisition of several of these clusters by the recent migrant strain *Pseudonocardia* sp. EC080619-08 may suggest several ecological and evolutionary processes, which are not mutually exclusive. It may suggest acquisition of ecologically relevant genes can be relatively rapid, if migration of the ant host was followed by gene acquisition soon after. It may also suggest that strong selection from pathogens or other environmental conditions prevents the colonization of BCI by ant hosts whose *Pseudonocardia* lack the ability to produce particular small molecules. Higher numbers of samples from both mainland and BCI populations would help address this question by providing a more comprehensive view of BGC family conservation and diversity in both locations. Additional studies on migration and survival of new *Apterostigma* ant colonies would also shed light on the dynamics of host dispersal and survival in new geographic areas.

Model organisms with discrete, physically separated niches, such as host-associated microbes or those that live in unusual environments, provide opportunities for investigating population dynamics and biogeography in microbes by separating microbial strains and communities on physical scales that are amenable to sampling. Deeper understanding of population-level biology for microbes generally will require applying theoretical and experimental knowledge from this scale to the microscopic physical scale at which most microbial communities interact and evolve. Although population dynamics in mesophilic microbes can be observed at the macro scale, there is likely a vast reservoir of micro-scale diversity

within complex environments that has yet to be studied. These ant-associated *Pseudonocardia* strains show variable population dynamics and secondary metabolite biosynthesis potential, structured by their geography. Therefore, they do not appear to be ecologically coherent groups. Approaches to identify ecologically coherent groups of microbes by sequence similarity alone (Caro-Quintero and Konstantinidis, 2012) would not be able to distinguish these fine-scale microbial groups, as they share extremely high core gene nucleotide similarity. Further analysis of this diversity is important not only for targeted sampling strategies in enzyme discovery and drug development, but for our understanding of microbial ecology and evolution as a whole.

4.5 Methods

Genome assembly and annotation. *Pseudonocardia* strains were isolated from the cuticle of fungus-growing ants and sequenced using either Pacific Biosciences technology at Duke University (EC080625-04) or Illumina at Washington University in St. Louis. PacBio assemblies were performed using HGAP 1.4 (Chin et al., 2013), while Illumina genomes were assembled using Velvet (Zerbino and Birney, 2008) by Wash. U. Protein coding genes for all genomes were predicted *de-novo* using Prodigal v2.60 (Hyatt et al., 2010), while ribosomal RNAs were predicted using RFAM (Nawrocki et al., 2015) hidden markov models and Infernal 1.1.1 (Nawrocki and Eddy, 2013). Protein coding genes were annotated using TIGRFam v15 (Haft et al., 2013) , PFAM v29, KEGG, and actNOG (Powell et al., 2012) hidden markov models and HMMer 3.1 (Eddy, 2011). Secondary metabolite clusters were identified in each genome by antiSMASHv3 (Blin et al., 2013) followed by manual curration

of cluster boundaries. Secondary metabolite gene clusters were grouped into families by 80% nucleotide identity via nucmer alignment and 50% coverage for each segment of a cluster that matched another.

Population clustering. SNP identification was performed by aligning all Illumina genome assemblies to the PacBio assembly of EC080625-04 using nucmer. For fineSTRUCTURE, only reference SNP positions that were covered by a contig for every genome were used, i.e. no missing data for any SNP positions. Multiple runs with varying values of c and estimated population size had little effect on overall strain clustering, except for very high values of c merging neighboring clusters together.

Conserved gene analysis. Tajima's D and monophyletic gene analyses were conducted using genes from EC080625-04 and the matching regions in contigs from other genomes aligned to this reference. Tajima's D was calculated for genes that had at least 1% polymorphic nucleotide sites within each population. Monophyletic genes were identified by generating nucleotide alignment using MAFFT v7.221 (Katoh and Standley, 2013) followed by gene phylogenies generated using FastTree 2.0 (Price et al., 2010).

4.6 Figures

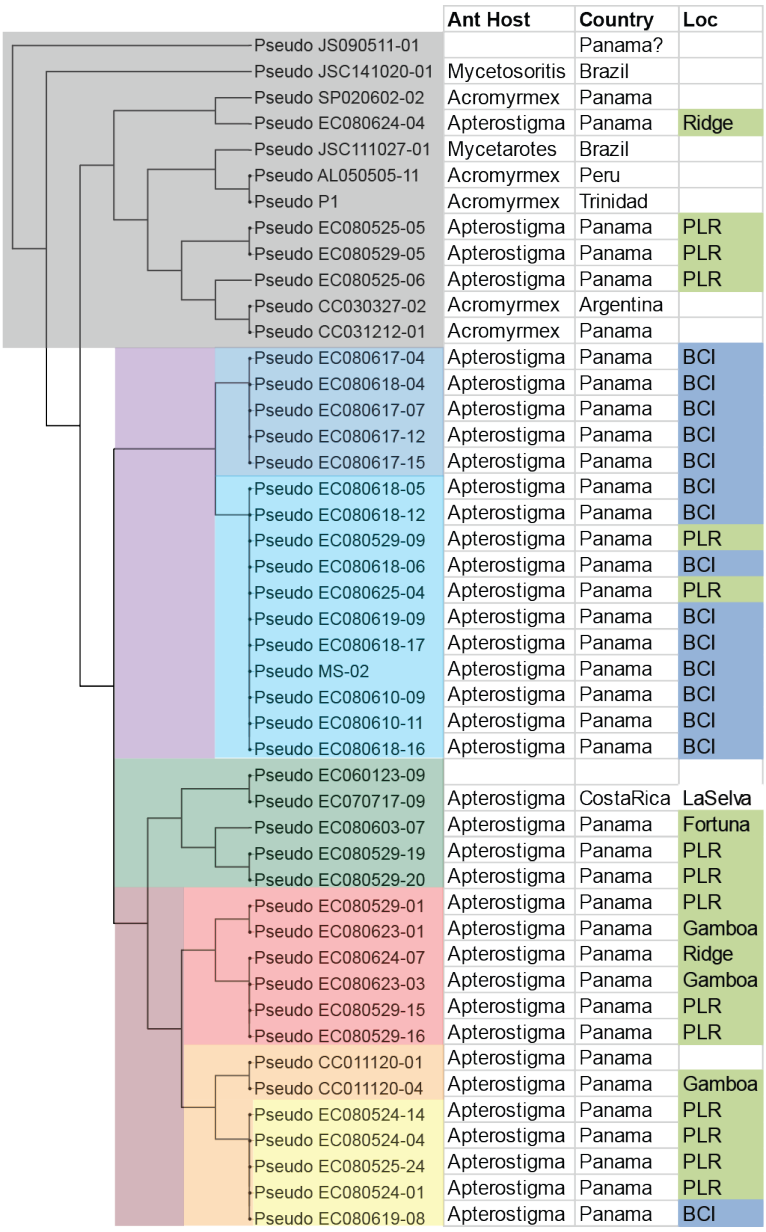

Figure 4.1: **SNP-based population clustering of ant-associated *Pseudonocardia*** fineSTRUCTURE clustering of *Pseudonocardia* strains using 270,000 polymorphic SNP positions. Ant host and isolation location are shown to the right. Lineage colors are used throughout the study.

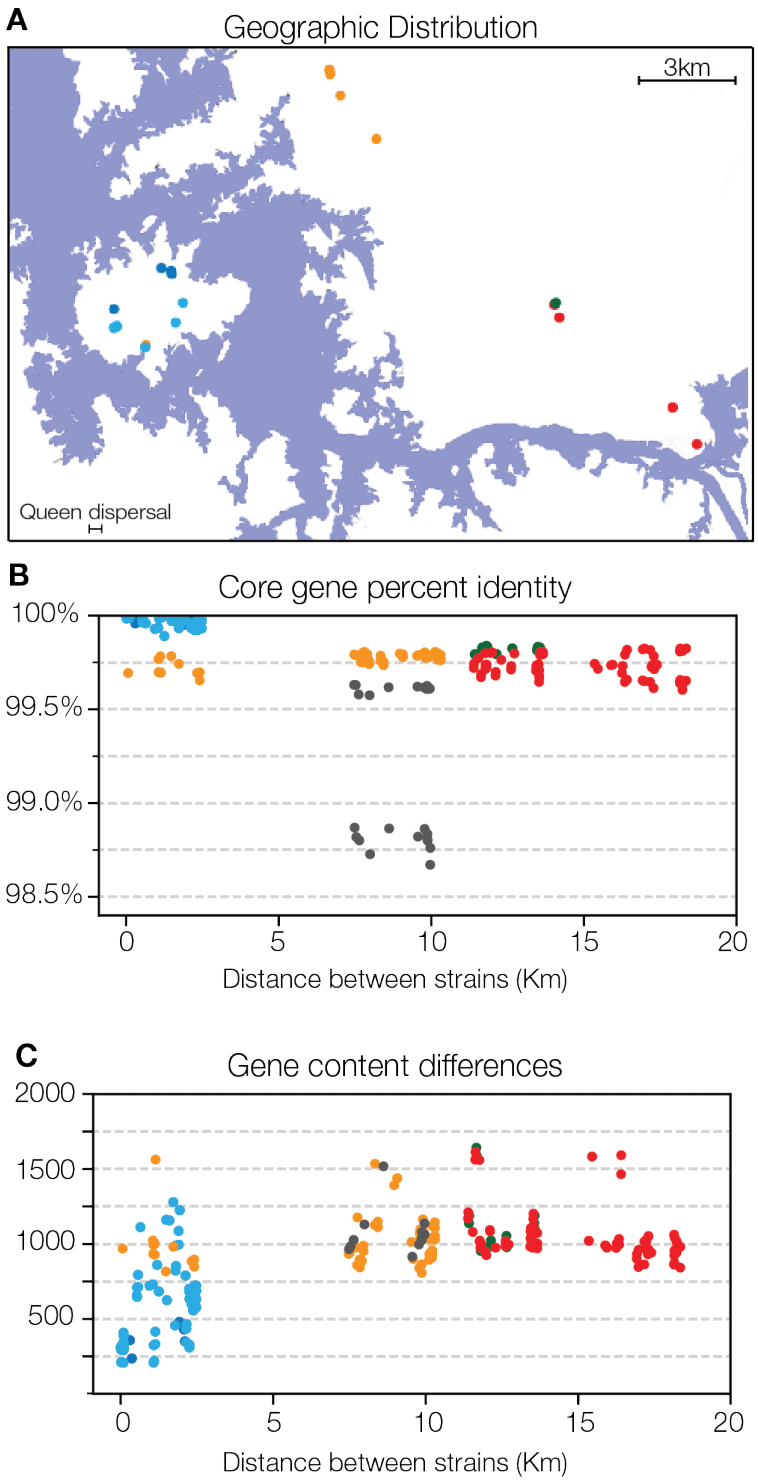

Figure 4.2: **Geographic distribution and genetic diversity A.** Isolation locations based on available GPS data are shown for *Pseudonocardia* strains in the area around Barro Colorado Island. Strains are colored by fineSTRUCTURE population (see Figure 4.1). **B.** Core gene percent identity between BCI population strains (purple in Figure 4.1) and all strains with GPS coordinates. Points are colored by strain population. **C.** Raw number of gene content differences between BCI population strains and all others with GPS coordinates. Points are by strain population.

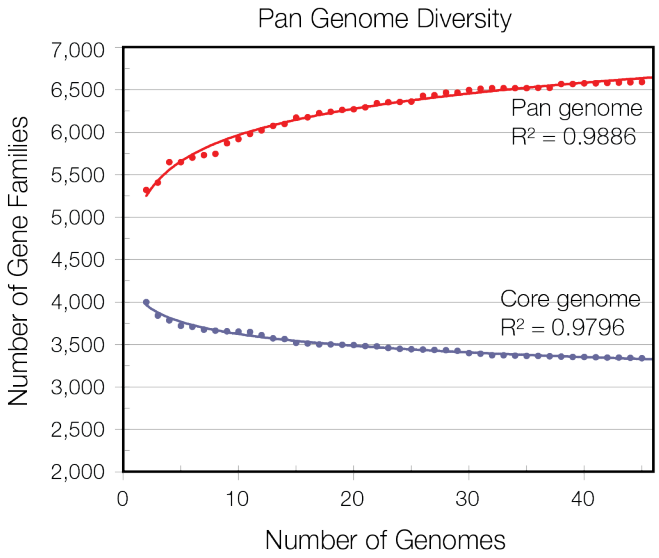

Figure 4.3: **Pan-genome and core-genome size in closely related ant-associated *Pseudonocardia*.** The total number of actNOG gene families present in the genome dataset (red) and the number of actNOG gene families conserved in all genomes (blue), as more genomes are added to the dataset. The Pan-genome size fits a logarithmic function, while the core-genome size fits a power-law function.

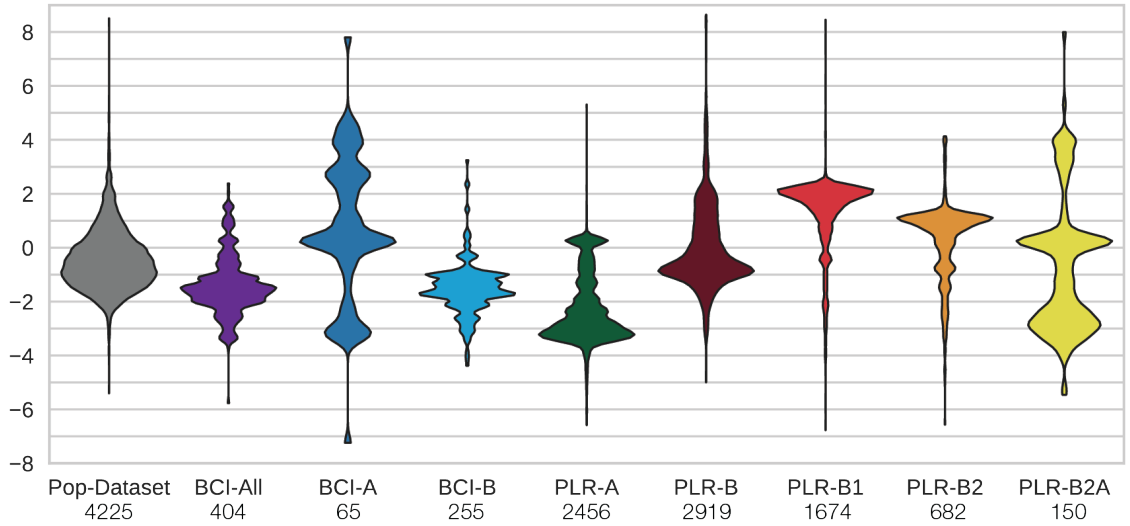

Figure 4.4: **Genetic diversity of conserved genes** Each violin plot shows the distribution of Tajima's D values for the specified strain group in genes with at least 1% polymorphic sites. The number below each strain group name indicates the number of genes analyzed.

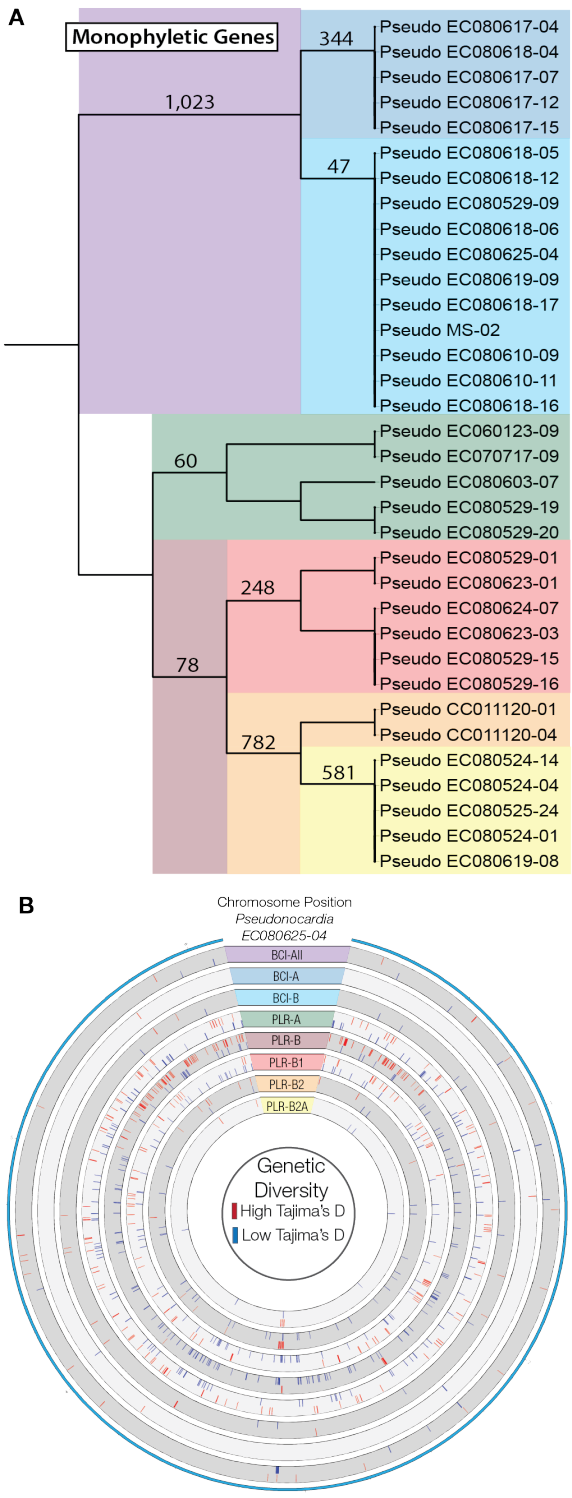

Figure 4.5: **Population-specific gene divergence.** **A.** The number of genes that exhibit monophyly for the strains in each lineage are shown on the appropriate branch. **B.** Conserved genes in the top or bottom 5% of Tajima's D values in each population are marked at their chromosomal location in *Pseudonocardia* EC080625-04, in red or blue respectively.

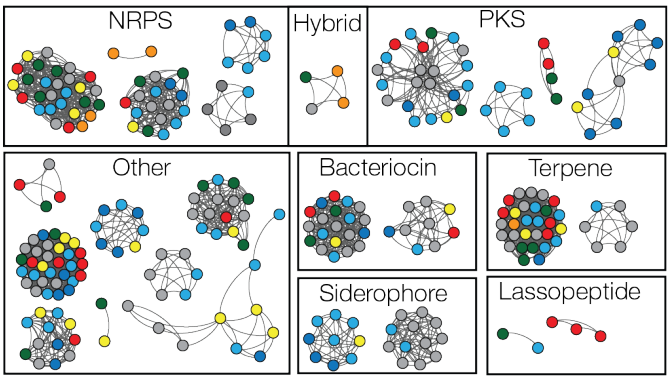

Figure 4.6: **Conserved secondary metabolite biosynthetic gene clusters.** Each node represents a contiguous set of genes that are part of a secondary metabolite biosynthesis cluster. Edges represent at least 80% nucleotide sequence identity and 50% coverage between gene sets. Nodes are colored by their population of origin (see Figure 4.1).

Chapter 5

Conclusions and Future Directions

Ultimately, integrating microbes into an evolutionary framework that addresses species concepts, phylogenetics, and diversification dynamics will require a deeper understanding of genetic exchange. Together the studies included in this dissertation demonstrate the power of investigating genetic exchange across a range of phylogenetic scales. Both short term dynamics and long term consequences of genetic exchange have significant impact on the diversification and adaptation of microbial lineages. Analysis of lineages at a variety of scales provides insight into different aspects of genetic exchange and microbial evolution.

In Chapter 1 I argue that the temporal and phylogenetic scale of sampling and analysis is critically important for our understanding of ecological and evolutionary patterns in microbes, and the processes that generate these patterns. Genome sequencing technologies have enabled much deeper insights into microbial ecology and evolution, but a few key aspects of microbial life in nature are poorly understood. The ecological diversity of microbial lineages, the rates of genetic exchange that constrain evolutionary independence, and the diversity of niches in seemly homogenous environments are each critical parameters that deserve intense study. Continued development of new technologies, both computational and laboratory-based, provide great opportunity to clarify these aspects of microbiology.

I analyze general metabolic and genomic properties across multiple cultivated phyla in Chapter 2. This analysis reveals conservation of both core metabolism and genomic functions that are generally thought to vary according to ecological niche, including membrane transporters and signaling proteins. The evolution of high 16S copy number is also limited primarily to subgroups of the Gammaproteobacteria and Firmicutes. These patterns strengthen the idea that barriers to lateral gene transfer constrain the genomic diversification of bacterial lineages. Although such barriers have been studied in a few model organisms, further work can provide significant insight by identifying physical mechanisms, epistatic interactions, and selective pressures that structure the distribution of laterally transferred genes at both large and small phylogenetic scales.

In Chapter 3 I investigate the rate and distribution of laterally transferred genes in the genus *Streptomyces*, interpreting LGT events in a strong temporal framework. This temporal framework reveals that the rate of successful LGT events, genes that are acquired and subsequently retained over long evolutionary time periods, is surprisingly low. Many genes acquired through LGT are lost relatively quickly, suggesting that neutral turnover of transferred genes is a significant but understudied aspect of LGT dynamics. This observation indicates the need for a neutral theory of LGT that would provide an expected number of gene acquisitions and an expected age of acquired genes if the acquired genes had no selective benefit. The observed dynamics of natural product biosynthesis gene clusters also suggest that different patterns may be observed at different phylogenetic scales. I found that most biosynthesis clusters were a mix of vertically inherited and laterally acquired genes. I also found that laterally acquired genes in a single cluster were rarely acquired from one donor. This is in contrast to analysis of LGT dynamics in *Salinispora* biosynthetic

clusters, which were primarily acquired as whole clusters from a single source. Because of differences in age between the two genera, I hypothesize that the *Salinispora* study captures short-term evolutionary dynamics while my analysis of *Streptomyces* reveals longer term evolution of these clusters.

Chapter 4 provides a fine scale view of evolutionary dynamics by investigating genome content, selection, and population dynamics in ant-associated *Pseudonocardia*. Spatial structuring of *Pseudonocardia* populations provided by the dispersal of their ant host provides a unique opportunity to investigate the impact of migration and adaptation in a microbe that does not live in an extreme environment such as a hot spring or hydrothermal vent. The limited amount of gene families found within the local population suggests that LGT does not provide as large a source of genetic potential in these organisms as has been suggested for more diverse free living organisms such as E. coli. The distribution and diversity of natural product biosynthesis gene clusters, structured by both population and geographic factors, indicates that a targeted approach for small molecule discovery may increase efficiency. Further studies will need to correlate this biosynthesis potential to the diversity of small molecules that are actually produced. In addition, analysis of other microbes such as *Streptomyces* at a similarly fine phylogenetic scale are critical for understanding how generalizable the *Pseudonocardia* observations may be.

Appendix A

Chapter 2 Supplementary Materials

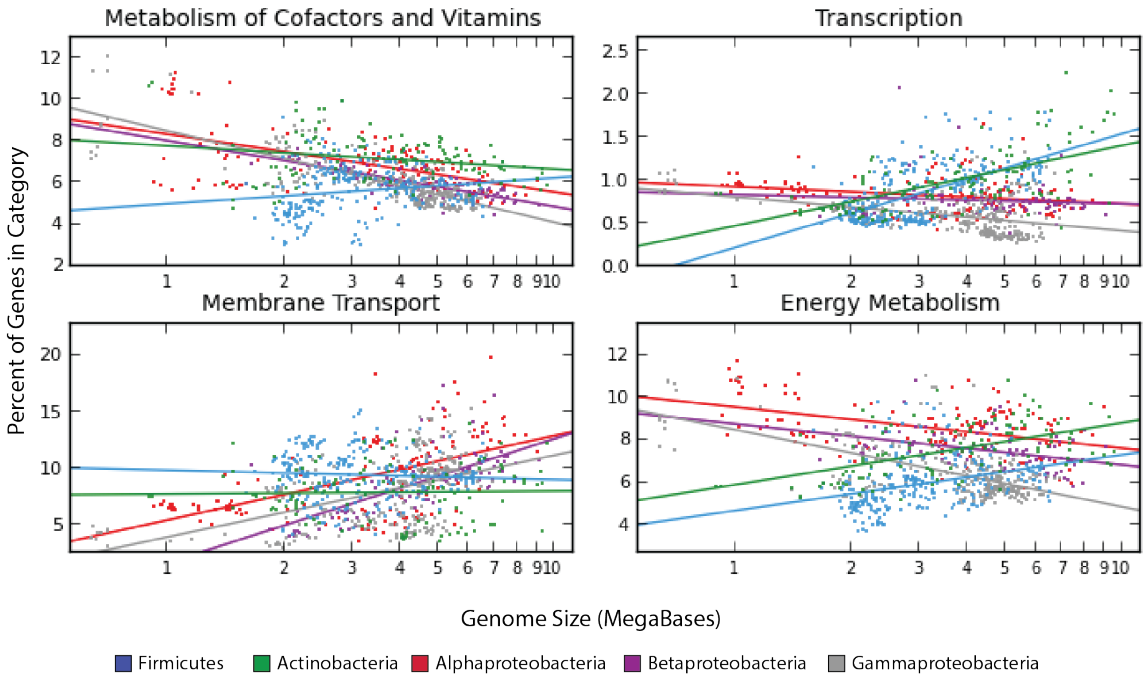

Figure A.1: **Correlation between genome size and genome content across well sampled bacterial phyla.** In genomes with more than 600 genes, the genomic proportion dedicated to each KEGG gene category is plotted versus genome size.

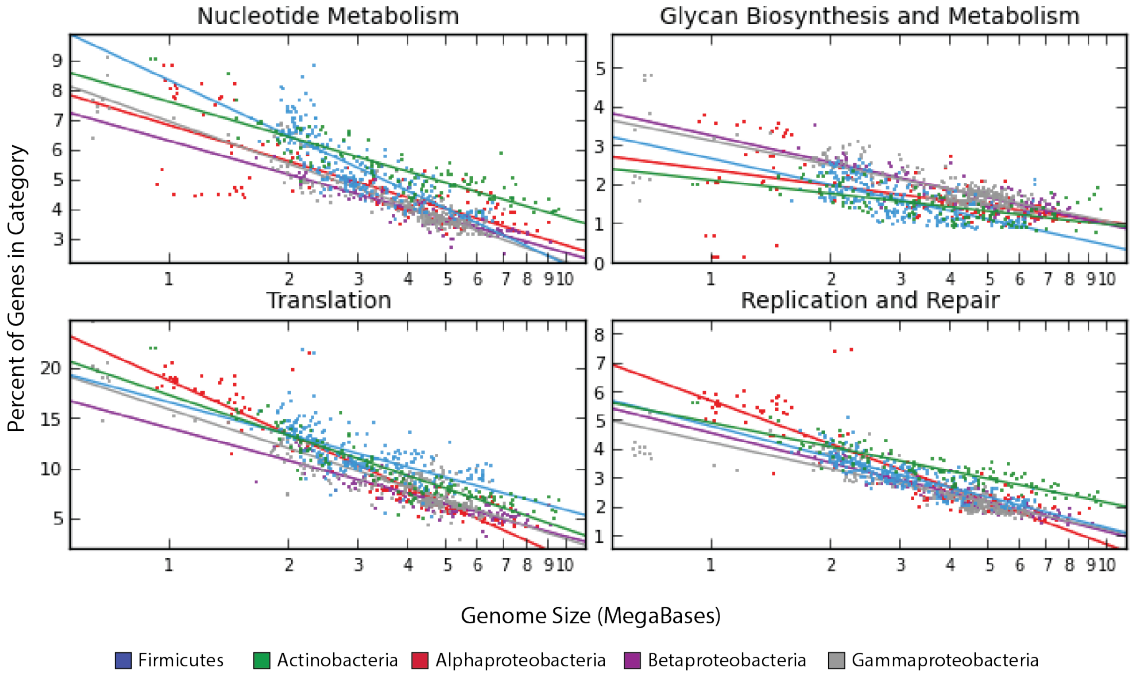

Figure A.2: **Correlation between genome size and genome content across well sampled bacterial phyla.** In genomes with more than 600 genes, the genomic proportion dedicated to each KEGG gene category is plotted versus genome size.

Appendix B

Chapter 3 Supplementary Materials

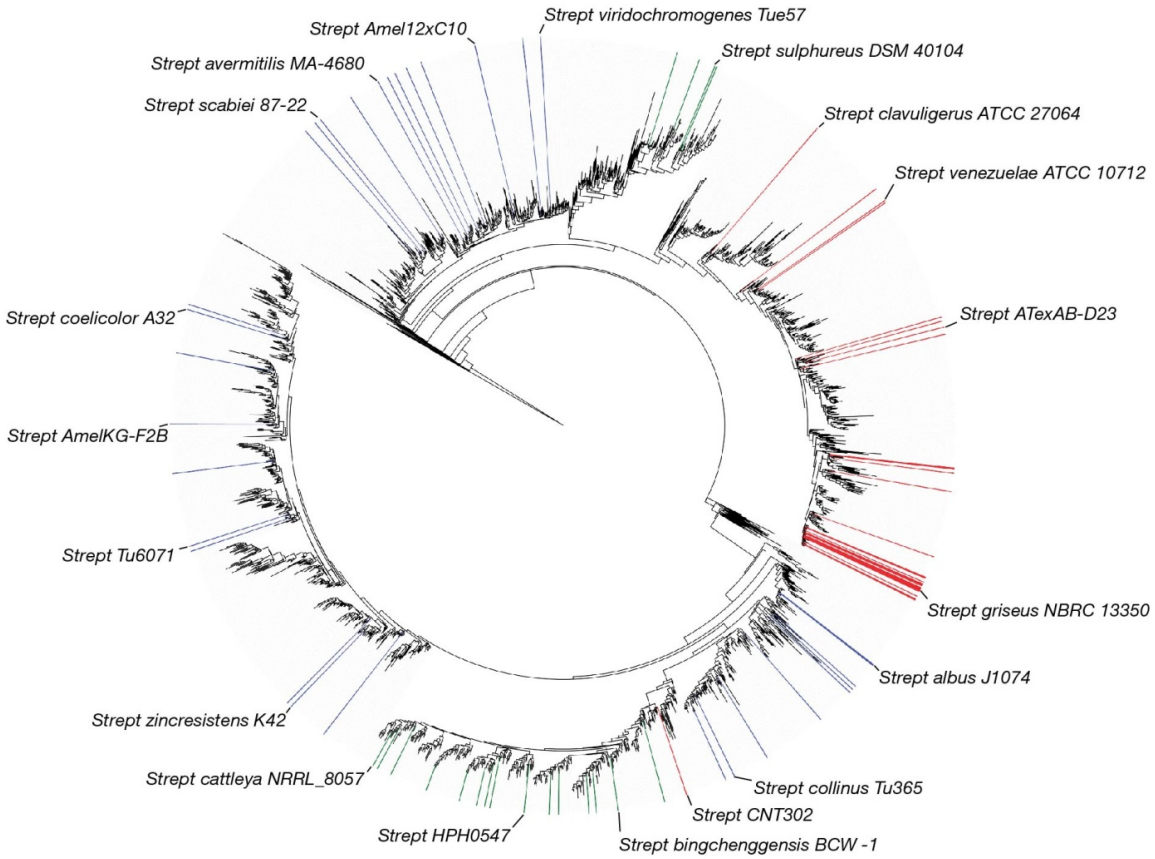

Figure B.1: **16S rRNA gene phylogeny of the genus *Streptomyces*.** Genome strains are colored by phylogenetic clade in the Figure 1 multilocus phylogeny, and several are labeled for reference. *Streptomyces* is abbreviated as Strept.

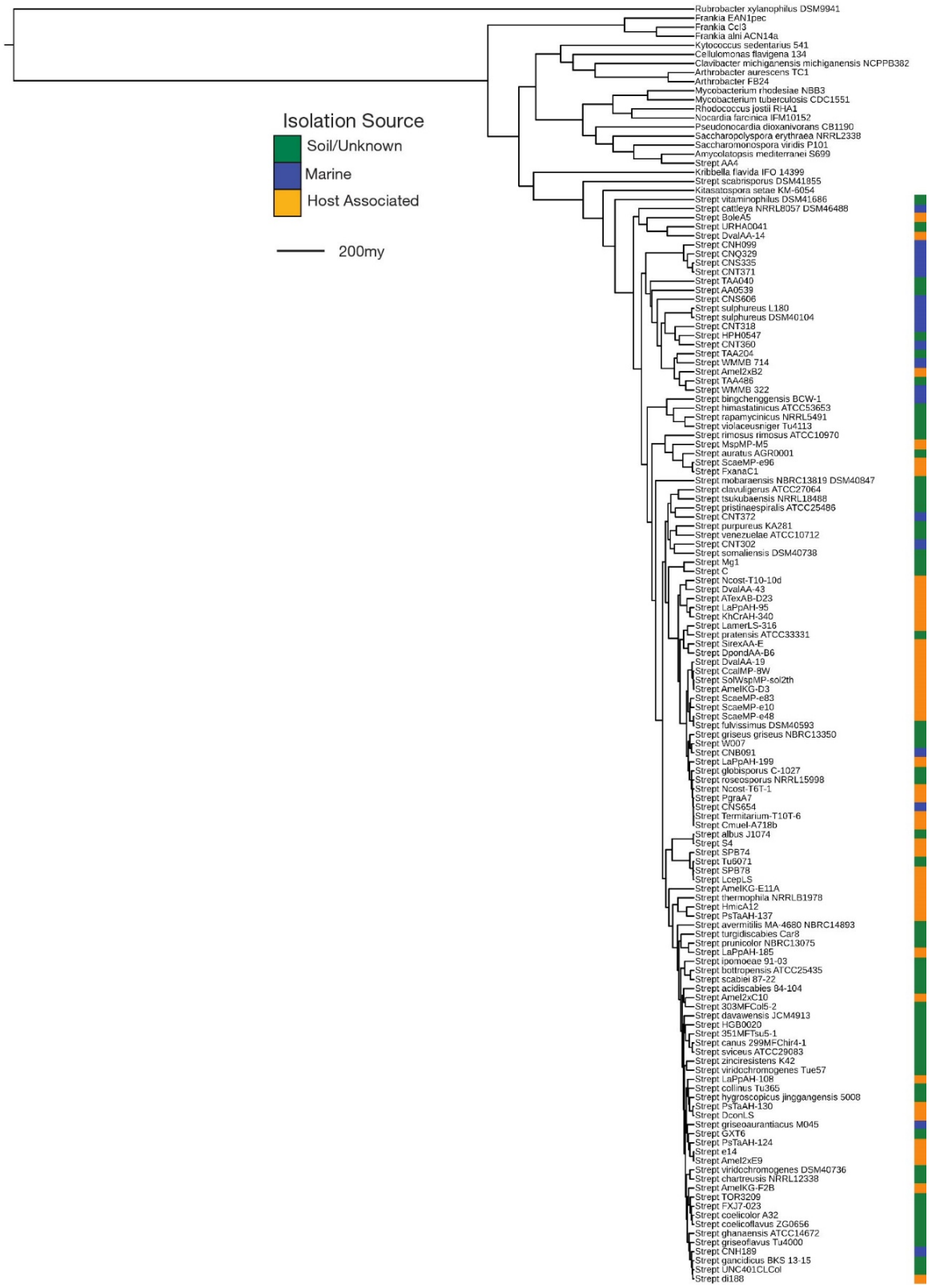

Figure B.2: **Full multilocus tree of *Streptomyces* and outgroup Actinobacteria.** This phylogeny includes the full set of genomes used for AnGST analysis. *Streptomyces* strains are colored by isolation habitat. Scale bar indicates molecular clock divergence times estimated by Reltime. *Streptomyces* is abbreviated as Strept.

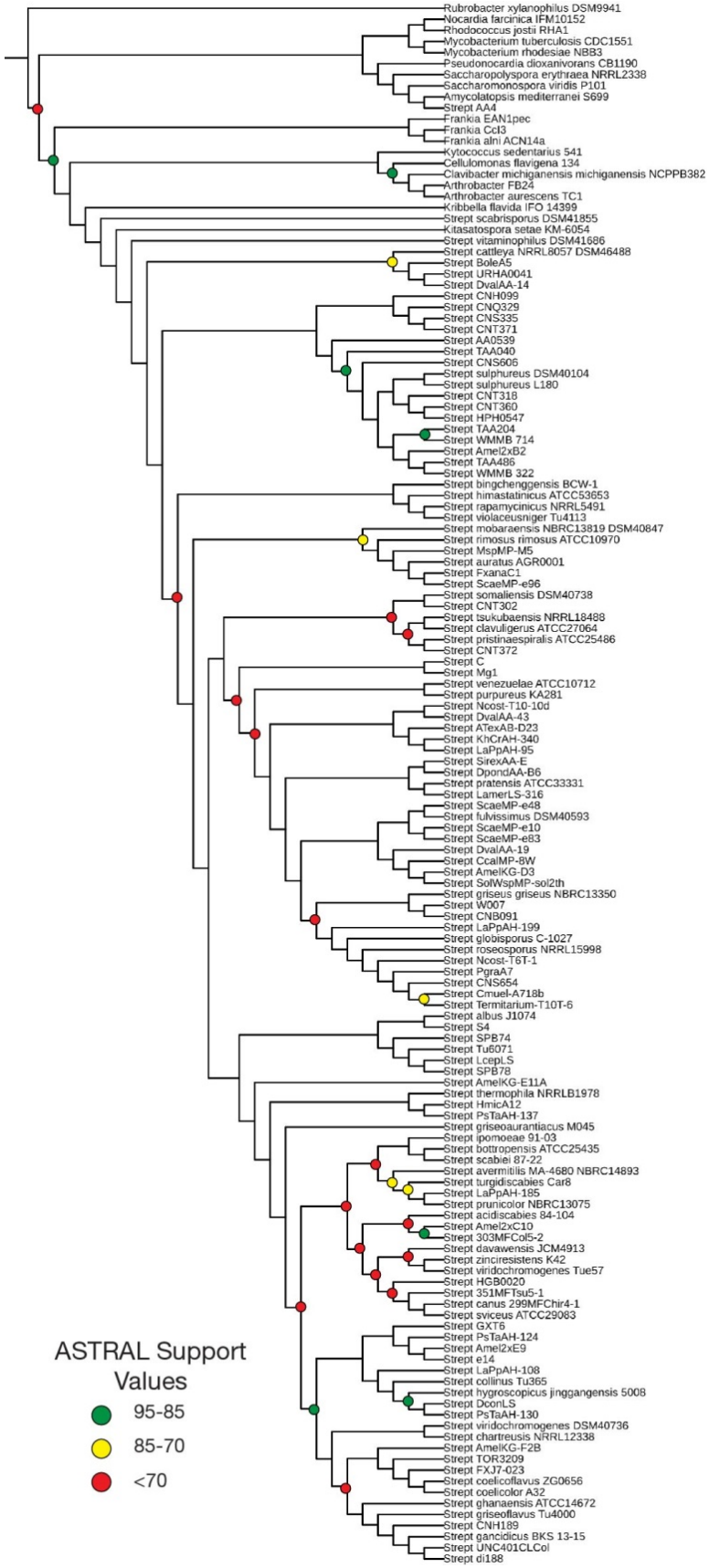

Figure B.3: **Gene-tree reconciliation based *Streptomyces* phylogeny generated with ASTRAL-II.** TIGRFAM gene families used for gene tree construction were the same as those used for the multilocus analysis. *Streptomyces* is abbreviated as Strept.

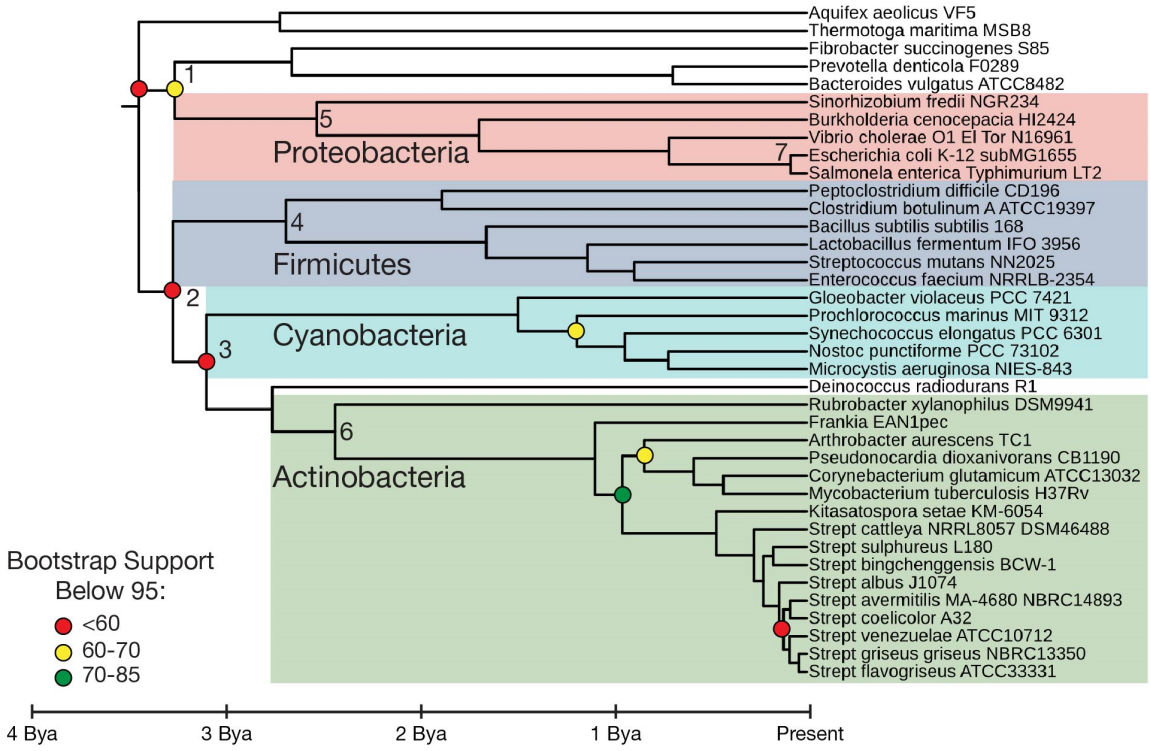

Figure B.4: **Molecular clock phylogeny of all Bacteria.** Phylogeny of all Bacteria used to calibrate *Streptomyces* divergence times. All unlabeled nodes have a bootstrap value $\geqslant$95. Numbered nodes refer to Reltime divergence times that are compared to the molecular clock analysis performed by Battistuzzi et al. (2004), found in Supplementary Table B.1. *Streptomyces* is abbreviated as Strept.

Node	Reltime	BFU et. al.	Reltime Confidence Interval	BFU et. al. Credibility Interval
1	3463	3054	1806-5121	2697-3490
2	3454	3051	1806-5121	2738-3434
3	3281	2784	1710-4852	2490-3203
4	2847	2508	1487-4208	2154-2928
5	2677	1924	1396-3958	1809-2525
6	2578	2743	1345-3811	2512-3076
7	87	102	45-130	57-176

Table B.1: Comparison of molecular clock estimated divergence times

KEGG category	P value	Odds ratio
Metabolism of Terpenoids and Polyketides	4.70E-108	1.421
Xenobiotics Biodegradation and Metabolism	2.12E-127	1.351
Biosynthesis of Other Secondary Metabolites	1.50E-015	1.183
Lipid Metabolism	7.01E-017	1.12

Table B.2: KEGG gene classes over-represented in LGT events

KEGG category	P value	Odds ratio
Replication and Repair	9.80E-095	0.638
Transcription	2.57E-020	0.652
Nucleotide Metabolism	4.61E-100	0.724
Cell Growth and Death	1.92E-013	0.762
Glycan Biosynthesis and Metabolism	5.43E-026	0.783
Metabolism of Cofactors and Vitamins	1.23E-050	0.839
Energy Metabolism	8.67E-051	0.842
Translation	6.41E-009	0.911
Metabolism of Other Amino Acids	3.96E-007	0.917
Carbohydrate Metabolism	1.22E-018	0.925
Membrane Transport	3.77E-007	0.943
Signal Transduction	0.038	0.973
Amino Acid Metabolism	0.012	0.981

Table B.3: KEGG gene classes under-represented in LGT events

LGT Source	Genes acquired per million years Clade I	Clade II
Clade I *Streptomyces*	3.816	1.082
Clade II *Streptomyces*	1.434	7.065
Basal *Streptomyces*	0.49	0.651
Other Actinobacteria	0.189	0.277

Table B.4: Rate of LGT into Clade I and Clade II

Organism	Data source	NCBI Bioproject
Streptomyces sp AmelKG-F2B	Currie lab	in prep
Streptomyces sp thermophila NRRLB1978	Currie lab	in prep
Streptomyces sp LaPpAH-185	Currie lab	PRJNA302566
Streptomyces sp Ame12xE9	Currie lab	PRJNA201126
Streptomyces sp Amel2xB2	Currie lab	PRJNA195845
Streptomyces sp Amel2xC10	Currie lab	PRJNA195842
Streptomyces sp AmelKG-D3	Currie lab	PRJNA163015
Streptomyces sp AmelKG-E11A	Currie lab	PRJNA163025
Streptomyces sp ATexAB-D23	Currie lab	PRJNA199252
Streptomyces sp BoleA5	Currie lab	PRJNA169757
Streptomyces sp CcalMP-8W	Currie lab	PRJNA199236
Streptomyces sp Cmuel-A718b	Currie lab	PRJNA319231
Streptomyces sp DconLS	Currie lab	PRJNA319222
Streptomyces sp di188	Currie lab	PRJNA319209
Streptomyces sp DpondAA-B6	Currie lab	PRJNA195846
Streptomyces sp DvalAA-14	Currie lab	PRJNA319210
Streptomyces sp DvalAA-19	Currie lab	PRJNA319214
Streptomyces sp DvalAA-43	Currie lab	PRJNA319215
Streptomyces sp e14	Currie lab	PRJNA47353
Streptomyces sp e83	Currie lab	PRJNA319216
Streptomyces sp FxanaC1	Currie lab	PRJNA199074
Streptomyces sp KhCrAH-340	Currie lab	PRJNA199243
Streptomyces sp LamerLS-316	Currie lab	PRJNA169718
Streptomyces sp LaPpAH-108	Currie lab	PRJNA199251
Streptomyces sp LaPpAH-199	Currie lab	PRJNA302568
Streptomyces sp LaPpAH-95	Currie lab	PRJNA199076
Streptomyces sp LcepLS	Currie lab	PRJNA319218
Streptomyces sp MspMP-M5	Currie lab	PRJNA199249
Streptomyces sp Ncost-T10-10d	Currie lab	PRJNA319230
Streptomyces sp Ncost-T6T-1	Currie lab	PRJNA163023
Streptomyces sp PgraA7	Currie lab	PRJNA169756
Streptomyces sp PsTaAH-124	Currie lab	PRJNA199254
Streptomyces sp PsTaAH-130	Currie lab	PRJNA195843
Streptomyces sp PsTaAH-137	Currie lab	PRJNA195844

Streptomyces sp ScaeMP-e10	Currie lab	PRJNA199241
Streptomyces sp ScaeMP-e48	Currie lab	PRJNA163017
Streptomyces sp ScaeMP-e96	Currie lab	PRJNA163037
Streptomyces sp SirexAA-E	Currie lab	PRJNA72627
Streptomyces sp So1WspMP-so12th	Currie lab	PRJNA169755
Streptomyces sp SPB74	Currie lab	PRJNA48415
Streptomyces sp SPB78	Currie lab	PRJNA55819
Streptomyces sp Termitarium-T10T-6	Currie lab	PRJNA319232
Streptomyces sp WMMB 322	Currie lab	PRJNA187213
Streptomyces sp WMMB 714	Currie lab	PRJNA187214
Streptomyces albus J1074	NCBI	PRJNA196849
Streptomyces auratus AGR0001	NCBI	PRJNA171646
Streptomyces avermitilis MA4680	NCBI	PRJNA57739
Streptomyces bingchenggensis BCW-1	NCBI	PRJNA82931
Streptomyces coelicoflavus ZG0656	NCBI	PRJNA180030
Streptomyces coelicolor A3-2	NCBI	PRJNA57801
Streptomyces ghanaensis ATCC14672	NCBI	PRJNA55543
Streptomyces griseoaurantiacus M045	NCBI	PRJNA66149
Streptomyces griseoflavus Tu4000	NCBI	PRJNA55831
Streptomyces griseus NBRC13350	NCBI	PRJNA58983
Streptomyces hygroscopicus ATCC53653	NCBI	PRJNA33605
Streptomyces hygroscopicus jinggangensis 5008	NCBI	PRJNA89409
Streptomyces ipomoeae 91-03	NCBI	PRJNA183480
Streptomyces pristinaespiralis ATCC25486	NCBI	PRJNA59511
Streptomyces roseosporus NRRL15998	NCBI	PRJNA55545
Streptomyces scabiei 87-22	NCBI	PRJNA46531
Streptomyces scabrisporus DSM41855	NCBI	PRJNA199206
Streptomyces sp 303MFCol5-2	NCBI	PRJNA187949
Streptomyces sp 351MFTsu5-1	NCBI	PRJNA187950
Streptomyces sp AA0539	NCBI	PRJNA199548
Streptomyces sp AA4	NCBI	PRJNA33599
Streptomyces sp acidiscabies 84-104	NCBI	PRJNA77031
Streptomyces sp bottropensis ATCC25435	NCBI	PRJNA176092
Streptomyces sp C	NCBI	PRJNA55823
Streptomyces sp canus 299MFChir4-1	NCBI	PRJNA187948
Streptomyces sp cattleya NRRL8057 DSM46488	NCBI	PRJNA78941

Streptomyces sp chartreusis NRRL12338	NCBI	PRJNA72673
Streptomyces sp clavuligerus ATCC27064	NCBI	PRJNA19249
Streptomyces sp CNB091	NCBI	PRJNA199379
Streptomyces sp CNH099	NCBI	PRJNA169791
Streptomyces sp CNH189	NCBI	PRJNA169772
Streptomyces sp CNQ329	NCBI	PRJNA190863
Streptomyces sp CNS335	NCBI	PRJNA199338
Streptomyces sp CNS606	NCBI	PRJNA195771
Streptomyces sp CNS654	NCBI	PRJNA239507
Streptomyces sp CNT302	NCBI	PRJNA199358
Streptomyces sp CNT318	NCBI	PRJNA187951
Streptomyces sp CNT360	NCBI	PRJNA187952
Streptomyces sp CNT371	NCBI	PRJNA169776
Streptomyces sp CNT372	NCBI	PRJNA199339
Streptomyces sp collinus Tu365	NCBI	PRJNA171216
Streptomyces sp davawensis JCM4913	NCBI	PRJEB184
Streptomyces sp fulvissimus DSM40593	NCBI	PRJNA192408
Streptomyces sp FXJ7-023	NCBI	PRJNA189794
Streptomyces sp gancidicus BKS 13-15	NCBI	PRJNA186841
Streptomyces sp globisporus C-1027	NCBI	PRJNA158251
Streptomyces sp GXT6	NCBI	PRJNA178392
Streptomyces sp HGB0020	NCBI	PRJNA72491
Streptomyces sp HmicA12	NCBI	PRJNA169744
Streptomyces sp HPH0547	NCBI	PRJNA169487
Streptomyces sp Mg1	NCBI	PRJNA207881
Streptomyces sp mobaraensis NBRC13819 DSM40847	NCBI	PRJNA188290
Streptomyces sp pratensis ATCC33331	NCBI	PRJNA33771
Streptomyces sp prunicolor NBRC13075	NCBI	PRJDB1071
Streptomyces sp purpureus KA281	NCBI	PRJNA157921
Streptomyces sp rapamycinicus NRRL5491	NCBI	PRJNA207502
Streptomyces sp rimosus rimosus ATCC10970	NCBI	PRJNA182749
Streptomyces sp S4	NCBI	PRJNA78151
Streptomyces sp somaliensis DSM40738	NCBI	PRJNA81125
Streptomyces sp sulphureus DSM40104	NCBI	PRJNA182442
Streptomyces sp TAA040	NCBI	PRJNA187955
Streptomyces sp TAA204	NCBI	PRJNA188326

Streptomyces sp TAA486	NCBI	PRJNA190864
Streptomyces sp TOR3209	NCBI	PRJNA198964
Streptomyces sp Tu6071	NCBI	PRJNA66919
Streptomyces sp turgidiscabies Car8	NCBI	PRJNA42361
Streptomyces sp UNC401CLCol	NCBI	PRJNA234927
Streptomyces sp URHA0041	NCBI	PRJNA213783
Streptomyces sp viridochromogenes Tue57	NCBI	PRJNA89157
Streptomyces sp vitaminophilus DSM41686	NCBI	PRJNA199207
Streptomyces sp W007	NCBI	PRJNA80699
Streptomyces sulphureus L180	NCBI	PRJNA200371
Streptomyces sviceus ATCC29083	NCBI	PRJNA59513
Streptomyces tsukubaensis NRRL18488	NCBI	PRJNA162933
Streptomyces venezuelae ATCC10712	NCBI	PRJNA177080
Streptomyces violaceusniger Tu4113	NCBI	PRJNA52609
Streptomyces viridochromogenes DSM40736	NCBI	PRJNA55829
Streptomyces zinciresistens K42	NCBI	PRJNA72955
Amycolatopsis mediterranei S699	NCBI	PRJNA170006
Aquifex aeolicusVF5	NCBI	PRJNA215
Arthrobacter aurescens TC1	NCBI	PRJNA12512
Arthrobacter FB24	NCBI	PRJNA12640
Bacillus subtilis subtilis 168	NCBI	PRJNA57675
Bacteroides vulgatus ATCC 8482	NCBI	PRJNA13378
Burkholderia cenocepacia HI2424	NCBI	PRJNA13918
Cellulamonas flavigena 134	NCBI	PRJNA19707
Clavibacter michiganensis NCPPB382	NCBI	PRJNA19643
Clostridium botulinum A ATCC19397	NCBI	PRJNA19517
Deinococcus radiodurans R1	NCBI	PRJNA57665
Enterococcus faecium NRRLB-2354	NCBI	PRJNA74725
Escherichia coli K-12 subMG1655	NCBI	PRJNA57779
Fibrobacter succinogenes S85	NCBI	PRJNA32617
Frankia alni ACN14a	NCBI	PRJNA17403
Frankia CcI3	NCBI	PRJNA250957
Frankia EAN1pec	NCBI	PRJNA13915
Gloeobacter violaceus PCC 7421	NCBI	PRJNA58011
Kitasatospora setae KM-6054	NCBI	PRJNA77027
Kribbella flavida IFO 14399	NCBI	PRJNA21089

Kytococcus sedentarius 541	NCBI	PRJNA21067
Lactobacillus fermentum IFO 3956	NCBI	PRJDA18979
Microcystis aruginosa NIES-843	NCBI	PRJDA27835
Mycobacterium rhodesiae NBB3	NCBI	PRJNA60027
Mycobacterium tuberculosis CDC1551	NCBI	PRJNA223
Nocardia farcinica IFM10152	NCBI	PRJNA13117
Nostoc punctiforme PCC 73102	NCBI	PRJNA216
Peptoclostridium difficile CD196	NCBI	PRJNA38037
Prevotella denticola F0289	NCBI	PRJNA49293
Prochlorococcus marinus MIT 9312	NCBI	PRJNA13910
Pseudonocardia dioxanivorans CB1190	NCBI	PRJNA40557
Rhodococcus jostii RHA1	NCBI	PRJNA13693
Rubrobacter xylanophilus DSM9941	NCBI	PRJNA10670
Saccharomonospora viridis P101	NCBI	PRJNA20835
Saccharopolyspora erythraea NRRL2338	NCBI	PRJEA18489
Salmonela enterica Typhimurium LT2	NCBI	PRJNA57799
Sinorhizobium fredii NGR234	NCBI	PRJNA59081
Streptococcus mutans NN2025	NCBI	PRJDA28997
Synechococcus elongatus PCC 6301	NCBI	PRJNA58235
Thermotoga maritima MSB8	NCBI	PRJNA57723
Vibrio cholerae O1 El Tor N16961	NCBI	PRJNA57623

Table B.5: NCBI accession numbers for all genomes

Appendix C

Chapter 4 Supplementary Materials

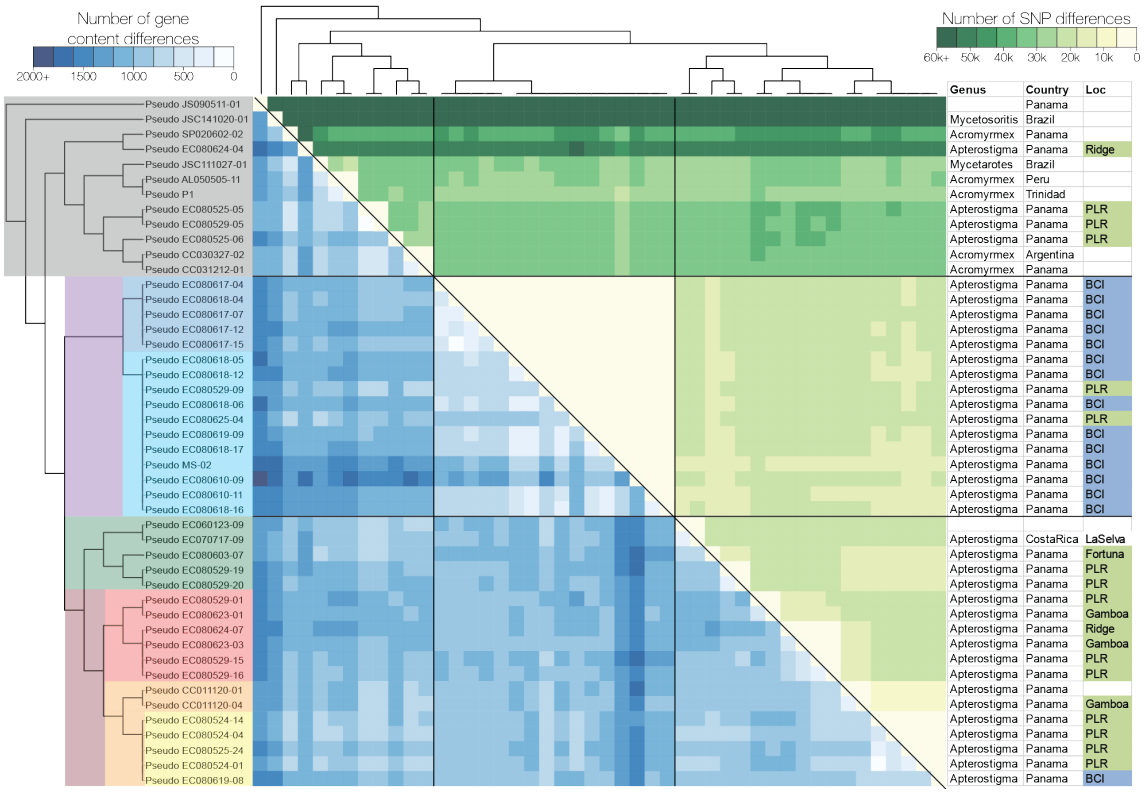

Figure C.1: **Genome content and SNP diversity in ant-associated *Pseudonocardia*** The number of actNOG gene content differences (blue) or SNP diffrences (green) are shown between every pair of genomes. Isolation hosts and locations are shown on the right.

References

Achtman, Mark, and Michael Wagner. 2008. Microbial diversity and the genetic nature of microbial species. *Nature reviews. Microbiology* 6(6):431–440.

Agashe, Deepa, Mrudula Sane, Kruttika Phalnikar, Gaurav D Diwan, Alefiyah Habibullah, N Cecilia Martinez-Gomez, Vinaya Sahasrabuddhe, William Polachek, Jue Wang, Lon M Chubiz, and Christopher J Marx. 2016. Large-effect beneficial synonymous mutations mediate rapid and parallel adaptation in a bacterium. *Molecular Biology and Evolution* msw035.

Ahmer, Brian M M, and John S. Gunn. 2011. Interaction of Salmonella spp. With the intestinal microbiota. *Frontiers in Microbiology* 2(MAY):1–9.

Akashi, H. 1994. Synonymous codon usage in Drosophila melanogaster: Natural selection and translational accuracy. *Genetics* 136(3):927–935.

Allison, Steven D, and Jennifer B H Martiny. 2008. Colloquium paper: resistance, resilience, and redundancy in microbial communities. *Proceedings of the National Academy of Sciences of the United States of America* 105 Suppl(Supplement_1):11512–9. NIHMS150003.

Altschul, S F, W Gish, W Miller, E W Myers, and D J Lipman. 1990. Basic local alignment search tool. *Journal of molecular biology* 215(3):403–10.

Amann, Rudolf, and Ramon Rosselló-Móra. 2016. After All, Only Millions? *mBio* 7(4): e00999–16.

Andam, Cheryl P., James R. Doroghazi, Ashley N. Campbell, Peter J. Kelly, Mallory J. Choudoir, and Daniel H. Buckley. 2016. A Latitudinal Diversity Gradient in Terrestrial Bacteria of the Genus Streptomyces. *mBio* 7(2):1–9.

Andam, Cheryl P, and J Peter Gogarten. 2011. Biased gene transfer in microbial evolution. *Nature reviews Microbiology* 9(7):543–555.

Bailey, Susan F, Aaron Hinz, and Rees Kassen. 2014. Adaptive synonymous mutations in an experimentally evolved Pseudomonas fluorescens population. *Nature communications* 5(May):4076.

Bakker, Henk C Den, Xavier Didelot, Esther D Fortes, Kendra K Nightingale, and Martin Wiedmann. 2008. Lineage specific recombination rates and microevolution in Listeria monocytogenes. *BMC evolutionary biology* 13:1–13.

Battistuzzi, Fabia U, Andreia Feijao, and S Blair Hedges. 2004. A genomic timescale of prokaryote evolution: insights into the origin of methanogenesis, phototrophy, and the colonization of land. *BMC evolutionary biology* 4:44.

Beiko, Robert G, Timothy J Harlow, and Mark a Ragan. 2005. Highways of gene sharing in prokaryotes. *Proceedings of the National Academy of Sciences of the United States of America* 102(40):14332–14337.

Bergey, David H. 1923. *Bergey's manual of determinative bacteriology*. Williams & Wilkins Co.

Blainey, Paul C. 2013. The future is now: Single-cell genomics of bacteria and archaea. *FEMS Microbiology Reviews* 37(3):407–427. 3878092.

Blin, Kai, Marnix H. Medema, Daniyal Kazempour, Michael a. Fischbach, Rainer Breitling, Eriko Takano, and Tilmann Weber. 2013. antiSMASH 2.0–a versatile platform for genome mining of secondary metabolite producers. *Nucleic acids research* 41(Web Server issue): 1–9.

Boedicker, James Q., Meghan E. Vincent, and Rustem F. Ismagilov. 2009. Microfluidic confinement of single cells of bacteria in small volumes initiates high-density behavior of quorum sensing and growth and reveals its variability. *Angewandte Chemie - International Edition* 48(32):5908–5911.

Book, Adam J., Gina R. Lewin, Bradon R. McDonald, Taichi E. Takasuka, Evelyn Wendt-Pienkowski, Drew T. Doering, Steven Suh, Kenneth F. Raffa, Brian G. Fox, and Cameron R. Currie. 2016. Evolution of High Cellulolytic Activity in Symbiotic Streptomyces through Selection of Expanded Gene Content and Coordinated Gene Expression. *PLOS Biology* 14(6):e1002475.

Bos, Kirsten I, Verena J Schuenemann, G Brian Golding, Hernán a Burbano, Nicholas Waglechner, Brian K Coombes, Joseph B McPhee, Sharon N DeWitte, Matthias Meyer,

Sarah Schmedes, James Wood, David J D Earn, D Ann Herring, Peter Bauer, Hendrik N Poinar, and Johannes Krause. 2011. A draft genome of Yersinia pestis from victims of the Black Death. *Nature* 478(7370):506–10.

Boucher, Yan, Christophe J Douady, R Thane Papke, David a Walsh, Mary Ellen R Boudreau, Camilla L Nesbø, Rebecca J Case, and W Ford Doolittle. 2003. Lateral gene transfer and the origins of prokaryotic groups. *Annual review of genetics* 37:283–328.

Brocks, J. J. 1999. Archean Molecular Fossils and the Early Rise of Eukaryotes. *Science* 285(5430):1033–1036.

Cadillo-Quiroz, Hinsby, Xavier Didelot, Nicole L. Held, Alfa Herrera, Aaron Darling, Michael L. Reno, David J. Krause, and Rachel J. Whitaker. 2012. Patterns of Gene Flow Define Species of Thermophilic Archaea. *PLoS Biology* 10(2):e1001265.

Cafaro, Matías J, Michael Poulsen, Ainslie E F Little, Shauna L Price, Nicole M Gerardo, Bess Wong, Alison E Stuart, Bret Larget, Patrick Abbot, and Cameron R Currie. 2011. Specificity in the symbiotic association between fungus-growing ants and protective Pseudonocardia bacteria. *Proceedings. Biological sciences / The Royal Society* 278(1713):1814–22.

Caldera, Eric J., and Cameron R. Currie. 2012. The Population Structure of Antibiotic-Producing Bacterial Symbionts of Apterostigma dentigerum Ants: Impacts of Coevolution and Multipartite Symbiosis. *The American Naturalist* 180(5):604–617.

Caro-Quintero, Alejandro, and Konstantinos T. Konstantinidis. 2012. Bacterial species may exist, metagenomics reveal. *Environmental Microbiology* 14(2):347–355.

Chater, Keith F., Sandor Biro, Kye Joon Lee, Tracy Palmer, and Hildgund Schrempf. 2010. The complex extracellular biology of Streptomyces. *FEMS Microbiology Reviews* 34(2): 171–198.

Chin, Chen-Shan, David H Alexander, Patrick Marks, Aaron A Klammer, James Drake, Cheryl Heiner, Alicia Clum, Alex Copeland, John Huddleston, Evan E Eichler, Stephen W Turner, and Jonas Korlach. 2013. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nature methods* 10(6):563–9.

Cho, HoJung, Henrik Jonsson, Kyle Campbell, Pontus Melke, Joshua W. Williams, Bruno Jedynak, Ann M. Stevens, Alex Groisman, and Andre Levchenko. 2007. Self-organization in high-density bacterial colonies: Efficient crowd control. *PLoS Biology* 5(11):2614–2623. 0611087.

Cimermancic, Peter, Marnix H. Medema, Jan Claesen, Kenji Kurita, Laura C. Wieland Brown, Konstantinos Mavrommatis, Amrita Pati, Paul a. Godfrey, Michael Koehrsen, Jon Clardy, Bruce W. Birren, Eriko Takano, Andrej Sali, Roger G. Linington, and Michael a. Fischbach. 2014. Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters. *Cell* 158(2):412–421.

Cohan, Frederick M. 1994. The Effects of Rare but Promiscuous Genetic Exchange on Evolutionary Divergence in Prokaryotes. *The American Naturalist* 143(6):965–986.

Cole, S T. 1999. Learning from the genome sequence of Mycobacterium tuberculosis H37Rv. *FEBS letters* 452(1-2):7–10.

Copeland, Julia K, Lijie Yuan, Mehdi Layeghifard, Pauline W Wang, and David S Guttman. 2015. Seasonal community succession of the phyllosphere microbiome. *Mol Plant-Microbe Interact* 28(3):274–285.

Cordero, Otto X, and Paulien Hogeweg. 2009. The impact of long-distance horizontal gene transfer on prokaryotic genome size. *Proceedings of the National Academy of Sciences of the United States of America* 106:21748–21753.

Cotter, Paul D, R Paul Ross, and Colin Hill. 2013. Bacteriocins - a viable alternative to antibiotics? *Nature reviews. Microbiology* 11(2):95–105.

Currie, Cameron R, Bess Wong, Alison E Stuart, Ted R Schultz, Stephen A Rehner, Ulrich G Mueller, Gi-Ho Sung, Joseph W Spatafora, and Neil A Straus. 2003. Ancient tripartite coevolution in the attine ant-microbe symbiosis. *Science (New York, N.Y.)* 299(5605):386–388.

David, Lawrence a, and Eric J Alm. 2011. Rapid evolutionary innovation during an Archaean genetic expansion. *Nature* 469(7328):93–96.

Davies, J. 1994. Inactivation of antibiotics and the dissemination of resistance genes. *Science* 264(5157):375–382.

Davis, Elaine O., Edith M. Dullaghan, and Lucinda Rand. 2002. Definition of the mycobacterial SOS box and use to identify LexA-regulated genes in Mycobacterium tuberculosis. *Journal of Bacteriology* 184(12):3287–3295.

Doolittle, W. Ford, and Olga Zhaxybayeva. 2009. On the origin of prokaryotic species. *Genome Research* (19):744–756. `arXiv:1104.0048v2`.

Eddy, Sean R. 2011. Accelerated profile HMM searches. *PLoS Computational Biology* 7(10).

Embley, T M. 1994. The molecular Phylogeny and Systematics of the Actinomycetes. *Annual Reviews* 48:257–289.

Ferris, M J, G Muyzer, and D M Ward. 1996. Denaturing gradient gel electrophoresis profiles of 16S rRNA-defined populations inhabiting a hot spring microbial mat Denaturing Gradient Gel Electrophoresis Profiles of 16S rRNA-Defined Populations Inhabiting a Hot Spring Microbial Mat Community. *Applied and Environmental Microbiology* 62(2):340–346.

Finn, Robert D, Alex Bateman, Jody Clements, Penelope Coggill, Ruth Y Eberhardt, Sean R Eddy, Andreas Heger, Kirstie Hetherington, Liisa Holm, Jaina Mistry, Erik L L Sonnhammer, John Tate, and Marco Punta. 2014. Pfam: the protein families database. *Nucleic acids research* 42(Database issue):D222–30.

Fleischmann, Robert D, Mark D Adams, Owen White, Rebecca A Clayton, F Ewen, Anthony R Kerlavage, Carol J Bult, Jean-francois Tomb, Brian A Dougherty, Joseph M Merrick, Keith Mckenney, Granger Sutton, Will Fitzhugh, Chris Fields, D Jeannie, John Scott, Robert Shirley, Li-ing Liu, Anna Glodek, Jenny M Kelley, F Janice, Cheryl A Phillips, Tracy Spriggs, Eva Hedblom, and Matthew D Cotton. 1995. Whole-Genome Random Sequencing and Assembly of Haemophilus Influenzae Rd Published by : American Association for the Advancement of Science Stable URL : http://www.jstor.org/stable/2887657. *Science* 269(5223):496–512.

Gao, Z, C H Tseng, Z Pei, and M J Blaser. 2007. Molecular analysis of human forearm superficial skin bacterial biota. *Proc Natl Acad Sci U S A* 104(8):2927–2932.

Garvin, Jessica, Roger Buick, Ariel D Anbar, Gail L Arnold, and Alan J Kaufman. 2009. Isotopic evidence for an aerobic nitrogen cycle in the latest Archean. *Science (New York, N.Y.)* 323(5917):1045–8.

Gawad, Charles, Winston Koh, and Stephen R. Quake. 2016. Single-cell genome sequencing: current state of the science. *Nature Reviews Genetics* 17(3):175–188.

Ge, Fan, Li S. Wang, and Junhyong Kim. 2005. The cobweb of life revealed by genome-scale estimates of horizontal gene transfer. *PLoS Biology* 3(10).

Gibb, Ewan A., and David R. Edgell. 2007. Multiple controls regulate the expression of mobE, an HNH homing endonuclease gene embedded within a ribonucleotide reductase gene of phage Aeh1. *Journal of Bacteriology* 189(13):4648–4661.

Gilbert, Jack A., Robert A. Quinn, Justine Debelius, Zhenjiang Z. Xu, James Morton, Neha Garg, Janet K. Jansson, Pieter C. Dorrestein, and Rob Knight. 2016. Microbiome-wide association studies link dynamic microbial consortia to disease. *Nature* 535(7610):94–103.

Gilbert, Jack A., Joshua A. Steele, J. Gregory Caporaso, Lars Steinbrück, Jens Reeder, Ben Temperton, Susan Huse, Alice C. McHardy, Rob Knight, Ian Joint, Paul Somerfield, Jed A. Fuhrman, and Dawn Field. 2012. Defining seasonal marine microbial community dynamics. *The ISME Journal* 6(2):298–308.

Gogarten, J Peter, W Ford Doolittle, and Jeffrey G Lawrence. 2002. Prokaryotic evolution in light of gene transfer. *Molecular biology and evolution* 19(12):2226–2238.

Gonzalez, I L, and R D Schmickel. 1986. The human 18S ribosomal RNA gene: evolution and stability. *American journal of human genetics* 38(4):419–27.

Green, Jessica, and Brendan J M Bohannan. 2006. Spatial scaling of microbial biodiversity. *Trends in Ecology and Evolution* 21(9):501–507.

Green, Jessica L, Brendan J M Bohannan, and Rachel J Whitaker. 2008. Microbial biogeography: from taxonomy to traits. *Science* 320(May):1039–1043.

Grice, Elizabeth A, Heidi H Kong, Sean Conlan, Clayton B Deming, Joie Davis, Alice C Young, NISC Comparative Sequencing Program, Gerard G Bouffard, Robert W Blakesley, Patrick R Murray, Eric D Green, Maria L Turner, and Julia A Segre. 2009. Topographical and temporal diversity of the human skin microbiome. *Science (New York, N.Y.)* 324(5931): 1190–2.

Haft, Daniel H., Jeremy D. Selengut, Roland a. Richter, Derek Harkins, Malay K. Basu, and Erin Beck. 2013. TIGRFAMs and genome properties in 2013. *Nucleic Acids Research* 41(D1):387–395.

Handelsman, Jo. 2004. Metagenomics : Application of Genomics to Uncultured Microorganisms Metagenomics : Application of Genomics to Uncultured Microorganisms. *Microbiology and Molecular Biology Reviews* 68(4):669–685.

Hao, Weilong, and G. Brian Golding. 2006. The fate of laterally transferred genes: Life in the fast lane to adaptation or death. *Genome Research* 16(5):636–643.

Hedges, S Blair, Julie Marin, Michael Suleski, Madeline Paymer, and Sudhir Kumar. 2015. Tree of Life Reveals Clock-Like Speciation and Diversification. *Molecular Biology and Evolution* 32(4):835–845.

Hess, Matthias. 2011. Metagenomic Discovery of. *Science* 463(6016):463–467. `arXiv:1011.1669v3`.

Hilario, Elena, and Johann Peter Gogarten. 1993. Horizontal transfer of ATPase genes - the tree of life becomes a net of life. *BioSystems* 31(2-3):111–119.

Hopwood, David A. 2007. *Streptomyces in Nature and Medicine*. New York: Oxford University Press.

Hugenholtz, P, P Hugenholtz, B M Goebel, B M Goebel, N R Pace, and N R Pace. 1998. Impact of culture independent studies on the emerging phylogenetic view of bacterial diversity. *Journal of Bacteriology* v(18):180p4765–4774.

Huttenhower, Curtis, Dirk Gevers, Rob Knight, Sahar Abubucker, Jonathan H. Badger, Asif T. Chinwalla, Heather H. Creasy, Ashlee M. Earl, Michael G. FitzGerald, Robert S. Fulton, Michelle G. Giglio, Kymberlie Hallsworth-Pepin, Elizabeth A. Lobos, Ramana Madupu, Vincent Magrini, John C. Martin, Makedonka Mitreva, Donna M. Muzny, Erica J. Sodergren, James Versalovic, Aye M. Wollam, Kim C. Worley, Jennifer R. Wortman, Sarah K. Young, Qiandong Zeng, Kjersti M. Aagaard, Olukemi O. Abolude, Emma Allen-Vercoe, Eric J. Alm, Lucia Alvarado, Gary L. Andersen, Scott Anderson, Elizabeth Appelbaum, Harindra M. Arachchi, Gary Armitage, Cesar A. Arze, Tulin Ayvaz, Carl C. Baker, Lisa Begg, Tsegahiwot Belachew, Veena Bhonagiri, Monika Bihan, Martin J. Blaser, Toby Bloom, Vivien Bonazzi, J. Paul Brooks, Gregory A. Buck, Christian J. Buhay, Dana A. Busam, Joseph L. Campbell, Shane R. Canon, Brandi L. Cantarel, Patrick S. G. Chain, I-Min A. Chen, Lei Chen, Shaila Chhibba, Ken Chu, Dawn M. Ciulla, Jose C. Clemente, Sandra W. Clifton, Sean Conlan, Jonathan Crabtree, Mary A. Cutting, Noam J. Davidovics, Catherine C. Davis, Todd Z. DeSantis, Carolyn Deal, Kimberley D. Delehaunty, Floyd E. Dewhirst, Elena Deych, Yan Ding, David J. Dooling, Shannon P. Dugan, Wm Michael Dunne, A. Scott Durkin, Robert C. Edgar, Rachel L. Erlich, Candace N. Farmer, Ruth M. Farrell, Karoline Faust, Michael Feldgarden, Victor M. Felix, Sheila Fisher, Anthony A. Fodor, Larry J. Forney, Leslie Foster, Valentina Di Francesco, Jonathan Friedman, Dennis C. Friedrich, Catrina C. Fronick, Lucinda L. Fulton, Hongyu Gao, Nathalia Garcia, Georgia Giannoukos, Christina Giblin, Maria Y. Giovanni, Jonathan M. Goldberg, Johannes Goll, Antonio Gonzalez, Allison Griggs, Sharvari Gujja, Susan Kinder Haake, Brian J. Haas, Holli A. Hamilton, Emily L. Harris, Theresa A. Hepburn, Brandi Herter, Diane E. Hoffmann, Michael E. Holder, Clinton Howarth, Katherine H. Huang, Susan M. Huse, Jacques Izard, Janet K. Jansson, Huaiyang Jiang, Catherine Jordan, Vandita Joshi, James A. Katancik, Wendy A. Keitel, Scott T. Kelley, Cristyn Kells, Nicholas B. King, Dan Knights, Heidi H.

Kong, Omry Koren, Sergey Koren, Karthik C. Kota, Christie L. Kovar, Nikos C. Kyrpides, Patricio S. La Rosa, Sandra L. Lee, Katherine P. Lemon, Niall Lennon, Cecil M. Lewis, Lora Lewis, Ruth E. Ley, Kelvin Li, Konstantinos Liolios, Bo Liu, Yue Liu, Chien-Chi Lo, Catherine A. Lozupone, R. Dwayne Lunsford, Tessa Madden, Anup A. Mahurkar, Peter J. Mannon, Elaine R. Mardis, Victor M. Markowitz, Konstantinos Mavromatis, Jamison M. McCorrison, Daniel McDonald, Jean McEwen, Amy L. McGuire, Pamela McInnes, Teena Mehta, Kathie A. Mihindukulasuriya, Jason R. Miller, Patrick J. Minx, Irene Newsham, Chad Nusbaum, Michelle O'Laughlin, Joshua Orvis, Ioanna Pagani, Krishna Palaniappan, Shital M. Patel, Matthew Pearson, Jane Peterson, Mircea Podar, Craig Pohl, Katherine S. Pollard, Mihai Pop, Margaret E. Priest, Lita M. Proctor, Xiang Qin, Jeroen Raes, Jacques Ravel, Jeffrey G. Reid, Mina Rho, Rosamond Rhodes, Kevin P. Riehle, Maria C. Rivera, Beltran Rodriguez-Mueller, Yu-Hui Rogers, Matthew C. Ross, Carsten Russ, Ravi K. Sanka, Pamela Sankar, J. Fah Sathirapongsasuti, Jeffery A. Schloss, Patrick D. Schloss, Thomas M. Schmidt, Matthew Scholz, Lynn Schriml, Alyxandria M. Schubert, Nicola Segata, Julia A. Segre, William D. Shannon, Richard R. Sharp, Thomas J. Sharpton, Narmada Shenoy, Nihar U. Sheth, Gina A. Simone, Indresh Singh, Christopher S. Smillie, Jack D. Sobel, Daniel D. Sommer, Paul Spicer, Granger G. Sutton, Sean M. Sykes, Diana G. Tabbaa, Mathangi Thiagarajan, Chad M. Tomlinson, Manolito Torralba, Todd J. Treangen, Rebecca M. Truty, Tatiana A. Vishnivetskaya, Jason Walker, Lu Wang, Zhengyuan Wang, Doyle V. Ward, Wesley Warren, Mark A. Watson, Christopher Wellington, Kris A. Wetterstrand, James R. White, Katarzyna Wilczek-Boney, YuanQing Wu, Kristine M. Wylie, Todd Wylie, Chandri Yandava, Liang Ye, Yuzhen Ye, Shibu Yooseph, Bonnie P. Youmans, Lan Zhang, Yanjiao Zhou, Yiming Zhu, Laurie Zoloth, Jeremy D. Zucker, Bruce W. Birren, Richard A. Gibbs, Sarah K. Highlander, Barbara A. Methé, Karen E. Nelson, Joseph F. Petrosino, George M. Weinstock, Richard K. Wilson, and Owen White. 2012. Structure, function and diversity of the healthy human microbiome. *Nature* 486(7402):207–214. NIHMS150003.

Hyatt, Doug, Gwo-Liang Chen, Philip F Locascio, Miriam L Land, Frank W Larimer, and Loren J Hauser. 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC bioinformatics* 11(1):119.

Jansen, J L, J Nienhuis, E W Triplett, James Borneman, Paul W Skroch, Katherine M O Sullivan, James a Palus, Norma G Rumjanek, Jennifer L Jansen, and James Nienhuis. 1996. Molecular microbial diversity of an agricultural soil in Wisconsin . Molecular Microbial Diversity of an Agricultural Soil in Wisconsin 62(6):1935–1943.

Jochmann, Nina, Anna Katharina Kurze, Lisa F. Czaja, Karina Brinkrolf, Iris Brune, Andrea T. Hüser, Nicole Hansmeier, Alfred Pühler, Ilya Borovok, and Andreas Tauch. 2009.

Genetic makeup of the Corynebacterium glutamicum LexA regulon deduced from comparative transcriptomics and in vitro DNA band shift assays. *Microbiology* 155(5):1459–1477.

Kanehisa, Minoru, Susumu Goto, Yoko Sato, Masayuki Kawashima, Miho Furumichi, and Mao Tanabe. 2014. Data, information, knowledge and principle: Back to metabolism in KEGG. *Nucleic Acids Research* 42(D1):199–205.

Katoh, Kazutaka, and Daron M Standley. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution* 30(4):772–80.

Keeling, Patrick J. 2009. Functional and ecological impacts of horizontal gene transfer in eukaryotes. *Current Opinion in Genetics and Development* 19(6):613–619.

Keeling, Patrick J, and Jeffrey D Palmer. 2008. Horizontal gene transfer in eukaryotic evolution. *Nature reviews. Genetics* 9(8):605–618.

Kelsic, Eric D., Jeffrey Zhao, Kalin Vetsigian, and Roy Kishony. 2015. Counteraction of antibiotic production and degradation stabilizes microbial communities. *Nature* 521(7553): 516–519.

Keymer, Juan E, Peter Galajda, Guillaume Lambert, David Liao, and Robert H. Austin. 2008. Computation of mutual fitness by competing bacteria. *Proceedings of the National Academy of Sciences of the United States of America* 105(51):20269–20273.

Klappenbach, Joel, John Dunbar, Thomas Schmidt, J Klappenbach, J Dunbar, and T Schmidt. 2000. rRNA Operon Copy Number Reflects Ecological Strategies of Bacteria. *Applied and Environmental Microbiology* 66(4):1328–1333.

Kloesges, Thorsten, Ovidiu Popa, William Martin, and Tal Dagan. 2011. Networks of gene sharing among 329 proteobacterial genomes reveal differences in lateral gene transfer frequency at different phylogenetic depths. *Molecular Biology and Evolution* 28(2):1057–1074.

Knöppel, Anna, Joakim Näsvall, and Dan I Andersson. 2016. Compensating the fitness costs of synonymous mutations 1–46.

Koeppel, Alexander F., and Martin Wu. 2013. Surprisingly extensive mixed phylogenetic and ecological signals among bacterial Operational Taxonomic Units. *Nucleic Acids Research* 41(10):5175–5188.

Konstantinidis, Konstantinos T, and James M Tiedje. 2004. Trends between gene content and genome size in prokaryotic species with larger genomes. *Proceedings of the National Academy of Sciences of the United States of America* 101(9):3160–5.

———. 2005. Genomic insights that advance the species definition for prokaryotes. *Proceedings of the National Academy of Sciences of the United States of America* 102(7):2567–72.

Kroiss, Johannes, Martin Kaltenpoth, Bernd Schneider, Maria-Gabriele Schwinger, Christian Hertweck, Ravi Kumar Maddula, Erhard Strohm, and Ales Svatos. 2010. Symbiotic Streptomycetes provide antibiotic combination prophylaxis for wasp offspring. *Nature chemical biology* 6(4):261–3.

Kuo, Chih Horng, and Howard Ochman. 2009. The fate of new bacterial genes. *FEMS Microbiology Reviews* 33(1):38–43.

Kurtz, Stefan, Adam Phillippy, Arthur L Delcher, Michael Smoot, Martin Shumway, Corina Antonescu, and Steven L Salzberg. 2004. Versatile and open software for comparing large genomes. *Genome biology* 5(2):R12.

Lawrence, J G, and H Ochman. 1998. Molecular archaeology of the Escherichia coli genome. *Proceedings of the National Academy of Sciences of the United States of America* 95(16): 9413–9417.

Lawson, Daniel John. 2012. Populations in statistical genetic modelling and inference 1–29.

Lechner, Marcus, Sven Findeiss, Lydia Steiner, Manja Marz, Peter F Stadler, and Sonja J Prohaska. 2011. Proteinortho: detection of (co-)orthologs in large-scale analysis. *BMC bioinformatics* 12(1):124.

Lederberg, Joshua, and E L Tatum. 1946. Gene Recombination in Escherichia coli. *Nature* 158(4016):558.

Lerat, Emmanuelle, Vincent Daubin, Howard Ochman, and Nancy a. Moran. 2005. Evolutionary origins of genomic repertoires in bacteria. *PLoS Biology* 3(5):0807–0814.

Lercher, Martin J., and Csaba Pál. 2008. Integration of horizontally transferred genes into regulatory interaction networks takes many million years. *Molecular Biology and Evolution* 25(3):559–567.

Levy, Michael B. 1992. the Problem of Pattern and Scale in Ecology. *Ecology* 73(6):1943–1967.

Lewin, Gina R., Camila Carlos, Marc G. Chevrette, Heidi A. Horn, Bradon R. McDonald, Robert J. Stankey, Brian G. Fox, and Cameron R. Currie. 2016. Evolution and Ecology of Actinobacteria and their Bioenergy Applications. *Annual Reviews Microbiology* In press.

Lin, Kui, Lei Zhu, and Da Yong Zhang. 2006. An initial strategy for comparing proteins at the domain architecture level. *Bioinformatics* 22(17):2081–2086.

Long, Lijuan, and Jingfa Xiao. 2016. Comparative Genomics Analysis of Streptomyces Species Reveals Their Adaptation to the Marine Environment and Their Diversity at the Genomic Level 7(June):1–16.

Long, Tao, and Dani Or. 2005. Aquatic habitats and diffusion constraints affecting microbial coexistence in unsaturated porous media. *Water Resources Research* 41(8):1–10.

———. 2009. Dynamics of microbial growth and coexistence on variably saturated rough surfaces. *Microbial Ecology* 58(2):262–275.

Ma, Bin, Haizhen Wang, Melissa Dsouza, Jun Lou, Yan He, Zhongmin Dai, Philip C Brookes, Jianming Xu, and Jack A Gilbert. 2016. Geographic patterns of co-occurrence network topological features for soil microbiota at continental scale in eastern China. *The ISME journal* 10(8):1–11.

MacLean, R. Craig. 2005. Adaptive radiation in microbial microcosms. *Journal of Evolutionary Biology* 18(6):1376–1386.

Mao, Hanbin, Paul S Cremer, and Michael D Manson. 2003. A sensitive, versatile microfluidic assay for bacterial chemotaxis. *Proceedings of the National Academy of Sciences of the United States of America* 100(9):5449–54.

Marsh, Sarah E., Michael Poulsen, Adri??n Pinto-Tom??s, and Cameron R. Currie. 2014. Interaction between workers during a short time window is required for bacterial symbiont transmission in acromyrmex leaf-cutting ants. *PLoS ONE* 9(7).

Martiny, Jennifer B Hughes, Brendan J M Bohannan, James H Brown, Robert K Colwell, Jed a Fuhrman, Jessica L Green, M Claire Horner-Devine, Matthew Kane, Jennifer Adams Krumins, Cheryl R Kuske, Peter J Morin, Shahid Naeem, Lise Ovreås, Anna-Louise Reysenbach, Val H Smith, and James T Staley. 2006. Microbial biogeography: putting microorganisms on the map. *Nature reviews. Microbiology* 4(February):102–112.

Marttinen, Pekka, William P Hanage, Nicholas J Croucher, Thomas R Connor, Simon R Harris, Stephen D Bentley, and Jukka Corander. 2012. Detection of recombination events in bacterial genomes from large population samples. *Nucleic acids research* 40(1):e6.

McCallum, F S, and B E Maden. 1985. Human 18 S ribosomal RNA sequence inferred from DNA sequence. Variations in 18 S sequences and secondary modification patterns between vertebrates. *The Biochemical journal* 232(3):725–33.

McFall-Ngai, Margaret J. 2015. Giving microbes their due–animal life in a microbially dominant world. *The Journal of experimental biology* 218(Pt 12):1968–73.

Mirarab, Siavash, and Tandy Warnow. 2015. ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. *Bioinformatics (Oxford, England)* 31(12):i44–52.

Mojzsis, S J, G Arrhenius, K D McKeegan, T M Harrison, A P Nutman, and C R Friend. 1996. Evidence for life on Earth before 3,800 million years ago. *Nature* 384(6604):55–9.

Moran, Nancy A., John P. McCutcheon, and Atsushi Nakabachi. 2008. Genomics and evolution of heritable bacterial symbionts. *Ann Rev Genet* 42(1):165–190.

Moran, Nancy a, Heather J McLaughlin, and Rotem Sorek. 2009. The dynamics and time scale of ongoing genomic erosion in symbiotic bacteria. *Science (New York, N.Y.)* 323(5912): 379–382.

Moriya, Yuki, Masumi Itoh, Shujiro Okuda, Akiyasu C. Yoshizawa, and Minoru Kanehisa. 2007. KAAS: An automatic genome annotation and pathway reconstruction server. *Nucleic Acids Research* 35(SUPPL.2):182–185.

Moyer, Craig L, Fred C Dobbs, and David M Karl. 1994. Estimation of Diversity and Community Structure through Restriction Fragment Length Polymorphism Distribution Analysis of Bacterial 16S rRNA Genes from a Microbial Mat at an Active , Hydrothermal Vent System , Loihi Seamount , Hawaiit 60(3):871–879.

Nawrocki, Eric P., Sarah W. Burge, Alex Bateman, Jennifer Daub, Ruth Y. Eberhardt, Sean R. Eddy, Evan W. Floden, Paul P. Gardner, Thomas A. Jones, John Tate, and Robert D. Finn. 2015. Rfam 12.0: Updates to the RNA families database. *Nucleic Acids Research* 43(D1): D130–D137.

Nawrocki, Eric P., and Sean R. Eddy. 2013. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* 29(22):2933–2935.

Nowell, Reuben W, Sarah Green, Bridget E Laue, and Paul M Sharp. 2014. The extent of genome flux and its role in the differentiation of bacterial lineages. *Genome biology and evolution* 6(6):1514–29.

Ochman, H, S Elwyn, and N a Moran. 1999. Calibrating bacterial evolution. *Proceedings of the National Academy of Sciences of the United States of America* 96(22):12638–12643.

Ochman, H, J G Lawrence, and E A Groisman. 2000. Lateral gene transfer and the nature of bacterial innovation. *Nature* 405(6784):299–304.

Ochman, Howard, and Allan C. Wilson. 1987. Evolution in bacteria: Evidence for a universal substitution rate in cellular genomes. *Journal of Molecular Evolution* 26(1-2): 74–86.

Or, Dani, Sachin Phutane, and Arnaud Dechesne. 2007. Extracellular Polymeric Substances Affecting Pore-Scale Hydrologic Conditions for Bacterial Activity in Unsaturated Soils. *Vadose Zone Journal* 6(2):298.

Paez-Espino, David, Emiley A. Eloe-Fadrosh, Georgios A. Pavlopoulos, Alex D. Thomas, Marcel Huntemann, Natalia Mikhailova, Edward Rubin, Natalia N. Ivanova, and Nikos C. Kyrpides. 2016. Uncovering Earth's virome. *Nature* 536(7617):425–430. NIHMS150003.

Park, Seongyong, Dasol Kim, Robert J Mitchell, and Taesung Kim. 2011. A microfluidic concentrator array for quantitative predation assays of predatory microbes. *Lab on a chip* 11(17):2916–2923.

Powell, Sean, Damian Szklarczyk, Kalliopi Trachana, Alexander Roth, Michael Kuhn, Jean Muller, Roland Arnold, Thomas Rattei, Ivica Letunic, Tobias Doerks, Lars J. Jensen, Christian Von Mering, and Peer Bork. 2012. eggNOG v3.0: Orthologous groups covering 1133 organisms at 41 different taxonomic ranges. *Nucleic Acids Research* 40(D1):284–289.

Press, Maximilian Oliver, Christine Queitsch, and Elhanan Borenstein. 2016. Evolutionary assembly patterns of prokaryotic genomes. *Genome R* 26:826–833.

Price, Morgan N., Paramvir S. Dehal, and Adam P. Arkin. 2010. FastTree 2 - Approximately maximum-likelihood trees for large alignments. *PLoS ONE* 5(3).

Quast, Christian, Elmar Pruesse, Pelin Yilmaz, Jan Gerken, Timmy Schweer, Pablo Yarza, Jörg Peplies, and Frank Oliver Glöckner. 2013. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic acids research* 41(Database issue):D590–6.

de Queiroz, Kevin. 2005. Ernst Mayr and the modern concept of species. *Proceedings of the National Academy of Sciences of the United States of America* 102 Suppl:6600–6607.

———. 2007. Species concepts and species delimitation. *Systematic Botany* 56(6):879–886.

Rainey, Paul B., and Michael Travisano. 1998. Adaptive radiation in a heterogeneous environment. *Nature* 394(6688):69–72.

Raynaud, Xavier, and Naoise Nunan. 2014. Spatial ecology of bacteria at the microscale in soil. *PLoS ONE* 9(1).

Retchless, Adam C, and Jeffrey G Lawrence. 2010. Phylogenetic incongruence arising from fragmented speciation in enteric bacteria. *Proceedings of the National Academy of Sciences of the United States of America* 107(15):11453–11458.

Rinke, Christian, Patrick Schwientek, Alexander Sczyrba, Natalia N Ivanova, Iain J Anderson, Jan-Fang Cheng, Aaron E Darling, Stephanie Malfatti, Brandon K Swan, Esther a Gies, Jeremy a Dodsworth, Brian P Hedlund, George Tsiamis, Stefan M Sievert, Wen-Tso Liu, Jonathan a Eisen, Steven J Hallam, Nikos C Kyrpides, Ramunas Stepanauskas, Edward M Rubin, Philip Hugenholtz, and Tanja Woyke. 2013. Insights into the phylogeny and coding potential of microbial dark matter. *Nature* 499(7459):431–437.

Romano, Kymberleigh A, Eugenio I Vivas, Daniel Amador-noguez, and Federico E Rey. 2015. Intestinal Microbiota Composition Modulates Choline Bioavailability. *mBio* 6(2): 1–8.

Rosing, M. T. 1999. 13C-Depleted Carbon Microparticles in 3700-Ma Sea-Floor Sedimentary Rocks from West Greenland. *Science* 283(5402):674–676.

Scherlach, Kirstin, and Christian Hertweck. 2009. Triggering cryptic natural product biosynthesis in microorganisms. *Organic & biomolecular chemistry* 7(9):1753–1760.

Schloss, Patrick D., Rene A. Girard, Thomas Martin, Joshua Edwards, and J. Cameron Thrash. 2016. Status of the Archaeal and Bacterial Census: an Update. *mBio* 7(3):e00201–16.

Shapiro, B. J., J. Friedman, O. X. Cordero, S. P. Preheim, S. C. Timberlake, G. Szabo, M. F. Polz, and E. J. Alm. 2012. Population Genomics of Early Events in the Ecological Differentiation of Bacteria. *Science* 336(6077):48–51.

Shapiro, B. Jesse, Jean-Baptiste Leducq, and James Mallet. 2016. What Is Speciation? *PLOS Genetics* 12(3):e1005860.

Sit, Clarissa S., Antonio C. Ruzzini, Ethan B. Van Arnam, Timothy R. Ramadhar, Cameron R. Currie, and Jon Clardy. 2015. Variable genetic architectures produce virtually identical molecules in bacterial symbionts of fungus-growing ants. *Proceedings of the National Academy of Sciences* 112(43):13150–13154.

Smillie, Chris S, Mark B Smith, Jonathan Friedman, Otto X Cordero, Lawrence A David, and Eric J Alm. 2011. Ecology drives a global network of gene exchange connecting the human microbiome. *Nature* 480(7376):241–4.

Sorek, Rotem, Yiwen Zhu, Christopher J. Creevey, M. Pilar Francino, Peer Bork, and Edward M. Rubin. 2007. Genome-Wide Experimental Determination of Barriers to Horizontal Gene Transfer. *Science* 318(5855):1449–1452.

Soucy, Shannon M., Jinling Huang, and Johann Peter Gogarten. 2015. Horizontal gene transfer: building the web of life. *Nature Reviews Genetics* 16(8):472–482.

Stamatakis, Alexandros. 2006. RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22(21):2688–2690.

———. 2014. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30(9):1312–1313. `bioinformatics/btu033`.

Stewart, Eric J., Richard Madden, Gregory Paul, and Franc Taddei. 2005. Aging and death in an organism that reproduces by morphologically symmetric division. *PLoS Biology* 3(2):0295–0300. `1001.0057`.

Stocker, Roman, Justin R Seymour, Azadeh Samadani, Dana E Hunt, and Martin F Polz. 2008. Rapid chemotactic response enables marine bacteria to exploit ephemeral microscale nutrient patches. *Proceedings of the National Academy of Sciences of the United States of America* 105(11):4209–4214.

Stoeckli, M, P Chaurand, D E Hallahan, and R M Caprioli. 2001. Imaging mass spectrometry: a new technology for the analysis of protein expression in mammalian tissues. *Nature medicine* 7(4):493–496.

Stubbendieck, Reed M, Carol Vargas-Bautista, and Paul D Straight. 2016. Bacterial Communities: Interactions to Scale. *Frontiers in Microbiology* 7(August):1–19.

Syvanen, Michael. 2012. Evolutionary implications of horizontal gene transfer. *Annual review of genetics* 46:341–58.

Sze, Marc, and Patrick D Schloss. 2016. Looking for a Signal in the Noise: Revisiting Obesity and the Microbiome. *mBio* 7(4):1–9.

Tajima, F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123(3):585–595.

Tamura, Koichiro, Fabia Ursula Battistuzzi, Paul Billing-Ross, Oscar Murillo, Alan Filipski, and Sudhir Kumar. 2012. Estimating divergence times in large molecular phylogenies. *Proceedings of the National Academy of Sciences of the United States of America* 109(47):19333–8.

Taoka, Masato, Yoshio Yamauchi, Takashi Shinkawa, Hiroyuki Kaji, Wakana Motohashi, Hiroshi Nakayama, Nobuhiro Takahashi, and Toshiaki Isobe. 2004. Only a Small Subset of the Horizontally Transferred Chromosomal Genes in Escherichia coli Are Translated into Proteins 780–787.

Tatusova, Tatiana, Stacy Ciufo, Boris Fedorov, Kathleen O'Neill, and Igor Tolstoy. 2014. RefSeq microbial genomes database: new representation and annotation strategy. *Nucleic acids research* 42(Database issue):D553–9.

Thomas, Christopher M, and Kaare M Nielsen. 2005. Mechanisms of, and barriers to, horizontal gene transfer between bacteria. *Nature reviews. Microbiology* 3(9):711–721.

Treves, D. S., B. Xia, J. Zhou, and J. M. Tiedje. 2003. A two-species test of the hypothesis that spatial isolation influences microbial diversity in soil. *Microbial Ecology* 45(1):20–28.

Turnbaugh, Peter J, Micah Hamady, Tanya Yatsunenko, Brandi L Cantarel, Alexis Duncan, Ruth E Ley, Mitchell L Sogin, William J Jones, Bruce A Roe, Jason P Affourtit, Michael Egholm, Bernard Henrissat, Andrew C Heath, Rob Knight, and Jeffrey I Gordon. 2009. A core gut microbiome in obese and lean twins. *Nature* 457(7228):480–484. NIHMS150003.

Van Arnam, Ethan B., Antonio C. Ruzzini, Clarissa S. Sit, Cameron R. Currie, and Jon Clardy. 2015. A Rebeccamycin Analog Provides Plasmid-Encoded Niche Defense. *Journal of the American Chemical Society* 137(45):14272–14274.

Van Passel, M. W J, Pradeep Reddy Marri, and Howard Ochman. 2008. The emergence and fate of horizontally acquired genes in Escherichia coli. *PLoS Computational Biology* 4(4).

Vega, Nicole M, Kyle R Allison, Ahmad S Khalil, and James J Collins. 2012. Signaling-mediated bacterial persister formation. *Nature chemical biology* 8(5):431–3.

Vetsigian, Kalin, and Nigel Goldenfeld. 2005. Global divergence of microbial genome sequences mediated by propagating fronts. *Proceedings of the National Academy of Sciences of the United States of America* 102(20):7332–7337. 0501033.

Vos, Michiel, and Xavier Didelot. 2008. A comparison of homologous recombination rates in bacteria and archaea. *The ISME Journal* 3:199–208.

Vos, Michiel, Matthijn C. Hesselman, Tim A. te Beek, Mark W J van Passel, and Adam Eyre-Walker. 2015. Rates of Lateral Gene Transfer in Prokaryotes: High but Why? *Trends in Microbiology* 23(10):598–605.

Vos, Michiel, Alexandra B. Wolf, Sarah J. Jennings, and George A. Kowalchuk. 2013. Micro-scale determinants of bacterial diversity in soil. *FEMS Microbiology Reviews* 37(6): 936–954.

Vulić, M, F Dionisio, F Taddei, and M Radman. 1997. Molecular keys to speciation: DNA polymorphism and the control of genetic exchange in enterobacteria. *Proceedings of the National Academy of Sciences of the United States of America* 94(18):9763–7.

Watrous, Jeramie D, and Pieter C Dorrestein. 2011. Imaging mass spectrometry in microbiology. *Nature reviews. Microbiology* 9(9):683–694. NIHMS150003.

Welch, R A, V Burland, G Plunkett, P Redford, P Roesch, D Rasko, E L Buckles, S-R Liou, A Boutin, J Hackett, D Stroud, G F Mayhew, D J Rose, S Zhou, D C Schwartz, N T Perna, H L T Mobley, M S Donnenberg, and F R Blattner. 2002. Extensive mosaic structure revealed by the complete genome sequence of uropathogenic Escherichia coli. *Proceedings of the National Academy of Sciences of the United States of America* 99(26):17020–4.

Whitaker, Rachel J, Dennis W Grogan, and John W Taylor. 2003. Geographic barriers isolate endemic populations of hyperthermophilic archaea. *Science (New York, N.Y.)* 301(5635): 976–8.

Woese, C R, and G E Fox. 1977. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proceedings of the National Academy of Sciences of the United States of America* 74(11):5088–5090.

Woese, Carl R. 1987. Bacterial Evolution. *Microbiology* 51(2):221–271. NIHMS150003.

Yang, Ziheng. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Molecular biology and evolution* 24(8):1586–91.

Yawata, Yutaka, Otto X Cordero, Filippo Menolascina, Jan-Hendrik Hehemann, Martin F Polz, and Roman Stocker. 2014. Competition-dispersal tradeoff ecologically differentiates recently speciated marine bacterioplankton populations. *Proceedings of the National Academy of Sciences of the United States of America* 111(15):5622–7.

Zerbino, Daniel R., and Ewan Birney. 2008. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research* 18(5):821–829. 0209100.

Zhang, D, R F de Souza, V Anantharaman, L M Iyer, and L Aravind. 2012. Polymorphic toxin systems: Comprehensive characterization of trafficking modes, processing, mechanisms of action, immunity and ecology using comparative genomics. *Biol Direct* 7: 18.

Ziemert, Nadine, Anna Lechner, Matthias Wietz, Natalie Millán-Aguiñaga, Krystle L Chavarria, and Paul Robert Jensen. 2014. Diversity and evolution of secondary metabolism in the marine actinomycete genus Salinispora. *Proceedings of the National Academy of Sciences of the United States of America* 111(12):E1130–9.